1 One parameter exponential families

The world of exponential families bridges the gap between the Gaussian family and general distributions. Many properties of Gaussians carry through to exponential families in a fairly precise sense.

- In the Gaussian world, there exact small sample distributional results (i.e. t, F, χ^2).
- In the exponential family world, there are approximate distributional results (i.e. deviance tests).
- In the general setting, we can only appeal to asymptotics.

A one-parameter exponential family, $\mathcal{F}$ is a one-parameter family of distributions of the form

$$\mathbb{P}_\eta(dx) = \exp\left(\eta \cdot t(x) - \Lambda(\eta)\right) \mathbb{P}_0(dx)$$

for some probability measure $\mathbb{P}_0$. The parameter η is called the *natural* or *canonical* parameter and the function Λ is called the *cumulant generating function*, and is simply the normalization needed to make

$$f_\eta(x) = \frac{d\mathbb{P}_\eta}{d\mathbb{P}_0}(x) = \exp\left(\eta \cdot t(x) - \Lambda(\eta)\right)$$

a proper probability density. The random variable $t(X)$ is the *sufficient statistic* of the exponential family.

Note that $\mathbb{P}_0$ does not have to be a distribution on $\mathbb{R}$, but these are of course the simplest examples.

1.0.1 A first example: Gaussian with linear sufficient statistic

Consider the standard normal distribution

$$\mathbb{P}_0(A) = \int_A \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz$$

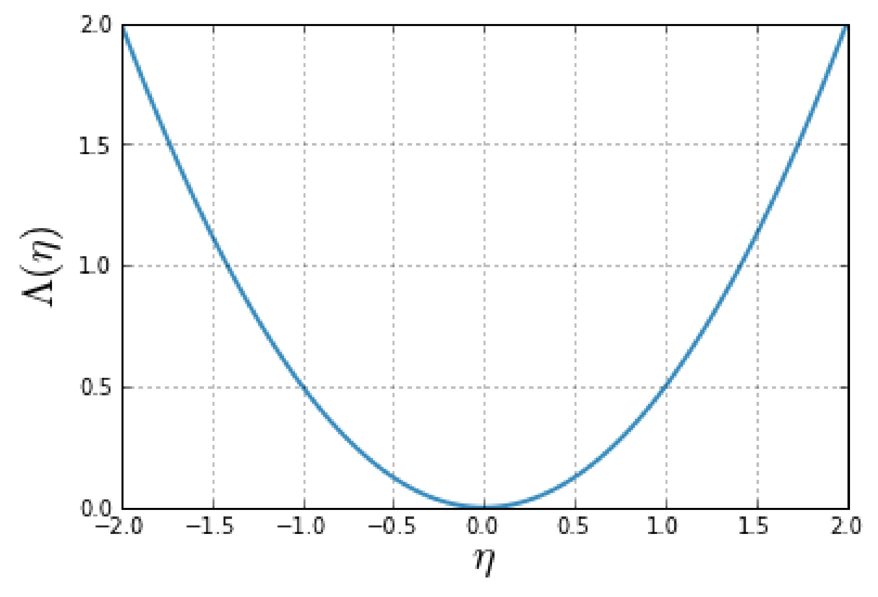
and let $t(x) = x$. Then, the exponential family is

$$\mathbb{P}_\eta(dx) \propto \frac{e^{\eta \cdot x - x^2/2}}{\sqrt{2\pi}}$$

and we see that

$$\Lambda(\eta) = \eta^2/2.$$

```
eta = np.linspace(-2,2,101)
CGF = eta**2/2.
plt.plot(eta, CGF)
A = plt.gca()
A.set_xlabel(r'$\eta$', size=20)
A.set_ylabel(r'$\Lambda(\eta)$', size=20)
f = plt.gcf()
```



Thus, the exponential family in this setting is the collection

$$\mathcal{F} = \{N(\eta, 1) : \eta \in \mathbb{R}\}.$$

1.0.2 Normal with quadratic sufficient statistic on $\mathbb{R}^d$

As a second example, take $\mathbb{P}_0 = N(0, I_{d\times d})$, i.e. the standard normal distribution on $\mathbb{R}^d$. As sufficient statistic, we take $t(x) = \|x\|_2^2/2$. Then, the exponential family is

$$\mathbb{P}_\eta(dx) \propto e^{\eta\cdot\|x\|_2^2/2 - \|x\|_2^2/2}$$

and we see that the family is only defined for $\eta < 1$. For $\eta < 1$,

$$\Lambda(\eta) = -\frac{d}{2}\log(1-\eta).$$

We see that not all exponential families have all of $\mathbb{R}$ as their parameter space.

We might as well define Λ over all of $\mathbb{R}$:

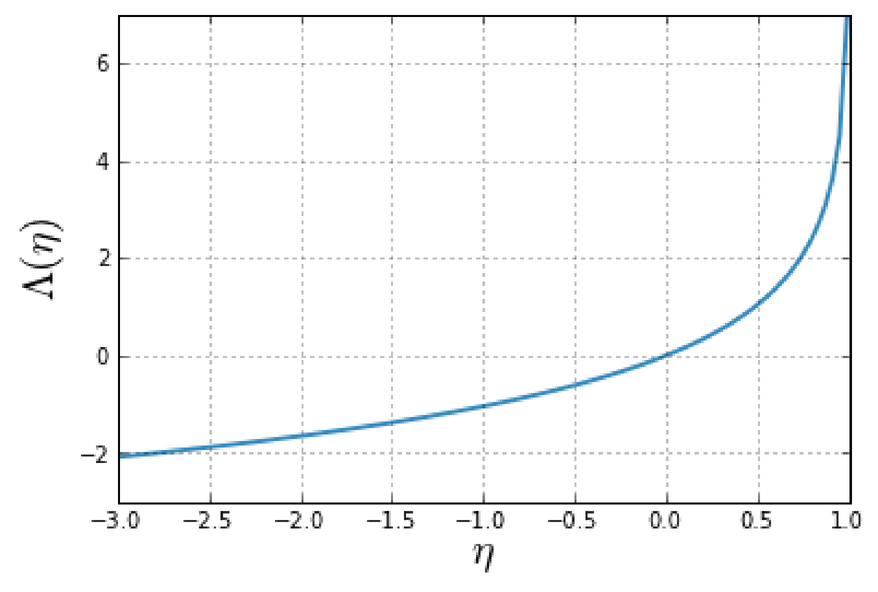
$$\Lambda(\eta) = \begin{cases} -\frac{d}{2}\log(1-\eta) & \eta < 1 \\ \infty & \eta \geq 1. \end{cases}$$

The exponential family here is

$$\mathcal{F} = \left\{N(0_{d\times 1}, (1-\eta)^{-1} \cdot I_{d\times d}), \eta < 1\right\}.$$

```
eta = np.linspace(-3,0.99,101)
d = 3
CGF = -d * np.log(1-eta)/2.
plt.plot(eta, CGF)
A = plt.gca()
A.set_xlabel(r'$\eta$', size=20)
A.set_ylabel(r'$\Lambda(\eta)$', size=20)
```

```
<matplotlib.text.Text at 0x10fcae0d0>
```



1.0.3 Tilts of triangular distribution

The previous two examples, we could express Λ explicitly by simple integration. This is not always possible, though we can use the computer to do some calculations for us. Set $\mathbb{P}_0$ to be the triangular distribution on $(-1,1)$ with sufficient statistic $t(x)=x$ so that

$$\mathbb{P}_\eta(dx) = \exp(\eta \cdot x - \Lambda(\eta))\mathbb{P}_0(dx)$$

with

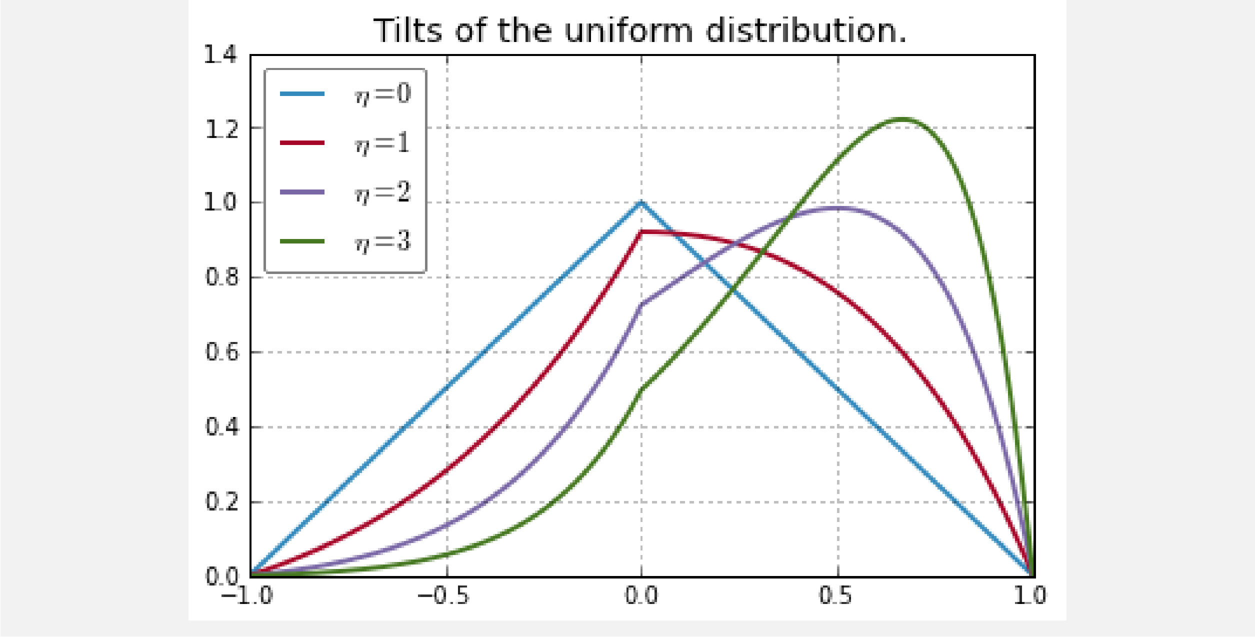
$$\Lambda(\eta) = \log\left(\int_{-1}^1 e^{\eta x}\, dx\right).$$

```
X = np.linspace(-1,1,501)
dX = X[1]-X[0]

def tilted_density(eta):
    D = np.exp(eta*X) * np.minimum((1 + X), (1 - X))
    CGF = np.log((np.exp(eta*X) * np.minimum((1 + X), (1 - X)) * dX).sum())
    return D / np.exp(CGF)

[plt.plot(X, tilted_density(eta), label=r’$\eta=%d$’ % eta) for eta in [0,1,2,3]]
plt.gca().set_title(’Tilts of the uniform distribution.’)
plt.legend(loc=’upper left’)
```

```
<matplotlib.legend.Legend at 0x10fcaea50>
```



1.0.4 Carrier measure

More generally, $\mathbb{P}_0$ could be replaced by some measure m_0 that is not a probability density.

For example, if m_0 is Lebesgue measure on $\mathbb{R}$ and $t(x) = x^2/2$. Then, for all $\eta < 0$

$$\frac{d\mathbb{P}_\eta}{dm_0}(x) = \frac{e^{\eta x^2/2}}{\sqrt{-2\pi/\eta}}$$

corresponds to a $N(0, -\eta^{-1})$ density.

To find $\Lambda(\eta)$, note that

$$e^{\Lambda(\eta)} = \int_{\mathbb{R}} e^{\eta x^2/2} dm_0(x) = \begin{cases} \sqrt{2\pi/-\eta} & \eta < 0 \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, for $\eta < 0$

$$\Lambda(\eta) = -\frac{1}{2}\log(-\eta) + \frac{1}{2}\log(2\pi).$$

The exponential family is therefore

$$\left\{N(0, -\eta^{-1}), \eta < 0\right\}.$$

1.1 Reparametrizing the family

Note that the exponential family is determined by the pair $(t(X), m_0)$. The choice of m_0 is somewhat arbitrary. We could fix some η_0 and consider a new family with carrier measure $\mathbb{P}_{\eta_0} \in \mathcal{F}$:

$$\widetilde{\mathcal{F}} = \left\{\widetilde{\mathbb{P}}_{\widetilde{\eta}} = \exp\left(\widetilde{\eta} \cdot t(x) - \widetilde{\Lambda}(\widetilde{\eta})\right) \mathbb{P}_{\eta_0}(dx)\right\}$$

But, a simple manipulation shows that

$$\begin{aligned}\mathbb{P}_\eta(dx) &= \exp\left(\eta \cdot t(x) - \Lambda(\eta)\right) m_0(dx) \\ &= \exp\left((\eta - \eta_0) \cdot t(x) - (\Lambda(\eta) - \Lambda(\eta_0)\right) \mathbb{P}_{\eta_0}(dx).\end{aligned}$$

This shows that there is a 1:1 correspondence between $\mathcal{F}$ and $\widetilde{\mathcal{F}}$. Namely

$$\begin{aligned}\widetilde{\mathcal{F}} \ni \widetilde{P}_{\widetilde{\eta}} &\mapsto \mathbb{P}_{\widetilde{\eta}+\eta_0} \in \mathcal{F} \\ \widetilde{\Lambda}(\widetilde{\eta}) &= \Lambda(\widetilde{\eta} + \eta_0) - \Lambda(\eta_0).\end{aligned}$$

1.2 Domain of an exponential family

In the examples above, we saw that not all values of η lead to a probability distribution due to the sufficient statistic not being integrable with respect to m_0. The *domain* $\mathcal{D}(\mathcal{F})$ can be thought of as the set of all natural parameters which lead to a probablity distribution.

Formally, we define the domain as

$$\mathcal{D}(\mathcal{F}) = \mathcal{D}((t(X), m_0)) = \left\{\eta : \Lambda(\eta) < \infty\right\}.$$

The domain is also defined relative to the carrier measure m_0. As in the previous section on reparametrization, we see

$$\begin{aligned}\mathcal{D}(\widetilde{\mathcal{F}}) &= \mathcal{D}((t(X), \mathbb{P}_{\eta_0}) \\ &= \{\widetilde{\eta} : \widetilde{\eta} + \eta_0 \in \mathcal{D}(\mathcal{F})\} \\ &= \{\widetilde{\eta} : \Lambda(\widetilde{\eta} + \eta_0) - \Lambda(\eta_0) < \infty\} = \mathcal{D}(\mathcal{F}) \oplus (-\eta_0).\end{aligned}$$

Hence, the domain of two exponential families with different parametrizations determined by different canonical parameters are related by a simple translation.

1.2.1 *Exercise: convexity of* $\mathcal{D}(\mathcal{F})$

1. Show that Λ is a (possibly infinite) convex function on $\mathbb{R}$.

2. Use this to show that $\mathcal{D}(\mathcal{F})$ is convex, i.e. a (possibly infinite) interval.

1.2.2 *Exercise: half-Gaussian density*

Consider the half-Gaussian distribution with density

$$f(x) = \frac{2e^{-x^2/2}}{\sqrt{2\pi}}, \qquad x \geq 0.$$

1. Use f as carrier measure to create an exponential family with sufficient statistic $t(x) = -x$.

2. Plot the density for $\eta \in [0, 2, 4, 6]$.

3. What is $\mathcal{D}(\mathcal{F})$?

4. What happens as $\eta \to \infty$? What about $\eta \to -\infty$?

5. Can you renormalize the random variables with distributions in $\mathcal{F}$ to get a "nice" limit at either $\pm\infty$? That is, suppose $Z_n \sim \mathbb{P}_{\eta_n}$ with $\eta_n \to \pm\infty$ Can you define $W_n = c_n(Z_n - \mu_n)$ so that W_n converges in distribution?

1.3 Example: the Poisson family

An important example of a one-parameter family that we will revisit often is the Poisson family on the non-negative integers $\mathbb{Z}_{\geq 0} = \{0, 1, 2, \dots\}$. The carrier measure is

$$m_0(dx) = \frac{1}{x!} m(dx)$$

with m the counting measure on $\mathbb{Z}_{\geq 0}$.

Poisson random variables are usually parametrized by their expectation λ. This is different than the canonical parametrization.

Let's write this parameterization as

$$\mathbb{Q}_\lambda(dx) = \frac{e^{-\lambda}\lambda^x}{x!} \, m_0(dx).$$

We see, then

$$\mathbb{P}_\eta(dx) = \exp(\eta \cdot x - \Lambda(\eta)) m_0(dx).$$

The two parametrizations are related by

$$\eta = \log(\lambda)$$
$$\Lambda(\eta) = e^\eta = \lambda$$

1.3.1 *Exercise: reparametrizing the Poisson family*

Let $\mathcal{F} = (x, m_0)$ denote the Poisson family.

1. What is $\mathcal{D}(\mathcal{F})$?
2. Rewrite the Poisson family $\mathbb{P}_\eta$ so that the carrier measure is a Poisson distribution with mean 2. Call the exponential family with this carrier measure $\mathcal{F}_2$. What is $\mathcal{D}(\mathcal{F}_2)$?
3. Write the Poisson distribution with mean 6 as a point in $\mathcal{D}(\mathcal{F})$ and as a point in $\mathcal{D}(\mathcal{F}_2)$. That is, in each case, find the canonical parameter such the corresponding distribution is a Poisson with mean 6.

1.4 Expectation and variances

The function Λ is the *cumulant generating function* of the family and differentiating it yields the cumulants of the random variable $t(X)$. Specifically, if the carrier measure is a probability measure, it is the logarithm of the *moment generating function* of $t(X)$ under $\mathbb{P}_0$. More generally, if the carrier measure is not a probability measure but just a measure on some sample space Ω, then for any $\eta \in \mathcal{D}(\mathcal{F})$

$$\mathbb{E}_\eta(e^{\theta \cdot t(X)}) = \int_\Omega e^{(\theta+\eta)t(x) - \Lambda(\eta)} \, m_0(dx) = e^{\Lambda(\theta+\eta) - \Lambda(\eta)}.$$

Note that

$$e^{\Lambda(\eta)} = \int_\Omega e^{\eta \cdot t(x)} \, m_0(dx).$$

Differentiating yields with respect to η

$$\begin{aligned}\dot{\Lambda}(\eta)e^{\Lambda(\eta)} &= \int_\Omega t(x)e^{\eta\cdot t(x)}\, m_0(dx) \\ &= e^{\Lambda(\eta)}\cdot \int_\Omega t(x)\mathbb{P}_\eta(dx) \\ &= e^{\Lambda(\eta)}\cdot \mathbb{E}_\eta(t(X)).\end{aligned}$$

Differentiating a second time yields

$$\begin{aligned}\left(\ddot{\Lambda}(\eta)+\dot{\Lambda}(\eta)^2\right)e^{\Lambda(\eta)} &= \int_\Omega t(x)^2 e^{\eta\cdot t(x)}\, m_0(dx) \\ &= e^{\Lambda(\eta)}\mathbb{E}_\eta(t(X)^2).\end{aligned}$$

Summarizing,

$$\begin{aligned}&\dot{\Lambda}(\eta) = \mathbb{E}_\eta(t(X)) \\ &\ddot{\Lambda}(\eta) = \mathbb{E}_\eta[(t(X)-\mathbb{E}_\eta(t(X))^2] = \mathrm{Var}_\eta(t(X))\end{aligned}$$

The above also motivates definition of another space related to $\mathcal{F}$, the set of realizable expected values

$$\mathcal{M}(\mathcal{F}) = \left\{\dot{\Lambda}(\eta) : \eta\in\mathcal{D}(\mathcal{F})\right\} = \dot{\Lambda}(\mathcal{D}(\mathcal{F})).$$

1.4.1 Parametrization by the mean

The above calculation yields a parameterization

$$\mu(\eta) = \mathbb{E}_\eta(t(X)) = \dot{\Lambda}(\eta).$$

As

$$\frac{d\mu}{d\eta} = \ddot{\Lambda}(\eta) = \mathrm{Var}_\eta(t(X)) \geq 0$$

we see that the mapping is 1:1 and non-decreasing and is invertible as long as the random variable $t(X)$ is not constant under $\mathbb{P}_\eta$.

Further, as the moment generating function of $t(X)$ under $\mathbb{P}_\eta$ is defined for all $\eta\in\mathcal{D}(\mathcal{F})$. This map is infinitely differentiable on $\mathcal{D}(\mathcal{F})$.

1.5 Skewness and kurtosis

We saw above that the moment generating function of $t(X)$ under $\mathbb{P}_\eta$ can be expressed as

$$\mathbb{E}_\eta(e^{\theta\cdot t(X)}) = e^{\Lambda(\theta+\eta)-\Lambda(\eta)}.$$

Taking the logs and expanding yields the cumulants of $t(X)$ under $\mathbb{P}_\eta$. That is,

$$\Lambda(\theta+\eta)-\Lambda(\eta) = k_1\theta + k_2\frac{\theta^2}{2} + k_3\frac{\theta^3}{6} + \dots$$

where

$$k_i = \Lambda^{(i)}(\eta)$$

are the cumulants of $t(X)$ under $\mathbb{P}_\eta$.

The skewness of a random variable defined as

$$\mathrm{Skew}(Y) = \mathrm{Skew}(Y, \mathcal{L}_Y) = \frac{\mathbb{E}[(Y - \mathbb{E}(Y))^3]}{\mathrm{Var}(Y)^{3/2}} \triangleq \gamma = \frac{k_3}{k_2^{3/2}}$$

and kurtosis is defined as

$$\mathrm{Kurtosis}(Y) = \mathrm{Kurtosis}(Y, \mathcal{L}_Y) = \frac{\mathbb{E}[(Y - \mathbb{E}(Y))^4]}{\mathrm{Var}(Y)^2} - 3 \triangleq \delta = \frac{k_4}{k_2^2}.$$

For a one-parameter exponential family, we see that

$$\begin{aligned}\gamma(\eta) &= \mathrm{Skew}(t(X), \mathbb{P}_\eta) \\ &= \frac{\Lambda^{(3)}(\eta)}{\ddot{\Lambda}(\eta)^{3/2}} \\ \delta(\eta) &= \mathrm{Kurtosis}(t(X), \mathbb{P}_\eta) \\ &= \frac{\Lambda^{(4)}(\eta)}{\ddot{\Lambda}(\eta)^2}\end{aligned}$$

1.5.1 *Exercise: skewness and kurtosis of the Poisson family*

1. Plot the skewness and kurtosis of the Poisson family as a function of η.

2. What happens as $\eta \to \infty$? Is this expected?

1.6 Cumulants in the mean parametrization

Above, we see that the most natural parameterization of cumulants (and skewness, kurtosis) is in terms of the canonical parameter η. However, we now have this alternative parametrization in terms of the mean μ. Given a quantity like $\gamma(\eta)$ we can reparametrize this as

$$\tilde{\gamma}(\mu) = \mathrm{Skew}(t(X), \mathbb{P}_{\eta(\mu)}) = \gamma(\eta(\mu)) = \frac{\Lambda^{(3)}(\eta(\mu))}{\mathrm{Var}_{\eta(\mu)}(t(X))^{3/2}}.$$

This new parametrization yields some interesting relations.

1.6.1 *Exercise: reparametrization of skewness*

Show that

$$\tilde{\gamma}(\mu) = 2\frac{d}{d\mu}\left(\mathrm{Var}_{\eta(\mu)}\right)^{1/2}.$$

1.6.2 *Exercise: estimation of μ and η*

Suppose $X \sim \mathbb{P}_\eta$, then, from the calculations we have seen above

$$\mathbb{E}_\eta(t(X)) = \dot{\Lambda}(\eta) = \mu(\eta).$$

Can we find an unbiased estimate of η? For this exercise, assume $\Omega = \mathbb{R}, t(x) = x$ and the carrier measure has a density with respect to Lebesgue measure and write $\mathcal{D}(\mathcal{F}) = [m, M]$ (with one or both of m, M possibly infinite). That is,

$$\mathbb{P}_\eta(dx) = \begin{cases} e^{\eta \cdot x - \Lambda(\eta)} g_0(x)\ dx & m \leq x \leq M \\ 0 & \text{otherwise.} \end{cases}$$

with $\Lambda(0) = 0$.

Define

$$\ell_0(x) = \log(g_0(x)).$$

Show that

$$\mathbb{E}_\eta \left[-\frac{d}{dx} \ell_0(x) \right] = \eta - \left[g_\eta(M) - g_\eta(m) \right].$$

Try this estimator out numerically for the half-Gaussian family.

1.7 Repeated sampling

One of the very nice properties of exponential families is the behaviour under IID sampling. Specifically, let

$$X_1, \dots, X_n \overset{IID}{\sim} \mathbb{P}_\eta$$

with

$$\frac{d\mathbb{P}_\eta}{dm_0}(x) = \exp(\eta \cdot t(x) - \Lambda(\eta)).$$

The joint density has a very simple expression:

$$\prod_{i=1}^n \left[\exp(\eta \cdot t(x_i) - \Lambda(\eta))\ m_0(dx_i) \right] = \exp\left(n \cdot \left[\eta \cdot \overline{t(X)} - \Lambda(\eta) \right] \right) \prod_{i=1}^n m_0(dx_i)$$

with

$$\overline{t(X)} = \frac{1}{n} \sum_{i=1}^n t(X_i).$$

This is a one-parameter exponential family with parameter η, sufficient statistic

$$n \cdot \overline{t(X)} = \sum_{i=1}^n t(X_i)$$

and carrier measure $\prod_{i=1}^n m_0(dx_i)$ defined on Ω^n where Ω is the sample space for each of the X_is.

1.7.1 *Exercise: cumulants of sum*

1. What is the cumulant generating function of the above exponential family? Call it Λ_n.

2. Relate the first 4 cumulants, as a function of η, of Λ_n to the mean, variance, skewness and kurtosis of $\sum_{i=1}^n t(X_i)$.

3. Repeat 2. for the random variable $\overline{t(X)}$.

1.7.2 *Exercise: sufficiency*

For this question, assume $m_0 = \mathbb{P}_0$ is a probability distribution.

1. Relate the one-parameter exponential family of distributions on Ω^n above, to a one-parameter exponential family of distributions on $\mathbb{R}$.
2. What is the carrier measure of this exponential family? What is its Λ, $\mathcal{D}(\mathcal{F})$?
3. How does this relate to sufficiency?

1.8 Examples

1.8.1 Binomial: Bin(n, p)

Suppose that $X \sim \text{Binomial}(n, \pi)$. Then,

$$\mathbb{P}(X = j) = \binom{n}{j} \pi^j (1-\pi)^{n-j} = \binom{n}{j} \exp\left(\log\left(\frac{\pi}{1-\pi}\right) \cdot j + n \cdot \log(1-\pi)\right)$$

This is a one-parameter family whose carrier measure we can take as having a density

$$m_0(j) = \begin{cases} \binom{n}{j} & 0 \leq j \leq n \\ 0 & \text{otherwise.} \end{cases}$$

with respect to counting measure on $\mathbb{Z}$.

The natural parameter is $\eta = \log\left(\frac{\pi}{1-\pi}\right)$ with $\mathcal{D} = (-\infty, \infty)$ and cumulant generating function

$$\Lambda(\eta) = n \log(1 + e^\eta).$$

We see that

$$\dot{\Lambda}(\eta) = \frac{ne^\eta}{1+e^\eta} = n\pi$$

and

$$\ddot{\Lambda}(\eta) = n \cdot \frac{e^\eta(1+e^\eta) - e^{2\eta}}{(1+\eta)^2} = n \cdot \pi(1-\pi).$$

1.8.2 *Exercise: Gamma with fixed shape*

Suppose we consider the Gamma family with fixed shape parameter k and unknown scale. The Gamma density is

$$f_{\lambda,k}(x) = \frac{1}{\Gamma(k)\lambda^k} x^{k-1} e^{-x/\lambda}, \qquad x \geq 0.$$

1. Write this as a one-parameter exponential family. What is the canonical parameter?
2. Compute the mean, variance, skewness and kurtosis of this family.

1.8.3 *Exercise: Gamma with a fixed scale*

Suppose that, instead of fixing the shape of the Gamma family, we fix the scale at some value λ. Can you write this as a one-parameter exponential family?

1.8.4 *Exercise: Negative binomial*

The negative binomial arises when waiting for a fixed number, k of failures in IID Bernoulli(π) trials. Specifically,

$$\mathbb{P}(X = n) = \binom{n+k-1}{k}(1-\pi)^k \pi^n.$$

1. Write this as a one-parameter exponential family.

2. Compute the mean, variance and skewness as a function of π.

1.8.5 Inverse Gaussian

This distribution arises from the hitting time of standard Brownian motion with drift to cross the boundary 1. If the drift is $1/\mu$, then the density has the form

$$g_\mu(x) = \frac{1}{\sqrt{2\pi x^3}} e^{-\frac{(x-\mu)^2}{2\mu^2 x}}.$$

2 Basic results on exponential families

2.1 MLE

Recall the joint density under repeated IID sampling

$$\prod_{i=1}^n \mathbb{P}_\eta(dx_i) = \prod_{i=1}^n \left[\exp(\eta \cdot t(x_i) - \Lambda(\eta)) m_0(dx_i)\right] = \exp\left(n \cdot \left[\eta \cdot \overline{t(X)} - \Lambda(\eta)\right]\right) \prod_{i=1}^n m_0(dx_i).$$

From this, we see that the *log-likelihood* has a very compact expression

$$\ell(\eta) = \log\left(\prod_{i=1}^n \mathbb{P}_\eta(dx_i)\right) = n \cdot \left[\eta \cdot \overline{t(X)} - \Lambda(\eta)\right].$$

The *score function* is defined as

$$\dot{\ell}(\eta) = \frac{d}{d\eta}\ell(\eta).$$

2.2 The MLE map

The *maximum likelihood estimator* of the canonical parameter is

$$\widehat{\eta} = \operatorname{argmax}_{\eta \in \mathcal{D}} \ell(\eta)$$

and they satisfy (assuming the maximum is achieved in the interior of $\mathcal{D}(\mathcal{F})$)

$$0 = \dot{\ell}(\widehat{\eta}) = n \cdot \left[\overline{t(X)} - \dot{\Lambda}(\widehat{\eta})\right].$$

In words, the MLE of η is chosen so that, under $\mathbb{P}_{\widehat{\eta}}$ the expected value of the sufficient statistic is the observed value $\overline{t(X)}$.

Solving the MLE equations therefore determine a map from $\mathcal{M}(\mathcal{F})$, the set of all possible mean values for $\mathcal{F}$ to $\mathcal{D}(\mathcal{F})$ the canonical parameter space. This map is effectively the inverse of $\dot{\Lambda}$. That is,

$$\widehat{\eta}(t(X)) = \dot{\Lambda}^{-1}(t(X)) = \text{argmax}_{\eta} \left(\eta \cdot \left[\frac{1}{n} \sum_{i=1}^{n} t(X_i) \right] - \Lambda(\eta) \right).$$

2.2.1 Fenchel-Legendre transform

Consider the maximized log-likelihood, as a function of $\overline{t(X)}$

$$\sup_{\eta \in \mathcal{D}} \left(\eta \cdot t - \Lambda(\eta) \right).$$

In convex analysis, this function is called the *Fenchel-Legendre* transform of Λ and is often denoted by Λ^*. That is,

$$\Lambda^*(t) = \sup_{\eta \in \mathcal{D}} \left(\eta \cdot t - \Lambda(\eta) \right).$$

Another general fact from convex analysis says that

$$\frac{d}{dt} \Lambda^*(t) = \text{argmax}_{\eta \in D} \left(\eta \cdot t - \Lambda(\eta) \right).$$

Sometimes, the Fenchel-Legendre transform may fail to be differentiable. In this case, the argmax above is a set, called the *subdifferential* of Λ^* at t. This would correspond to there being more than one MLE, which will not happen for one-parameter exponential families.

In any case, Λ^* provides the MLE map. That is,

$$\eta(\mu) = \dot{\Lambda}^*(\mu).$$

Another property of this map is

$$\Lambda^*(\mu) = \eta(\mu) \cdot \mu - \Lambda(\eta(\mu)).$$

In turn, this implies

$$\dot{\Lambda}(\dot{\Lambda}^*(\mu)) = \dot{\Lambda}(\eta(\mu)) = \mathbb{E}_{\eta(\mu)}[t(X)] = \mu.$$

This implies that $\dot{\Lambda} \circ \dot{\Lambda}^*$ is the identity on $\mathcal{M}$. And therefore, $\dot{\Lambda}^* \circ \dot{\Lambda}$ is the identity on $\mathcal{D}$. In other words, $\dot{\Lambda}^{-1} = \dot{\Lambda}^*$.

2.2.2 Likelihood as a function of μ

Alternatively, we might try computing the MLE in the mean parametrization. In this case, we write the likelihood as

$$\tilde{\ell}(\mu) = \ell(\dot{\Lambda}^*(\mu))$$

Differentiating

$$\begin{aligned} \frac{d}{d\mu} \tilde{\ell}(\mu) &= n \frac{\frac{d}{d\eta} \left(\eta \cdot \overline{t(X)} - \Lambda(\eta) \right)}{\frac{d\mu}{d\eta}} \\ &= n \cdot \frac{\overline{t(X)} - \mu}{\text{Var}_{\eta(\mu)}(t(X))} \\ &= n \cdot \ddot{\Lambda}^*(\mu) \cdot \left[\overline{t(X)} - \mu \right] \end{aligned}$$

which shows $\widehat{\mu} = \overline{t(X)}$.

2.2.3 *Exercise: density in $\mathcal{M}$ parametrization*

1. Show that

$$\frac{d\mathbb{P}_{\eta(\mu)}}{d\mathbb{P}_0}(x) = \exp\Big(\Lambda^*(\mu) - (\mu - t(x))\dot{\Lambda}^*(\mu)\Big)$$

2. Rederive the score for μ

$$n\ddot{\Lambda}^*(\hat{\mu}) \cdot \Big[\overline{t(X)} - \hat{\mu}\Big] = 0$$

   directly with this formula.

2.2.4 *Exercise: computing* Λ^*

1. Compute Λ^* for the Poisson family.
2. Knowing the relationship between μ and λ implied by $\mu(\eta) = \dot{\Lambda}(\eta)$, show $\widehat{\eta}$ computed with Λ^* agrees with the the usual MLE rule by plugging in $\widehat{\mu}$ into this relationship.

2.2.5 Score for arbitrary parameters

In general, the score function for some (invertible) function of η, i.e. $\xi = h(\eta)$ is

$$\frac{d}{d\xi}\ell(h^{-1}(\xi)) = \frac{\frac{d}{d\eta}\ell(\eta)\Big|_{\eta=h^{-1}(\xi)}}{\frac{dh}{d\eta}\Big|_{\eta=h^{-1}(\xi)}}.$$

One key property of the score function is

$$\begin{aligned}\mathbb{E}_\eta\left[\frac{d}{d\zeta}\ell(\zeta)\Big|_{\zeta=\eta}\right] &= n\left[\mathbb{E}_\eta[\overline{t(X)}] - \dot{\Lambda}(\eta)\right] \\ &= n\left[\mu(\eta) - \mu(\eta)\right] \\ &= 0\end{aligned}$$

This is also true for the score of any $\xi = h(\eta)$.

2.3 Fisher information

The Fisher information in $(X_1, \ldots, X_n)$ for $\xi = h(\eta)$ at $\eta \in \mathcal{D}$ is given by $\mathrm{Var}_{\eta(\xi)}\left(\frac{d}{d\xi}\ell(h^{-1}(\xi))\right)$. That is,

$$\begin{aligned}I_\eta^{(n)}(h(\eta)) &= \mathbb{E}_\eta\left(\frac{d}{d\xi}\ell(h^{-1}(\xi))^2\Big|_{\xi=h(\eta)}\right) \\ &= \frac{I_\eta^{(n)}}{\dot{h}(\eta)^2}.\end{aligned}$$

Above,

$$I_\eta^{(n)} = I_\eta^{(n)}(\eta) = n\mathbb{E}_\eta\left((t(X) - \dot{\Lambda}(\eta))^2\right) = n \cdot \mathrm{Var}_\eta(t(X))$$

where $\mathrm{Var}_\eta(t(X))$ is the Fisher information for η in one observation.

On reinspection of the loglikelihood $\ell(\eta)$ we see that

$$-\ddot{\ell}(\eta) = n\mathrm{Var}_\eta(t(X)).$$

So, the second derivative of the likelihood is in fact not random and we can write

$$I_\eta^{(n)} = -\ddot{\ell}(\eta).$$

2.3.1 *Exercise: Fisher information for the mean*

Compute the Fisher information for the mean parameter in an IID sample $(X_1, \dots, X_n)$.

2.3.2 *Exercise: Fisher information as a pullback*

Consider the map

$$\mathbb{R} \ni \eta \overset{\Psi}{\mapsto} \left(\frac{d\mathbb{P}_\eta}{dm_0}\right)^{1/2} = \exp\left(\frac{1}{2}\left(\eta \cdot t(x) - \Lambda(\eta)\right)\right) \in L^2(\Omega, m_0).$$

1. Show that the derivative of this map is

$$\frac{\partial \Psi}{\partial \eta} = \frac{1}{2}\left(t(x) - \dot{\Lambda}(\eta)\right) \cdot \exp\left(\frac{1}{2}\left(\eta \cdot t(x) - \Lambda(\eta)\right)\right) \in L^2(\Omega, m_0).$$

2. Show that

$$I_\eta = 4\left\|\frac{\partial \Psi}{\partial \eta}\right\|^2_{L^2(\Omega, m_0)}.$$

Geometrically, this says that the Fisher information is the *pull-back* of the inner product determined by Hellinger distance. In other words, the *length* of a curve

$$\int_a^b \sqrt{I_\eta}\, d\eta$$

in a one-parameter exponential family is, up to the constant factor 2, equal to *arclength* in the Hilbert structure induced by Hellinger distance.

2.3.3 Cramer-Rao lower bound

The Cramer-Rao lower bound for an unbiased estimator $\hat{\xi}$ of $\xi = h(\eta)$ based on an IID sample $(X_1, \dots, X_n)$ from $\mathbb{P}_\eta$ is

$$\mathrm{Var}_\eta(\hat{\xi}) \geq \frac{1}{I_\eta^{(n)}(h(\eta))} = \frac{\dot{h}(\eta)^2}{n \cdot \mathrm{Var}_\eta(t(X))}.$$

Applying this to $\mu = \dot{\Lambda}(\eta)$ yields

$$\mathrm{Var}_\eta(\hat{\mu}) \geq \frac{\ddot{\Lambda}(\eta)^2}{n \cdot \mathrm{Var}_\eta(t(X))} = \frac{\mathrm{Var}_\eta(t(X))}{n}$$

and

$$\hat{\mu} = \overline{t(X)}$$

achieves the Cramer-Rao lower bound. This happens for linear functions of μ but generally not for η.

The Cramer-Rao bound applies for unbiased estimators. In general the MLE $\hat{\xi} = h(\hat{\eta})$ is not unbiased and the bias should be included. Nevertheless, the delta rule approximation

$$\mathrm{Var}(\hat{\xi}) \approx \frac{\dot{h}(\eta)^2}{n\mathrm{Var}_\eta(t(X))}$$

is usually reasonable. In practice, of course we don't know η so we must use

$$\widehat{\mathrm{Var}(\hat{\xi})} \approx \frac{\dot{h}(\widehat{\eta})^2}{n\mathrm{Var}_{\widehat{\eta}}(t(X))}$$

2.3.4 Deviance

The *deviance* (also known as *mutual information*, *Kullback Leibler (KL) divergence*) between probability measures is defined as

$$D(\mathbb{P}_1;\mathbb{P}_2) = \begin{cases} 2 \cdot \mathbb{E}_1\left(\log \frac{d\mathbb{P}_1}{d\mathbb{P}_2}\right) & \mathbb{P}_2 \ll \mathbb{P}_1 \\ \infty & \text{otherwise.} \end{cases}$$

This notation is slightly different then the usual notation for KL divergence:

$$D_{KL}(\mathbb{P}||\mathbb{Q}) = \mathbb{E}_{\mathbb{P}}\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right).$$

I have used the same indexing as Brad Efron's notes: the first parameter is the one the integral is computed with respect to.

2.3.5 *Exercise: deviance in exponential families*

1. Show that, in a one-parameter exponential family $\mathcal{F}$, for any $\eta_1, \eta_2 \in \mathcal{D}(\mathcal{F})$

$$\begin{aligned} D(\eta_1;\eta_2) &\stackrel{\Delta}{=} D(\mathbb{P}_{\eta_1}, \mathbb{P}_{\eta_2}) \\ &= 2 \cdot \left[\Lambda(\eta_2) - \Lambda(\eta_1) + (\eta_1 - \eta_2) \cdot \dot{\Lambda}(\eta_1)\right]. \end{aligned}$$

2. Give a direct argument based on convexity that shows that $D(\eta_1;\eta_2) \geq 0$.

2.3.6 Convexity picture

The form of the deviance in one-parameter exponential families shows that it is in fact a remainder in a Taylor series

$$\Lambda(\eta_2) = \Lambda(\eta_1) + (\eta_2 - \eta_1) \cdot \dot{\Lambda}(\eta_1) + R(\eta_1; \eta_2)$$

with

$$R(\eta_1; \eta_2) = \frac{1}{2} D(\eta_1; \eta_2) = \frac{1}{2} \ddot{\Lambda}(\theta)(\eta_1 - \eta_2)^2$$

for some $\theta = \theta(\eta_1, \eta_2) \in [\eta_1, \eta_2]$.

Let's make a simple exponential family, which we suppose has

$$\begin{aligned} \Lambda(\eta) &= 5\eta^2 - \log(\eta) \\ \dot{\Lambda}(\eta) &= 10\eta - \eta^{-1} \end{aligned}$$

```
# Two points in the domain

eta1 = 1.5
d_eta1 = 1
eta2 = eta1 + d_eta1

# Our CGF

def CGF(eta):
    return 5*eta**2 - np.log(eta)

# Derivative of the CGF

def dotCGF(eta):
    return 10*eta - 1/eta
```

Here is our deviance function on the η scale

```
# Deviance on the natural parameter scale

def deviance(eta2, eta1=eta1):
    return 2 * (CGF(eta2) - CGF(eta1) + (eta1-eta2) * dotCGF(eta1))

deviance(eta2, eta1)
```

```
10.311682085801348
```

We can plot the deviance as the difference, at η_2 between the tangent approximation to the graph of Λ at η_1 and the true $\Lambda(\eta_2)$.

```
# Plotting points

eta = np.linspace(eta1- d_eta1,eta1+1.5 * d_eta1,101)
plt.figure(figsize=(8,8))

# Plot the CGF
```

```
plt.plot(eta, CGF(eta), label=r’$\Lambda(\eta)$’, linewidth=4)

# First order Taylor approximaton at eta1

plt.plot(eta, CGF(eta1) + (eta-eta1) * dotCGF(eta1), label=r’$\Lambda(\eta_1) + \dot{\
    Lambda}(\eta_1) (\eta-\eta_1)$’, linewidth=4)

# Difference between CGF and approximation: half the deviance

plt.plot([eta2,eta2],[CGF(eta2), CGF(eta1) + d_eta1 * dotCGF(eta1)], label=r’$D(\eta_1;\
    eta_2)/2$’, linewidth=4)

# Markers for where the two points are

plt.plot([eta1,eta1],[0,CGF(eta1)], linestyle=’--’, color=’gray’, linewidth=4)
plt.plot([eta2,eta2],[0,CGF(eta2)-deviance(eta2,eta1)/2], linestyle=’--’, color=’gray’,
    linewidth=4)

# Labelling

a = plt.gca()
a.set_xticks([eta1,eta2])
a.set_xticklabels([r’$\eta_1$’, r’$\eta_2$’], size=15)
a.set_xlim(sorted([eta1-.8*d_eta1, eta1+1.2*d_eta1]))
a.set_ylim([0,40])
a.set_xlabel(r’$\eta \in {\cal D}$’, size=20)
a.set_ylabel(r’$\Lambda(\eta)$’, size=20)

# Add a legend and title

plt.legend(loc=’upper left’)
a.set_title(r’Convexity picture on ${\cal D}$ at $(\eta_1,\eta_2)=(%0.1f,%0.1f)$’ % (
    eta1,eta2), size=20)
f = plt.gcf()
plt.close()
```

Finally, here is our rendered figure.

```
f
```

```
<matplotlib.figure.Figure at 0x109f1cd50>
```

2.3.7 *Exercise: convexity picture for Poisson family*

1. Plot the convexity picture for the Poisson family with $\eta_1$ corresponding to a mean of 10 and $\eta_2$ to a mean of 20.

2.3.8 Hoeffding's formula

Let $\widehat{\eta}(t(x)) = \dot{\Lambda}^*(t(x))$ denote the MLE of η having observed $t(x)$ as sufficient statistic. Then, using the identity

$$t(x) = \dot{\Lambda}(\widehat{\eta}(t(x))),$$

we arrive at

$$\frac{d\mathbb{P}_\eta}{d\mathbb{P}_{\widehat{\eta}}}(x) = e^{-D(\widehat{\eta}(t(x));\eta)/2}.$$

This leads to a scaled version of the likelihood having maximum value 1.

$$L(\eta) = \exp\left(-D(\widehat{\eta}(t(X));\eta)/2\right) = \exp\left(\eta \cdot t(X) - \Lambda(\eta) - \Lambda^*(t(X))\right).$$

In the $\mathcal{M}$ parametrization, this reads as

$$\tilde{L}(\mu) = L(\eta(\mu)) = \exp\left(-D(\eta(\widehat{\mu});\eta(\mu))/2\right) = \exp\left(-\tilde{D}(t(X);\mu)/2\right).$$

The other point we see here is that *maximum likelihood estimation* is the same as *minimum deviance* or *minimum KL estimation.* That is,

$$\text{argmax}_\eta \left[\eta \cdot t(X) - \Lambda(\eta)\right] = \underset{\eta}{\text{argmin}}\, D(\eta(t(X));\eta).$$

2.3.9 *Exercise: the Normal deviance*

1. When the family is the Normal family with unknown mean and known variance σ^2, show that

$$\eta(\mu) = \dot{\Lambda}^*(\mu) = \frac{\mu}{2\sigma^2}.$$

2. Conclude that

$$\tilde{D}(t(X);\mu) = \frac{(t(X)-\mu)^2}{\sigma^2}$$

2.3.10 *Exercise: familiar deviances*

In both the $\mathcal{M}$ and $\mathcal{D}$ parameterization, compute the deviances for the following families:

1. Poisson with mean parameter $\mu \in \mathcal{M}$.

2. Binomial with n trials having probability of success $n \cdot \pi \in \mathcal{M}$ (n is fixed).

3. Gamma with fixed shape parameter k and mean $k \cdot \mu \in \mathcal{M}$.

2.3.11 Deviance under IID sampling

Suppose we observe n IID samples from a one-parameter exponential family $\mathcal{F}$. Then, the total deviance is

$$D^{(n)}(\eta_1;\eta_2) = 2nD(\eta_1;\eta_2).$$

2.3.12 Deviance and Fisher information

So, as with Fisher information, the deviance grows with n. The first and second order versions of Taylor's theorem imply

$$\begin{aligned}
\Lambda(\eta_2) &= \Lambda(\eta_1) + \dot{\Lambda}(\eta_1) \cdot (\eta_1 - \eta_2) - \frac{1}{2} D(\eta_1; \eta_2) \\
&= \Lambda(\eta_1) + \dot{\Lambda}(\eta_1) \cdot (\eta_1 - \eta_2) + \frac{1}{2} \ddot{\Lambda}(\eta_1 - \eta_2)^2 + \Lambda^{(3)}(\theta) \frac{(\eta_1 - \eta_2)^3}{6}
\end{aligned}$$

for some $\theta \in [\eta_1, \eta_2]$. We see, then, that

$$\left| D(\eta_1; \eta_2) - I_{\eta_1}(\eta_1 - \eta_2)^2 \right| \leq C(\eta_1; |\eta_1 - \eta_2|) \cdot \frac{(\eta_1 - \eta_2)^3}{6}$$

where

$$C(\eta; r) = \sup_{\theta : |\theta - \eta| \leq r} \frac{|\ddot{\Lambda}(\theta) - \ddot{\Lambda}(\eta)|}{|\theta - \eta|} = \sup_{\theta : |\theta - \eta| \leq r} \frac{|I_\theta - I_\eta|}{|\theta - \eta|}$$

is a local Lipschitz constant or modulus of continuity for I at η.

2.3.13 *Exercise: dual version of convexity picture*

In this exercise, we parameterize the deviance by μ instead of η.

1. Show that
$$\Lambda^*(\mu) = \dot{\Lambda}^*(\mu) \cdot \mu - \Lambda(\dot{\Lambda}^*(\mu)).$$

2. Verify Λ^* in the code below.

3. Use this to verify the convexity picture below for
$$\tilde{D}(\mu_1; \mu_2) \triangleq D(\eta(\mu_1); \eta(\mu_2))$$
as above in the μ parametrization.

4. From the picture below, give an explicit formula for $\tilde{D}(\mu_1; \mu_2)$.

Based on our previous exponential family, here are the conjugate and its derivative. Note that we only need to compute the MLE map to compute Λ^*.

```
# Based on the previous family, the MLE map

def dotCGFstar(mu):
    return (mu + np.sqrt(mu**2+40)) / 20.

# A general formula for CGF^*

def CGFstar(mu):
    return dotCGFstar(mu)*mu-CGF(dotCGFstar(mu))
```

```
# Find the two corresponding points from natural parameter scale

mu1 = dotCGF(eta1)
mu2 = dotCGF(eta2)
d_mu2 = mu1 - mu2

mu = np.linspace(mu2- d_mu2,mu2+1.5 * d_mu2,101)
plt.figure(figsize=(8,8))

# Plot CGF^*

plt.plot(mu, CGFstar(mu), label=r’$\Lambda^*(\mu)$’, linewidth=4)

# The first order Taylor approximation

plt.plot(mu, CGFstar(mu2) + (mu-mu2) * dotCGFstar(mu2), label=r’$\Lambda^*(\mu_2) + \dot
    {\Lambda}^*(\mu_2) (\mu-\mu_2)$’, linewidth=4)

# The difference between CGF^* and the Taylor approximation at mu1

plt.plot([mu1,mu1],[CGFstar(mu1), CGFstar(mu2) + d_mu2 * dotCGFstar(mu2)], label=r’$\
    tilde{D}(\mu_1;\mu_2)/2$’, linewidth=4)

# Mark where the points are

plt.plot([mu1,mu1],[0,CGFstar(mu1)-deviance(eta1,eta2)/2], linestyle=’--’, color=’gray’,
     linewidth=4)
plt.plot([mu2,mu2],[0,CGFstar(mu2)], linestyle=’--’, color=’gray’, linewidth=4)

# Labelling

a = plt.gca()
a.set_xticks([mu1,mu2])
a.set_xticklabels([r’$\mu_1$’, r’$\mu_2$’], size=15)
a.set_xlim(sorted([mu2-1.2*d_mu2, mu2+1.2*d_mu2]))
a.set_xlabel(r’$\mu \in {\cal M}$’, size=20)
a.set_ylabel(r’$\Lambda^*(\mu)$’, size=20)

plt.legend(loc=’upper left’)

# Add a legend and title

a.set_title(r"""Convexity picture on ${\cal M}$ at
$(\mu_1,\mu_2)=(\dot{\Lambda}^*(%0.1f), \dot{\Lambda}^*(%0.1f)\approx(%0.2f,%0.2f))$"""
    % (eta1,eta2,mu1,mu2), size=20)

f = plt.gcf()
plt.close()
```

Here is our rendered figure.

```
f
```

```
<matplotlib.figure.Figure at 0x10fcafa10>
```

2.4 Deviance residuals

For the normal family, you showed in your homework that, in the mean parameter, the deviance has the form

$$\tilde{D}(\hat{\mu};\mu) = \frac{(\hat{\mu}-\mu)^2}{\sigma^2}.$$

Hence, it is like a normalized residual squared. To recover the original residual, one would compute

$$r(\hat{\mu};\mu) = \sqrt{\tilde{D}(\hat{\mu};\mu)} \cdot \text{sign}(\hat{\mu}-\mu).$$

This is the general form of a *deviance residual*.

In the repeated sampling setting, we observe a *sample* of deviance residuals. Namely,

$$r_i = r(\overline{t(X)}; t(X_i)).$$

There is also a deviance residual for $\overline{t(X)}$ with respect to μ

$$\begin{aligned} R_D &= \text{sign}(\overline{t(X)}-\mu) \cdot \sqrt{\tilde{D}_n(\overline{t(X)};\mu)} \\ &= \text{sign}(\overline{t(X)}-\mu) \cdot \sqrt{n \cdot \tilde{D}(\overline{t(X)};\mu)} \end{aligned}$$

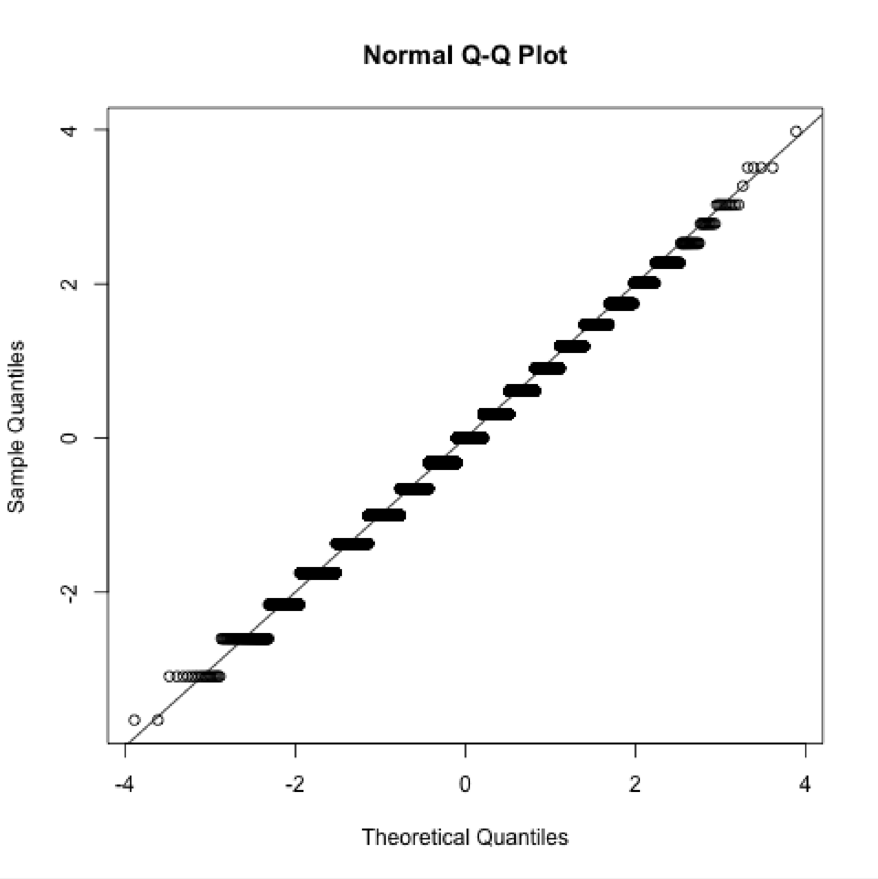
We might compare this with the *Pearson residual*

$$R_P = \frac{\overline{t(X)}-\mu}{\sqrt{\text{Var}_\eta(t(X))/n}}.$$

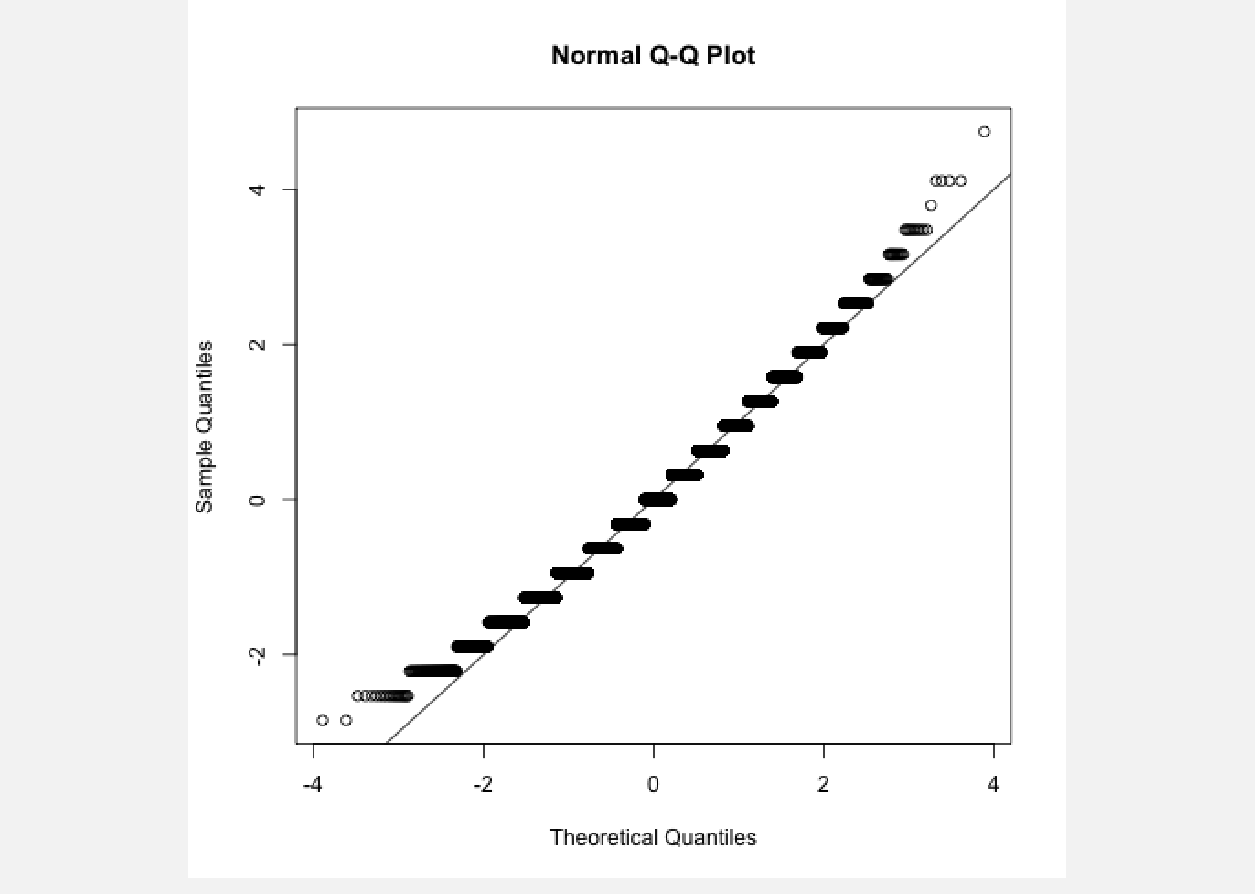
Below are comparisons of the two residuals for a sample of size 10 from Poisson(1). Generally speaking, the deviance residuals are closer to normal than Pearson residuals.

```
%%R
mu = 10
nsim = 10000
Y = rpois(nsim, mu)

dev.resid = sign(Y-mu)*sqrt(2*(Y*log(Y/mu)-(Y-mu)))
qqnorm(dev.resid)
abline(0,1)
```



```
%%R
pearson.resid = (Y-mu) / sqrt(mu)
qqnorm(pearson.resid)
abline(0,1)
```



Of course, for Poisson the residuals can't be exactly normal: there is some lattice effect. We do see that the deviance residuals are closer to the diagonal than the Pearson residuals in the tails, though in the center, the approximation looks reasonable.

2.4.1 Bias and variance correction of deviance residuals

(From Appendix C, McCullagh and Nelder)

The deviance residual of $\overline{t(X)}$ for μ is asymptotically

$$N(B_n, V_n)$$

with

$$B_n = -\frac{\rho_3}{6}$$
$$V_n = \left(1 + \frac{7}{36}\rho_3^2 - \frac{\rho_4}{8}\right)^2$$

where

$$\rho_3 = \text{Skew}(\overline{t(X)}, \mathbb{P}_\eta)$$
$$\rho_4 = \text{Kurtosis}(\overline{t(X)}, \mathbb{P}_\eta)$$

Hence, families whose sufficient statistics are highly skewed will have poorer approximations by the normal distribution.

The corresponding result for the deviance itself is

$$D_n(\overline{t(X)};\mu) \approx \left(1 + \frac{5\rho_3^2 - 3\rho_4}{12}\right) \cdot \chi_1^2.$$

These results extend to multiparameter versions as well.

Here is a brief sketch of how you might begin to prove some of these asymptotic bias formulae for D_n. The convexity picture tells us that the deviance is a residual in a Taylor series expansion. This residual has an exact form in terms of an infinite Taylor series

$$\frac{1}{2} D_n(\overline{t(X)};\mu) = \frac{1}{2}\ddot{\Lambda}_n{}^*(\mu)(\overline{t(X)}-\mu)^2 + \sum_{j=3}^{\infty} \frac{(\Lambda_n^*)^{(j)}}{j!}(\overline{t(X)}-\mu)^j$$

where Λ_n^* is the Fenchel-Legendre transform of the CGF of $\overline{t(X)}$.

The first term is recognized as a random variable with mean 0 and variance 1. The remaining terms can be analyzed to approximate the bias and variance for a fixed n.

2.4.2 *Exercise*

In this exercise, we use the Gamma exponential family with fixed shape parameter set to 1. That is, $t(x) = x$ and

$$\mathbb{P}_\eta(dx) = e^{\eta \cdot x - \Lambda(\eta)}\, dx \qquad x \geq 0.$$

1. Compute the skewness and kurtosis of $\overline{t(X)} = \bar{X}_n$ for an IID sample of size n.

2. Set $\eta = -1$ and simulate the deviance residual for $\bar{X}_n$ with respect to η. Compare this distribution to $N(B_n, V_n)$ for various values of n.

3. Try to choose the sample size n, so that B_n is approximately $1/100$.

2.5 References

- Efron, B. Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency)

- Efron, B. The geometry of exponential families

- McCullagh, P. and Nelder, J., Appendix C Generalized Linear Models