Chapter 2
Exponential Dispersion Family

We introduce the exponential family (EF) and the exponential dispersion family (EDF) in this chapter. The single-parameter EF has been introduced in 1934 by the British statistician Sir Fisher [128], and it has been extended to vector-valued parameters by Darmois [88], Koopman [223] and Pitman [306] between 1935 and 1936. It is the most commonly used family of distribution functions in statistical modeling; among others, it contains the Gaussian distribution, the gamma distribution, the binomial distribution and the Poisson distribution. Its parametrization is taken in a special form that is convenient for statistical modeling. The EF can be introduced in a constructive way providing the main properties of this family of distribution functions. In this chapter we follow Jørgensen [201–203] and Barndorff-Nielsen [23], and we state the most important results based on this constructive introduction. This gives us a unified notation which is going to be useful for our purposes.

2.1 Exponential Family

2.1.1 *Definition and Properties*

We define the EF w.r.t. a σ-finite measure ν on $\mathbb{R}$. The results in this section can be generalized to σ-finite measures on $\mathbb{R}^m$, but such an extension is not necessary for our purposes. Select an integer $k \in \mathbb{N}$, and choose measurable functions $a : \mathbb{R} \to \mathbb{R}$ and $T : \mathbb{R} \to \mathbb{R}^k$.[1] Consider for a *canonical parameter* $\boldsymbol{\theta} \in \mathbb{R}^k$ the Laplace

[1] We could also use boldface notation for T because $T(y) \in \mathbb{R}^k$ is vector-valued, but we prefer to not use boldface notation for (vector-valued) functions.

M. V. Wüthrich, M. Merz, *Statistical Foundations of Actuarial Learning and its Applications*, Springer Actuarial, https://doi.org/10.1007/978-3-031-12409-9_2

transform

$$\mathfrak{L}(\boldsymbol{\theta}) = \int_{\mathbb{R}} \exp\left\{\boldsymbol{\theta}^\top T(y) + a(y)\right\} d\nu(y).$$

Assume that this Laplace transform is not identically equal to $+\infty$. The *effective domain* is defined by

$$\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} \in \mathbb{R}^k;\ \mathfrak{L}(\boldsymbol{\theta}) < \infty\right\} \ \subseteq\ \mathbb{R}^k. \tag{2.1}$$

Lemma 2.1 *The effective domain $\boldsymbol{\Theta} \subseteq \mathbb{R}^k$ is a convex set.*

The effective domain $\boldsymbol{\Theta}$ is not necessarily an open set, but in many applications it is open. Counterexamples are given in Problem 4.1 of Chapter 1 in Lehmann [244], and in the inverse Gaussian example in Sect. 2.1.3, below.

Proof of Lemma 2.1 Choose $\boldsymbol{\theta}_i \in \mathbb{R}^k$, $i = 1, 2$, with $\mathfrak{L}(\boldsymbol{\theta}_i) < \infty$. Set $\boldsymbol{\theta} = c\boldsymbol{\theta}_1 + (1-c)\boldsymbol{\theta}_2$ for $c \in (0, 1)$. We use Hölder's inequality, applied to the norms $p = 1/c$ and $q = 1/(1-c)$,

$$\begin{aligned}
\mathfrak{L}(\boldsymbol{\theta}) &= \int_{\mathbb{R}} \exp\left\{(c\boldsymbol{\theta}_1 + (1-c)\boldsymbol{\theta}_2)^\top T(y) + a(y)\right\} d\nu(y) \\
&= \int_{\mathbb{R}} \exp\left\{\boldsymbol{\theta}_1^\top T(y) + a(y)\right\}^c \exp\left\{\boldsymbol{\theta}_2^\top T(y) + a(y)\right\}^{1-c} d\nu(y) \\
&\le \mathfrak{L}(\boldsymbol{\theta}_1)^c \mathfrak{L}(\boldsymbol{\theta}_2)^{1-c} < \infty.
\end{aligned}$$

This implies $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and proves the claim. □

We define *the cumulant function* on the effective domain $\boldsymbol{\Theta}$

$$\kappa : \boldsymbol{\Theta} \to \mathbb{R}, \qquad \boldsymbol{\theta} \mapsto \kappa(\boldsymbol{\theta}) = \log\mathfrak{L}(\boldsymbol{\theta}).$$

Definition 2.2 The EF with σ-finite measure ν on $\mathbb{R}$ and cumulant function $\kappa : \boldsymbol{\Theta} \to \mathbb{R}$ is given by the distribution functions F on $\mathbb{R}$ with

$$dF(y;\boldsymbol{\theta}) \ = \ f(y;\boldsymbol{\theta})d\nu(y) \ = \ \exp\left\{\boldsymbol{\theta}^\top T(y) - \kappa(\boldsymbol{\theta}) + a(y)\right\} d\nu(y), \tag{2.2}$$

for canonical parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$.

Remarks 2.3

- The definition of the EF (2.2) assumes that the effective domain $\boldsymbol{\Theta} \subseteq \mathbb{R}^k$ has been constructed from the choices $a : \mathbb{R} \to \mathbb{R}$ and $T : \mathbb{R} \to \mathbb{R}^k$ as described in (2.1). This is not explicitly stated in the surrounding text of (2.2).
- The support of any random variable $Y \sim F(\cdot; \boldsymbol{\theta})$ of this EF does *not* depend on the explicit choice of the canonical parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, but solely on the choice of the σ-finite measure ν on $\mathbb{R}$, and the distribution functions $F(\cdot; \boldsymbol{\theta})$ are mutually absolutely continuous (equivalent) w.r.t. ν.
- In statistics, the main object of interest is the canonical parameter $\boldsymbol{\theta}$. Importantly for parameter estimation, the function $a(\cdot)$ does not involve the canonical parameter. Therefore, it is irrelevant for parameter estimation and (only) serves as a normalization so that F in (2.2) is a proper distribution function. In fact, this is the way how the EF is often introduced in the statistical and actuarial literature, but in this latter introduction we lose the deeper interpretation of the cumulant function κ, nor is it immediately clear what properties it possesses.
- The case $k \geq 2$ gives a *vector-valued* canonical parameter $\boldsymbol{\theta}$. The case $k = 1$ gives a *single-parameter* EF, and, if additionally $T(y) = y$, it is called a *single-parameter linear EF*.

Theorem 2.4 *Assume the effective domain* $\boldsymbol{\Theta}$ *has a non-empty interior* $\mathring{\boldsymbol{\Theta}}$. *Choose* $Y \sim F(\cdot; \boldsymbol{\theta})$ *for fixed* $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$. *The moment generating function of* $T(Y)$ *for sufficiently small* $\boldsymbol{r} \in \mathbb{R}^k$ *is given by*

$$M_{T(Y)}(\boldsymbol{r}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\exp\left\{\boldsymbol{r}^\top T(Y)\right\}\right] = \exp\left\{\kappa(\boldsymbol{\theta}+\boldsymbol{r}) - \kappa(\boldsymbol{\theta})\right\},$$

where the expectation operator $\mathbb{E}_{\boldsymbol{\theta}}$ *illustrates the selected canonical parameter* $\boldsymbol{\theta}$ *for* Y.

Proof Choose $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$ and $\boldsymbol{r} \in \mathbb{R}^k$ so small that $\boldsymbol{\theta} + \boldsymbol{r} \in \mathring{\boldsymbol{\Theta}}$. We receive

$$\begin{aligned} M_{T(Y)}(\boldsymbol{r}) &= \int_{\mathbb{R}} \exp\left\{(\boldsymbol{\theta}+\boldsymbol{r})^\top T(y) - \kappa(\boldsymbol{\theta}) + a(y)\right\} d\nu(y) \\ &= \exp\left\{\kappa(\boldsymbol{\theta}+\boldsymbol{r}) - \kappa(\boldsymbol{\theta})\right\} \int_{\mathbb{R}} \exp\left\{(\boldsymbol{\theta}+\boldsymbol{r})^\top T(y) - \kappa(\boldsymbol{\theta}+\boldsymbol{r}) + a(y)\right\} d\nu(y) \\ &= \exp\left\{\kappa(\boldsymbol{\theta}+\boldsymbol{r}) - \kappa(\boldsymbol{\theta})\right\}, \end{aligned}$$

where the last identity follows from the fact that the support of the EF does not depend on the explicit choice of the canonical parameter. □

Theorem 2.4 has a couple of immediate implications. First, in any interior point $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$ both the moment generating function $\boldsymbol{r} \mapsto M_{T(Y)}(\boldsymbol{r})$ (in the neighborhood of the origin) and the cumulant function $\boldsymbol{\theta} \mapsto \kappa(\boldsymbol{\theta})$ have derivatives of all orders, and, similarly to Sect. 1.2, moments of all orders of $T(Y)$ exist, see also (1.1). Existence of moments of all orders implies that the distribution function of $T(Y)$ cannot have a regularly varying tails.

Corollary 2.5 *Assume* $\mathring{\boldsymbol{\Theta}}$ *is non-empty. The cumulant function* $\boldsymbol{\theta} \mapsto \kappa(\boldsymbol{\theta})$ *is convex, and for* $Y \sim F(\cdot; \boldsymbol{\theta})$ *with* $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$

$$\mu = \mathbb{E}_{\boldsymbol{\theta}}\left[T(Y)\right] = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) \qquad \textit{and} \qquad \textit{Var}_{\boldsymbol{\theta}}\left(T(Y)\right) = \nabla^2_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}),$$

where $\nabla_{\boldsymbol{\theta}}$ *is the gradient and* $\nabla^2_{\boldsymbol{\theta}}$ *the Hessian w.r.t. vector* $\boldsymbol{\theta}$*.*

Similarly to $T : \mathbb{R} \to \mathbb{R}^k$, we will not use boldface notation for the (multi-dimensional) mean because later on we will understand the mean $\mu = \mu(\boldsymbol{\theta}) \in \mathbb{R}^k$ as a function of the canonical parameter $\boldsymbol{\theta}$; see Footnote 1 on page 13 on boldface notation.

Proof Existence of the moment generating function for all sufficiently small $\boldsymbol{r} \in \mathbb{R}^k$ (around the origin) implies that we have first and second moments. For the first moment we receive

$$\mu = \mathbb{E}_{\boldsymbol{\theta}}\left[T(Y)\right] = \left.\nabla_{\boldsymbol{r}} M_{T(Y)}(\boldsymbol{r})\right|_{\boldsymbol{r}=0} = \exp\left\{\kappa(\boldsymbol{\theta}+\boldsymbol{r}) - \kappa(\boldsymbol{\theta})\right\} \left.\nabla_{\boldsymbol{r}}\kappa(\boldsymbol{\theta}+\boldsymbol{r})\right|_{\boldsymbol{r}=0} = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}).$$

Denote component j of $T(Y) \in \mathbb{R}^k$ by $T_j(Y)$. We have for $1 \le j, l \le k$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}\left[T_j(Y)T_l(Y)\right] &= \left.\frac{\partial^2}{\partial r_j \partial r_l} M_{T(Y)}(\boldsymbol{r})\right|_{\boldsymbol{r}=0} \\
&= \exp\left\{\kappa(\boldsymbol{\theta}+\boldsymbol{r}) - \kappa(\boldsymbol{\theta})\right\} \left.\left(\frac{\partial^2}{\partial r_j \partial r_l}\kappa(\boldsymbol{\theta}+\boldsymbol{r}) + \frac{\partial}{\partial r_j}\kappa(\boldsymbol{\theta}+\boldsymbol{r})\frac{\partial}{\partial r_l}\kappa(\boldsymbol{\theta}+\boldsymbol{r})\right)\right|_{\boldsymbol{r}=0} \\
&= \left(\frac{\partial^2}{\partial \theta_j \partial \theta_l}\kappa(\boldsymbol{\theta}) + \frac{\partial}{\partial \theta_j}\kappa(\boldsymbol{\theta})\frac{\partial}{\partial \theta_l}\kappa(\boldsymbol{\theta})\right).
\end{aligned}$$

This implies for the covariance

$$\mathrm{Cov}_{\boldsymbol{\theta}}(T_j(Y), T_l(Y)) = \frac{\partial^2}{\partial \theta_j \partial \theta_l}\kappa(\boldsymbol{\theta}).$$

The convexity of κ follows because $\nabla^2_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta})$ is the positive semi-definite covariance matrix of $T(Y)$, for all $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$. This finishes the proof. □

Assumption 2.6 (Minimal Representation) *We assume that the interior* $\mathring{\boldsymbol{\Theta}}$ *of the effective domain* $\boldsymbol{\Theta}$ *is non-empty and that the cumulant function* κ *is strictly convex on this interior* $\mathring{\boldsymbol{\Theta}}$*.*

Remarks 2.7

- Throughout these notes we will work under Assumption 2.6 without making explicit reference. This assumption strengthens the properties of the cumulant function κ from being convex, see Corollary 2.5, to being strictly convex. This strengthening implies that the mean function $\boldsymbol{\theta} \mapsto \mu = \mu(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \kappa(\boldsymbol{\theta})$ can be inverted; this is needed for the canonical link, see Definition 2.8, below.
- The strict convexity of κ means that the covariance matrix $\nabla^2_{\boldsymbol{\theta}} \kappa(\boldsymbol{\theta})$ of $T(Y)$ is positive definite and has full rank k for all $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$, see Corollary 2.5. This property is important, otherwise we do not have identifiability in the canonical parameter $\boldsymbol{\theta}$ because we have a linear dependence between the components of $T(Y)$.
- Mathematically, this strict convexity is not a restriction because it can be obtained by working under a so-called *minimal representation*. If the covariance matrix $\nabla^2_{\boldsymbol{\theta}} \kappa(\boldsymbol{\theta})$ does not have full rank k, the choice k is "non-optimal" because the problem lives in a smaller dimension. Thus, w.l.o.g., we may and will assume to work in this smaller dimension, called minimal representation; for a rigorous derivation of a minimal representation we refer to Section 8.1 in Barndorff-Nielsen [23].

Definition 2.8 The canonical link is defined by $h = (\nabla_{\boldsymbol{\theta}} \kappa)^{-1}$.

The application of the canonical link h to the mean implies under Assumption 2.6

$$h(\mu) = h\left(\mathbb{E}_{\boldsymbol{\theta}}\left[T(Y)\right]\right) = \boldsymbol{\theta},$$

for mean $\mu = \mathbb{E}_{\boldsymbol{\theta}}[T(Y)]$ of $Y \sim F(\cdot; \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$.

Remarks 2.9 (Dual Parameter Space) Assumption 2.6 provides that the canonical link h is well-defined, and we can either work with the canonical parameter representation $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}} \subseteq \mathbb{R}^k$ or with its dual (mean) parameter representation $\mu = \mathbb{E}_{\boldsymbol{\theta}}\left[T(Y)\right] \in \mathcal{M}$ with

$$\mathcal{M} \overset{\text{def.}}{=} \nabla_{\boldsymbol{\theta}} \kappa(\mathring{\boldsymbol{\Theta}}) = \{\nabla_{\boldsymbol{\theta}} \kappa(\boldsymbol{\theta});\ \boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}\} \subseteq \mathbb{R}^k. \tag{2.3}$$

Strict convexity of κ implies that there is a one-to-one correspondence between these two parametrizations. $\boldsymbol{\Theta}$ is called the *effective domain* and $\mathcal{M}$ is called the *dual parameter space* or the *mean parameter space*.

In Sect. 2.2.4, below, we introduce one more property called *steepness* that the cumulant function κ should satisfy. This additional property gives a relationship between the support $\mathfrak{T}$ of the random variables $T(Y)$ of the given EF and the boundary of the dual parameter space $\mathcal{M}$. This steepness property is important for parameter estimation.

2.1.2 Single-Parameter Linear EF: Count Variable Examples

We start by giving single-parameter discrete linear EF examples based on counting measures on $\mathbb{N}_0$. Since we work in one dimension $k = 1$, we replace boldface $\boldsymbol{\theta}$ by scalar $\theta \in \boldsymbol{\Theta} \subseteq \mathbb{R}$ in this section.

Bernoulli Distribution as a Single-Parameter Linear EF

For the Bernoulli distribution with parameter $p \in (0, 1)$ we choose as ν the counting measure on $\{0, 1\}$. We make the following choices: $T(y) = y$,

$$a(y) = 0, \quad \kappa(\theta) = \log(1 + e^{\theta}), \quad p = \kappa'(\theta) = \frac{e^{\theta}}{1 + e^{\theta}}, \quad \theta = h(p) = \log\left(\frac{p}{1 - p}\right),$$

for effective domain $\boldsymbol{\Theta} = \mathbb{R}$, dual parameter space $\mathcal{M} = (0, 1)$ and support $\mathfrak{T} = \{0, 1\}$ of $Y = T(Y)$. With these choices we have

$$dF(y; \theta) = \exp\left\{\theta y - \log(1 + e^{\theta})\right\} d\nu(y) = \left(\frac{e^{\theta}}{1 + e^{\theta}}\right)^{y} \left(\frac{1}{1 + e^{\theta}}\right)^{1-y} d\nu(y).$$

$\theta \mapsto \kappa'(\theta)$ is the logistic or sigmoid function, and the canonical link $p \mapsto h(p)$ is the logit function. Mean and variance are given by

$$\mu = \mathbb{E}_{\theta}\left[Y\right] = \kappa'(\theta) = p \quad \text{and} \quad \mathrm{Var}_{\theta}\left(Y\right) = \kappa''(\theta) = \frac{e^{\theta}}{(1 + e^{\theta})^2} = p(1 - p),$$

and the probability weights satisfy for $y \in \mathfrak{T} = \{0, 1\}$

$$\mathbb{P}_{\theta}[Y = y] = p^{y}(1 - p)^{1-y}.$$

Binomial Distribution as a Single-Parameter Linear EF

For the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in (0, 1)$ we choose as ν the counting measure on $\{0, \ldots, n\}$. We make the following choices: $T(y) = y$,

$$a(y) = \log\binom{n}{y}, \quad \kappa(\theta) = n\log(1{+}e^{\theta}), \quad \mu = \kappa'(\theta) = \frac{ne^{\theta}}{1 + e^{\theta}}, \quad \theta = h(\mu) = \log\left(\frac{\mu}{n - \mu}\right),$$

for effective domain $\boldsymbol{\Theta} = \mathbb{R}$, dual parameter space $\mathcal{M} = (0, n)$ and support $\mathfrak{T} = \{0, \ldots, n\}$ of $Y = T(Y)$. With these choices we have

$$dF(y; \theta) = \binom{n}{y} \exp\left\{\theta y - n\log(1 + e^{\theta})\right\} d\nu(y) = \binom{n}{y} \left(\frac{e^{\theta}}{1 + e^{\theta}}\right)^{y} \left(\frac{1}{1 + e^{\theta}}\right)^{n-y} d\nu(y).$$

Mean and variance are given by

$$\mu = \mathbb{E}_\theta[Y] = \kappa'(\theta) = np \qquad \text{and} \qquad \mathrm{Var}_\theta(Y) = \kappa''(\theta) = n\frac{e^\theta}{(1+e^\theta)^2} = np(1-p),$$

where we set $p = e^\theta/(1+e^\theta)$. The probability weights satisfy for $y \in \mathfrak{T} = \{0, \dots, n\}$

$$\mathbb{P}_\theta[Y = y] = \binom{n}{y} p^y (1-p)^{n-y}.$$

Poisson Distribution as a Single-Parameter Linear EF

For the Poisson distribution with parameter $\lambda > 0$ we choose as ν the counting measure on $\mathbb{N}_0$. We make the following choices: $T(y) = y$,

$$a(y) = \log\left(\frac{1}{y!}\right), \quad \kappa(\theta) = e^\theta, \quad \mu = \kappa'(\theta) = e^\theta, \quad \theta = h(\mu) = \log(\mu),$$

for effective domain $\boldsymbol{\Theta} = \mathbb{R}$, dual parameter space $\mathcal{M} = (0, \infty)$ and support $\mathfrak{T} = \mathbb{N}_0$ of $Y = T(Y)$. With these choices we have

$$dF(y;\theta) = \frac{1}{y!}\exp\left\{\theta y - e^\theta\right\} d\nu(y) = e^{-\mu}\frac{\mu^y}{y!} d\nu(y). \tag{2.4}$$

The canonical link $\mu \mapsto h(\mu)$ is the log-link. Mean and variance are given by

$$\mu = \mathbb{E}_\theta[Y] = \kappa'(\theta) = \lambda \qquad \text{and} \qquad \mathrm{Var}_\theta(Y) = \kappa''(\theta) = \lambda = \mu = \mathbb{E}_\theta[Y],$$

where we set $\lambda = e^\theta$. The probability weights in the Poisson case satisfy for $y \in \mathfrak{T} = \mathbb{N}_0$

$$\mathbb{P}_\theta[Y = y] = e^{-\lambda}\frac{\lambda^y}{y!}.$$

Negative-Binomial (Pólya) Distribution as a Single-Parameter Linear EF

For the negative-binomial distribution with $\alpha > 0$ and $p \in (0, 1)$ we choose as ν the counting measure on $\mathbb{N}_0$; α plays the role of a nuisance parameter or hyper-parameter. We make the following choices: $T(y) = y$,

$$a(y) = \log\binom{y+\alpha-1}{y}, \quad \kappa(\theta) = -\alpha\log(1-e^\theta),$$

$$\mu = \kappa'(\theta) = \alpha \frac{e^\theta}{1-e^\theta}, \quad \theta = h(\mu) = \log\left(\frac{\mu}{\mu+\alpha}\right),$$

for effective domain $\boldsymbol{\Theta} = (-\infty, 0)$, dual parameter space $\mathcal{M} = (0, \infty)$ and support $\mathfrak{T} = \mathbb{N}_0$ of $Y = T(Y)$. With these choices we have

$$\begin{aligned} dF(y;\theta) &= \binom{y+\alpha-1}{y} \exp\left\{\theta y + \alpha \log(1-e^\theta)\right\} d\nu(y) \\ &= \binom{y+\alpha-1}{y} p^y \,(1-p)^\alpha \, d\nu(y), \end{aligned}$$

with $p = e^\theta$. Parameter $\alpha > 0$ is treated as nuisance parameter, otherwise we drop out of the EF framework. We have first the two moments

$$\mu = \mathbb{E}_\theta[Y] = \alpha \frac{e^\theta}{1-e^\theta} = \alpha \frac{p}{1-p} \qquad \text{and} \qquad \text{Var}_\theta(Y) = \mathbb{E}_\theta[Y]\left(1 + \frac{e^\theta}{1-e^\theta}\right) > \mathbb{E}_\theta[Y].$$

This model allows us to model over-dispersion, in contrast to the Poisson model. In fact, the negative-binomial model is a mixed Poisson model with a gamma mixing distribution, for details see Sect. 5.3.5, below. Typically, one uses a different parametrization. Set $e^\theta = \lambda/(\alpha+\lambda)$, for $\lambda > 0$. This implies

$$\mu = \mathbb{E}_\theta[Y] = \lambda \qquad \text{and} \qquad \text{Var}_\theta(Y) = \lambda\left(1 + \frac{\lambda}{\alpha}\right) > \lambda.$$

For $\alpha \in \mathbb{N}$ this model can also be interpreted as the waiting time until we observe α successful trials among i.i.d. trials, for instance, for $\alpha = 1$ we have the geometric distribution (with a small reparametrization).

The probability weights of the negative-binomial model satisfy for $y \in \mathfrak{T} = \mathbb{N}_0$

$$\mathbb{P}_\theta[Y = y] = \binom{y+\alpha-1}{y} p^y \,(1-p)^\alpha \,. \tag{2.5}$$

2.1.3 Vector-Valued Parameter EF: Absolutely Continuous Examples

We give vector-valued parameter absolutely continuous EF examples with $k = 2$, and being based on the Lebesgue measure on (subsets of) $\mathbb{R}$, in this section.

Gaussian Distribution as a Vector-Valued Parameter EF

For the Gaussian distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ we choose as ν the Lebesgue measure on $\mathbb{R}$, and we make the following choices: $T(y) = (y, y^2)^\top$,

$$a(y) = -\frac{1}{2}\log(2\pi), \qquad \kappa(\boldsymbol{\theta}) \;=\; -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2),$$

$$(\mu, \sigma^2 + \mu^2)^\top = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) \;=\; \left(\frac{\theta_1}{-2\theta_2}, (-2\theta_2)^{-1} + \frac{\theta_1^2}{4\theta_2^2}\right)^\top,$$

for effective domain $\boldsymbol{\Theta} = \mathbb{R} \times (-\infty, 0)$, dual parameter space $\mathcal{M} = \mathbb{R} \times (0, \infty)$ and support $\mathfrak{T} = \mathbb{R} \times [0, \infty)$ of $T(Y) = (Y, Y^2)^\top$. With these choices we have

$$\begin{aligned} dF(y; \boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi}} \exp\left\{\boldsymbol{\theta}^\top T(y) + \frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log(-2\theta_2)\right\} d\nu(y) \\ &= \frac{1}{\sqrt{2\pi}(-2\theta_2)^{-1/2}} \exp\left\{-\frac{1}{2}\frac{1}{(-2\theta_2)^{-1}}\left(y - \frac{\theta_1}{-2\theta_2}\right)^2\right\} d\nu(y). \end{aligned}$$

This is the Gaussian model with mean $\mu = \theta_1/(-2\theta_2)$ and variance $\sigma^2 = (-2\theta_2)^{-1}$.

If we treat $\sigma > 0$ as a nuisance parameter, we obtain the Gaussian model as a single-parameter EF. This is the most common example of an EF. Set $T(y) = y/\sigma$ and

$$a(y) = -\frac{1}{2}\log(2\pi\sigma^2) - y^2/(2\sigma^2), \quad \kappa(\theta) = \theta^2/2, \quad \mu = \kappa'(\theta) = \theta, \quad \theta = h(\mu) = \mu,$$

for effective domain $\boldsymbol{\Theta} = \mathbb{R}$, dual parameter space $\mathcal{M} = \mathbb{R}$ and support $\mathfrak{T} = \mathbb{R}$ of $T(Y) = Y/\sigma$. With these choices we have

$$\begin{aligned} dF(y; \theta) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\theta y/\sigma - y^2/(2\sigma^2) - \theta^2/2\right\} d\nu(y) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y - \sigma\theta)^2\right\} d\nu(y), \end{aligned}$$

and, in particular, the canonical link is the *identity link* $\mu \mapsto \theta = h(\mu) = \mu$ in this single-parameter EF example.

Gamma Distribution as a Vector-Valued Parameter EF

For the gamma distribution with parameters $\alpha, \beta > 0$ we choose as ν the Lebesgue measure on $\mathbb{R}_+$. Then we make the following choices: $T(y) = (y, \log y)^\top$,

$$a(y) = -\log y, \qquad \kappa(\boldsymbol{\theta}) \ = \ \log\Gamma(\theta_2) - \theta_2\log(-\theta_1),$$

$$\left(\alpha/\beta, \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log(\beta)\right)^\top = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) \ = \ \left(\frac{\theta_2}{-\theta_1}, \frac{\Gamma'(\theta_2)}{\Gamma(\theta_2)} - \log(-\theta_1)\right)^\top,$$

for effective domain $\boldsymbol{\Theta} = (-\infty, 0) \times (0, \infty)$, and setting $\beta = -\theta_1 > 0$ and $\alpha = \theta_2 > 0$. The dual parameter space is $\mathcal{M} = (0, \infty) \times \mathbb{R}$, and we have support $\mathfrak{T} = (0, \infty) \times \mathbb{R}$ of $T(Y) = (Y, \log Y)^\top$. With these choices we obtain

$$\begin{aligned} dF(y; \boldsymbol{\theta}) &= \exp\left\{\boldsymbol{\theta}^\top T(y) - \log\Gamma(\theta_2) + \theta_2\log(-\theta_1) - \log y\right\} d\nu(y) \\ &= \frac{(-\theta_1)^{\theta_2}}{\Gamma(\theta_2)} y^{\theta_2 - 1} \exp\{-(-\theta_1)y\}\, d\nu(y) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-\beta y\}\, d\nu(y). \end{aligned}$$

This is a vector-valued parameter EF with $k = 2$, and the first moment is given by

$$\mathbb{E}_{\boldsymbol{\theta}}\left[(Y, \log Y)^\top\right] = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) = \left(\alpha/\beta, \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log(\beta)\right)^\top.$$

Parameter α is called *shape parameter* and parameter β is called *scale parameter*.[2]

If we treat the shape parameter $\alpha > 0$ as a nuisance parameter we can turn the gamma distribution into a single-parameter linear EF. Set $T(y) = y$ and

$$a(y) = (\alpha - 1)\log y - \log\Gamma(\alpha),\ \kappa(\theta) = -\alpha\log(-\theta),\ \mu = \kappa'(\theta) = \frac{\alpha}{-\theta},\ \theta = h(\mu) = -\frac{\alpha}{\mu},$$

for effective domain $\boldsymbol{\Theta} = (-\infty, 0)$, dual parameter space $\mathcal{M} = (0, \infty)$ and support $\mathfrak{T} = (0, \infty)$. With these choices we have for $\beta = -\theta > 0$

$$dF(y; \theta) = \frac{(-\theta)^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-(-\theta)y\}\, d\nu(y). \tag{2.6}$$

This provides us with mean and variance

$$\mu = \mathbb{E}_\theta[Y] = \frac{\alpha}{\beta} \qquad \text{and} \qquad \sigma^2 = \mathrm{Var}_\theta(Y) = \frac{\alpha}{\beta^2} = \frac{1}{\alpha}\mu^2.$$

[2] The function $\Psi(x) = \frac{d}{dx}\log\Gamma(x) = \Gamma'(x)/\Gamma(x)$ is called digamma function.

For parameter estimation one often needs to invert these identities which gives us

$$\alpha = \frac{\mu^2}{\sigma^2} \qquad \text{and} \qquad \beta = \frac{\mu}{\sigma^2}.$$

Remarks 2.10

- The gamma distribution contains as special cases the exponential distribution for $\alpha = \theta_2 = 1$ and $\beta = -\theta_1 > 0$, and the χ_r^2-distribution with r degrees of freedom for $\alpha = \theta_2 = r/2$ and $\beta = -\theta_1 = 1/2$.
- The distributions of the EF are all light-tailed in the sense that all moments of $T(Y)$ exist. Therefore, the EF does not allow for regularly varying survival functions, see (1.3). If Y is gamma distributed, then $Z = \exp\{Y\}$ is log-gamma distributed (with the special case of the Pareto distribution for the exponential case $\alpha = \theta_2 = 1$). For an example we refer to Sect. 2.2.5. However, this log-transformation is not always recommended because it may provide accurate models on the transformed log-scale, but back-transformation to the original scale may not necessarily provide a good predictive model on that original scale.
- The gamma density (2.6) may be a bit tricky in applications because the effective domain $\boldsymbol{\Theta} = (-\infty, 0)$ is one-sided bounded (we come back to this below). For this reason, in practice, one often uses links different from the canonical link $h(\mu) = -\alpha/\mu$. For instance, a parametrization $\theta = -\exp\{-\vartheta\}$ for $\vartheta \in \mathbb{R}$, see Ohlsson–Johansson [290], leads to the following model

$$dF(y; \vartheta) = \frac{y^{\alpha-1}}{\Gamma(\alpha)} \exp\left\{-e^{-\vartheta} y - \alpha\vartheta\right\} d\nu(y). \tag{2.7}$$

We will study the gamma model in more depth below, and parametrization (2.7) will correspond to the log-link choice, see Example 5.5, below.

Figure 2.1 gives examples of gamma densities for shape parameters $\alpha \in \{1/2, 1, 3/2, 2\}$ and scale parameters $\beta \in \{1/2, 1, 3/2, 2\}$ with $\alpha = \beta$ all providing the same mean $\mu = \mathbb{E}_\theta[Y] = \alpha/\beta = 1$. The crucial observation is that these gamma densities can have two different shapes, for $\alpha \le 1$ we have a strictly decreasing shape and for $\alpha > 1$ we have a unimodal density with mode in $(\alpha - 1)/\beta$.

Inverse Gaussian Distribution as a Vector-Valued Parameter EF

For the inverse Gaussian distribution with parameters $\alpha, \beta > 0$ we choose as ν the Lebesgue measure on $\mathbb{R}_+$. Then we make the following choices: $T(y) = (y, 1/y)^\top$,

$$a(y) = -\frac{1}{2}\log(2\pi y^3), \qquad \kappa(\boldsymbol{\theta}) \; = \; -2(\theta_1\theta_2)^{1/2} - \frac{1}{2}\log(-2\theta_2),$$

$$\left(\alpha/\beta, \beta/\alpha + 1/\alpha^2\right)^\top = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) \; = \; \left(\left(\frac{-2\theta_2}{-2\theta_1}\right)^{1/2}, \left(\frac{-2\theta_1}{-2\theta_2}\right)^{1/2} + \frac{1}{-2\theta_2}\right)^\top,$$

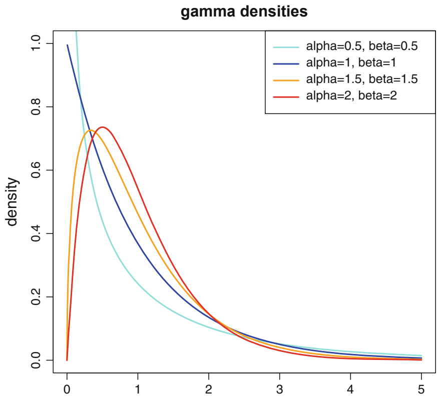

Fig. 2.1 Gamma densities for shape parameters $\alpha \in \{1/2, 1, 3/2, 2\}$ and scale parameters $\beta \in \{1/2, 1, 3/2, 2\}$ all providing the same mean $\mu = \alpha/\beta = 1$

for $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top \in (-\infty, 0)^2$, and setting $\beta = (-2\theta_1)^{1/2}$ and $\alpha = (-2\theta_2)^{1/2}$. The dual parameter space is $\mathcal{M} = (0, \infty)^2$, and we have support $\mathfrak{T} = (0, \infty)^2$ of $T(Y) = (Y, 1/Y)^\top$. With these choices we obtain

$$\begin{aligned} dF(y;\boldsymbol{\theta}) &= \exp\left\{\boldsymbol{\theta}^\top T(y) + 2(\theta_1\theta_2)^{1/2} + \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi y^3)\right\} d\nu(y) \\ &= \frac{1}{(2\pi y^3)^{1/2}}(-2\theta_2)^{1/2}\exp\left\{-\frac{1}{2y}\left((-2\theta_1)y^2 + (-2\theta_2) - 4(\theta_1\theta_2)^{1/2}y\right)\right\} d\nu(y) \\ &= \frac{\alpha}{(2\pi y^3)^{1/2}}\exp\left\{-\frac{\alpha^2}{2y}\left(1 - \frac{\beta}{\alpha}y\right)^2\right\} d\nu(y). \end{aligned} \tag{2.8}$$

This is a vector-valued parameter EF with $k = 2$ and with first moment

$$\mathbb{E}_{\boldsymbol{\theta}}\left[(Y, 1/Y)^\top\right] = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) = \left(\alpha/\beta, \beta/\alpha + 1/\alpha^2\right)^\top.$$

For receiving (2.8) we have chosen canonical parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top \in (-\infty, 0)^2$. Interestingly, we can close this parameter space for $\theta_1 = 0$, i.e., the effective domain $\boldsymbol{\Theta}$ is not open in this example. The choice $\theta_1 = 0$ gives us cumulant function $\kappa(\boldsymbol{\theta}) = -\frac{1}{2}\log(-2\theta_2)$ and boundary case

$$\begin{aligned} dF(y;\boldsymbol{\theta}) &= \exp\left\{\boldsymbol{\theta}^\top T(y) + \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi y^3)\right\} d\nu(y) \\ &= \frac{1}{(2\pi y^3)^{1/2}}(-2\theta_2)^{1/2}\exp\left\{-\frac{-2\theta_2}{2y}\right\} d\nu(y) \\ &= \frac{\alpha}{(2\pi y^3)^{1/2}}\exp\left\{-\frac{\alpha^2}{2y}\right\} d\nu(y). \end{aligned} \tag{2.9}$$

This is the distribution of the first-passage time of level $\alpha > 0$ of a standard Brownian motion, see Bachelier [20]; this distribution is also known as Lévy distribution.

If we treat $\alpha > 0$ as a nuisance parameter, we can turn the inverse Gaussian distribution into a single-parameter linear EF by setting $T(y) = y$,

$$a(y) = \log\left(\frac{\alpha}{(2\pi y^3)^{1/2}}\right) - \frac{\alpha^2}{2y}, \quad \kappa(\theta) = -\alpha(-2\theta)^{1/2},$$

$$\mu = \kappa'(\theta) = \frac{\alpha}{(-2\theta)^{1/2}}, \quad \theta = h(\mu) = -\frac{1}{2}\frac{\alpha^2}{\mu^2},$$

for $\theta \in (-\infty, 0)$, dual parameter space $\mathcal{M} = (0, \infty)$ and support $\mathfrak{T} = (0, \infty)$. With these choices we have the inverse Gaussian model for $\beta = (-2\theta)^{1/2} > 0$

$$\begin{aligned} dF(y;\theta) &= \exp\{a(y)\}\exp\left\{-\frac{1}{2y}\left((-2\theta)y^2 - 2\alpha(-2\theta)^{1/2}y\right)\right\} d\nu(y) \\ &= \frac{\alpha}{(2\pi y^3)^{1/2}}\exp\left\{-\frac{\alpha^2}{2y}\left(1 - \frac{\beta}{\alpha}y\right)^2\right\} d\nu(y). \end{aligned}$$

This provides us with mean and variance

$$\mu = \mathbb{E}_\theta[Y] = \frac{\alpha}{\beta} \qquad \text{and} \qquad \sigma^2 = \mathrm{Var}_\theta(Y) = \frac{\alpha}{\beta^3} = \frac{1}{\alpha^2}\mu^3.$$

For parameter estimation one often needs to invert these identities, which gives us

$$\alpha = \frac{\mu^{3/2}}{\sigma} \qquad \text{and} \qquad \beta = \frac{\mu^{1/2}}{\sigma}.$$

Figure 2.2 gives examples of inverse Gaussian densities for parameter choices $\alpha = \beta \in \{1/2, 1, 3/2, 2\}$ all providing the same mean $\mu = \mathbb{E}_\theta[Y] = \alpha/\beta = 1$.

Generalized Inverse Gaussian Distribution as a Vector-Valued Parameter EF

For the generalized inverse Gaussian distribution with parameters $\alpha, \beta > 0$ and $\gamma \in \mathbb{R}$ we choose as ν the Lebesgue measure on $\mathbb{R}_+$. We combine the terms of the gamma and the inverse Gaussian models to the vector-valued choice: $T(y) = (y, \log y, 1/y)^\top$ with $k = 3$. Moreover, we choose $a(y) = -\log y$ and cumulant function

$$\kappa(\boldsymbol{\theta}) = \log\left(2K_{\theta_2}(2\sqrt{\theta_1\theta_3})\right) - \frac{\theta_2}{2}\log(\theta_1/\theta_3),$$

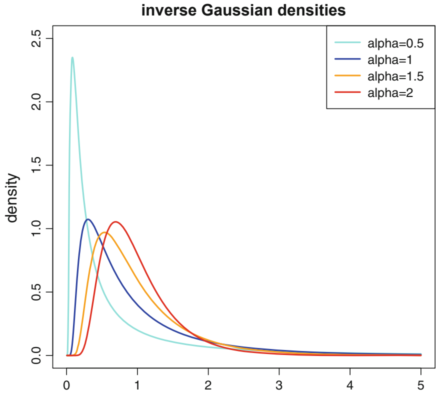

Fig. 2.2 Inverse Gaussian densities for parameters $\alpha = \beta \in \{1/2, 1, 3/2, 2\}$ all providing the same mean $\mu = \alpha/\beta = 1$

for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^\top \in (-\infty, 0) \times \mathbb{R} \times (-\infty, 0)$, and where K_{θ_2} denotes the modified Bessel function of the second kind with index $\gamma = \theta_2 \in \mathbb{R}$. With these choices we obtain generalized inverse Gaussian density

$$\begin{aligned} dF(y; \boldsymbol{\theta}) &= \exp\left\{\boldsymbol{\theta}^\top T(y) - \log\left(2K_{\theta_2}(2\sqrt{\theta_1\theta_3})\right) + \frac{\theta_2}{2}\log(\theta_1/\theta_3) - \log y\right\} d\nu(y) \\ &= \frac{(\alpha/\beta)^{\gamma/2}}{2K_\gamma(\sqrt{\alpha\beta})}\, y^{\gamma-1} \exp\left\{-\frac{1}{2}\left(\alpha y + \beta y^{-1}\right)\right\} d\nu(y), \end{aligned} \tag{2.10}$$

setting $\alpha = -2\theta_1$ and $\beta = -2\theta_3$. This is a vector-valued parameter EF with $k = 3$, and the first moment is given by

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\theta}}\left[\left(Y, \log Y, \frac{1}{Y}\right)^\top\right] = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) \\ &= \left(\frac{K_{\gamma+1}(\sqrt{\alpha\beta})}{K_\gamma(\sqrt{\alpha\beta})}\sqrt{\frac{\beta}{\alpha}},\ \log\sqrt{\frac{\beta}{\alpha}} + \frac{\partial}{\partial\gamma}\log K_\gamma(\sqrt{\alpha\beta}),\ \frac{K_{\gamma+1}(\sqrt{\alpha\beta})}{K_\gamma(\sqrt{\alpha\beta})}\sqrt{\frac{\alpha}{\beta}} - \frac{2\gamma}{\beta}\right)^\top. \end{aligned}$$

The effective domain $\boldsymbol{\Theta}$ is a bit complicated because the possible choices of (θ_1, θ_3) depend on $\theta_2 \in \mathbb{R}$, namely, for $\theta_2 < 0$ the negative half-line $(-\infty, 0]$ can be closed at the origin for θ_1, and for $\theta_2 > 0$ it can be closed at the origin for θ_3. The inverse Gaussian model is obtained for $\theta_2 = -1/2$ and the gamma model is obtained for $\theta_3 = 0$. For further properties of the generalized inverse Gaussian distribution we refer to the textbook of Jørgensen [200].

2.1.4 Vector-Valued Parameter EF: Count Variable Example

We close our EF examples by giving a discrete example with a vector-valued parameter.

Categorical Distribution as a Vector-Valued Parameter EF

For the categorical distribution with $k \in \mathbb{N}$ and $\boldsymbol{p} \in (0, 1)^k$ such that $\sum_{i=1}^k p_i < 1$, we choose as ν the counting measure on the finite set $\{1, \dots, k+1\}$. Then we make the following choices: $T(y) = (\mathbb{1}_{\{y=1\}}, \dots, \mathbb{1}_{\{y=k\}})^\top \in \mathbb{R}^k$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^\top$, $\boldsymbol{e^{\theta}} = (e^{\theta_1}, \dots, e^{\theta_k})^\top$ and

$$a(y) = 0, \qquad \kappa(\boldsymbol{\theta}) = \log\left(1 + \sum_{i=1}^k e^{\theta_i}\right), \qquad \boldsymbol{p} = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) = \frac{\boldsymbol{e^{\theta}}}{1 + \sum_{i=1}^k e^{\theta_i}},$$

for effective domain $\boldsymbol{\Theta} = \mathbb{R}^k$, dual parameter space $\mathcal{M} = (0, 1)^k$, and the support $\mathfrak{T}$ of $T(Y)$ are the $k+1$ corners of the unit simplex in $\mathbb{R}^k$. This representation is minimal, see Assumption 2.6. With these choices we have (set $\theta_{k+1} = 0$)

$$dF(y; \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}^\top T(y) - \log\left(1 + \sum_{i=1}^k e^{\theta_i}\right)\right\} d\nu(y) = \prod_{j=1}^{k+1}\left(\frac{e^{\theta_j}}{\sum_{i=1}^{k+1} e^{\theta_i}}\right)^{\mathbb{1}_{\{y=j\}}} d\nu(y).$$

This is a vector-valued parameter EF with $k \in \mathbb{N}$. The canonical link is slightly more complicated. Set vectors $\boldsymbol{v} = \exp\{\boldsymbol{\theta}\} \in \mathbb{R}^k$ and $\boldsymbol{w} = (1, \dots, 1)^\top \in \mathbb{R}^k$. This provides $\boldsymbol{p} = \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) = \frac{1}{1+\boldsymbol{w}^\top\boldsymbol{v}}\boldsymbol{v} \in \mathbb{R}^k$. Set matrix $A_{\boldsymbol{p}} = \mathbb{1} - \boldsymbol{p}\boldsymbol{w}^\top \in \mathbb{R}^{k\times k}$, the latter gives us $\boldsymbol{p} = A_{\boldsymbol{p}}\boldsymbol{v}$, and since $A_{\boldsymbol{p}}$ has full rank k, we obtain canonical link

$$\boldsymbol{p} \mapsto \boldsymbol{\theta} = h(\boldsymbol{p}) = \log\left(A_{\boldsymbol{p}}^{-1}\boldsymbol{p}\right) = \log\left(\frac{\boldsymbol{p}}{1 - \boldsymbol{w}^\top\boldsymbol{p}}\right).$$

The last identity can be verified by explicit calculation

$$\log\left(\frac{\boldsymbol{p}}{1 - \boldsymbol{w}^\top\boldsymbol{p}}\right) = \log\left(\frac{\boldsymbol{e^{\theta}}/(1 + \sum_{j=1}^k e^{\theta_j})}{1 - \sum_{i=1}^k e^{\theta_i}/(1 + \sum_{j=1}^k e^{\theta_j})}\right) = \log\left(\boldsymbol{e^{\theta}}\right) = \boldsymbol{\theta}.$$

Remarks 2.11

- There are many more examples that belong to the EF. From Theorem 2.4, we know that all examples of the EF are light-tailed in the sense that all moments of $T(Y)$ exist. If we want to model heavy-tailed distributions within the EF, we first need to apply a suitable transformation. We could model the Pareto distribution using transformation $T(y) = \log y$, and assuming that the transformed random variable has an exponential distribution. Different light-tailed examples are obtained by, e.g., using transformation $T(y) = y^\tau$ for the Weibull distribution or $T(y) = (\log y, \log(1-y))^\top$ for the beta distribution. We refrain from giving explicit formulas for these or other examples.

- Observe that in all examples above we have $\mathfrak{T} \subset \overline{\mathcal{M}}$, i.e., the support of $T(Y)$ is contained in the closure of the dual parameter space $\mathcal{M}$, we come back to this observation in Sect. 2.2.4, below.

2.2 Exponential Dispersion Family

In the previous section we have introduced the EF, and we have explicitly studied the vector-valued parameter EF examples of the Gaussian, the gamma and the inverse Gaussian models. We have highlighted that these three vector-valued parameter EFs can be turned into single-parameter EFs by declaring one parameter to be a nuisance parameter that is not modeled (and acts as a hyper-parameter). This changes these three models into single-parameter EFs. These three single-parameter EFs with nuisance parameter can also be interpreted as EDF models. In this section we discuss the single-parameter EDF; this is sufficient for our purposes, and vector-valued parameter extensions can be obtained in a canonical way.

2.2.1 Definition and Properties

The EFs of Sect. 2.1 can be extended to EDFs. In the single-parameter case this is achieved by a transformation $Y = X/\omega$, where $\omega > 0$ is a scaling and where X belongs to a single-parameter linear EF, i.e., with $T(x) = x$. We restrict ourselves to the single-parameter case $k = 1$ throughout this section. Choose a σ-finite measure ν_1 on $\mathbb{R}$ and a measurable function $a_1 : \mathbb{R} \to \mathbb{R}$. These choices give a single-parameter linear EF, directly modeling a real-valued random variable $T(X) = X$. By (2.2) we have distribution for the single-parameter linear EF random variable X

$$dF(x;\theta,1) = f(x;\theta,1)d\nu_1(x) = \exp\Big\{\theta x - \kappa(\theta) + a_1(x)\Big\} d\nu_1(x),$$

on the effective domain

$$\boldsymbol{\Theta} = \left\{\theta \in \mathbb{R};\ \int_{\mathbb{R}} \exp\left\{\theta x + a_1(x)\right\} d\nu_1(x) < \infty\right\}, \tag{2.11}$$

and with cumulant function

$$\theta \in \boldsymbol{\Theta} \;\mapsto\; \kappa(\theta) = \log\left(\int_{\mathbb{R}} \exp\left\{\theta x + a_1(x)\right\} d\nu_1(x)\right). \tag{2.12}$$

Throughout, we assume that the effective domain $\boldsymbol{\Theta}$ has a non-empty interior $\mathring{\boldsymbol{\Theta}}$. Thus, since $\boldsymbol{\Theta}$ is convex, we assume that $\mathring{\boldsymbol{\Theta}}$ is a non-empty (possibly infinite) open interval in $\mathbb{R}$.

Following Jørgensen [201, 202], we extend this linear EF to an EDF as follows. Choose a family of σ-finite measures ν_ω on $\mathbb{R}$ and measurable functions $a_\omega : \mathbb{R} \to \mathbb{R}$ for a given index set $\mathcal{W} \ni \omega$ with $\{1\} \subset \mathcal{W} \subset \mathbb{R}_+$. Assume that we have an ω-independent scaled cumulant function κ on this index set $\mathcal{W}$, that is,

$$\theta \in \boldsymbol{\Theta} \;\mapsto\; \kappa(\theta) = \frac{1}{\omega}\left(\log \int_{\mathbb{R}} \exp\left\{\theta x + a_\omega(x)\right\} d\nu_\omega(x)\right) \qquad \text{for all } \omega \in \mathcal{W},$$

with effective domain $\boldsymbol{\Theta}$ defined by (2.11), i.e., for $\omega = 1$. This allows us to consider the distribution functions

$$\begin{aligned} dF(x;\theta,\omega) = f(x;\theta,\omega)d\nu_\omega(x) \;&=\; \exp\Big\{\theta x - \omega\kappa(\theta) + a_\omega(x)\Big\} d\nu_\omega(x) \\ &= \exp\Big\{\omega\left(\theta y - \kappa(\theta)\right) + a_\omega(\omega y)\Big\} d\nu_\omega(\omega y), \end{aligned} \tag{2.13}$$

in the third identity we did a change of variable $x \mapsto y = x/\omega$. By reparametrizing the function $a_\omega(\omega\ \cdot)$ and the σ-finite measures $\nu_\omega(\omega\ \cdot)$ slightly differently, depending on the particular structure of the chosen σ-finite measures, we arrive at the following single-parameter EDF.

Definition 2.12 The (single-parameter) EDF is given by densities of the form

$$Y \;\sim\; f(y;\theta, v/\varphi) = \exp\left\{\frac{y\theta - \kappa(\theta)}{\varphi/v} + a(y; v/\varphi)\right\}, \tag{2.14}$$

with

$\kappa : \boldsymbol{\Theta} \to \mathbb{R}$ is the cumulant function (2.12),

$\theta \in \boldsymbol{\Theta}$ is the canonical parameter in the effective domain (2.11),

$v > 0$ is a given weight (exposure, volume),

$\varphi > 0$ is the dispersion parameter,

$a(\cdot;\cdot)$ is the normalization, *not* depending on the canonical parameter θ.

Remarks 2.13

- Exposure $v > 0$ and dispersion parameter $\varphi > 0$ provide the parametrization usually used for $\omega = v/\varphi \in \mathcal{W}$. Their meaning and interpretation will become clear below, and they will always appear as a ratio $\omega = v/\varphi$.
- The support of these EDF distributions does not depend on the explicit choice of the canonical parameter $\theta \in \boldsymbol{\Theta}$, but it may depend on $\omega = v/\varphi \in \mathcal{W}$ through the choices of the σ-finite measures ν_ω, for $\omega \in \mathcal{W}$. Consequently, $a(y;\omega)$ is a normalization such that $f(y;\theta,\omega)$ integrates to 1 w.r.t. the chosen σ-finite measure ν_ω to receive a proper distributional model.
- The transformation $x \mapsto y = x/\omega$ in (2.13) is called duality transformation, see Section 3.1 in Jørgensen [203]. It provides the duality between the *additive form* (in variable x in (2.13)) and the *reproductive form* (in variable y in (2.13)) of the EDF; Definition 2.12 is the reproductive form.
- Lemma 2.1 tells us that $\boldsymbol{\Theta}$ is convex, thus, it is a possibly infinite interval in $\mathbb{R}$. To exclude trivial cases we will always assume that the σ-finite measure ν_1 is not concentrated in one single point (this relates to the minimal representation for $k = 1$ in the linear EF case, see Assumption 2.6), and that the interior $\mathring{\boldsymbol{\Theta}}$ of the effective domain $\boldsymbol{\Theta}$ is non-empty.

Corollary 2.14 *Assume $\mathring{\boldsymbol{\Theta}}$ is non-empty and that ν_1 is not concentrated in one single point. Choose $Y \sim F(\cdot;\theta, v/\varphi)$ for fixed $\theta \in \mathring{\boldsymbol{\Theta}}$. The moment generating function of Y for small $r \in \mathbb{R}$ satisfies*

$$M_Y(r) = \mathbb{E}_\theta\left[\exp\{rY\}\right] = \exp\left\{\frac{v}{\varphi}\left[\kappa(\theta + r\varphi/v) - \kappa(\theta)\right]\right\}.$$

The first two moments of Y are given by

$$\mu = \mathbb{E}_\theta\left[Y\right] = \kappa'(\theta) \qquad \text{and} \qquad \mathrm{Var}_\theta\left(Y\right) = \frac{\varphi}{v}\kappa''(\theta) > 0.$$

The cumulant function κ is smooth and strictly convex on $\mathring{\boldsymbol{\Theta}}$ with canonical link $h = (\kappa')^{-1}$. The variance function is defined by $\mu \mapsto V(\mu) = (\kappa'' \circ h)(\mu)$ and, consequently, for the variance of Y we have $\mathrm{Var}_\mu(Y) = \frac{\varphi}{v}V(\mu)$ for $\mu \in \mathcal{M}$.

Proof This follows analogously to Theorem 2.4. The linear case $T(y) = y$ with ν_1 not being concentrated in one single point guarantees that the minimal dimension is $k = 1$, providing a minimal representation in this dimension, see Assumption 2.6. □

Before giving explicit examples we state the so-called convolution formula.

Corollary 2.15 (Convolution Formula) *Assume $\mathring{\boldsymbol{\Theta}}$ is non-empty and that v_1 is not concentrated in one single point. Assume that $Y_i \sim F(\cdot; \theta, v_i/\varphi)$ are independent, for $1 \le i \le n$, with fixed $\theta \in \mathring{\boldsymbol{\Theta}}$. Set $v_+ = \sum_{i=1}^n v_i$. Then*

$$Y_+ = \frac{1}{v_+} \sum_{i=1}^n v_i Y_i \ \sim \ F(\cdot; \theta, v_+/\varphi).$$

Proof The proof immediately follows from calculating the moment generating function $M_{Y_+}(r)$ and from using the independence between the Y_i's. □

2.2.2 *Exponential Dispersion Family Examples*

The single-parameter linear EF examples introduced above can be reformulated as EDF examples.

Binomial Distribution as a Single-Parameter EDF

For the binomial distribution with parameters $p \in (0, 1)$ and $n \in \mathbb{N}$ we choose the counting measure on $\{0, 1/n, \dots, 1\}$ with $\omega = n$. Then we make the following choices

$$a(y) = \log\binom{n}{ny}, \quad \kappa(\theta) = \log(1+e^\theta), \quad p = \kappa'(\theta) = \frac{e^\theta}{1+e^\theta}, \quad \theta = h(p) = \log\left(\frac{p}{1-p}\right),$$

for effective domain $\boldsymbol{\Theta} = \mathbb{R}$ and dual parameter space $\mathcal{M} = (0, 1)$. With these choices we have

$$f(y; \theta, n) = \binom{n}{ny} \exp\left\{n\left(\theta y - \log(1+e^\theta)\right)\right\} = \binom{n}{ny} \left(\frac{e^\theta}{1+e^\theta}\right)^{ny} \left(\frac{1}{1+e^\theta}\right)^{n-ny}.$$

This is a single-parameter EDF. The canonical link $p \mapsto h(p)$ gives the logit function. Mean and variance are given by

$$p = \mathbb{E}_\theta[Y] = \kappa'(\theta) = \frac{e^\theta}{1+e^\theta} \quad \text{and} \quad \mathrm{Var}_\theta(Y) = \frac{1}{n}\kappa''(\theta) = \frac{1}{n}\frac{e^\theta}{(1+e^\theta)^2} = \frac{1}{n}p(1-p),$$

and the variance function is given by $V(\mu) = \mu(1-\mu)$. The binomial random variable is obtained by setting $X = nY \sim \mathrm{Binom}(n, p)$.

Poisson Distribution as a Single-Parameter EDF

For the Poisson distribution with parameters $\lambda > 0$ and $v > 0$ we choose the counting measure on $\mathbb{N}_0/v$ for exposure $\omega = v$. Then we make the following choices

$$a(y) = \log\left(\frac{v^{vy}}{(vy)!}\right), \quad \kappa(\theta) = e^{\theta}, \quad \lambda = \kappa'(\theta) = e^{\theta}, \quad \theta = h(\lambda) = \log(\lambda),$$

for effective domain $\boldsymbol{\Theta} = \mathbb{R}$ and dual parameter space $\mathcal{M} = (0, \infty)$. With these choices we have

$$f(y; \theta, v) = \frac{v^{vy}}{(vy)!} \exp\left\{v\left(\theta y - e^{\theta}\right)\right\} = e^{-v\lambda}\frac{(v\lambda)^{vy}}{(vy)!}. \tag{2.15}$$

This is a single-parameter EDF. The canonical link $\lambda \mapsto h(\lambda)$ is the log-link. Mean and variance are given by

$$\lambda = \mathbb{E}_{\theta}[Y] = \kappa'(\theta) = e^{\theta} \qquad \text{and} \qquad \mathrm{Var}_{\theta}(Y) = \frac{1}{v}\kappa''(\theta) = \frac{1}{v}e^{\theta} = \frac{1}{v}\lambda,$$

and the variance function is given by $V(\lambda) = \lambda$, that is, the variance function is linear in the mean parameter λ. The Poisson random variable is obtained by setting $X = vY \sim \mathrm{Poi}(v\lambda)$. We choose $\varphi = 1$, here, meaning that we have neither under- nor over-dispersion. Thus, the choices v and φ in $\omega = v/\varphi$ have the interpretation of an exposure and a dispersion parameter, respectively. This interpretation is going to be important in claim counts modeling, below.

Gamma Distribution as a Single-Parameter EDF

For the gamma distribution with parameters $\alpha, \beta > 0$ we choose the Lebesgue measure on $\mathbb{R}_+$ and shape parameter $\omega = v/\varphi = \alpha$. We make the following choices

$$a(y) = (\alpha - 1)\log y + \alpha\log\alpha - \log\Gamma(\alpha), \quad \kappa(\theta) = -\log(-\theta),$$
$$\mu = \kappa'(\theta) = -1/\theta, \quad \theta = h(\mu) = -1/\mu,$$

for effective domain $\boldsymbol{\Theta} = (-\infty, 0)$ and dual parameter space $\mathcal{M} = (0, \infty)$. With these choices we have

$$f(y; \theta, \alpha) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} \exp\left\{\alpha\left(y\theta + \log(-\theta)\right)\right\} = \frac{(-\theta\alpha)^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} \exp\left\{-(-\theta\alpha)y\right\}.$$

This is analogous to (2.6) with shape parameter $\alpha > 0$ and scale parameter $\beta = -\theta > 0$. Mean and variance are given by

$$\mu = \mathbb{E}_\theta[Y] = \kappa'(\theta) = -\theta^{-1} \qquad \text{and} \qquad \mathrm{Var}_\theta(Y) = \frac{1}{\alpha}\kappa''(\theta) = \frac{1}{\alpha}\theta^{-2},$$

and the variance function is given by $V(\mu) = \mu^2$, that is, the variance function is quadratic in the mean parameter μ. The gamma random variable is obtained by setting $X = \alpha Y \sim \Gamma(\alpha, \beta)$. This gives us for the first two moments of X

$$\mu_X = \mathbb{E}_\theta[X] = \frac{\alpha}{\beta} \qquad \text{and} \qquad \mathrm{Var}_\theta(X) = \frac{\alpha}{\beta^2} = \frac{1}{\alpha}\mu_X^2.$$

Suppose $v = 1$, for shape parameter $\alpha > 1$, we have under-dispersion $\varphi = 1/\alpha < 1$ and the gamma density is unimodal; for shape parameter $\alpha < 1$, we have over-dispersion $\varphi = 1/\alpha > 1$ and the gamma density is strictly decreasing, we refer to Fig. 2.1.

Inverse Gaussian Distribution as a Single-Parameter EDF

For the inverse Gaussian distribution with parameters $\alpha, \beta > 0$ we choose the Lebesgue measure on $\mathbb{R}_+$ and we set $\omega = v/\varphi = \alpha$. We make the following choices

$$a(y) = \log\left(\frac{\alpha^{1/2}}{(2\pi y^3)^{1/2}}\right) - \frac{\alpha}{2y}, \quad \kappa(\theta) = -(-2\theta)^{1/2},$$

$$\mu = \kappa'(\theta) = \frac{1}{(-2\theta)^{1/2}}, \quad \theta = h(\mu) = -\frac{1}{2\mu^2},$$

for $\theta \in (-\infty, 0)$ and dual parameter space $\mathcal{M} = (0, \infty)$. With these choices we have

$$\begin{aligned} f(y; \theta, \alpha)dy &= \frac{\alpha^{1/2}}{(2\pi y^3)^{1/2}} \exp\left\{\alpha\left(\theta y + (-2\theta)^{1/2}\right) - \frac{\alpha}{2y}\right\} dy \\ &= \frac{\alpha^{1/2}}{(2\pi y^3)^{1/2}} \exp\left\{-\frac{\alpha}{2y}\left(1 - (-2\theta)^{1/2} y\right)^2\right\} dy \\ &= \frac{\alpha}{(2\pi x^3)^{1/2}} \exp\left\{-\frac{\alpha^2}{2x}\left(1 - \frac{(-2\theta)^{1/2}}{\alpha} x\right)^2\right\} dx, \end{aligned}$$

where in the last step we did a change of variable $y \mapsto x = \alpha y$. This is exactly (2.8). Mean and variance are given by

$$\mu = \mathbb{E}_\theta[Y] = \kappa'(\theta) = (-2\theta)^{-1/2} \quad \text{and} \quad \mathrm{Var}_\theta(Y) = \frac{1}{\alpha}\kappa''(\theta) = \frac{1}{\alpha}(-2\theta)^{-3/2},$$

and the variance function is given by $V(\mu) = \mu^3$, that is, the variance function is cubic in the mean parameter μ. The inverse Gaussian random variable is obtained by setting $X = \alpha Y$. The mean and variance of X are given by, set $\beta = (-2\theta)^{1/2} > 0$,

$$\mu_X = \mathbb{E}_\theta[X] = \frac{\alpha}{\beta} \qquad \text{and} \qquad \mathrm{Var}_\theta(X) = \frac{\alpha}{\beta^3} = \frac{1}{\alpha^2}\mu_X^3.$$

This inverse Gaussian density is illustrated in Fig. 2.2.

Similarly to (2.9), we can extend the inverse Gaussian model to the boundary case $\theta = 0$, i.e., the effective domain $\boldsymbol{\Theta} = (-\infty, 0]$ is not open. This provides us with density

$$f(y; \theta = 0, \alpha)dy = \frac{\alpha}{(2\pi x^3)^{1/2}} \exp\left\{-\frac{\alpha^2}{2x}\right\} dx, \tag{2.16}$$

using, as above, the change of variable $y \mapsto x = \alpha y$. An additional transformation $x \mapsto 1/x$ gives a gamma distribution with shape parameter 1/2 and scale parameter $\alpha^2/2$.

Remark 2.16 The inverse Gaussian case gives an example of a non-open effective domain $\boldsymbol{\Theta} = (-\infty, 0]$. It is worth noting that for the boundary parameter $\theta = 0$, the first moment does not exist, i.e., Corollary 2.14 only makes statements in the interior $\mathring{\boldsymbol{\Theta}}$ of the effective domain $\boldsymbol{\Theta}$. This also relates to Remarks 2.9 on the dual parameter space $\mathcal{M}$.

2.2.3 *Tweedie's Distributions*

Tweedie's compound Poisson (CP) model was introduced in 1984 by Tweedie [358], and it has been studied in detail in Jørgensen [202], Jørgensen–de Souza [204], Smyth–Jørgensen [342] and in the review paper of Delong et al. [94]. Tweedie's CP model belongs to the EDF. We spend more time on explaining Tweedie's CP model because it plays an important role in actuarial modeling.

Tweedie's CP model is received by choosing as σ-finite measure ν_1 a mixture of the Lebesgue measure on $(0, \infty)$ and a point measure in 0. Furthermore, we choose *power variance parameter* $p \in (1, 2)$ and cumulant function

$$\kappa(\theta) = \kappa_p(\theta) = \frac{1}{2-p}\left((1-p)\theta\right)^{\frac{2-p}{1-p}}, \tag{2.17}$$

on the effective domain $\theta \in \boldsymbol{\Theta} = (-\infty, 0)$. This provides us with Tweedie's CP model

$$Y \sim f(y; \theta, v/\varphi) = \exp\left\{ \frac{y\theta - \kappa_p(\theta)}{\varphi/v} + a(y; v/\varphi) \right\},$$

with exposure $v > 0$ and dispersion parameter $\varphi > 0$; the normalizing function $a(\cdot; v/\varphi)$ does not have any simple closed form, we refer to Section 2.1 in Jørgensen–de Souza [204] and Section 4.2 in Jørgensen [203].

The first two moments of Tweedie's CP random variable Y are given by

$$\mu = \mathbb{E}_\theta [Y] = \kappa_p'(\theta) = ((1-p)\theta)^{\frac{1}{1-p}} \in \mathcal{M} = (0, \infty), \tag{2.18}$$

$$\mathrm{Var}_\theta (Y) = \frac{\varphi}{v}\kappa_p''(\theta) = \frac{\varphi}{v} ((1-p)\theta)^{\frac{p}{1-p}} = \frac{\varphi}{v}\mu^p > 0. \tag{2.19}$$

The parameter $p \in (1, 2)$ determines the power variance functions $V(\mu) = \mu^p$ between the Poisson $p = 1$ and the gamma $p = 2$ cases, see Sect. 2.2.2.

The moment generating function of Tweedie's CP random variable $X = vY/\varphi = \omega Y$ in its additive form is given by, we use Corollary 2.14,

$$M_X(r) = M_{vY/\varphi}(r) = \exp\left\{ \frac{v}{\varphi}\, \kappa_p(\theta) \left(\left(\frac{-\theta}{-\theta - r} \right)^{\frac{2-p}{p-1}} - 1 \right) \right\} \qquad \text{for } r < -\theta.$$

Some readers will notice that this is the moment generating function of a CP distribution having i.i.d. gamma claim sizes. This is exactly the statement of the next proposition which is found, e.g., in Smyth–Jørgensen [342].

Proposition 2.17 *Assume $S = \sum_{i=1}^N Z_i$ is CP distributed with Poisson claim counts $N \sim \mathrm{Poi}(\lambda v)$ and i.i.d. gamma claim sizes $Z_i \sim \Gamma(\alpha, \beta)$ being independent of N. We have $S \overset{(d)}{=} vY/\varphi$ by identifying the parameters as follows*

$$p = \frac{\alpha + 2}{\alpha + 1} \in (1, 2), \qquad \beta = -\theta > 0 \qquad \textit{and} \qquad \lambda = \frac{1}{\varphi}\, \kappa_p(\theta) > 0.$$

Proof of Proposition 2.17 Assume S is CP distributed with i.i.d. gamma claim sizes. From Proposition 2.11 and Section 3.2.1 in Wüthrich [387] we receive that the moment generating function of S is given by

$$M_S(r) = \exp\left\{ \lambda v \left(\left(\frac{\beta}{\beta - r} \right)^\alpha - 1 \right) \right\} \qquad \text{for } r < \beta.$$

Using the proposed parameter identification, the claim immediately follows. □

Proposition 2.17 gives us a second interpretation of Tweedie's CP model which was introduced in an EDF fashion, above. This second interpretation explains the name of this EDF model, it explains the mixture of the Lebesgue measure and the point measure in 0, and it also highlights why the Poisson model and the gamma model are the boundary cases in terms of power variance functions.

An interesting question is whether the EDF can be extended beyond power variance functions $V(\mu) = \mu^p$ with $p \in [1, 2]$. The answer to this question is yes, and the full answer is provided in Theorem 2 of Jørgensen [202]:

Theorem 2.18 (Jørgensen [202], Without Proof) *Only power variance parameters* $p \in (0, 1)$ *do not allow for EDF models.*

Table 2.1 gives the EDF distributions that have a power variance function. These distributions are called *Tweedie's distributions*, with the special case of Tweedie's CP distributions for $p \in (1, 2)$. The densities for $p \in \{0, 1, 2, 3\}$ have a closed form, but the other Tweedie's distributions do not have a closed-form density. Thus, they cannot explicitly be constructed as suggested in Sect. 2.2.1. Besides the constructive approach presented above, there is a uniqueness theorem saying that the variance function $V(\cdot)$ on the domain $\mathcal{M}$ characterizes the single-parameter linear EF, see Theorem 2.11 in Jørgensen [203]. This uniqueness theorem is the basis of the proof of Theorem 2.18. Tweedie's distributions for $p \notin [0, 1] \cup \{2, 3\}$ involve infinite sums for the normalization $\exp\{a(\cdot, \cdot)\}$, we refer to formulas (4.19), (4.20) and (4.31) in Jørgensen [203], this is the reason that one has to go via the uniqueness theorem to prove Theorem 2.18. Dunn–Smyth [112] provide methods of fast calculation of some of these infinite sums; in Sect. 5.5.2, below, we present an approximation (saddlepoint approximation). The uniqueness theorem is also useful to construct new examples within the EF, see, e.g., Section 2 of Awad et al. [15].

Table 2.1 Power variance function models $V(\mu) = \mu^p$ within the EDF (taken from Table 4.1 in Jørgensen [203])

p	Distribution	Support of Y	$\boldsymbol{\Theta}$	$\mathcal{M}$
$p < 0$	Generated by extreme stable distributions	$\mathbb{R}$	$[0, \infty)$	$(0, \infty)$
$p = 0$	Gaussian distribution	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
$p = 1$	Poisson distribution	$\mathbb{N}_0$	$\mathbb{R}$	$(0, \infty)$
$1 < p < 2$	Tweedie's CP distribution	$[0, \infty)$	$(-\infty, 0)$	$(0, \infty)$
$p = 2$	Gamma distribution	$(0, \infty)$	$(-\infty, 0)$	$(0, \infty)$
$p > 2$	Generated by positive stable distributions	$(0, \infty)$	$(-\infty, 0]$	$(0, \infty)$
$p = 3$	Inverse Gaussian distribution	$(0, \infty)$	$(-\infty, 0]$	$(0, \infty)$

2.2.4 Steepness of the Cumulant Function

Assume we have a fixed EF satisfying Assumption 2.6. All random variables $T(Y)$ belonging to this EF have the same support, not depending on the particular choice of the canonical parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We denote this support of $T(Y)$ by $\mathfrak{T}$.

Below, we are going to estimate the canonical parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ from data using maximum likelihood estimation. For this it is advantageous to have the property $\mathfrak{T} \subset \mathcal{M}$, because, intuitively, this allows us to directly select $\widehat{\mu} = T(Y)$ as the parameter estimate in the dual parameter space $\mathcal{M}$, for a given observation $T(Y) \in \mathfrak{T}$. This then translates to a canonical parameter $\widehat{\boldsymbol{\theta}} = h(\widehat{\mu}) = h(T(Y)) \in \boldsymbol{\Theta}$, using the canonical link h; this estimation approach will be better motivated in Chap. 3, below. Unfortunately, many examples of the EF do not satisfy this property $\mathfrak{T} \subset \mathcal{M}$. For instance, in the Poisson model the observation $T(Y) = Y = 0$ is not included in $\mathcal{M}$, see Table 2.1. This poses some challenges in parameter estimation, and the purpose of this small discussion is to be prepared for these challenges.

A cumulant function κ is called *steep* if for all $\boldsymbol{\theta} \in \mathring{\boldsymbol{\Theta}}$ and all $\widetilde{\boldsymbol{\theta}}$ in the boundary of $\boldsymbol{\Theta}$

$$\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^\top \nabla_{\boldsymbol{\theta}} \kappa \left(\alpha \boldsymbol{\theta} + (1-\alpha)\widetilde{\boldsymbol{\theta}}\right) \;\to\; \infty \qquad \text{for } \alpha \downarrow 0, \tag{2.20}$$

we refer to Formula (20) in Section 8.1 of Barndorff-Nielsen [23]. Define the convex closure of the support $\mathfrak{T}$ by $\mathfrak{C} = \overline{\mathrm{conv}}(\mathfrak{T})$.

Theorem 2.19 (Theorem 9.2 in Barndorff-Nielsen [23], Without Proof) *Assume we have a fixed EF satisfying Assumption 2.6. The cumulant function κ is steep if and only if $\mathring{\mathfrak{C}} = \mathcal{M} = \nabla_{\boldsymbol{\theta}} \kappa(\mathring{\boldsymbol{\Theta}})$.*

Theorem 2.19 tells us that for a steep cumulant function we have $\mathfrak{C} = \overline{\mathcal{M}} = \overline{\nabla_{\boldsymbol{\theta}} \kappa(\mathring{\boldsymbol{\Theta}})}$. In this case parameter estimation can be extended to observations $T(Y) \notin \mathcal{M}$ such that we may obtain a degenerate model at the boundary of $\mathcal{M}$. Coming back to our Poisson example from above, in this case we set $\widehat{\mu} = 0$, which gives a degenerate Poisson model.

Throughout this book we will work under the assumption that κ is steep. The classical examples satisfy this assumption: the examples with power variance parameter p in $\{0\} \cup [1, \infty)$ satisfy Theorem 2.19; this includes the Gaussian, the Poisson, the gamma, the inverse Gaussian and Tweedie's CP models, see Table 2.1. Moreover, the examples we have met in Sect. 2.1 fulfill this assumption; these are the single-parameter linear EF models of the Bernoulli, the binomial and the negative binomial distributions, as well as the vector-valued parameter examples of the Gaussian, the gamma and the inverse Gaussian models and of the categorical distribution. The only models we have seen that do not have a steep cumulant function are the power variance models with $p < 0$, see Table 2.1.

Remark 2.20 Working within the EDF needs some additional thoughts because the support $\mathfrak{T} = \mathfrak{T}_\omega$ of the single-parameter linear EDF random variable $Y = T(Y)$ may depend on the specific choice of the dispersion parameter $\omega \in \mathcal{W} \supset \{1\}$ through the σ-finite measure $d\nu_\omega(\omega\,\cdot)$, see (2.13). For instance, in the binomial case the support of Y is given by $\mathfrak{T}_\omega = \{0, 1/n, \ldots, 1\}$ with $\omega = n$, see Sect. 2.2.2.

Assume that the cumulant function κ is steep for the single-parameter linear EF that corresponds to the single-parameter EDF with $\omega = 1$. Theorem 2.19 then implies that for this choice we have $\mathring{\mathfrak{C}}_{\omega=1} = \nabla_{\boldsymbol{\theta}}\kappa(\mathring{\boldsymbol{\Theta}})$ with convex closure $\mathfrak{C}_{\omega=1} = \overline{\mathrm{conv}}(\mathfrak{T}_{\omega=1})$.

Consider $\omega \in \mathcal{W}\setminus\{1\}$ which corresponds to the choice ν_ω of the σ-finite measure on $\mathbb{R}$. This choice belongs to the cumulant function $\theta \mapsto \omega\kappa(\theta)$ in the additive form (x-parametrization in (2.13)). Since steepness (2.20) holds for any $\omega > 0$ we receive that the convex closure of the support of this distribution in the x-parametrization in (2.13) is given by $\overline{\nabla_{\boldsymbol{\theta}}\omega\kappa(\mathring{\boldsymbol{\Theta}})} = \omega\overline{\nabla_{\boldsymbol{\theta}}\kappa(\mathring{\boldsymbol{\Theta}})}$. The duality transformation $x \mapsto y = x/\omega$ leads to the change of measure $d\nu_\omega(x) \mapsto d\nu_\omega(\omega y)$ and to the corresponding change of support, see (2.13). The latter implies that in the reproductive form (y-parametrization) the convex closure of the support does not depend on the specific choice of $\omega \in \mathcal{W}$. Since the EDF representation given in (2.14) corresponds to the y-parametrization (reproductive form), we can use Theorem 2.19 without limitation also for the single-parameter linear EDF given by (2.14), and $\mathfrak{C}$ does not depend on $\omega \in \mathcal{W}$.

2.2.5 Lab: Large Claims Modeling

From Corollary 2.14 we know that the moment generating function exists around the origin for all examples belonging to the EDF. This implies that the moments of all orders exist, and that we have an exponentially decaying survival function $\mathbb{P}_\theta[Y > y] = 1 - F(y;\theta,\omega) \sim \exp\{-\varrho y\}$ for some $\varrho > 0$ as $y \to \infty$, see (1.2). In many applied situations the data is more heavy-tailed and, thus, cannot be modeled by such an exponentially decaying survival function. In such cases one often chooses a distribution function with a regularly varying survival function; regular variation with tail index $\beta > 0$ has been introduced in (1.3). A popular choice is a log-gamma distribution which can be obtained from the gamma distribution (belonging to the EDF). We briefly explain how this is done and how it relates to the Pareto and the Lomax [256] distributions.

We start from the gamma density (2.6). The random variable Z has a log-gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta = -\theta > 0$ if $\log(Z) = Y$ has a gamma distribution with these parameters. Thus, the gamma density of $Y = \log(Z)$ is given by

$$f(y;\beta,\alpha)dy = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-\beta y\}\, dy \qquad \text{for } y > 0.$$

We do a change of variable $y \mapsto z = \exp\{y\}$ to receive the density of the log-gamma distributed random variable $Z = \exp\{Y\}$

$$f(z;\beta,\alpha)dz = \frac{\beta^\alpha}{\Gamma(\alpha)}(\log z)^{\alpha-1} z^{-(\beta+1)}dz \qquad \text{for } z > 1.$$

This log-gamma density has support $(1,\infty)$. The distribution function of this log-gamma distributed random variable needs to be calculated numerically, and its survival function is regularly varying with tail index $\beta > 0$.

A special case of the log-gamma distribution is the Pareto distribution. The Pareto distribution is more tractable and it is obtained by setting shape parameter $\alpha = 1$ in the log-gamma density. This gives us the Pareto density

$$f(z;\beta)dz = f(z;\beta,\alpha=1)dz = \beta z^{-(\beta+1)}dz \qquad \text{for } z > 1.$$

The distribution function in this Pareto case is for $z \geq 1$ given by

$$F(z;\beta) = 1 - z^{-\beta}.$$

Obviously, this provides a regularly varying survival function with tail index $\beta > 0$; in fact, in this case we do not need to go over to the limit in (1.3) because we have an exact identity. The Pareto distribution has the nice property that it is closed under thresholding (lower-truncation) with M, that is, we remain within the family of Pareto distributions with the same tail index β by considering lower-truncated claims: for $1 \leq M \leq z$ we have

$$F(z;\beta,M) = \mathbb{P}[Z \leq z | Z > M] = \frac{\mathbb{P}[M < Z \leq z]}{\mathbb{P}[Z > M]} = 1 - \left(\frac{z}{M}\right)^{-\beta}.$$

This is the classical definition of the Pareto distribution, and it allows to preserve full flexibility in the choice of the threshold $M > 0$.

The disadvantage of the Pareto distribution is that it does not provide a continuous density on $\mathbb{R}_+$ as there is a discontinuity in threshold M. For this reason, one sometimes explores another change of variable $Z \mapsto X = Z - M$ for a Pareto distributed random variable $Z \sim F(\cdot;\beta,M)$. This provides the Lomax distribution, also called Pareto Type II distribution. X has the following distribution function on $(0,\infty)$

$$\mathbb{P}[X \leq x] = 1 - \left(\frac{x+M}{M}\right)^{-\beta} \qquad \text{for } x \geq 0.$$

This distribution has again a regularly varying survival function with tail index $\beta > 0$. Moreover, we have

$$\lim_{x\to\infty} \frac{\left(\frac{x+M}{M}\right)^{-\beta}}{\left(\frac{x}{M}\right)^{-\beta}} = \lim_{x\to\infty}\left(1 + \frac{M}{x}\right)^{-\beta} = 1.$$

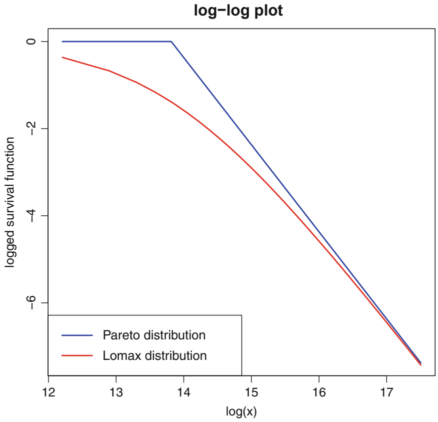

Fig. 2.3 Log-log plot of a Pareto and a Lomax distribution with tail index $\beta = 2$ and threshold $M = 1'000'000$

This says that we should choose the same threshold $M > 0$ for both the Pareto and the Lomax distribution to receive the same asymptotic tail behavior, and this also quantifies the rate of convergence between the two survival functions. Figure 2.3 illustrates this convergence in a log-log plot choosing tail index $\beta = 2$ and threshold $M = 1'000'000$.

For completeness we provide the density of the Pareto distribution

$$f(z;\beta,M) = \frac{\beta}{M}\left(\frac{z}{M}\right)^{-(\beta+1)} \qquad \text{for } z \geq M,$$

and of the Lomax distribution

$$f(x;\beta,M) = \frac{\beta}{M}\left(\frac{x+M}{M}\right)^{-(\beta+1)} \qquad \text{for } x \geq 0.$$

2.3 Information Geometry in Exponential Families

We do a short excursion to information geometry. This excursion may look a bit disconnected from what we have done so far, but it provides us with important background information for the chapter on forecast evaluation, see Chap. 4, below.

2.3.1 Kullback–Leibler Divergence

There is literature in information geometry which uses techniques from differential geometry to study EFs as Riemannian manifolds with points corresponding to EF densities parametrized by their canonical parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we refer to Amari [10], Ay et al. [16] and Nielsen [285] for an extended treatment of these mathematical concepts.

Choose a fixed EF (2.2) with cumulant function κ on the effective domain $\boldsymbol{\Theta} \subseteq \mathbb{R}^k$ and with σ-finite measure ν on $\mathbb{R}$. We define the Kullback–Leibler (KL) divergence (relative entropy) from model $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$ to model $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ within this EF by

$$D_{\mathrm{KL}}(f(\cdot;\boldsymbol{\theta}_0)||f(\cdot;\boldsymbol{\theta}_1)) = \int_{\mathbb{R}} f(y;\boldsymbol{\theta}_0) \log\left(\frac{f(y;\boldsymbol{\theta}_0)}{f(y;\boldsymbol{\theta}_1)}\right) d\nu(y) \;\geq\; 0.$$

Recall that the support of the EF does not depend on the specific choice of the canonical parameter $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$, see Remarks 2.3; this implies that the KL divergence is well-defined, here. The positivity of the KL divergence is obtained from Jensen's inequality; this is proved in Lemma 2.21, below.

The KL divergence has the interpretation of having a data model that is characterized by the distribution $f(\cdot;\boldsymbol{\theta}_0)$, and we would like to measure how close another model $f(\cdot;\boldsymbol{\theta}_1)$ is to the data model. Note that the KL divergence is not a distance function because it is neither symmetric nor does it satisfy the triangle inequality.

We calculate the KL divergence within the chosen EF

$$\begin{aligned} D_{\mathrm{KL}}(f(\cdot;\boldsymbol{\theta}_0)||f(\cdot;\boldsymbol{\theta}_1)) &= \int_{\mathbb{R}} f(y;\boldsymbol{\theta}_0) \Big[(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_1)^\top T(y) - \kappa(\boldsymbol{\theta}_0) + \kappa(\boldsymbol{\theta}_1)\Big] d\nu(y) \\ &= (\boldsymbol{\theta}_0 - \boldsymbol{\theta}_1)^\top \nabla_{\boldsymbol{\theta}} \kappa(\boldsymbol{\theta}_0) - \kappa(\boldsymbol{\theta}_0) + \kappa(\boldsymbol{\theta}_1) \;\geq\; 0, \qquad (2.21) \end{aligned}$$

where we have used Corollary 2.5, and the positivity of the KL divergence can be seen from the convexity of κ. This allows us to consider the following (Taylor) expansion

$$\kappa(\boldsymbol{\theta}_1) = \kappa(\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} \kappa(\boldsymbol{\theta}_0)^\top (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) + D_{\mathrm{KL}}(f(\cdot;\boldsymbol{\theta}_0)||f(\cdot;\boldsymbol{\theta}_1)). \qquad (2.22)$$

This illustrates that the KL divergence corresponds to second and higher order differences between the cumulant value $\kappa(\boldsymbol{\theta}_0)$ and another cumulant value $\kappa(\boldsymbol{\theta}_1)$. The gradients of the KL divergence w.r.t. $\boldsymbol{\theta}_1$ in $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$ and w.r.t. $\boldsymbol{\theta}_0$ in $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1$ are given by

$$\begin{aligned} &\nabla_{\boldsymbol{\theta}_1} D_{\mathrm{KL}}(f(\cdot;\boldsymbol{\theta}_0)||f(\cdot;\boldsymbol{\theta}_1))\big|_{\boldsymbol{\theta}_1=\boldsymbol{\theta}_0} \qquad (2.23) \\ &\quad = \nabla_{\boldsymbol{\theta}_0} D_{\mathrm{KL}}(f(\cdot;\boldsymbol{\theta}_0)||f(\cdot;\boldsymbol{\theta}_1))\big|_{\boldsymbol{\theta}_0=\boldsymbol{\theta}_1} \;=\; \mathbf{0}. \end{aligned}$$

This emphasizes that the KL divergence reflects second and higher-order terms in cumulant function κ; and that the data model $\boldsymbol{\theta}_0$ forms the minimum of this KL divergence (as a function of $\boldsymbol{\theta}_1$) as we will just see. We calculate the Hessian (second order term) w.r.t. $\boldsymbol{\theta}_1$ in $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$

$$\nabla^2_{\boldsymbol{\theta}_1} D_{\mathrm{KL}}(f(\cdot;\boldsymbol{\theta}_0)||f(\cdot;\boldsymbol{\theta}_1))\Big|_{\boldsymbol{\theta}_1=\boldsymbol{\theta}_0} = \nabla^2_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \overset{\text{def.}}{=} \mathcal{I}(\boldsymbol{\theta}_0).$$

The positive definite matrix $\mathcal{I}(\boldsymbol{\theta}_0)$ (in a minimal representation) is called *Fisher's information*. Fisher's information is an important tool in statistics that we will meet in Theorem 3.13 of Sect. 3.3, below. A function satisfying (2.21) (with being zero if and only if $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1$), fulfilling (2.23) and having positive definite Fisher's information is called *divergence*, see Definition 5 in Nielsen [285]. Fisher's information $\mathcal{I}(\boldsymbol{\theta}_0)$ measures the curvature of the KL divergence in $\boldsymbol{\theta}_0$ and we have the second order Taylor approximation

$$\kappa(\boldsymbol{\theta}_1) \;\approx\; \kappa(\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}_0)^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^\top \mathcal{I}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0).$$

Next-order terms are obtained from the so-called Amari–Chentsov tensor, see Amari [10] and Section 4.2 in Ay et al. [16]. In information geometry one studies the (possibly degenerate) Riemannian metric on the effective domain $\boldsymbol{\Theta}$ induced by Fisher's information; we refer to Section 3.7 in Nielsen [285].

Lemma 2.21 *Consider two densities p and q w.r.t. a given σ-finite measure ν. We have $D_{KL}(p||q) \geq 0$, and $D_{KL}(p||q) = 0$ if and only if $p = q$, ν-a.s.*

Proof Assume $Y \sim p d\nu$, then we can rewrite the KL divergence, using Jensen's inequality,

$$\begin{aligned} D_{\mathrm{KL}}(p||q) = \int p(y)\log\left(\frac{p(y)}{q(y)}\right)d\nu(y) \;&=\; -\mathbb{E}_p\left[\log\left(\frac{q(Y)}{p(Y)}\right)\right] \\ \geq -\log\mathbb{E}_p\left[\frac{q(Y)}{p(Y)}\right] \;&=\; -\log\int q(y)d\nu(y) \;\geq\; 0. \end{aligned} \tag{2.24}$$

Equality holds if and only if $p = q$, ν-a.s. The last inequality of (2.24) considers that q does not necessarily need to be a density w.r.t. ν, i.e., we can also have $\int q(y)d\nu(y) < 1$. □

2.3.2 Unit Deviance and Bregman Divergence

In the next chapter we are going to introduce maximum likelihood estimation for parameters, see Definition 3.4, below. Maximum likelihood estimators are obtained by maximizing likelihood functions (evaluated in the observations). Maximizing likelihood functions within the EDF is equivalent to minimizing deviance loss functions. Deviance loss functions are based on unit deviances, which, in turn, correspond to KL divergences. The purpose of this small section is to discuss this relation. This should be viewed as a preparation for Chap. 4.

Assume we work within a single-parameter linear EDF, i.e., $T(y) = y$. Using the canonical link h we obtain the canonical parameter $\theta = h(\mu) \in \boldsymbol{\Theta} \subseteq \mathbb{R}$ from the mean parameter $\mu \in \mathcal{M}$. If we replace the (typically unknown) mean parameter μ by an observation Y, supposed $Y \in \mathcal{M}$, we get the specific model that is exactly calibrated to this observation. This provides us with the canonical parameter estimate $\widehat{\theta}_Y = h(Y)$ for θ. We can now measure the KL divergence from any model represented by θ to the observation calibrated model $\widehat{\theta}_Y = h(Y)$. This KL divergence is given by (we use (2.21) and we set $\omega = v/\varphi = 1$)

$$\begin{aligned} D_{\mathrm{KL}}\left(f(\cdot; h(Y), 1) || f(\cdot; \theta, 1)\right) &= \int_{\mathbb{R}} f(y; \widehat{\theta}_Y, 1) \log\left(\frac{f(y; \widehat{\theta}_Y, 1)}{f(y; \theta, 1)}\right) d\nu(y) \\ &= (h(Y) - \theta)\, Y - \kappa(h(Y)) + \kappa(\theta) \;\geq\; 0. \end{aligned}$$

This latter object is the unit deviance (up to factor 2) of the chosen EDF. It plays a crucial role in predictive modeling.

We define the *unit deviance* under the assumption that κ is steep as follows:

$$\begin{aligned} &\mathfrak{d} : \mathring{\mathfrak{C}} \times \mathcal{M} \to \mathbb{R}_+ && (2.25) \\ &\qquad (y, \mu) \mapsto \mathfrak{d}(y, \mu) = 2\Big(y h(y) - \kappa\left(h(y)\right) - y h(\mu) + \kappa\left(h(\mu)\right)\Big) \;\geq\; 0, \end{aligned}$$

where $\mathfrak{C}$ is the convex closure of the support $\mathfrak{T}$ of Y and $\mathcal{M}$ is the dual parameter space of the chosen EDF. Steepness of κ implies $\mathring{\mathfrak{C}} = \mathcal{M}$, see Theorem 2.19.

This unit deviance $\mathfrak{d}$ is received from the KL divergence, and it is (twice) the difference of two log-likelihood functions, one using canonical parameter $h(y)$ and the other one having any canonical parameter $\theta = h(\mu) \in \mathring{\boldsymbol{\Theta}}$. That is, for $\mu = \kappa'(\theta)$,

$$\begin{aligned} \mathfrak{d}(y, \mu) &= 2\, D_{\mathrm{KL}}(f(\cdot; h(y), 1) || f(\cdot; \theta, 1)) && (2.26) \\ &= 2\, \frac{\varphi}{v}\, \left(\log f(y; h(y), v/\varphi) - \log f(y; \theta, v/\varphi)\right), \end{aligned}$$

for general $\omega = v/\varphi \in \mathcal{W}$. The latter can be rewritten as

$$f(y; \theta, v/\varphi) = f(y; h(y), v/\varphi)\, \exp\left\{-\frac{1}{2\varphi/v}\mathfrak{d}(y, \kappa'(\theta))\right\}. \tag{2.27}$$

This looks like a generalization of the Gaussian distribution, where the square difference $(y - \mu)^2$ in the exponent is replaced by the unit deviance $\mathfrak{d}(y, \mu)$ with $\mu = \kappa'(\theta)$. This interpretation gets further support by the following lemma.

Lemma 2.22 *Under Assumption 2.6 and the assumption that the cumulant function κ is steep, the unit deviance $\mathfrak{d}\,(y,\mu) \geq 0$ of the chosen EDF is zero if and only if $y = \mu$. Moreover, the unit deviance $\mathfrak{d}\,(y,\mu)$ is twice continuously differentiable w.r.t. (y,μ) in $\mathring{\mathfrak{C}} \times \mathcal{M}$, and*

$$\left.\frac{\partial^2 \mathfrak{d}\,(y,\mu)}{\partial \mu^2}\right|_{y=\mu} = \left.\frac{\partial^2 \mathfrak{d}\,(y,\mu)}{\partial y^2}\right|_{y=\mu} = -\left.\frac{\partial^2 \mathfrak{d}\,(y,\mu)}{\partial \mu \partial y}\right|_{y=\mu} = 2/V(\mu) > 0.$$

Proof The positivity and the if and only if statement follows from Lemma 2.21 and the strict convexity of κ. Continuous differentiability follows from the smoothness of κ in the interior of $\boldsymbol{\Theta}$. Moreover we have

$$\left.\frac{\partial^2 \mathfrak{d}\,(y,\mu)}{\partial \mu^2}\right|_{y=\mu} = \left.\frac{\partial}{\partial \mu} 2\left(-yh'(\mu) + \mu h'(\mu)\right)\right|_{y=\mu} = 2h'(\mu) = 2/\kappa''(h(\mu)) = 2/V(\mu) > 0,$$

where $V(\mu)$ is the variance function of the chosen EDF introduced in Corollary 2.14. The remaining second derivatives are received by similar (straightforward) calculations. □

Remarks 2.23

- Lemma 2.22 shows that the unit deviance definition of $\mathfrak{d}(y,\mu)$ provides a so-called regular unit deviance according to Definition 1.1 in Jørgensen [203]. Moreover, any model that can be brought into the form (2.27) for a (regular) unit deviance is called (regular) reproductive dispersion model, see Definition 1.2 of Jørgensen [203].
- In general the unit deviance $\mathfrak{d}(y,\mu)$ is not symmetric in its two arguments y and μ, we come back to this in Fig. 11.1, below.

More generally, the KL divergence and the unit deviance can be embedded into the framework of Bregman loss functions [50]. We restrict to the single-parameter EDF case. Assume that $\psi : \mathring{\mathfrak{C}} \to \mathbb{R}$ is a strictly convex function. The *Bregman divergence* w.r.t. ψ between y and μ is defined by

$$D_\psi(y,\mu) = \psi(y) - \psi(\mu) - \psi'(\mu)\,(y-\mu) \;\geq\; 0, \tag{2.28}$$

where ψ' is a (sub-)gradient of ψ. The lower bound holds because of convexity of ψ. Consider the specific choice $\psi(\mu) = \mu h(\mu) - \kappa(h(\mu))$ for the chosen EDF. Similar to Lemma 2.22 we have $\psi''(\mu) = h'(\mu) = 1/V(\mu) > 0$, which says that this choice is strictly convex. Using this choice for ψ gives us unit deviance (up to factor $1/2$)

$$D_\psi(y,\mu) = yh(y) - \kappa(h(y)) + \kappa(h(\mu)) - h(\mu)y = \frac{1}{2}\mathfrak{d}(y,\mu). \tag{2.29}$$

Thus, the unit deviance $\mathfrak{d}$ can be understood as a difference of log-likelihoods (2.26), as a KL divergence D_{KL} and as a Bregman divergence D_ψ.

Example 2.24 (Poisson Model) We start with a single-parameter EF example. Consider cumulant function $\kappa(\theta) = \exp\{\theta\}$ for canonical parameter $\theta \in \boldsymbol{\Theta} = \mathbb{R}$, this gives us the Poisson model. For the KL divergence from model θ_1 to model θ_0 we receive

$$D_{\mathrm{KL}}(f(\cdot;\theta_0)||f(\cdot;\theta_1)) = \exp\{\theta_1\} - \exp\{\theta_0\} - (\theta_1 - \theta_0)\exp\{\theta_0\} \;\ge\; 0,$$

which is zero if and only if $\theta_0 = \theta_1$. Fisher's information is given by

$$\mathcal{I}(\theta) = \kappa''(\theta) = \exp\{\theta\} > 0.$$

If we have observation $Y > 0$ we receive a model described by canonical parameter $\widehat{\theta}_Y = h(Y) = \log(Y)$. This gives us unit deviance, see (2.26),

$$\begin{aligned}\mathfrak{d}(Y,\mu) &= 2D_{\mathrm{KL}}(f(\cdot;h(Y),1)||f(\cdot;\theta,1))\\ &= 2\left(e^{\theta} - Y - (\theta - \log(Y))Y\right)\\ &= 2\left(\mu - Y - Y\log\left(\frac{\mu}{Y}\right)\right) \;\ge\; 0,\end{aligned}$$

with $\mu = \kappa'(\theta) = \exp\{\theta\}$. This Poisson unit deviance will commonly be used for model fitting and forecast evaluation, see, e.g., (5.28). ■

Example 2.25 (Gamma Model) The second example considers a vector-valued parameter EF example. We consider the cumulant function $\kappa(\boldsymbol{\theta}) = \log\Gamma(\theta_2) - \theta_2\log(-\theta_1)$ for $\boldsymbol{\theta} = (\theta_1,\theta_2)^\top \in \boldsymbol{\Theta} = (-\infty,0)\times(0,\infty)$; this gives us the gamma model, see Sect. 2.1.3. For the KL divergence from model $\boldsymbol{\theta}_1$ to model $\boldsymbol{\theta}_0$ we receive

$$\begin{aligned}D_{\mathrm{KL}}(f(\cdot;\boldsymbol{\theta}_0)||f(\cdot;\boldsymbol{\theta}_1)) &= \left(\theta_{0,2} - \theta_{1,2}\right)\frac{\Gamma'(\theta_{0,2})}{\Gamma(\theta_{0,2})} - \log\left(\frac{\Gamma(\theta_{0,2})}{\Gamma(\theta_{1,2})}\right)\\ &\quad + \theta_{1,2}\log\left(\frac{-\theta_{0,1}}{-\theta_{1,1}}\right) + \theta_{0,2}\left(\frac{-\theta_{1,1}}{-\theta_{0,1}} - 1\right) \;\ge\; 0.\end{aligned}$$

Fisher's information matrix is given by

$$\mathcal{I}(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) = \begin{pmatrix}\frac{\theta_2}{(-\theta_1)^2} & \frac{1}{-\theta_1}\\ \frac{1}{-\theta_1} & \frac{\Gamma''(\theta_2)\Gamma(\theta_2)-\Gamma'(\theta_2)^2}{\Gamma(\theta_2)^2}\end{pmatrix}.$$

The off-diagonal terms in Fisher's information matrix $\mathcal{I}(\boldsymbol{\theta})$ are non-zero which means that the two components of the canonical parameter $\boldsymbol{\theta}$ interact. Choosing a different parametrization $\mu = \theta_2/(-\theta_1)$ (dual mean parametrization) and $\alpha = \theta_2$ we receive diagonal Fisher's information in (μ, α)

$$\mathcal{I}(\mu, \alpha) = \begin{pmatrix} \frac{\alpha}{\mu^2} & 0 \\ 0 & \frac{\Gamma''(\alpha)\Gamma(\alpha)-\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} - \frac{1}{\alpha} \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{\mu^2} & 0 \\ 0 & \Psi'(\alpha) - \frac{1}{\alpha} \end{pmatrix}, \tag{2.30}$$

where Ψ is the digamma function, see Footnote 2 on page 22. This transformation is obtained by using the corresponding Jacobian matrix for variable transformation; more details are provided in (3.16) below. In this new representation, the parameters μ and α are orthogonal; the term $\Psi'(\alpha) - \frac{1}{\alpha}$ is further discussed in Remarks 5.26 and Remarks 5.28, below.

Using this second parametrization based on mean μ and dispersion $1/\alpha$, we arrive at the EDF representation of the gamma model. This allows us to calculate the corresponding unit deviance (within the EDF), which in the gamma case is given by

$$\mathfrak{d}(Y, \mu) = 2\left(\frac{Y}{\mu} - 1 + \log\left(\frac{\mu}{Y}\right)\right) \ \geq \ 0.$$

■

Example 2.26 (Inverse Gaussian Model) Our final example considers the inverse Gaussian vector-valued parameter EF case. We consider the cumulant function $\kappa(\boldsymbol{\theta}) = -2(\theta_1\theta_2)^{1/2} - \frac{1}{2}\log(-2\theta_2)$ for $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top \in \boldsymbol{\Theta} = (-\infty, 0] \times (-\infty, 0)$, see Sect. 2.1.3. For the KL divergence from model $\boldsymbol{\theta}_1$ to model $\boldsymbol{\theta}_0$ we receive

$$\begin{aligned} D_{\mathrm{KL}}(f(\cdot;\boldsymbol{\theta}_0)||f(\cdot;\boldsymbol{\theta}_1)) = &-\theta_{1,1}\sqrt{\frac{-\theta_{0,2}}{-\theta_{0,1}}} - \theta_{1,2}\sqrt{\frac{-\theta_{0,1}}{-\theta_{0,2}}} - 2\sqrt{\theta_{1,1}\theta_{1,2}} \\ &+ \frac{\theta_{0,2} - \theta_{1,2}}{-2\theta_{0,2}} + \frac{1}{2}\log\left(\frac{-\theta_{0,2}}{-\theta_{1,2}}\right) \ \geq \ 0. \end{aligned}$$

Fisher's information matrix is given by

$$\mathcal{I}(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) = \begin{pmatrix} \frac{(-2\theta_2)^{1/2}}{(-2\theta_1)^{3/2}} & -\frac{1}{2(\theta_1\theta_2)^{1/2}} \\ -\frac{1}{2(\theta_1\theta_2)^{1/2}} & \frac{(-2\theta_1)^{1/2}}{(-2\theta_2)^{3/2}} + \frac{2}{(-2\theta_2)^2} \end{pmatrix}.$$

Again the off-diagonal terms in Fisher's information matrix $\mathcal{I}(\boldsymbol{\theta})$ are non-zero in the canonical parametrization. We switch to the mean parametrization by setting

$\mu = (-2\theta_2/(-2\theta_1))^{1/2}$ and $\alpha = -2\theta_2$. This provides us with diagonal Fisher's information

$$\mathcal{I}(\mu, \alpha) = \begin{pmatrix} \frac{\alpha}{\mu^3} & 0 \\ 0 & \frac{1}{2\alpha^2} \end{pmatrix}. \tag{2.31}$$

This transformation is again obtained by using the corresponding Jacobian matrix for variable transformation, see (3.16), below. We compare the lower-right entries of (2.30) and (2.31). Remark that we have first order approximation of the digamma function

$$\Psi(\alpha) \approx \log\alpha - \frac{1}{2\alpha},$$

and taking derivatives says that these entries of Fisher's information are first order equivalent; this is also used in the saddlepoint approximation in Sect. 5.5.2, below. Using this second parametrization based on mean μ and dispersion $1/\alpha$, we arrive at the EDF representation of the inverse Gaussian model with unit deviance

$$\mathfrak{d}(Y, \mu) = \frac{(Y-\mu)^2}{\mu^2 Y} \geq 0.$$

■

More examples will be given in Chap. 4, below.