On Rényi Divergence Measures for Continuous Alphabet Sources

by

Manuel Gil

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
the degree of Masters of Science

Queen's University
Kingston, Ontario, Canada
August 2011

Copyright © Manuel Gil, 2011

Abstract

The idea of 'probabilistic distances' (also called divergences), which in some sense assess how 'close' two probability distributions are from one another, has been widely employed in probability, statistics, information theory, and related fields. Of particular importance due to their generality and applicability are the Rényi divergence measures. While the closely related concept of Rényi entropy of a probability distribution has been studied extensively, and closed-form expressions for the most common univariate and multivariate continuous distributions have been obtained and compiled [57, 45, 62], the literature currently lacks the corresponding compilation for continuous Rényi divergences. The present thesis addresses this issue for the analytically tractable cases. Closed-form expressions for Kullback-Leibler divergences are also derived and compiled, as they can be seen as an extension by continuity of the Rényi divergences. Additionally, we establish a connection between Rényi divergence and the variance of the log-likelihood ratio of two distributions, which extends the work of Song [57] on the relation between Rényi entropy and the log-likelihood function, and which becomes practically useful in light of the Rényi divergence expressions we have derived. Lastly, we consider the Rényi divergence rate between two stationary Gaussian processes.

Acknowledgments

I would firstly like to express my gratitude towards my two supervisors, Professors Fady Alajaji and Tamas Linder, for accepting me as their graduate student and for their guidance during the completion of this thesis.

I am also grateful to the School of Graduate Studies, the Department of Mathematics and Statistics, and NSERC for their funding; to the staff in the Department of Mathematics and Statistics for their always friendly assistance, in particular Jennifer Read in her role of Graduate Secretary; and to those graduate students who have in one way or another been supportive and instructive, especially Nevroz Sen, whose support and advice has been very valuable during challenging times, as well as Caroline Seguin for her friendly encouragement.

I am always immensely grateful to my parents for their unconditional support.

Table of Contents

List of Tables

Chapter 1

Introduction

In this chapter we give an overview of important measures of information and introduce the concept of probabilistic distances[1]. In Section 1.1 we provide a general overview of these notions, while the Rényi and Kullback-Leibler divergences are discussed in more detail in Section 1.2. Finally, a description of the main results of the present work as well as a literature review of the relevant topics are given in Section 1.3.

1.1 Probability Distances and Measures of Information

Claude Shannon's 1948 paper '*A Mathematical Theory of Communication*' [56] introduced a powerful mathematical framework to quantify our intuitive notion of information, laying the foundations for the field of information theory and originating a major revolution in communications and related fields. The power of the paradigm introduced by Shannon is reflected in the two results known as the *Source Coding Theorem*

[1] Also called statistical distances, probabilistic divergences, or divergence measures.

and the *Channel Coding Theorem*.

With the Source Coding Theorem, Shannon demonstrated that all discrete alphabet random processes possess an irreducible complexity below which a signal cannot be compressed without loss of information; such amount of complexity is known as the source's *entropy*. In the case of a discrete distribution with probability mass function (pmf) $p(x)$ over an alphabet $\mathscr{X}$, the entropy is defined as

$$H(p) = -\sum_{x\in\mathscr{X}} p(x)\log p(x) = -E_p\left[\log p(x)\right] .$$

For a continuous distribution with a density $f(x)$ one considers the *differential entropy*

$$h(f) = -\int_{\mathscr{X}} f(x)\ln f(x)\,dx = -E_f\left[\ln f(x)\right] ,$$

but unlike in the discrete case, the entropy (i.e.,. the irreducible complexity of the source) is not given by the differential entropy. However, other operational interpretations similar to those holding in the discrete case do extend to differential entropy.[2]

Shannon's axiomatic derivation of the entropy functional as a measure of information was ensued by the introduction of a myriad of other information measures following a similar approach, where the specific axioms to be introduced would have some commonality with Shannon's but would be motivated within sometimes very specialized settings [30]. A survey of axiomatic characterizations of information measures can be found for example in [18, 2].

In the same way that entropy-like functionals have been widely investigated as measures of the amount of information intrinsic to a given probability distribution, it is natural to investigate similarly defined functionals which allow one to somehow

[2]For example, in the context of the Asymptotic Equipartition Property, the differential entropy also provides bounds for the size of a typical sets, $A_\epsilon^{(n)}$. See for example [15] for a discussion of this idea as well as a more detailed introduction to the coding theorems, in particular the Channel Coding Theorem which we omit here for brevity.

quantify how much information is shared between two probability distributions. The extent of this shared information may also be seen as providing a certain measure of how 'close'[3] two distributions are from one-another. As pointed out by Liese and Vajda [40], the origins of these ideas go back to the early 1900s literature in the works of Pearson [47] and Hellinger [31], although research in this area became much more prolific after the publication of Shannon's 1948 paper.

Motivated by Shannon's notion of mutual information [56], Kullback and Leibler [38] introduced the information measure now known as the Kullback-Leibler Divergence (KLD) within the context of hypothesis testing. The authors consider two probability spaces $(\mathscr{X}, \mathscr{A}, \mu_i)$, $i = 1, 2$, such that $\mu_1 \equiv \mu_2$[4] and λ a probability measure such that $\lambda \equiv \{\mu_1, \mu_2\}$.[5] Denote the corresponding Radon-Nikodym derivatives by $f_i(x)$, and let H_i be the hypothesis that an observation x came from μ_i. Kullback and Leibler define the *mean information for discrimination between H_1 and H_2 per observation from μ_1*[6] as

$$I(\mu_1 : \mu_2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} \, d\lambda(x) \, ,$$

[3]It is worth noting that although the terms *distance* and even sometimes *metric* are used within this context in the literature, these functionals do not generally satisfy all the properties required of a mathematical metric; in particular, symmetry is often not met. For example, according to [20], the general requirements for a probabilistic distance is that it be 'positive, zero if the values of the two functions coincide, and correlated to their absolute difference'.

[4]Given two measures μ and ν over the same $\sigma-$algebra $\mathscr{A}$, ν is said to be *absolutely continuous with respect to* μ (or equivalently μ *dominates* ν) if $\forall A \in \mathscr{A}, \; \mu(A) = 0 \Rightarrow \nu(A) = 0$, and this is denoted by $\mu \gg \nu$. Whenever $\mu \gg \nu$ and $\nu \gg \mu$ this is denoted by $\mu \equiv \nu$.

[5]It is worth noting that these requirements vary slightly from the modern definition given in the literature (e.g. [34]), which we provide in Section 1.2.

[6]This original nomenclature now replaced by the more concise 'Kullback-Leibler divergence', and the notation usually replaced by $D(f_1||f_2)$ or $H(\mu_1, \mu_2)$. Also, although the word *between* is generally used, this divergence is directed and not symmetric, so that it would be more correct to say 'the divergence *from f_1 to f_2*'.

and it is shown to generalize Shannon's original notion of mutual information.

Further generalization came about in the 1961 work of Rényi [53], who introduced an indexed family of generalized information and divergence measures akin to the Shannon entropy and Kullback-Leibler divergence. Originally considering discrete probability distributions, Rényi introduced the *entropy of order* α of a distribution $P = \{p_1, ..., p_n\}$ as

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \left(\sum_{k=1}^{n} p_k^\alpha \right) ,$$

and for two discrete distributions P and Q, '*the information of order* α *obtained if the distribution* P *is replaced by the distribution* Q'[7] by

$$I_\alpha(P|Q) = \frac{1}{\alpha - 1} \log \left(\sum_{k=1}^{n} p_k^\alpha q_k^{1-\alpha} \right) , \alpha > 0 \ \text{ and } \ \alpha \neq 1 .$$

An important property of this family of information measures is that [53]

$$\lim_{\alpha \to 1} H_\alpha(P) = H(P) , \ \text{ and } \ \lim_{\alpha \to 1} I_\alpha(P|Q) = I(P : Q) .$$

It is worth pointing out that, prior to Rényi's paper, Chernoff introduced another measure of divergence which he derived by considering a certain class of hypothesis tests in his 1952 work [14]. His approach was similar as that of Kullback and Leibler in defining the information divergence. When considering two probabilities measures μ_i and μ_j the measure of divergence used by Chernoff was

$$D = -\log \left[\inf_{0<t<1} \int [f_i(x)]^t [f_j(x)]^{1-t} \, d\nu \right] ,$$

where f_i and f_j are the Radon-Nykodim derivatives of μ_i and μ_j with respect to a dominating measure ν. Some of the literature (e.g. [8, 52, 20]) identifies the Chernoff

[7] This is now known as the Rényi divergence of order α, and it is usually denoted by $D_\alpha(P||Q)$. We introduce the general definition for general probability spaces in Section 1.2.

distance as an indexed family of divergences

$$D_C(f_i||f_j;\lambda) = -\ln \int_{\mathscr{X}} f_i(x)^{\lambda} f_j(x)^{1-\lambda}\, d\nu(x)\ ,$$

where for a particular choice of λ the above is called the Chernoff distance of order λ. As a special case, the Bhattacharyya distance, $D_B(f_i||f_j)$, (also known as Bhattacharyya coefficient, ρ) [9] is given by

$$D_B(f_i||f_j) = D_C(f_i||f_j;\lambda = 1/2)\ .$$

Chernoff divergences are used in statistics, artificial intelligence, pattern recognition, and related fields (see for example [4, 20, 52]). We note that a definiton of Rényi divergence for general probability spaces (see Section 1.2) establishes the following relationship

$$D_C(f_i||f_j;\alpha) = (1-\alpha)D_{\alpha}(f_i||f_j)\ , \quad \alpha \in (0,1)\ ,$$

so that up to scaling the two divergences are measuring the same amount of 'information' between any two densities f_i and f_j.

Yet a higher level of generalization in the area of probabilistic divergences was achieved by the work of Csiszar [16] (and independently also Ali and Silvey [5]), who introduced the notion of $f-$divergences of probability distributions, a framework which encompasses a vast number of information measures used currently in the literature, including the Kullback divergence, and also divergences which are one-to-one functions of Rényi divergences. Liese and Vajda [40, 59, 41] have studied this formalism and its applications extensively. We omit a discussion of $f-$divergences here as it is not immediately relevant to the results of this work.

To this day a vast number of probability distance measures have been investigated [2, 6, 8, 42, 20, 24]. In Table 1.1 we present a brief sample of the most common probablistic distances, in particular those expressed as integrals of the corresponding densities. For more comprehensive overviews see the references above. We denote by X_i the support of $f_i(x)$, i.e., $X_i := \{x : f_i(x) \neq 0\}$. Note that the *Hellinger distance* in [24] is listed as *Jeffreys-Matusita* distance in [52, 20], and even different still are what Liese and Vajda [41] identify as the *Hellinger divergences*. Also, some authors (including Rényi [53]) restrict α to be a positive real number (not equal to one) in the definition of the Rényi information measures as a result of information theoretical considerations, although the definition can be extended mathematically to $\alpha \in \mathbb{R}$ [40].

Table 1.1: Probabilistic Divergences.

Divergence Name	**Mathematical Definition**
Bhattacharyya	$D_B(f_i\|\|f_j) = -\ln \int_{X_i} \sqrt{f_i(x) f_j(x)}\, dx$
Chernoff	$D_C(f_i\|\|f_j) = -\ln \int_{X_i} f_i(x)^{\lambda} f_j(x)^{1-\lambda}\, dx\ ,\ \ \lambda \in (0,1)$
χ^2	$D_{\chi^2}(f_i\|\|f_j) = \int_{X_i \cup X_j} \frac{(f_i(x) - f_j(x))^2}{f_j(x)}\, dx$
Jeffreys-Matusita	$D_J(f_i\|\|f_j) = \left[\int_{X_i} \left(\sqrt{f_i(x)} - \sqrt{f_j(x)} \right)^2 dx \right]^{1/2}$
Kullback-Liebler	$D_K(f_i\|\|f_j) = \int_{X_i} f_i(x) \ln \frac{f_i(x)}{f_j(x)}\, dx$
Generalized Matusita	$D_M(f_i\|\|f_j) = \left[\int_{X_i} \left\| f_i(x)^{1/r} - f_j(x)^{1/r} \right\|^r dx \right]^{1/r}\ ,\ r > 0$
Rényi	$D_\alpha(f_i\|\|f_j) = \frac{1}{\alpha - 1} \ln \int_{X_i} f_i(x)^{\alpha} f_j(x)^{1-\alpha} dx\ ,\ \alpha \in \mathbb{R}^+ \setminus \{1\}$
Varational	$V(f_i\|\|f_j) = \int_{X_i \cup X_j} \|f_i(x) - f_j(x)\|\, dx\ .$

A natural question is whether probabilistic distances can be generalized to stochastic processes. This leads to the consideration of *information measure rates*. For example, given a process $X = \{X_i\}_{i\in\mathbb{N}}$ over a discrete alphabet $\mathscr{X}$, the limit

$$H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, ..., X_n)$$

is defined as the *entropy rate* (or just entropy) of X, whenever it exists. For stationary processes, the entropy rate always exists [15]. The same idea can be applied to other information measures, such as the Kullback divergence or the Rényi information measures. We provide a more extensive discussion of this topic in Chapter 4.

1.2 Rényi and Kullback Divergence

In this section we give the general definitions of Rényi entropy and Rényi divergence, Shannon differential entropy, and the Kullback-Leibler divergence, and we also present some of the mathematical properties of the divergence measures.

Throughout this section, let $(\mathscr{X}, \mathscr{A})$ be a measurable space and P and Q be two probability measures on $\mathscr{A}$ with densities p and q relative to a $\sigma-$finite dominating measure μ (i.e., $p \ll \mu$ and $q \ll \mu$). In all of the above we use the conventions $p^\alpha q^{1-\alpha} = 0$ if $p = q = 0$, $x/0 = \infty$ for $x > 0$, and $0 \ln 0 = 0 \ln(0/0) = 0$, which are justified by continuity arguments. Also, from here onwards we denote the nonnegative real numbers by $\mathbb{R}^+$.

1.2.1 Shannon Entropy and Kullback-Leibler Divergence

This material can be found for example in chapter 1 of [34].

Definition 1.2.1. If P corresponds to a continuous probability distribution over $\mathbb{R}^n$ with density $p(\boldsymbol{x})$, the *differential (Shannon) entropy* of P, denoted $h(P)$[8] is defined by

$$h(P) = -\int_{\mathbb{R}^n} p(\boldsymbol{x}) \log p(\boldsymbol{x})\, d\boldsymbol{x} \ .$$

Definition 1.2.2. The *Kullback-Leibler Divergence* (KLD) between P and Q, denoted $D(P||Q)$ (or equivalently the Kullback-Leibler divergence between p and q, denoted $D(p||q)$) is defined by

$$D(P||Q) = \int_{\mathscr{X}} p(x) \ln \frac{p(x)}{q(x)}\, d\mu(x) \ .$$

Sometimes the literature refers to $D(P||Q)$ as the *relative entropy* of Q with respect to P. Since $D(p||q)$ is finite only when supp $p \subseteq$ supp q, $D(p||q)$ is sometimes written as

$$D(p||q) = \begin{cases} \int_{\mathscr{X}} p(x) \ln \frac{p(x)}{q(x)}\, d\mu(x) & P \ll Q \\ \infty & \text{otherwise} \ . \end{cases}$$

Proposition 1.2.3. *$D(P||Q) \geq 0$ and equality holds iff $P = Q$.*

This follows from the inequality $-\ln x \geq 1 - x \ , \forall x > 0$ where equality hold iff $x = 1$. Then we have

$$\begin{aligned} D(P||Q) &= \int_{\mathscr{X}} p(x) \ln \frac{p(x)}{q(x)}\, d\mu(x) \\ &= \int_{\mathscr{X}} p(x) \left[-\ln \frac{q(x)}{p(x)} \right] d\mu(x) \geq \int_{\mathscr{X}} p(x) \left[1 - \frac{q(x)}{p(x)} \right] d\mu(x) = 0 \ , \end{aligned}$$

with equality iff $p(x) = q(x)$ μ−almost everywhere.

[8] Sometimes called the *continuous entropy* of P, and also denoted as $h(p)$, or $h(\boldsymbol{X})$ if $\boldsymbol{X}$ is the continuous random vector having distribution P.

Remark 1.2.4. If $p(x)$ has finite differential entropy, then

$$D(P||Q) = \int_{\mathscr{X}} p(x) \ln p(x)\, d\mu(x) - \int_{\mathscr{X}} p(x) \ln q(x)\, d\mu(x) = -h(P) - E_p\left[\ln q(X)\right] .$$

If $\Delta = \{A_1, \ldots, A_n\}$ is a partition on $\mathscr{X}$, then *relative entropy associated with* Δ is defined as

$$D_\Delta(P||Q) = \sum_{i=1}^{n} P(A_i) \log \frac{P(A_i)}{Q(A_i)} .$$

Theorem 1.2.5. *Let $\mathscr{P}$ be the set of all finite partitions of X. Then*

$$D(P||Q) = \sup_{\Delta \in \mathscr{P}} D_\Delta(P||Q) .$$

Theorem 1.2.6. *$D(P||Q)$ is convex in the pair (P, Q), and for any fixed Q, $D(P||Q)$ is strictly convex in P.*

1.2.2 Rényi Information Measures

This material can be found in [53, 60, 41].

Definition 1.2.7. For $\alpha \in \mathbb{R}^+ \setminus \{1\}$[9] the *Rényi entropy* of order α of P, denoted $h_\alpha(P)$[10] is defined by

$$h_\alpha(P) = \frac{1}{1-\alpha} \ln \int_{\mathscr{X}} p(x)^\alpha\, d\mu(x) .$$

Definition 1.2.8. For $\alpha \in \mathbb{R}^+ \setminus \{1\}$ the *Rényi divergence* of order α between P and Q, denoted $D_\alpha(P||Q)$ (or equivalently the Rényi divergence of order α between p and q, denoted $D_\alpha(p||q)$) is defined by

$$D_\alpha(P||Q) = \frac{1}{\alpha-1} \ln \int_{\mathscr{X}} p(x)^\alpha q(x)^{1-\alpha} d\mu(x) .$$

[9]See comment at the end of Section 1.1 regarding the domain for α in the definition of the Rényi information measures.

[10] Equivalently the Rényi entropy of order α of p, denoted $h_\alpha(p)$.

By continuity and the following proposition the definition can be extended to $\alpha = 1$ and $\alpha = 0$:

Proposition 1.2.9.

$$D_0(P||Q) := \lim_{\alpha \downarrow 0} D_\alpha(f_i||f_j) = -\log Q(p > 0) \ ,$$

$$D_1(P||Q) := \lim_{\alpha \uparrow 1} D_\alpha(P||Q) = D(P||Q) \ .$$

Proposition 1.2.10. *(Data Processing Inequality) Let Σ be a $\sigma-$subalgebra of $\mathscr{A}$*[11] *and denote by P_Σ and Q_Σ the restrictions of P and Q to Σ. Then*

$$D_\alpha(P_\Sigma||Q_\Sigma) \leq D_\alpha(P||Q) \ .$$

Corollary 1.2.11. *If we set $\Sigma = \{\emptyset, \mathscr{X}\}$ then $P_\Sigma = Q_\Sigma$ and we obtain*

$$D_\alpha(P||Q) \geq 0 \ ,$$

and $D_\alpha(P||Q) = 0 \Leftrightarrow p = q$, $\mu-$almost surely.

Just like the Kullback-Leibler divergence, the Rényi divergence can be approximated arbitrarily closely by the corresponding divergence over finite partitions.

Theorem 1.2.12. *Let $\mathscr{P}$ be the set of all finite partitions of X. Then*

$$D_\alpha(P||Q) = \sup_{\Delta \in \mathscr{P}} D_\alpha(P_{\Sigma_\Delta}||Q_{\Sigma_\Delta}) \ ,$$

where Σ_Δ is the $\sigma-$algebra generated by a finite partition Δ.

Another important property of Rényi divergence is additivity in the following sense.

[11] *A subset of $\mathscr{A}$ which is itself a $\sigma-$algebra.*

Proposition 1.2.13. *For $i = 1, ..., n$ let $(X_i, \mathscr{A}_i)$ be a measurable space, P_i and Q_i be two probability measures on X_i, and denote the product measure*[12] *on $X_1 \times X_2 \times .... \times X_n$ by $\prod_{i=1}^{n} P_i$. Then*

$$D_\alpha\left(\prod_{i=1}^{n} P_i \middle|\middle| \prod_{i=1}^{n} Q_i\right) = \sum_{i=1}^{n} D_\alpha(P_i||Q_i) \; .$$

Proposition 1.2.14. *(Continuity) $D_\alpha(P||Q)$ is continuous in α on*

$$A = \{\alpha : 0 \leq \alpha \leq 1 \text{ or } D_\alpha(P||Q) < \infty\} \; .$$

Proposition 1.2.15. *(Joint Convexity) For $\alpha \in [0,1]$, $D_\alpha(P||Q)$ is convex in the pair (P,Q).*

While joint convexity is limited to $\alpha \in [0,1]$, the following holds for general $\alpha > 0$:

Proposition 1.2.16. *(Convexity in Q) For all positive α, $D_\alpha(P||Q)$ is convex in Q.*

Remark 1.2.17. The integral

$$\mathscr{H}_\alpha(P,Q) = \int_{\mathscr{X}} p(x)^\alpha q(x)^{1-\alpha} d\mu(x) \; , \; \alpha > 0$$

is usually known as the *Hellinger integral* of order α. Also, the *power divergence* [60] or *Hellinger divergence* [41] is defined as

$$H_\alpha(P||Q) = \frac{\mathscr{H}_\alpha(P,Q) - 1}{\alpha - 1} \quad \alpha > 0 \, , \alpha \neq 1 \; .$$

Since

$$\exp\big((\alpha-1)D_\alpha(P||Q)\big) = \mathscr{H}_\alpha(P,Q) \iff H_\alpha(P||Q) = \frac{\exp\big((\alpha-1)D_\alpha(P||Q)\big) - 1}{\alpha - 1} \; ,$$

H_α is a strictly increasing function of D_α for $\alpha > 0$, $\alpha \neq 1$, and also $D_\alpha = 0 \Leftrightarrow H_\alpha = 0$.

[12] *The unique measure P on the product $\sigma-$algebra generated by $\{E_1 \times E_2 \times ... \times E_n : E_i \in \mathscr{A}_i\}$ satisfying $P(E_1 \times ... \times E_n) = P_1(E_1) \times ... \times P_n(E_n)$.*

A more complete study of the mathematical properties of Rényi divergences is beyond the scope of this work. See [60, 53] and also [40] for a more in depth treatment.

1.2.3 Rényi Divergence for Natural Exponential Families

In Chapter 2 of their 1987 book *Convex Statistical Distances* [40], Liese and Vajda derive a closed-form expression for the Rényi divergence between two members of a canonical exponential family, which is presented below. Note that their definition of Rényi divergence, here denoted by $R_\alpha(f_i||f_j)$, differs by a factor of α from the one considered in this work, i.e., $D_\alpha(f_i||f_j) = \alpha R_\alpha(f_i||f_j)$.

Consider a natural exponential family (see Definition A.2.2) of probability measures P_τ on $\mathbb{R}^n$ having densities $p_\tau = \dfrac{1}{C(\boldsymbol{\tau})}\exp\langle \boldsymbol{\tau}, \boldsymbol{T}(\boldsymbol{x})\rangle$, and natural parameter space Θ (Definition A.2.3).

Proposition 1.2.18. *Let $D(\boldsymbol{\tau}) = \ln C(\boldsymbol{\tau})$. For every $\boldsymbol{\tau}_i, \boldsymbol{\tau}_j \in \Theta$ the limit*

$$\Delta(\boldsymbol{\tau}, \boldsymbol{\tau}_j) := \lim_{\alpha\downarrow 0} \frac{1}{\alpha}\left[\alpha D(\boldsymbol{\tau}_i) + (1-\alpha) D(\boldsymbol{\tau}_j) - D(\alpha\boldsymbol{\tau}_i + (1-\alpha)\boldsymbol{\tau}_j)\right]$$

exists in $[0, \infty]$.

Proof. See [40]. □

Theorem 1.2.19. *Let P_τ be an exponential family with natural parameters where $\boldsymbol{\tau}_i, \boldsymbol{\tau}_j \in \Theta$, having corresponding densities f_i and f_j. Then $R_\alpha(f_i||f_j)$ is given by the following cases:*

1. *If $\alpha \notin \{0, 1\}$ and $\alpha\boldsymbol{\tau}_i + (1-\alpha)\boldsymbol{\tau}_j \in \Theta$*

$$R_\alpha(f_i||f_j) = \frac{1}{\alpha(\alpha-1)} \ln \frac{C(\alpha\boldsymbol{\tau}_i + (1-\alpha)\boldsymbol{\tau}_j)}{C(\boldsymbol{\tau}_i)^\alpha C(\boldsymbol{\tau}_j)^{1-\alpha}}\ .$$

2. *If* $\alpha \notin \{0,1\}$ *and* $\alpha\boldsymbol{\tau}_i + (1-\alpha)\boldsymbol{\tau}_j \notin \Theta$

$$R_\alpha(f_i||f_j) = +\infty \ .$$

3. *If* $\alpha = 0$

$$R_\alpha(f_i||f_j) = \Delta(\boldsymbol{\tau}_i, \boldsymbol{\tau}_j)$$

4. *If* $\alpha = 1$

$$R_\alpha(f_i||f_j) = \Delta(\boldsymbol{\tau}_j, \boldsymbol{\tau}_i) \ ,$$

with $\Delta(\boldsymbol{\tau}_i, \boldsymbol{\tau}_j)$ *defined as in Proposition 1.2.18*

Proof. See [40]. □

Using this result we arrive at the corresponding expression for $D_\alpha(f_i||f_j)$, which we write in a form that facilitates the comparison to the expressions from Appendix B:

Corollary 1.2.20. *Let* $\boldsymbol{\tau}_i, \boldsymbol{\tau}_j \in \Theta$ *be the parameter vectors for two densities* f_i *and* f_j *of a given exponential family. For* $\alpha \in \mathbb{R} \setminus \{0,1\}$ *such that* $\alpha\boldsymbol{\tau}_i + (1-\alpha)\boldsymbol{\tau}_j \in \Theta$*,*

$$D_\alpha(f_i||f_j) = \ln \frac{C(\boldsymbol{\tau}_j)}{C(\boldsymbol{\tau}_i)} + \frac{1}{\alpha-1} \ln \frac{C(\alpha\boldsymbol{\tau}_i + (1-\alpha)\boldsymbol{\tau}_j)}{C(\boldsymbol{\tau}_i)} \ .$$

Proof.

$$\begin{aligned}
D_\alpha(f_i||f_j) &= \alpha R_\alpha(P_{\boldsymbol{\tau}_i}||P_{\boldsymbol{\tau}_j}) \\
&= \frac{1}{\alpha-1} \ln \left(\frac{C(\alpha\boldsymbol{\tau}_i + (1-\alpha)\boldsymbol{\tau}_j)}{C(\boldsymbol{\tau}_i)^\alpha C(\boldsymbol{\tau}_j)^{1-\alpha}} \right) \\
&= \frac{1}{\alpha-1} \ln \left(\left[\frac{C(\boldsymbol{\tau}_j)}{C(\boldsymbol{\tau}_i)} \right]^{\alpha-1} \frac{C(\alpha\boldsymbol{\tau}_i + (1-\alpha)\boldsymbol{\tau}_j)}{C(\boldsymbol{\tau}_i)} \right) \\
&= \ln \frac{C(\boldsymbol{\tau}_j)}{C(\boldsymbol{\tau}_i)} + \frac{1}{\alpha-1} \ln \frac{C(\alpha\boldsymbol{\tau}_i + (1-\alpha)\boldsymbol{\tau}_j)}{C(\boldsymbol{\tau}_i)} \ .
\end{aligned}$$

□

1.2.4 Applications of Rényi Divergence

As pointed out by Harremoës [30], Rényi entropies and divergences are particularly important as they possess an *operational definition* in the following sense

> An operational definition of a quantity means that the quantity is the natural way to answer a natural question and that the quantity can be estimated by feasible methods combined with a reasonable number of computations.

The operational definition of Rényi divergence given in [30] is that it 'measures how much a probabilistic mixture of two codes can be compressed'. This follows the observation that for any two codelength functions κ_1 and κ_2 for compact codes[13] with corresponding probability measures P_1 and P_2, and $\alpha \in (0,1)$

$$(1-\alpha)\kappa_1 + \alpha\kappa_2 - \alpha D_{1-\alpha}(P_1 \,||\, P_2)$$

is a codelength function of a compact code.

Another important operational definition of Rényi divergence was given by Csiszár [17] in terms of generalized cutoff rates related to the error exponent in hypothesis testing for identically distributed independent observations. A generalization of this result was presented by Alajaji *et al.* [3] by considering hypothesis testing for general sources with memory.

Additional applications of Rényi divergences include the derivation of a family of test statistics for the hypothesis that the coefficients of variation of k normal populations are equal [46], as well as their use in problems of classification, indexing and retrieval, for example [32].

[13] Codes for which Kraft's inequality is met as equality.

We close this section by pointing out an operational definition of Rényi entropy in the context of lossless source coding, which was established by Campbell in his 1965 work [12]. Considering an alphabet of D symbols, the author introduces $L(t)$, the *code length of order* t, defined as $L(t) = t^{-1}\log_D\left(\sum p_i D^{tn_i}\right)$, where $t > 0$, p_i is the probability of the ith symbol, and n_i is the length of code sequence for the ith symbol in an uniquely decipherable code. The following theorem is then established:

> Let $\alpha = (1+t)^{-1}$. By encoding sufficiently long sequences of input symbols, it is possible to make the average code length of order t per input symbol as close to H_α as desired. It is not possible to find a uniquely decipherable code whose average length of order t is less than H_α.

For $t = 0$ ($\alpha = 1$) the above becomes the standard source-coding theorem since H_α becomes the Shannon entropy and L_t becomes the (standard) average code length.

1.3 The Results of this Work

1.3.1 Rényi Divergence Expressions for Continuous Distributions

The applicability of Rényi entropy and Rényi divergence (either directly or via its relationship to the Chernoff and Bhattacharyya distance, and the Hellinger and Kullback-Leibler divergences), as well as the fact that they possess an operational definition in the sense given above, suggests the importance of establishing their general mathematical properties as well as having a compilation of readily available analytical expressions for commonly used distributions. The mathematical properties of the Rényi information measures have been studied both directly [53, 60], and indirectly as part of the f-divergence formalism [16, 40, 59, 41].

Closed-form formulas for differential Shannon and Rényi entropies for several univariate continuous distributions are presented in the work by Song [57]. The author also introduces an 'intrinsic loglikelihood-based distribution measure', $\mathscr{G}_f$, derived from the Rényi entropy, which we consider in detail in Chapter 3. Song's work was followed by [45] where the differential Shannon and Rényi entropy, as well as Song's intrinsic measure for 26 continuous univariate distribution families are presented. The same authors then expanded these results for several multivariate families in [62]. Differential entropy formulas for several continuous distributions can also be found in [15].

An initial review suggested that the literature was significantly less prolific for the case of Rényi divergences, and even for the case of Kullback Divergences, with only a few isolated results presented in separate works: The Rényi and Kullback-Leibler Divergence (KLD) between two univariate Pareto distributions is presented in [7]; the work [49] presents the KLD for two multivariate normal densities as well as for two univariate Gamma densities; in [58] the KLD between two univariate normal densities is also presented and numerical integration is used to estimate the KLD between a Gamma distribution and two approximating models with lognormal and normal distributions; finally the Rényi divergence for multivariate Dirichlet distributions (via the Chernoff distance expression) can also be found in [52], while the KLD is given [48].

Following these findings, one of the main objectives of this work was the computation and compilation of closed-form Rényi (and Kullback) divergences for a wide range of continuous probability distributions. Since most of the applications revolve around two distributions of the same family, this was the focus of the calculations as well. However, an expression for the Rényi divergence between two multivariate Gaussian distributions was found in[14] [33] (which itself cited [10] and [40]), soon after this expression (as well as all other results presented in Appendix B) had been independently derived. The work [40] of Liese and Vajda contains the closed-form expression for the Rényi divergence for exponential families presented in Section 1.2.3. While not all of the original derivations are covered by their result, most of what we obtained here can in fact be derived from this expression. In Chapter 2 we show that applying the expression given in [40] to the canonical parametrization of the exponential families yields expressions in agreement with what we obtained originally. Some expressions for Rényi divergences and/or Kullback divergences for the distributions not covered by their result are also presented in Chapter 2, namely the Rényi and Kullback divergence for general univariate Laplacian, general univariate Pareto, Cramér, and uniform distributions, as well as the Kullback divergence for general univariate Gumbel and Weibull densities. Other commonly used distributions were also originally considered but the computations appeared to be analytically intractable. For a given distribution having m parameters, the integrals involved in the divergence calculations carry $2m$ different parameters (excluding α itself), which is the main source of difficulty in these calculations when compared to differential and Rényi entropies; many of the natural variable transformations or reparametrizations involved in the latter fail in the former.

None of the works presenting Kullback-Leibler or Rényi divergences mentioned above make reference to the work of Liese and Vajda on divergences, while similar work by Vajda and Darbellay [19] on differential entropy for exponential families is cited in some of the works compiling the corresponding expressions. Providing an organized readily available compilation of Rényi and Kullback divergences is still something the

[14] Later on also the work [32] was found to contain this expression as well.

literature would benefit from, especially since the work of Liese and Vajda on Rényi divergences seems to be largely unknown. A summary table with all the collected results is presented in Section 2.4.

1.3.2 Rényi Information Spectrum and Rényi Divergence Spectrum

As mentioned above, Song [57] introduced the information measure $\mathscr{G}_f$, called 'the intrinsic loglikelihood-based distribution measure', which relates the derivative of Rényi divergence with respect to the parameter α to the variance of the log-likelihood function of the distribution. Following Song's approach we show that the variance of the log-likelihood ratio between two densities can be similarly derived from an analytic formula of their Rényi divergence of order α. Both results are discussed in Chapter 3. This connection between Rényi divergence and the loglikelihood ratio becomes practically useful in light of the Rényi divergence expressions presented in Section 2.4.

1.3.3 Rényi Divergence Rate between two Stationary Gaussian Sources

The study of information rates and the computation of expressions for special processes has been considered in the literature. Shannon proved that the entropy rate exists for stationary processes in [56]. Kolmogorov derived the differential entropy rate for stationary Gaussian sources in [37], which can also be found in p. 417 of [15] and p. 76 of [34]. The Rényi entropy and Rényi divergence rate for time-invariant, finite-alphabet Markov sources was obtained by Rached *et. al* in [50]. In [25], Golshani and

Pasha derive the entropy rate for stationary Gaussian processes, using a definition of conditional Rényi entropy [26], which is based on the axioms of Jizba and Arimitsu [35], and which they show to be more suitable than the definition of conditional Rényi entropy found in [11]. The case of Kullback-Leibler divergence rate for stationary Gaussian processes is considered in [61], and can also be found in p. 81 of [34]. Prior to discovering the work of Vajda [40], the literature review did not reveal any work on the Rényi divergence rate for stationary Gaussian sources. Having a closed-form expression for the Rényi divergence between two multivariate Gaussian distributions, it was natural to investigate this problem, and we arrived at the result presented in Chapter 4. The expression is obtained using the theory of Toeplitz Forms developed in [29], and presented in [28]. Following the work of Liese and Vajda led also to the discovery of Vajda's [59] book '*Theory of Statistical Inference and Information*', where the expression of the Rényi divergence rate is presented in p. 239[15].

---

[15]This result is largely unknown/unreferenced in the literature, just like the expression for Rényi divergence for exponential families in [40].

Chapter 2

Kullback and Renyi Divergences for Continuous Distributions

In this chapter we consider commonly used families of continuous distributions and present Rényi and Kullback divergence expressions between two members of a given family. The expressions for distributions belonging to exponential families are computed using the result obtained by Liese and Vajda [40] introduced in Section 1.2.3, and are shown to be in agreement with the original derivations presented in Appendix B. Note that that our original expressions assumed $\alpha \in \mathbb{R}^+ \setminus \{1\}$ while Liese and Vajda assumed the more general domain $\alpha \in \mathbb{R} \setminus \{0, 1\}$.

Definitions as well as many properties of the continuous distributions considered here can be found in any standard continuous distribution references, for example [36]. The distribution referred to as *Cramér* is cited in Song [57] [1].

[1] Correspondence with the author indicated that he gave it this name due to the fact it appears to have first been considered by Cramér.

2.1 Some words on notation

Throughout the calculations below terms of the form $\alpha x + (1-\alpha)y$ occur very frequently. When considering an expression for $D_\alpha(f_i||f_j)$ we use the notation $\theta_\alpha := \alpha\theta_i + (1-\alpha)\theta_j$, where θ_i and θ_j are parameters of the given family of densities, and the order corresponds naturally to the direction of the divergence $D_\alpha(f_i||f_j)$. In the cases where the order is reversed we will write $\theta^*_\alpha = \alpha\theta_j + (1-\alpha)\theta_i$. This notation is followed for both scalar and vector-valued parameters, in the latter case with the standard component-wise addition and scalar multiplication. The following properties are used in the calculations below.

Remark 2.1.1. Let θ_i, θ_j, ϕ_i, ϕ_j be scalar parameters. For fixed constants c_1, c_2, k_1, k_2

$$(k_1\theta + c_1)_\alpha + (k_2\phi + c_2)_\alpha = k_1\theta_\alpha + k_2\phi_\alpha + (c_1 + c_2)\,,$$

and if $\theta_i, \theta_j \neq 0$

$$\left(\frac{1}{\theta}\right)_\alpha = \frac{\theta^*_\alpha}{\theta_i\theta_j}\,.$$

2.2 Exponential Families

For clarity we restate the result from Corollary 1.2.20 which is used to obtain the expressions in this section: Given an exponential family in $\mathbb{R}^n$ satisfying

$$P_\tau(A) = \int_A \frac{1}{C(\boldsymbol{\tau})} \exp\left(\langle \boldsymbol{\tau}, \boldsymbol{T}(\boldsymbol{x})\rangle\right) d\mu(\boldsymbol{x})\,,$$

with natural parameter space Θ, then for $\boldsymbol{\tau}_\alpha = \alpha\boldsymbol{\tau}_i + (1-\alpha)\boldsymbol{\tau}_j \in \Theta$ and $\alpha \notin \{0,1\}$,

$$D_\alpha(f_i||f_j) = \ln\frac{C(\boldsymbol{\tau}_j)}{C(\boldsymbol{\tau}_i)} + \frac{1}{\alpha-1}\ln\frac{C(\boldsymbol{\tau}_\alpha)}{C(\boldsymbol{\tau}_i)}\,.$$

2.2.1 Univariate Gamma Distributions

Throughout this section let f_i and f_j be two univariate Gamma densities:

$$f_i(x) = \frac{x^{k_i-1}e^{-x/\theta_i}}{\theta_i^{k_i}\Gamma(k_i)} \quad k_i, \theta_i > 0; x \in \mathbb{R}^+ .$$

where $\Gamma(x)$ is the Gamma Function introduced in Section A.3.1. Let

$$\boldsymbol{\tau}_i = (\eta_i, \xi_i)^T = \left(-\frac{1}{\theta_i}, k_i - 1\right)^T \quad , \quad \text{and } \boldsymbol{T}(x) = (x, \ln x)^T .$$

We can rewrite the density in terms of its canonical parametrization:

$$f_i(x) = \frac{1}{C(\boldsymbol{\tau}_i)} e^{\langle \boldsymbol{\tau}_i, \boldsymbol{T}(x)\rangle} ,$$

where

$$C(\boldsymbol{\tau}_i) = \frac{\Gamma(\xi_i + 1)}{(-\eta_i)^{\xi_i+1}} = \theta_i^{k_i}\Gamma(k_i) .$$

Let $\alpha \in \mathbb{R} \setminus \{0,1\}$. If $\boldsymbol{\tau}_\alpha \in \Theta$, then by Corollary 1.2.20 we have

$$\begin{aligned} D_\alpha(f_i||f_j) &= \ln \frac{C(\boldsymbol{\tau}_j)}{C(\boldsymbol{\tau}_i)} + \frac{1}{\alpha - 1} \ln \frac{C(\boldsymbol{\tau}_\alpha)}{C(\boldsymbol{\tau}_i)} \\ &= \ln \left(\frac{\Gamma(\xi_j + 1)}{(-\eta_j)^{\xi_j+1}} \frac{(-\eta_i)^{\xi_i+1}}{\Gamma(\xi_i + 1)} \right) + \frac{1}{\alpha - 1} \ln \left(\frac{\Gamma(\xi_\alpha + 1)}{(-\eta_\alpha)^{\xi_\alpha+1}} \frac{(-\eta_i)^{\xi_i+1}}{\Gamma(\xi_i + 1)} \right) . \end{aligned}$$

Reverting to the original parametrization we note that

$$\xi_\alpha + 1 = (\xi + 1)_\alpha = k_\alpha \quad \text{and } -\eta_\alpha = (-\eta)_\alpha = \left(\frac{1}{\theta}\right)_\alpha = \frac{\theta_\alpha^*}{\theta_i\theta_j} ,$$

where we have made use of Remark 2.1.1. Then

$$D_\alpha(f_i||f_j) = \ln \left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}} \right) + \frac{1}{\alpha - 1} \ln \left(\frac{\Gamma(k_\alpha)}{(\theta_\alpha^*)^{k_\alpha}} \frac{(\theta_i\theta_j)^{k_\alpha}}{\theta_i^{k_i}\Gamma(k_i)} \right) .$$

In the notation of the original derivation $k_0 = k_\alpha$ and $\theta_0 = \theta_\alpha^*$, so that the expression above is the same as that obtained in Proposition B.1.6. Finally, note that $\boldsymbol{\tau}_\alpha \in \Theta \Longleftrightarrow$

$k_\alpha, (1/\theta)_a > 0$ and $(1/\theta)_a > 0 \Leftrightarrow \theta^*_\alpha > 0$, and so the constraints for finiteness also agree with those of Proposition B.1.6. The special cases of exponential and χ^2 densities, as well as the expressions for the case $\alpha = 1$ (Kullback-Leibler divergence), are both included in Section B.1 so we omit them here.

2.2.2 Univariate Chi Distributions

Throughout this section let f_i and f_j be two univariate Chi densities

$$f_i(x) = \frac{2^{1-k_i/2} x^{k_i - 1} e^{-x^2/2\sigma_i^2}}{\sigma_i^{k_i} \Gamma\left(\frac{k_i}{2}\right)} \ , \sigma_i > 0 \ , \ k_i \in \mathbb{N} \ ; \ x > 0 \ .$$

Let

$$\boldsymbol{\tau}_i = (\eta_i, \xi_i)^T = \left(-\frac{1}{2\sigma_i^2}, k_i - 1\right)^T \ , \quad \text{and } \boldsymbol{T}(x) = (x^2, \ln x)^T \ .$$

We can rewrite the density in terms of its canonical parametrization:

$$f_i(x) = \frac{1}{C(\boldsymbol{\tau}_i)} e^{\langle \boldsymbol{\tau}_i, \boldsymbol{T}(x)\rangle} \ ,$$

where

$$C(\boldsymbol{\tau}_i) = \frac{\Gamma\left(\frac{\xi_i+1}{2}\right) 2^{(\xi_i-1)/2}}{(-2\eta_i)^{(\xi_i+1)/2}} = \Gamma\left(\frac{k_i}{2}\right) \sigma_i^{k_i} 2^{k_i/2-1} \ .$$

Let $\alpha \in \mathbb{R} \setminus \{0, 1\}$. If $\boldsymbol{\tau}_\alpha \in \Theta$, then by Corollary 1.2.20 we have

$$\begin{aligned} D_\alpha(f_i||f_j) &= \ln \frac{C(\boldsymbol{\tau}_j)}{C(\boldsymbol{\tau}_i)} + \frac{1}{\alpha - 1} \ln \frac{C(\boldsymbol{\tau}_\alpha)}{C(\boldsymbol{\tau}_i)} \\ &= \ln \left(\frac{\Gamma\left(\frac{\xi_j+1}{2}\right) 2^{(\xi_j-1)/2}}{(-2\eta_j)^{(\xi_j+1)/2}} \frac{(-2\eta_i)^{(\xi_i+1)/2}}{\Gamma\left(\frac{\xi_i+1}{2}\right) 2^{(\xi_i-1)/2}} \right) \\ &\quad + \frac{1}{\alpha - 1} \ln \left(\frac{\Gamma\left(\frac{\xi_\alpha+1}{2}\right) 2^{(\xi_\alpha-1)/2}}{(-2\eta_\alpha)^{(\xi_\alpha+1)/2}} \frac{(-2\eta_i)^{(\xi_i+1)/2}}{\Gamma\left(\frac{\xi_i+1}{2}\right) 2^{(\xi_i-1)/2}} \right) \ . \end{aligned}$$

Next we revert to the original parametrization. Using Remark 2.1.1 it follows that

$$\xi_\alpha + 1 = (\xi+1)_\alpha = k_\alpha \ , \ \frac{\xi_\alpha - 1}{2} = \frac{k_\alpha}{2} - 1 \ ,$$

and

$$-2\eta_\alpha = (-2\eta)_\alpha = \left(\frac{1}{\sigma^2}\right)_\alpha = \frac{(\sigma^2)^*_\alpha}{\sigma_i^2\sigma_j^2} \ ;$$

hence

$$\begin{aligned}
D_\alpha(f_i||f_j) &= \ln\left(\frac{\Gamma\left(k_j/2\right)\sigma_j^{k_j}2^{k_j/2-1}}{\Gamma\left(k_i/2\right)\sigma_i^{k_i}2^{k_i/2-1}}\right) \\
&\quad + \frac{1}{\alpha-1}\ln\left(\left(\frac{\sigma_i^2\sigma_j^2}{(\sigma^2)^*_\alpha}\right)^{k_\alpha/2}\frac{\Gamma(k_\alpha/2)2^{k_\alpha/2-1}}{\Gamma\left(k_i/2\right)\sigma_i^{k_i}2^{k_i/2-1}}\right) \\
&= \ln\left(\frac{\sigma_j^{k_j}\Gamma\left(k_j/2\right)}{\sigma_i^{k_i}\Gamma\left(k_i/2\right)}\right) + \frac{1}{2}\left[(k_j-k_i) + \frac{k_\alpha - k_i}{\alpha-1}\right]\ln 2 \\
&\quad + \frac{1}{\alpha-1}\ln\left(\left(\frac{\sigma_i^2\sigma_j^2}{(\sigma^2)^*_\alpha}\right)^{k_\alpha/2}\frac{\Gamma(k_\alpha/2)}{\sigma_i^{k_i}\Gamma\left(k_i/2\right)}\right) \ .
\end{aligned}$$

Observe that

$$k_j - k_i + \frac{k_\alpha - k_i}{\alpha - 1} = \frac{1}{\alpha-1}\left[(\alpha-1)(k_j-k_i) + \alpha k_i + (1-\alpha)k_j - k_i\right] = 0 \ .$$

Thus,

$$D_\alpha(f_i||f_j) = \ln\left(\frac{\sigma_j^{k_j}\Gamma\left(k_j/2\right)}{\sigma_i^{k_i}\Gamma\left(k_i/2\right)}\right) + \frac{1}{\alpha-1}\ln\left(\left(\frac{\sigma_i^2\sigma_j^2}{(\sigma^2)^*_\alpha}\right)^{k_\alpha/2}\frac{\Gamma(k_\alpha/2)}{\sigma_i^{k_i}\Gamma\left(k_i/2\right)}\right) \ .$$

In the notation of the original derivation $k_0 = k_\alpha$ and $\sigma_0 = (\sigma^2)^*_\alpha$, so that the expression above is the same as that obtained in Proposition B.2.6. Finally, note that $\boldsymbol{\tau}_\alpha \in \Theta \iff k_\alpha, (1/(\sigma^2))_\alpha > 0$ and $(1/\sigma^2)_\alpha > 0 \iff (\sigma^2)^*_\alpha \geq 0$, and so the constraints for finiteness also agree with those of Proposition B.2.6. The special cases of half-normal, Rayleigh, and Maxwell-Boltzmann densities, as well as the expressions for the case $\alpha = 1$ (Kullback-Leibler divergence), are included in Section B.2 so we omit them here.

2.2.3 Dirichlet Distributions

Throughout this section let f_i and f_j be two Dirichlet densities of order n: [2]

$$f_i(\boldsymbol{x}, \boldsymbol{a}_i) = \frac{1}{B(\boldsymbol{a}_i)} \prod_{k=1}^{n} x_k^{a_{i_k}-1} \quad ; \boldsymbol{a}_i \in \mathbb{R}^n \ , \ ; \ \boldsymbol{x} \in \mathbb{R}^n \ , n \geq 2, \ n \in \mathbb{N} \ ,$$

where $\boldsymbol{x} = (x_1, ..., x_n)$ satisfies $\sum_{k=1}^{n} x_k = 1$, $\boldsymbol{a}_i = (a_{i_1}, ..., a_{i_n})$, $a_k > 0$, and $B(\boldsymbol{y})$ is the multinomial beta function defined in Definition A.3.10.

Let $\boldsymbol{\tau_i} = (a_1 - 1, ..., a_n - 1)^T$ and $\boldsymbol{T}(\boldsymbol{x}) = (\ln x_1, ..., \ln x_n)^T$. We can rewrite the density in terms of its canonical parametrization:

$$f_i(\boldsymbol{x}) = \frac{1}{C(\boldsymbol{\tau_i})} e^{\langle \boldsymbol{\tau_i}, \boldsymbol{T}(\boldsymbol{x}) \rangle} \ ,$$

where

$$C(\boldsymbol{\tau_i}) = B(\boldsymbol{\tau_i} + (1, 1, ..., 1)^T) = B(\boldsymbol{a}_i) \ .$$

Let $\alpha \in \mathbb{R} \setminus \{0, 1\}$. If $\boldsymbol{\tau}_\alpha \in \Theta$, then by Corollary 1.2.20 we have

$$D_\alpha(f_i||f_j) = \ln \frac{C(\boldsymbol{\tau_j})}{C(\boldsymbol{\tau_i})} + \frac{1}{\alpha - 1} \ln \frac{C(\boldsymbol{\tau}_\alpha)}{C(\boldsymbol{\tau_i})} = \ln \frac{B(\boldsymbol{a}_j)}{B(\boldsymbol{a}_i)} + \frac{1}{\alpha - 1} \ln \frac{B(\boldsymbol{a_\alpha})}{B(\boldsymbol{a}_i)} \ .$$

In the notation of the derivation given in Section B.3.2, $\boldsymbol{b_0} = \boldsymbol{b_\alpha}$ and $\boldsymbol{a_0} = \boldsymbol{a_\alpha}$, so that the expression above is the same as that obtained in Proposition B.3.6. Note also that

[2] It should be noted that the dimension of the underlying space is not n, but $n - 1$, as the distribution is over the hyperplane specified by the constraint $\sum_{k=1}^{n} x_k = 1$.

$\boldsymbol{\tau}_\alpha \in \Theta \iff \forall k \; a_k, b_k > 0$, so that the finiteness constraints are also in agreement. As mentioned in Section B.3.2, this result is in agreement with the Chernoff distance between two Dirichlet distributions derived in [52]. The special case of the Beta distributions as well as the expressions for the case $\alpha = 1$ (Kullback-Leibler divergence), are included in Section B.3 so we omit them here.

Also, the work [48] presents the KLD expression between two Dirichlet distributions. In our notation,

$$D(f_i||f_j) = \log \frac{\Gamma(a_{it})}{\Gamma(a_{jt})} + \sum_{k=1}^{d} \log \frac{\Gamma(a_{j_k})}{\Gamma(a_{i_k})} + \sum_{k=1}^{d} \left[a_{i_k} - a_{j_k}\right] \left[\psi(a_{i_k}) - \psi(a_{it})\right] ,$$

where

$$a_{it} = \sum_{k=1}^{d} a_{i_k} , \quad a_{jt} = \sum_{k=1}^{d} a_{j_k} .$$

This may be rewritten using the multivariate Beta function:

$$\begin{aligned} \log \frac{\Gamma(a_{it})}{\Gamma(a_{jt})} + \sum_{k=1}^{d} \log \frac{\Gamma(a_{j_k})}{\Gamma(a_{i_k})} &= \log \left(\frac{\Gamma\left(\sum_{k=1}^{d} a_{i_k}\right) \prod_{k=1}^{d} \Gamma(a_{j_k})}{\Gamma\left(\sum_{k=1}^{d} a_{j_k}\right) \prod_{k=1}^{d} \Gamma(a_{i_k})} \right) \\ &= \log \frac{B(\boldsymbol{a_j})}{B(\boldsymbol{a_i})} , \end{aligned}$$

hence

$$D(f_i||f_j) = \log \frac{B(\boldsymbol{a_j})}{B(\boldsymbol{a_i})} + \sum_{k=1}^{d} \left[a_{i_k} - a_{j_k}\right] \left[\psi(a_{i_k}) - \psi\left(\sum_{k=1}^{d} a_{i_k}\right)\right] .$$

Taking $\boldsymbol{a_i} = (a_i, b_i)$ and $\boldsymbol{a_j} = (a_j, b_j)$ we can see this agrees with the expression we have for the KLD of Beta distributions given in Section B.3.

2.2.4 Multivariate Gaussian Distributions

Throughout this section let f_i and f_j be two multivariate normal densities over $\mathbb{R}^n$:

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)} \ , \quad \boldsymbol{x} \in \mathbb{R}^n \ ,$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^n$, Σ_i is a symmetric positive-definite matrix, and $(.)'$ denotes transposition.

Obtaining an expression for the Rényi divergence between two multivariate normal densities has already been considered in the literature, for example [10, 32]. The work [10][3] presents the following expression:

$$\begin{aligned} R_\alpha(f_i\,||\,f_j) = {} & \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'(\alpha\Sigma_j + (1-\alpha)\Sigma_i)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\ & - \frac{1}{2\alpha(\alpha-1)} \ln \frac{|\alpha\Sigma_j + (1-\alpha)\Sigma_i|}{|\Sigma_i|^{1-\alpha}|\Sigma_j|^{\alpha}} \ . \end{aligned}$$

Note that in [10] $R_\alpha(f_i||f_j)$ is denoted as $B(i,j)$. In what follows we show that the Rényi divergence expression presented in Section B.4 is in agreement with the result above. The expression we obtained in Proposition B.4.10 is

$$D_\alpha(f_i||f_j) = \frac{1}{2}\ln\left(\frac{|\Sigma_j|}{|\Sigma_i|}\right) + \frac{1}{2(\alpha-1)}\ln\left(\frac{1}{|A||\Sigma_i|}\right) - \frac{F(\alpha)}{2(\alpha-1)} \ ,$$

with $A = \alpha\Sigma_i^{-1} + (1-\alpha)\Sigma_j^{-1}$ and

$$\begin{aligned} F(\alpha) := {} & \left[\alpha\boldsymbol{\mu}_i'\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\boldsymbol{\mu}_j'\Sigma_j^{-1}\boldsymbol{\mu}_j\right] \\ & - \left[\alpha\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\Sigma_j^{-1}\boldsymbol{\mu}_j\right]' A^{-1} \left[\alpha\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\Sigma_j^{-1}\boldsymbol{\mu}_j\right] \ . \end{aligned}$$

As we noted prior to introducing Corollary 1.2.20, the definitions of D_α and R_α differ

[3] This was the first explicit formula that we discovered in the literature, which obtains the result for $R_\alpha(f_i||f_j)$; hence the discussion below.

by a factor of α[4], i.e.,

$$R_\alpha(f||g) = \frac{1}{\alpha(\alpha-1)} \ln \int_{\mathscr{X}} f(x)^\alpha g(x)^{1-\alpha} d\mu(x) \,.$$

In order to compare the expressions for $D_\alpha(f_i||f_j)$ and $R_\alpha(f_i||f_j)$ we consider $\alpha R_\alpha(f_i||f_j)$. Examining the resulting logarithmic term we have

$$\begin{aligned}
-\frac{1}{2(\alpha-1)} \ln \frac{|\alpha\Sigma_j + (1-\alpha)\Sigma_i|}{|\Sigma_i|^{1-\alpha}|\Sigma_j|^\alpha} &= \frac{1}{2(\alpha-1)} \ln \frac{|\Sigma_i|^{1-\alpha}|\Sigma_j|^{\alpha-1}|\Sigma_j|}{|\alpha\Sigma_j + (1-\alpha)\Sigma_i|} \\
&= \frac{1}{2} \ln \frac{|\Sigma_j|}{|\Sigma_i|} + \frac{1}{2(\alpha-1)} \ln \frac{|\Sigma_j|}{|\alpha\Sigma_j + (1-\alpha)\Sigma_i|} \,.
\end{aligned}$$

Since $A = \alpha\Sigma_i^{-1} + (1-\alpha)\Sigma_j^{-1}$, we can write

$$B := \alpha\Sigma_j + (1-\alpha)\Sigma_i = \Sigma_i A \Sigma_j = \Sigma_j A \Sigma_i \,,$$

and

$$\frac{|\Sigma_j|}{|\alpha\Sigma_j + (1-\alpha)\Sigma_i|} = \frac{|\Sigma_j|}{|B|} = \frac{1}{|A||\Sigma_i|} \,,$$

so that the logarithmic terms for both expressions are in agreement. Examining the last term of αR_α it remains to show that

$$\begin{aligned}
F(\alpha) &= \alpha(1-\alpha) \left[(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'(\alpha\Sigma_j + (1-\alpha)\Sigma_i)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \right] \\
&= \alpha(1-\alpha) \left[(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' B^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \right] \,.
\end{aligned}$$

Note that

$$\begin{aligned}
F(\alpha) = \boldsymbol{\mu}_i' &\left[\alpha\Sigma_i^{-1} - \alpha^2 \Sigma_i^{-1} A^{-1} \Sigma_i^{-1} \right] \boldsymbol{\mu}_i \\
&+ \boldsymbol{\mu}_j' \left[(1-\alpha)\Sigma_j^{-1} - (1-\alpha)^2 \Sigma_j^{-1} A^{-1} \Sigma_j^{-1} \right] \boldsymbol{\mu}_j \\
&- \boldsymbol{\mu}_i' \left[\alpha(1-\alpha)\Sigma_i^{-1} A^{-1} \Sigma_j^{-1} \right] \boldsymbol{\mu}_j \\
&- \boldsymbol{\mu}_j' \left[\alpha(1-\alpha)\Sigma_j^{-1} A^{-1} \Sigma_i^{-1} \right] \boldsymbol{\mu}_i \,,
\end{aligned}$$

[4] The definition of Rényi divergence as given by R_α is also used in other, more recent, works in the statistical literature, e.g. [46].

which in turn can be written as

$$
\begin{aligned}
F(\alpha) = {} & \alpha(1-\alpha)\left[(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' B^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right] \\
& - \alpha(1-\alpha)\left[\boldsymbol{\mu}_i' B^{-1} \boldsymbol{\mu}_i + \boldsymbol{\mu}_j' B^{-1} \boldsymbol{\mu}_j\right] \\
& + \boldsymbol{\mu}_i' \left[\alpha \Sigma_i^{-1} - \alpha^2 \Sigma_i^{-1} A^{-1} \Sigma_i^{-1}\right] \boldsymbol{\mu}_i \\
& + \boldsymbol{\mu}_j' \left[(1-\alpha)\Sigma_j^{-1} - (1-\alpha)^2 \Sigma_j^{-1} A^{-1} \Sigma_j^{-1}\right] \boldsymbol{\mu}_j \ ,
\end{aligned}
$$

since

$$
B = \Sigma_i A \Sigma_j = \Sigma_j A \Sigma_i \Leftrightarrow B^{-1} = \Sigma_j^{-1} A^{-1} \Sigma_i^{-1} = \Sigma_i^{-1} A^{-1} \Sigma_j^{-1} \ .
$$

Collecting like terms,

$$
\begin{aligned}
F(\alpha) = {} & \alpha(1-\alpha)\left[(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' B^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right] \\
& + \boldsymbol{\mu}_i' \left[\alpha \Sigma_i^{-1} - \alpha^2 \Sigma_i^{-1} A^{-1} \Sigma_i^{-1} - \alpha(1-\alpha) B^{-1}\right] \boldsymbol{\mu}_i \\
& + \boldsymbol{\mu}_j' \left[(1-\alpha)\Sigma_j^{-1} - (1-\alpha)^2 \Sigma_j^{-1} A^{-1} \Sigma_j^{-1} - \alpha(1-\alpha) B^{-1}\right] \boldsymbol{\mu}_j \ .
\end{aligned}
$$

Finally observe that

$$
\begin{aligned}
& \alpha \Sigma_i^{-1} - \alpha^2 \Sigma_i^{-1} A^{-1} \Sigma_i^{-1} - \alpha(1-\alpha) B^{-1} \\
& = \alpha \Sigma_i^{-1} A^{-1} A - \alpha^2 \Sigma_i^{-1} A^{-1} \Sigma_i^{-1} - \alpha(1-\alpha) \Sigma_i^{-1} A^{-1} \Sigma_j^{-1} \\
& = \alpha \Sigma_i^{-1} A^{-1} \left[A - \alpha \Sigma_i^{-1} - (1-\alpha)\Sigma_j^{-1}\right] \\
& = \alpha \Sigma_i^{-1} A^{-1} [A - A] \\
& = 0 \ .
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& (1-\alpha)\Sigma_j^{-1} - (1-\alpha)^2 \Sigma_j^{-1} A^{-1} \Sigma_j^{-1} - \alpha(1-\alpha) B^{-1} \\
& = (1-\alpha)\Sigma_j^{-1} A^{-1} \left[A - (1-\alpha)\Sigma_j^{-1} - \alpha \Sigma_i^{-1}\right] \\
& = 0 \ .
\end{aligned}
$$

Thus

$$F(\alpha) = \alpha(1-\alpha)\left[(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' B^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\right] \text{ , and } \alpha R_\alpha(f_i||f_j) = D_\alpha(f_i||f_j) \text{ .}$$

We also showed in Proposition B.4.10 that above expression for the Rényi divergence is valid only when A is positive definite, which for $\alpha \in (0,1)$ is always the case given the positive-definiteness of Σ_i and Σ_j. When A is not positive-definite $D_\alpha(f_i||f_j) = +\infty$. Moreover, we derive expressions for the Kullback-Leibler divergence $D(f_i||f_j)$ and demonstrate that the expression for $D_\alpha(f_i||f_j)$ does indeed approach $D(f_i||f_j)$ as $\alpha \to 1$. We also consider the special cases of the Rényi divergence between two univariate Gaussian densities and the zero-mean, unit-variance bivariate case. We present these results as a remark below with the full derivation included in Section B.4:

Remark 2.2.1. Special Cases of $D_\alpha(f_i||f_j)$:

1. The Kullback Leibler divergence between f_i and f_j is

$$D(f_i||f_j) = \frac{1}{2}\left[\ln\frac{|\Sigma_j|}{|\Sigma_i|} + \operatorname{tr}\left(\Sigma_j^{-1}\Sigma_i\right) + \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)' \Sigma_j^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) - n\right] \text{ .}$$

2. For $n = 1$ $D_\alpha(f_i||f_j)$ reduces to

$$\begin{aligned} D_\alpha(f_i||f_j) = \ln\frac{\sigma_j}{\sigma_i} + \frac{1}{2(\alpha-1)}\ln\left(\frac{\sigma_j^2}{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2}\right) \\ + \frac{1}{2}\frac{\alpha(\mu_i - \mu_j)^2}{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2} \text{ ,} \end{aligned}$$

   and $D(f_i||f_j)$ reduces to

$$D(f_i||f_j) = \frac{1}{2\sigma_j^2}\left[(\mu_i - \mu_j)^2 + \sigma_i^2 - \sigma_j^2\right] + \ln\frac{\sigma_j}{\sigma_i}.$$

3. For two zero-mean, unit-variance, bivariate Gaussian densities f_i and f_j with correlation coefficients ρ_i and ρ_j,

$$D_\alpha(f_i||f_j) = \frac{1}{2}\ln\left(\frac{1-\rho_j^2}{1-\rho_i^2}\right) - \frac{1}{2(\alpha-1)}\ln\left(\frac{1-(\alpha\rho_j+(1-\alpha)\rho_i)^2}{(1-\rho_j^2)}\right) .$$

Proof. See Proposition B.4.8, Proposition B.4.4 and Proposition B.4.3, and Section B.4.3, respectively. □

2.2.5 Univariate Gumbel Distributions with Fixed Scale Parameter

Like Weibull distributions, a general family of univariate Gumbel distributions cannot be written as an exponential family, but we can again consider a special case, namely two densities f_i and f_j with fixed scale parameter $\beta_i = \beta_j = \beta$:

$$\begin{aligned} f_i(x) &= \beta^{-1}e^{-(x-\mu_i)/\beta}\exp\left(-e^{-(x-\mu_i)/\beta}\right) \\ &= \beta^{-1}e^{-x/\beta}e^{\mu_i/\beta}\exp\left(-e^{-x/\beta}e^{\mu_i/\beta}\right) , \quad \mu_i \in \mathbb{R},\ \beta > 0\ ; x \in \mathbb{R} . \end{aligned}$$

Let $\boldsymbol{\tau_i} = \eta_i = -e^{\mu_i/\beta}$ and $\boldsymbol{T}(x) = T(x) = e^{-x/\beta}$. If we consider a measure ν on X whose density with respect to the Lebesgue measure is $h(x) = T(x)/\beta$ we can rewrite the density above relative to ν in the canonical parametrization (see Definition A.2.2 and the discussion preceding it):

$$f_i(x) = \frac{1}{C(\boldsymbol{\tau_i})}e^{\langle \boldsymbol{\tau_i}, \boldsymbol{T}(x)\rangle} ,$$

where

$$C(\boldsymbol{\tau_i}) = \frac{1}{(-\eta_i)} = \frac{1}{e^{\mu_i/\beta}} = e^{-\mu_i/\beta} .$$

Let $\alpha \in \mathbb{R}\setminus\{0,1\}$. If $\boldsymbol{\tau_\alpha} \in \Theta$, then by Corollary 1.2.20 we have

$$D_\alpha(f_i||f_j) = \ln\frac{C(\boldsymbol{\tau_j})}{C(\boldsymbol{\tau_i})} + \frac{1}{\alpha-1}\ln\frac{C(\boldsymbol{\tau_\alpha})}{C(\boldsymbol{\tau_i})} = \ln\left(\frac{-\eta_i}{-\eta_j}\right) + \frac{1}{\alpha-1}\ln\left(\frac{-\eta_i}{-\eta_\alpha}\right)$$

where the natural parameter space is in this case $\Theta = \{\eta < 0\}$. Reverting to the original parametrization,

$$\begin{aligned} D_\alpha(f_i||f_j) &= \ln \frac{e^{-\mu_j/\beta}}{e^{-\mu_i/\beta}} + \frac{1}{\alpha-1} \ln \left(\frac{e^{\mu_i/\beta}}{\alpha e^{\mu_i/\beta} + (1-\alpha) e^{\mu_j/\beta}} \right) \\ &= \frac{\mu_i - \mu_j}{\beta} + \frac{1}{\alpha-1} \ln \left(\frac{e^{\mu_i/\beta}}{\alpha e^{\mu_i/\beta} + (1-\alpha) e^{\mu_j/\beta}} \right) , \end{aligned}$$

for $\alpha e^{\mu_i/\beta} + (1-\alpha) e^{\mu_j/\beta} > 0$.

2.2.6 Univariate Laplace Distributions with Location Parameter Equal to Zero

We consider the special case of two Laplace densities with location parameter $\theta = 0$. Throughout this section let f_i and f_j be two such densities:

$$f_i(x) = \frac{1}{2\lambda_i} e^{-|x|/\lambda_i} , \quad \lambda_i > 0; \; x \in \mathbb{R} .$$

Let $\boldsymbol{\tau_i} = \eta_i = -1/\lambda_i$ and $\boldsymbol{T}(x) = T(x) = |x|$. We can rewrite the density in terms of its canonical parametrization:

$$f_i(x) = \frac{1}{C(\boldsymbol{\tau_i})} e^{\langle \boldsymbol{\tau_i}, \boldsymbol{T}(x) \rangle} ,$$

where

$$C(\boldsymbol{\tau_i}) = -\frac{2}{\eta_i} = 2\lambda_i \; .$$

Let $\alpha \in \mathbb{R} \setminus \{0, 1\}$. If $\boldsymbol{\tau}_\alpha \in \Theta$, then by Corollary 1.2.20 we have

$$D_\alpha(f_i||f_j) = \ln \frac{C(\boldsymbol{\tau_j})}{C(\boldsymbol{\tau_i})} + \frac{1}{\alpha-1} \ln \frac{C(\boldsymbol{\tau}_\alpha)}{C(\boldsymbol{\tau_i})} = \ln \frac{\eta_i}{\eta_j} + \frac{1}{\alpha-1} \ln \frac{\eta_i}{\eta_\alpha} .$$

Since

$$\eta_\alpha = \left(-\frac{1}{\lambda}\right)_\alpha = -\frac{\lambda^*_\alpha}{\lambda_i\lambda_j}$$

then reverting to the original parametrization,

$$D_\alpha(f_i||f_j) = \ln\frac{\lambda_j}{\lambda_i} + \frac{1}{\alpha-1}\ln\frac{\lambda_i\lambda_j}{\lambda^*_\alpha}\frac{1}{\lambda_i} = \ln\frac{\lambda_j}{\lambda_i} + \frac{1}{\alpha-1}\ln\frac{\lambda_j}{\lambda^*_\alpha}\ .$$

We derive the Rényi divergence expression for general Laplacian distributions in Section 2.3.1, and we show in Remark 2.3.7 that it reduces to the above expression when $\theta_i = \theta_j = 0$.

2.2.7 Univariate Pareto Distributions with Fixed Scale Parameter

We consider the special case of two Pareto densities with equal scale parameter m. Throughout this section let f_i and f_j be two such densities:

$$f_i(x) = a_i m^{a_i} x^{-(a_i+1)}\ ,\quad a_i, m > 0\ ; x > m\ .$$

Let $\boldsymbol{\tau_i} = \eta_i = -(a_i+1)$ and $\boldsymbol{T}(x) = T(x) = \ln x$. We can rewrite the density in terms of its canonical parametrization:

$$f_i(x) = \frac{1}{C(\boldsymbol{\tau_i})} e^{\langle\boldsymbol{\tau_i},\boldsymbol{T}(x)\rangle}\ ,$$

where

$$C(\boldsymbol{\tau_i}) = \frac{m^{\eta_i+1}}{-(\eta_i+1)} = \frac{1}{a_i m^{a_i}}\ .$$

Let $\alpha \in \mathbb{R} \setminus \{0,1\}$. If $\tau_\alpha \in \Theta$, then by Corollary 1.2.20 we have

$$\begin{aligned}
D_\alpha(f_i||f_j) &= \ln \frac{C(\tau_j)}{C(\tau_i)} + \frac{1}{\alpha-1} \ln \frac{C(\tau_\alpha)}{C(\tau_i)} \\
&= \ln \left[\frac{m^{\eta_j+1}}{-(\eta_j+1)} \frac{-(\eta_i+1)}{m^{\eta_i+1}} \right] + \frac{1}{\alpha-1} \ln \left[\frac{m^{\eta_\alpha+1}}{-(\eta_\alpha+1)} \frac{-(\eta_i+1)}{m^{\eta_i+1}} \right] \\
&= \ln \left[m^{(\eta_j-\eta_i)} \frac{\eta_i+1}{\eta_j+1} \right] + \frac{1}{\alpha-1} \ln \left[m^{(\eta_\alpha-\eta_i)} \frac{\eta_i+1}{\eta_\alpha+1} \right] \\
&= \left(\eta_j - \eta_i + \frac{\eta_\alpha - \eta_i}{\alpha-1} \right) \ln m + \ln \frac{\eta_i+1}{\eta_j+1} + \frac{1}{\alpha-1} \ln \frac{\eta_i+1}{\eta_\alpha+1} \,.
\end{aligned}$$

But since

$$\begin{aligned}
\eta_j - \eta_i + \frac{\eta_\alpha - \eta_i}{\alpha-1} &= \frac{(\alpha-1)(\eta_j-\eta_i) + \eta_\alpha - \eta_i}{\alpha-1} \\
&= \frac{\alpha-1}{\alpha-1} \left[(\eta_j-\eta_i) + (\eta_i-\eta_j) \right] = 0 \,,
\end{aligned}$$

then

$$D_\alpha(f_i||f_j) = \ln \frac{\eta_i+1}{\eta_j+1} + \frac{1}{\alpha-1} \ln \frac{\eta_i+1}{\eta_\alpha+1} \,.$$

Reverting to the original parametrization,

$$D_\alpha(f_i||f_j) = \ln \frac{a_i}{a_j} + \frac{1}{\alpha-1} \ln \frac{a_i}{a_\alpha} \,,$$

noting that $\eta_\alpha + 1 = (\eta+1)_\alpha = (-a)_\alpha = -a_\alpha$. In the notation of the original derivation $a_0 = a_\alpha$, so that the expression above is the same as that obtained in Proposition B.5.4. As before, the constraints for finiteness also agree with those of Proposition B.5.4. As mentioned in Section B.5, this result is in agreement with that derived in [7].

2.2.8 Univariate Weibull Distributions with Fixed Shape Parameter

While a general family of univariate Weibull distributions cannot be written as an exponential family, we can consider the special case of Weibull densities f_i, f_j with fixed shape parameter $k_i = k_j = k$:

$$f_i(x) = k\lambda_i^{-k} x^{k-1} e^{-(x/\lambda_i)^k} \ , \quad k, \lambda_i > 0; x \in \mathbb{R}^+ \ .$$

Let $\boldsymbol{\tau_i} = \eta_i = -\lambda_i^{-k}$ and $\boldsymbol{T}(x) = T(x) = x^k$. If we consider a measure ν on X whose density with respect to the Lebesgue measure is $h(x) = kx^{k-1}$ we can rewrite the density above relative to ν in the canonical parametrization (see Definition A.2.2 and the discussion preceding it):

$$f_i(x) = \frac{1}{C(\boldsymbol{\tau_i})} e^{\langle \boldsymbol{\tau_i}, \boldsymbol{T}(x) \rangle} \ ,$$

where

$$C(\boldsymbol{\tau_i}) = \frac{1}{(-\eta_i)} = \frac{1}{\lambda_i^{-k}} = \lambda_i^k \ .$$

Let $\alpha \in \mathbb{R} \setminus \{0, 1\}$. If $\boldsymbol{\tau}_\alpha \in \Theta$, then by Corollary 1.2.20 we have

$$D_\alpha(f_i||f_j) = \ln \frac{C(\boldsymbol{\tau_j})}{C(\boldsymbol{\tau_i})} + \frac{1}{\alpha - 1} \ln \frac{C(\boldsymbol{\tau}_\alpha)}{C(\boldsymbol{\tau_i})} = \ln \left(\frac{-\eta_i}{-\eta_j} \right) + \frac{1}{\alpha - 1} \ln \left(\frac{-\eta_i}{-\eta_\alpha} \right) \ .$$

Reverting to the original parametrization we note that

$$-\eta_\alpha = \left(\frac{1}{\lambda^k} \right)_\alpha = \frac{(\lambda^k)^*_\alpha}{\lambda_i^k \lambda_j^k} \ .$$

Then

$$D_\alpha(f_i||f_j) = \ln \left(\frac{\lambda_j}{\lambda_i} \right)^k + \frac{1}{\alpha - 1} \ln \frac{\lambda_j^k}{(\lambda^k)^*_\alpha} \ .$$

In the notation of the original derivation $\lambda_0 = (\lambda^k)^*_\alpha$, so that the expression above is the same as that obtained in Proposition B.6.1. Finally, note that

$\boldsymbol{\tau}_\alpha \in \Theta \Longleftrightarrow (1/\lambda^k)_\alpha > 0 \Longleftrightarrow (\lambda^k)^*_\alpha > 0$, and so the constraints for finiteness also agree with those of Proposition B.6.1.

2.3 Other Distributions

2.3.1 Rényi and Kullback Divergence for General Univariate Laplace Distributions

Throughout this section let f_i and f_j be two univariate Laplace densities

$$f_i(x) = \frac{1}{2\lambda_i} e^{-|x-\theta_i|/\lambda_i} \ , \quad \lambda_i > 0;\ \theta_i \in \mathbb{R};\ x \in \mathbb{R} \ .$$

Proposition 2.3.1.

$$E_{f_i}\left[\ln f_j\right] = -\left[\ln 2\lambda_j + \frac{\lambda_i}{\lambda_j} e^{-|\theta_i-\theta_j|/\lambda_i} + \frac{|\theta_i-\theta_j|}{\lambda_j}\right] \ .$$

Proof.

$$E_{f_i}\left[\ln f_j\right] = E_{f_i}\left[-\ln 2\lambda_j - \frac{|X-\theta_j|}{\lambda_j}\right] = -\ln 2\lambda_j - \frac{1}{\lambda_j} E_{f_i}\left[|X-\theta_j|\right] \ .$$

Consider $E_{f_i}\left[|X-\theta_j|\right]$. Let $Y = X - \theta_i$. Then Y has a zero-mean Laplacian distribution[5] and $E_{f_i}\left[|X-\theta_j|\right] = E_{f_Y}\left[|Y-\Theta|\right]$, where $\Theta = \theta_j - \theta_i$. Then

$$\begin{aligned} E_{f_Y}\left[|Y-\Theta|\right] &= \int_{\mathbb{R}} |y-\Theta| \frac{1}{2\lambda_i} e^{-|y|/\lambda_i} dy \\ &= \int_{-\infty}^{\Theta} (\Theta - y)\frac{1}{2\lambda_i} e^{-|y|/\lambda_i} dy + \int_{\Theta}^{\infty} (y-\Theta)\frac{1}{2\lambda_i} e^{-|y|/\lambda_i} dy \ . \end{aligned}$$

Considering $\Theta > 0$ we can write the above as

$$\int_{\mathbb{R}} (\Theta - y)\frac{1}{2\lambda_i} e^{-|y|/\lambda_i} dy + 2\int_{\Theta}^{\infty} (y-\Theta)\frac{1}{2\lambda_i} e^{-y/\lambda_i} dy \ .$$

Note that

$$\int_{\mathbb{R}} (\Theta - y)\frac{1}{2\lambda_i} e^{-|y|/\lambda_i} dy = E_Y(\Theta - Y) = \Theta - E_Y[Y] = \Theta \ ,$$

[5] Since the Laplacian family is closed under this transformation and also here we are taking X to have mean θ_i.

and

$$\int_{\Theta}^{\infty} (y-\Theta)\frac{1}{2\lambda_i}e^{-y/\lambda_i}dy = \frac{1}{2}\lambda_i e^{-\Theta/\lambda_i}\int_0^{\infty} we^{-w}dw \ , \text{ with } w = \frac{y-\Theta}{\lambda_i}$$
$$= \frac{1}{2}\lambda_i e^{-\Theta/\lambda_i} \ ,$$

since the last integral can be interpreted as a Gamma pdf with $k=2$ and $\theta=1$ over its support. Thus, for $\Theta > 0$

$$E_Y\left[|Y-\Theta|\right] = \Theta + \lambda_i e^{-\Theta/\lambda_i} \ .$$

Similarly, considering $\Theta < 0$ we can write

$$\begin{aligned}
&\int_{-\infty}^{\Theta} (\Theta-y)\frac{1}{2\lambda_i}e^{-|y|/\lambda_i}dy + \int_{\Theta}^{\infty} (y-\Theta)\frac{1}{2\lambda_i}e^{-|y|/\lambda_i}dy \\
&= 2\int_{-\infty}^{\Theta} (\Theta-y)\frac{1}{2\lambda_i}e^{y/\lambda_i}dy + \int_{\mathbb{R}} (y-\Theta)\frac{1}{2\lambda_i}e^{-|y|/\lambda_i}dy \\
&= \lambda_i e^{\Theta/\lambda_i}\int_0^{\infty} we^{-w}dw - \Theta \\
&= \lambda_i e^{\Theta/\lambda_i} - \Theta \ .
\end{aligned}$$

Putting the two cases together we find

$$\begin{aligned}
E_Y\left[|Y-\Theta|\right] &= |\Theta| + \lambda_i e^{-|\Theta|/\lambda_i} \\
&= |\theta_i - \theta_j| + \lambda_i e^{-|\theta_i-\theta_j|/\lambda_i} \ ,
\end{aligned}$$

and so

$$\begin{aligned}
E_{f_i}\left[\ln f_j\right] &= -\ln 2\lambda_j - \frac{1}{\lambda_j}E_{f_i}\left[|X-\theta_j|\right] \\
&= -\ln 2\lambda_j - \frac{1}{\lambda_j}E_Y\left[|Y-\Theta|\right] \\
&= -\left[\ln 2\lambda_j + \frac{|\theta_i-\theta_j|}{\lambda_j} + \frac{\lambda_i}{\lambda_j}e^{-|\theta_i-\theta_j|/\lambda_i}\right] .
\end{aligned}$$

□

Corollary 2.3.2. *The differential entropy of f_i is*

$$h(f_i) = \ln 2\lambda_i e \ .$$

Proof. Setting $i = j$ in Proposition 2.3.1 we have

$$h(f_i) = -E_{f_i}\left[\ln f_i\right] = \ln 2\lambda_i + \frac{\lambda_i}{\lambda_i} e^{-|\theta_i-\theta_i|/\lambda_i} + \frac{|\theta_i - \theta_i|}{\lambda_i} = \ln 2\lambda_i e \ .$$

□

Proposition 2.3.3. *The Kullback-Liebler divergence between f_i and f_j is*

$$D(f_i||f_j) = \ln\frac{\lambda_j}{\lambda_i} + \frac{|\theta_i - \theta_j|}{\lambda_j} + \frac{\lambda_i}{\lambda_j} e^{-|\theta_i-\theta_j|/\lambda_i} - 1 \ .$$

Proof. Using Proposition 2.3.1 and Remark 1.2.4 we have

$$\begin{aligned}
D(f_i||f_j) &= E_{f_i}[\ln f_i] - E_{f_i}[\ln f_j] \\
&= -\ln 2\lambda_i e + \left[\ln 2\lambda_j + + \frac{\lambda_i}{\lambda_j} e^{-|\theta_i-\theta_j|/\lambda_i} + \frac{|\theta_i - \theta_j|}{\lambda_j}\right] \\
&= \ln\frac{\lambda_j}{\lambda_i} + \frac{\lambda_i}{\lambda_j} e^{-|\theta_i-\theta_j|/\lambda_i} + \frac{|\theta_i - \theta_j|}{\lambda_j} - 1 \ .
\end{aligned}$$

□

Proposition 2.3.4. *Let $\alpha \in \mathbb{R}^+ \setminus \{1\}$. Then the Rényi divergence between f_i and f_j is given by the following three cases*

1. *If $\alpha = \alpha_0 := \lambda_i/(\lambda_i + \lambda_j)$ then*

$$D_{\alpha_0}(f_i||f_j) = \ln\frac{\lambda_j}{\lambda_i} + \frac{|\theta_i - \theta_j|}{\lambda_j} + \frac{\lambda_i + \lambda_j}{\lambda_j} \ln\left(\frac{2\lambda_i}{\lambda_i + \lambda_j + |\theta_i - \theta_j|}\right) \ .$$

2. *If $\alpha \neq \lambda_i/(\lambda_i + \lambda_j)$ and $\alpha\lambda_j + (1-\alpha)\lambda_i > 0$ then*

$$\begin{aligned}
D_\alpha(f_i||f_j) &= \ln\frac{\lambda_j}{\lambda_i} + \frac{1}{\alpha - 1}\ln\left(\frac{\lambda_i\lambda_j^2}{\alpha^2\lambda_j^2 - (1-\alpha)^2\lambda_i^2}\right) \\
&+ \frac{1}{\alpha-1}\ln\left(\frac{\alpha}{\lambda_i}\exp\left(-\frac{(1-\alpha)|\theta_i - \theta_j|}{\lambda_j}\right) - \frac{1-\alpha}{\lambda_j}\exp\left(\frac{-\alpha|\theta_i-\theta_j|}{\lambda_i}\right)\right).
\end{aligned}$$

3. *If* $\alpha\lambda_j + (1-\alpha)\lambda_i \leq 0$

$$D_\alpha(f_i||f_j) = +\infty \; .$$

Proof. We have

$$\begin{aligned} f_i^\alpha f_j^{1-\alpha} &= \left(\frac{1}{2\lambda_i}\right)^\alpha e^{-(\alpha/\lambda_i)|x-\theta_i|} \left(\frac{1}{2\lambda_j}\right)^{1-\alpha} e^{-(1-\alpha)/\lambda_j|x-\theta_j|} \\ &= \left(\frac{\lambda_j}{\lambda_i}\right)^{\alpha-1} \frac{1}{2\lambda_i}\, e^{-\left[(\alpha/\lambda_i)|x-\theta_i|+(1-\alpha)/\lambda_j|x-\theta_j|\right]} \; . \end{aligned}$$

Let $\theta_M = \max\{\theta_i, \theta_j\}$ and $\theta_m = \min\{\theta_i, \theta_j\}$. Then

$$\begin{aligned} I :=& \int_{\mathbb{R}} e^{-\left[\alpha/\lambda_i|x-\theta_i|+(1-\alpha)/\lambda_j|x-\theta_j|\right]}\, dx \\ =& \int_{-\infty}^{\theta_m} e^{\alpha(x-\theta_i)/\lambda_i+(1-\alpha)(x-\theta_j)/\lambda_j}\, dx \\ &+ \int_{\theta_m}^{\theta_M} \exp\left(-\frac{\alpha(x-\theta_i)}{\lambda_i}\mathrm{sgn}(\theta_j-\theta_i) + \frac{(1-\alpha)(x-\theta_j)}{\lambda_j}\mathrm{sgn}(\theta_j-\theta_i)\right) dx \\ &+ \int_{\theta_M}^{\infty} e^{-[\alpha(x-\theta_i)/\lambda_i+(1-\alpha)(x-\theta_j)/\lambda_j]}\, dx \; . \end{aligned}$$

Note that,

$$I_1 := \int_{-\infty}^{\theta_m} e^{\alpha(x-\theta_i)/\lambda_i+(1-\alpha)(x-\theta_j)/\lambda_j}\, dx = e^{-\theta_0} \int_{-\infty}^{\theta_m} e^{\lambda_0 x}\, dx$$

where

$$\theta_0 = \frac{\alpha\lambda_j\theta_i + (1-\alpha)\lambda_i\theta_j}{\lambda_i\lambda_j} \; , \text{ and } \; \lambda_0 = \frac{\alpha\lambda_j + (1-\alpha)\lambda_i}{\lambda_i\lambda_j} \; ,$$

hence

$$I_1 = \begin{cases} \infty & \text{if } \lambda_0 \leq 0 \; , \\ \dfrac{\exp\left(\lambda_0\theta_m - \theta_0\right)}{\lambda_0} & \text{if } \lambda_0 > 0 \; . \end{cases}$$

Similarly,

$$I_3 := \int_{\theta_M}^{\infty} e^{-[\alpha(x-\theta_i)/\lambda_i+(1-\alpha)(x-\theta_j)/\lambda_j]}\,dx = \begin{cases} \infty & \text{if } \lambda_0 \leq 0\ , \\ \dfrac{\exp\left(\theta_0 - \lambda_0\theta_M\right)}{\lambda_0} & \text{if } \lambda_0 > 0\ , \end{cases}$$

Since

$$\begin{aligned} \lambda_0\theta_m - \theta_0 &= \frac{\alpha\lambda_j + (1-\alpha)\lambda_i}{\lambda_i\lambda_j}\theta_m - \frac{\alpha\lambda_j\theta_i + (1-\alpha)\lambda_i\theta_j}{\lambda_i\lambda_j} \\ &= \frac{\alpha\lambda_j(\theta_m - \theta_i) + (1-\alpha)\lambda_i(\theta_m - \theta_j)}{\lambda_i\lambda_j} \end{aligned}$$

and

$$\theta_0 - \lambda_0\theta_M = -\frac{\alpha\lambda_j(\theta_M - \theta_i) + (1-\alpha)\lambda_i(\theta_M - \theta_j)}{\lambda_i\lambda_j}\ .$$

Then for $\theta_i = \theta_m$ we have

$$\begin{aligned} &\exp\left(\lambda_0\theta_m - \theta_0\right) + \exp\left(\theta_0 - \lambda_0\theta_M\right) \\ &\quad = \exp\left(\frac{(1-\alpha)(\theta_i - \theta_j)}{\lambda_j}\right) + \exp\left(\frac{-\alpha(\theta_j - \theta_i)}{\lambda_i}\right) \\ &\quad = \exp\left(\frac{(1-\alpha)(\theta_i - \theta_j)}{\lambda_j}\right) + \exp\left(\frac{\alpha(\theta_i - \theta_j)}{\lambda_i}\right)\ , \end{aligned}$$

while for $\theta_i = \theta_M$ we have

$$\begin{aligned} &\exp\left(\lambda_0\theta_m - \theta_0\right) + \exp\left(\theta_0 - \lambda_0\theta_M\right) \\ &\quad = \exp\left(\frac{\alpha(\theta_j - \theta_i)}{\lambda_i}\right) + \exp\left(-\frac{(1-\alpha)(\theta_i - \theta_j)}{\lambda_j}\right) \\ &\quad = \exp\left(\frac{-\alpha(\theta_i - \theta_j)}{\lambda_i}\right) + \exp\left(-\frac{(1-\alpha)(\theta_i - \theta_j)}{\lambda_j}\right)\ , \end{aligned}$$

which together imply that

$$\begin{aligned} &\exp\left(\lambda_0\theta_m - \theta_0\right) + \exp\left(\theta_0 - \lambda_0\theta_M\right) \\ &\quad = \exp\left(-\alpha\frac{|\theta_i - \theta_j|}{\lambda_i}\right) + \exp\left(-(1-\alpha)\frac{|\theta_i - \theta_j|}{\lambda_j}\right)\ . \end{aligned}$$

Thus, for $\lambda_0 > 0$, we have

$$
\begin{aligned}
I_1 + I_3 &= \frac{\exp\left(\lambda_0\theta_m - \theta_0\right) + \exp\left(\theta_0 - \lambda_0\theta_M\right)}{\lambda_0} \\
&= \frac{\lambda_i\lambda_j}{\alpha\lambda_j + (1-\alpha)\lambda_i}\left[\exp\left(-\alpha\frac{|\theta_i - \theta_j|}{\lambda_i}\right) + \exp\left(-(1-\alpha)\frac{|\theta_i - \theta_j|}{\lambda_j}\right)\right] .
\end{aligned}
$$

Now,

$$
\begin{aligned}
I_2 &= \int_{\theta_m}^{\theta_M} \exp\left(-\frac{\alpha(x-\theta_i)}{\lambda_i}\mathrm{sgn}(\theta_j - \theta_i) + \frac{(1-\alpha)(x-\theta_j)}{\lambda_j}\mathrm{sgn}(\theta_j - \theta_i)\right) dx \\
&= \int_{\theta_m}^{\theta_M} \exp\left[\frac{\alpha(x-\theta_i)}{\lambda_i}\mathrm{sgn}(\theta_i - \theta_j) - \frac{(1-\alpha)(x-\theta_j)}{\lambda_j}\mathrm{sgn}(\theta_i - \theta_j)\right] dx \\
&= \int_{\theta_m}^{\theta_M} \exp\left(\mathrm{sgn}(\theta_i - \theta_j)\left[\frac{\alpha(x-\theta_i)}{\lambda_i} - \frac{(1-\alpha)(x-\theta_j)}{\lambda_j}\right]\right) dx \\
&= \int_{\theta_m}^{\theta_M} \exp\left[\mathrm{sgn}(\theta_i - \theta_j)\left(\tilde{\lambda}x - \tilde{\theta}\right)\right] dx ,
\end{aligned}
$$

where

$$
\tilde{\theta} = \frac{\alpha\lambda_j\theta_i + (\alpha-1)\lambda_i\theta_j}{\lambda_i\lambda_j} , \text{ and } \tilde{\lambda} = \frac{\alpha\lambda_j + (\alpha-1)\lambda_i}{\lambda_i\lambda_j} ,
$$

and so

$$
\begin{aligned}
I_2 &= \exp\left(-\mathrm{sgn}(\theta_i - \theta_j)\tilde{\theta}\right) \\
&\quad\cdot \begin{cases} \dfrac{\exp\left(\mathrm{sgn}(\theta_i - \theta_j)\tilde{\lambda}\theta_M\right) - \exp\left(\mathrm{sgn}(\theta_i - \theta_j)\tilde{\lambda}\theta_m\right)}{\mathrm{sgn}(\theta_i - \theta_j)\tilde{\lambda}} & \tilde{\lambda} \neq 0 \\ (\theta_M - \theta_m) & \tilde{\lambda} = 0 \end{cases} \\
&= \begin{cases} \dfrac{\exp\left(\mathrm{sgn}(\theta_i - \theta_j)(\tilde{\lambda}\theta_M - \tilde{\theta})\right) - \exp\left(\mathrm{sgn}(\theta_i - \theta_j)(\tilde{\lambda}\theta_m - \tilde{\theta})\right)}{\mathrm{sgn}(\theta_i - \theta_j)\tilde{\lambda}} & \tilde{\lambda} \neq 0 \\ |\theta_i - \theta_j| \exp\left(-\mathrm{sgn}(\theta_i - \theta_j)\tilde{\theta}\right) & \tilde{\lambda} = 0 \end{cases}
\end{aligned}
$$

with the obvious assumption of $\theta_i \neq \theta_j$.

- Consider first the case $\tilde{\lambda} = 0$. Note that

$$\tilde{\lambda} = 0 \Leftrightarrow \alpha\lambda_j + (\alpha - 1)\lambda_i = 0 \Leftrightarrow \alpha = \alpha_0 := \frac{\lambda_i}{\lambda_i + \lambda_j} ,$$

and we see that $\tilde{\lambda} = 0$ occurs only for $\alpha \in (0,1)$ since $\lambda_i, \lambda_j > 0$. Thus $\tilde{\lambda} = 0 \Rightarrow \lambda_0 > 0$ (being in this case the convex combination of two positive numbers) and all of I_1, I_2 and I_3 assume finite values. Hence

$$\begin{aligned} I &= \int_{\mathbb{R}} e^{-\left[\alpha/\lambda_i|x-\theta_i|+(1-\alpha)/\lambda_j|x-\theta_j|\right]}\, dx \\ &= I_1 + I_3 + I_2 \\ &= \frac{\lambda_i\lambda_j}{\alpha\lambda_j + (1-\alpha)\lambda_i}\left[\exp\left(-\alpha\frac{|\theta_i - \theta_j|}{\lambda_i}\right) + \exp\left(-(1-\alpha)\frac{|\theta_i - \theta_j|}{\lambda_j}\right)\right] \\ &\quad + |\theta_i - \theta_j| \exp\left(-\operatorname{sgn}(\theta_i - \theta_j)\tilde{\theta}\right) . \end{aligned}$$

Since

$$\alpha = \frac{\lambda_i}{\lambda_i + \lambda_j} \Leftrightarrow 1 - \alpha = \frac{\lambda_j}{\lambda_i + \lambda_j} ,$$

$$\tilde{\theta} = \frac{\alpha\lambda_j\theta_i + (\alpha - 1)\lambda_i\theta_j}{\lambda_i\lambda_j} = \frac{1}{\lambda_i\lambda_j}\left[\frac{\lambda_i\lambda_j\theta_i}{\lambda_i + \lambda_j} - \frac{\lambda_i\lambda_j\theta_j}{\lambda_i + \lambda_j}\right] = \frac{\theta_i - \theta_j}{\lambda_i + \lambda_j} ,$$

$$\alpha\lambda_j + (1-\alpha)\lambda_i = \frac{\lambda_i\lambda_j}{\lambda_i + \lambda_j} + \frac{\lambda_i\lambda_j}{\lambda_i + \lambda_j} = \frac{2\lambda_i\lambda_j}{\lambda_i + \lambda_j} ,$$

and

$$\frac{\alpha}{\lambda_i} = \frac{1}{\lambda_i + \lambda_j} = \frac{1-\alpha}{\lambda_j} ,$$

we have

$$\begin{aligned} I &= \frac{\lambda_i + \lambda_j}{2} 2\exp\left(-\frac{|\theta_i - \theta_j|}{\lambda_i + \lambda_j}\right) + |\theta_i - \theta_j| \exp\left(-\operatorname{sgn}(\theta_i - \theta_j)\frac{\theta_i - \theta_j}{\lambda_i + \lambda_j}\right) \\ &= \exp\left(-\frac{|\theta_i - \theta_j|}{\lambda_i + \lambda_j}\right)\left[\lambda_i + \lambda_j + |\theta_i - \theta_j|\right] . \end{aligned}$$

Finally,

$$
\begin{aligned}
&D_{\alpha_0}(f_i||f_j) \\
&\quad = \frac{1}{\alpha_0 - 1} \ln \int_{\mathbb{R}} f_i^{\alpha_0} f_j^{1-\alpha_0}\, dx \\
&\quad = \frac{1}{\alpha_0 - 1} \ln \left(\left[\frac{\lambda_j}{\lambda_i} \right]^{\alpha_0 - 1} \frac{1}{2\lambda_i} \int_{\mathbb{R}} e^{-\left[(\alpha_0/\lambda_i)|x-\theta_i| + (1-\alpha_0)/\lambda_j |x-\theta_j|\right]}\, dx \right) \\
&\quad = \ln \frac{\lambda_j}{\lambda_i} + \frac{1}{\alpha_0 - 1} \ln \frac{I}{2\lambda_i} \\
&\quad = \ln \frac{\lambda_j}{\lambda_i} - \frac{\lambda_i + \lambda_j}{\lambda_j} \ln \left(\frac{1}{2\lambda_i} \exp \left(-\frac{|\theta_i - \theta_j|}{\lambda_i + \lambda_j} \right) \left[\lambda_i + \lambda_j + |\theta_i - \theta_j| \right] \right) \\
&\quad = \ln \frac{\lambda_j}{\lambda_i} - \frac{\lambda_i + \lambda_j}{\lambda_j} \left[-\frac{|\theta_i - \theta_j|}{\lambda_i + \lambda_j} + \ln \left(\frac{\lambda_i + \lambda_j + |\theta_i - \theta_j|}{2\lambda_i} \right) \right] \\
&\quad = \ln \frac{\lambda_j}{\lambda_i} + \frac{|\theta_i - \theta_j|}{\lambda_j} + \frac{\lambda_i + \lambda_j}{\lambda_j} \ln \left(\frac{2\lambda_i}{\lambda_i + \lambda_j + |\theta_i - \theta_j|} \right) .
\end{aligned}
$$

- If $\tilde{\lambda} \neq 0$ and $\lambda_0 > 0$ then

$$
I_2 = \frac{\exp\left(\operatorname{sgn}(\theta_i - \theta_j)(\tilde{\lambda}\theta_M - \tilde{\theta})\right) - \exp\left(\operatorname{sgn}(\theta_i - \theta_j)(\tilde{\lambda}\theta_m - \tilde{\theta})\right)}{\operatorname{sgn}(\theta_i - \theta_j)\tilde{\lambda}} .
$$

Since

$$
\begin{aligned}
\tilde{\lambda}\theta_M - \tilde{\theta} &= \frac{\alpha\lambda_j + (\alpha - 1)\lambda_i}{\lambda_i \lambda_j} \theta_M - \frac{\alpha\lambda_j \theta_i + (\alpha - 1)\lambda_i \theta_j}{\lambda_i \lambda_j} \\
&= \frac{\alpha\lambda_j(\theta_M - \theta_i) + (\alpha - 1)\lambda_i(\theta_M - \theta_j)}{\lambda_i \lambda_j}
\end{aligned}
$$

and

$$
\tilde{\lambda}\theta_m - \tilde{\theta} = \frac{\alpha\lambda_j(\theta_m - \theta_i) + (\alpha - 1)\lambda_i(\theta_m - \theta_j)}{\lambda_i \lambda_j}
$$

then by considering the two cases $\theta_i = \theta_M$ and $\theta_j = \theta_M$ as before we see that

$$
I_2 = \frac{1}{\tilde{\lambda}} \left[\exp\left(-\frac{(1-\alpha)|\theta_i - \theta_j|}{\lambda_j} \right) - \exp\left(\frac{-\alpha|\theta_i - \theta_j|}{\lambda_i} \right) \right] .
$$

Thus

$$
\begin{aligned}
I &= I_1 + I_3 + I_2 \\
&= \frac{1}{\lambda_0}\left[\exp\left(-\alpha\frac{|\theta_i-\theta_j|}{\lambda_i}\right) + \exp\left(-(1-\alpha)\frac{|\theta_i-\theta_j|}{\lambda_j}\right)\right] \\
&\quad + \frac{1}{\tilde{\lambda}}\left[\exp\left(-\frac{(1-\alpha)|\theta_i-\theta_j|}{\lambda_j}\right) - \exp\left(\frac{-\alpha|\theta_i-\theta_j|}{\lambda_i}\right)\right] \\
&= \exp\left(-\alpha\frac{|\theta_i-\theta_j|}{\lambda_i}\right)\left[\frac{1}{\lambda_0}-\frac{1}{\tilde{\lambda}}\right] + \exp\left(-(1-\alpha)\frac{|\theta_i-\theta_j|}{\lambda_j}\right)\left[\frac{1}{\lambda_0}+\frac{1}{\tilde{\lambda}}\right] ,
\end{aligned}
$$

where

$$
\begin{aligned}
\frac{1}{\lambda_0}+\frac{1}{\tilde{\lambda}} &= \lambda_i\lambda_j\left[\frac{1}{\alpha\lambda_j+(1-\alpha)\lambda_i} + \frac{1}{\alpha\lambda_j+(\alpha-1)\lambda_i}\right] \\
&= \lambda_i\lambda_j\left[\frac{2\alpha\lambda_j}{\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2}\right] \\
&= \frac{2\lambda_i^2\lambda_j^2}{\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2}\frac{\alpha}{\lambda_i}
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{\lambda_0}-\frac{1}{\tilde{\lambda}} &= \lambda_i\lambda_j\left[\frac{2(1-\alpha)\lambda_i}{(1-\alpha)^2\lambda_i^2-a^2\lambda_j^2}\right] \\
&= -\frac{2\lambda_i^2\lambda_j^2}{\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2}\frac{1-\alpha}{\lambda_j} .
\end{aligned}
$$

Hence

$$
I = \frac{2\lambda_i^2\lambda_j^2}{\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2}\left[\frac{\alpha}{\lambda_i}\exp\left(-\frac{(1-\alpha)|\theta_i-\theta_j|}{\lambda_j}\right) - \frac{1-\alpha}{\lambda_j}\exp\left(\frac{-\alpha|\theta_i-\theta_j|}{\lambda_i}\right)\right] .
$$

Finally,

$$
\begin{aligned}
&D_\alpha(f_i||f_j) \\
&\quad = \frac{1}{\alpha-1}\ln\int_{\mathbb{R}} f_i^{\alpha} f_j^{1-\alpha}\,dx \\
&\quad = \frac{1}{\alpha-1}\ln\left(\left[\frac{\lambda_j}{\lambda_i}\right]^{\alpha-1}\frac{1}{2\lambda_i}\int_{\mathbb{R}} e^{-\left[(\alpha/\lambda_i)|x-\theta_i|+(1-\alpha)/\lambda_j|x-\theta_j|\right]}\,dx\right) \\
&\quad = \ln\frac{\lambda_j}{\lambda_i}+\frac{1}{\alpha-1}\ln\frac{I}{2\lambda_i} \\
&\quad = \ln\frac{\lambda_j}{\lambda_i}+\frac{1}{\alpha-1}\ln\left(\frac{\lambda_i\lambda_j^2}{\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2}\right) \\
&\quad + \frac{1}{\alpha-1}\ln\left(\frac{\alpha}{\lambda_i}\exp\left(-\frac{(1-\alpha)|\theta_i-\theta_j|}{\lambda_j}\right)-\frac{1-\alpha}{\lambda_j}\exp\left(\frac{-\alpha|\theta_i-\theta_j|}{\lambda_i}\right)\right)\ .
\end{aligned}
$$

- If $\lambda_0 \le 0$ then $I_1 = I_3 = \infty$, and since this case can only happen for $\alpha > 1$ (given λ_i and λ_j are positive numbers), then

$$
D_\alpha(f_i||f_j) = \frac{1}{\alpha-1}\ln\int_{\mathbb{R}} f_i^{\alpha} f_j^{1-\alpha}\,dx = \infty\ .
$$

□

Remark 2.3.5.

$$
\lim_{\alpha\uparrow 1} D_\alpha(f_i||f_j) = D(f_i||f_j)\ .
$$

Proof. Note that the term

$$
\frac{\lambda_i\lambda_j^2}{\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2}\left[\frac{\alpha}{\lambda_i}\exp\left(-(1-\alpha)\frac{|\theta_i-\theta_j|}{\lambda_j}\right)-\frac{(1-\alpha)}{\lambda_j}\exp\left(-\alpha\frac{|\theta_i-\theta_j|}{\lambda_i}\right)\right]
$$

approaches 1 as $\alpha \to 1$. Grouping the second and third logarithms in the expression for D_α above we see this attains an indeterminate limit. Applying l'Hospital's rule we can rewrite the limit as

$$\lim_{\alpha \uparrow 1} \left[-\frac{2\alpha\lambda_j^2 + 2(1-\alpha)\lambda_i^2}{\alpha^2\lambda_j^2 - (1-\alpha)^2\lambda_i^2} + \frac{g'(\alpha)}{g(\alpha)} \right] ,$$

where

$$g(\alpha) := \frac{\alpha}{\lambda_i} \exp\left(-(1-\alpha)\frac{|\theta_i - \theta_j|}{\lambda_j} \right) - \frac{(1-\alpha)}{\lambda_j} \exp\left(-\alpha \frac{|\theta_i - \theta_j|}{\lambda_i} \right)$$

and

$$\begin{aligned} g'(\alpha) = \exp\left(-(1-\alpha)\frac{|\theta_i - \theta_j|}{\lambda_j} \right) \left[\frac{1}{\lambda_i} + \frac{\alpha|\theta_i - \theta_j|}{\lambda_i\lambda_j} \right] \\ + \exp\left(-\alpha\frac{|\theta_i - \theta_j|}{\lambda_i} \right) \left[\frac{1}{\lambda_j} + \frac{(1-\alpha)|\theta_i - \theta_j|}{\lambda_i\lambda_j} \right] . \end{aligned}$$

Then, with $\alpha \to 1$ we have $g(\alpha) \to 1/\lambda_i$ and

$$g'(\alpha) \to \frac{1}{\lambda_i} + \frac{|\theta_i - \theta_j|}{\lambda_i\lambda_j} + \frac{1}{\lambda_j} \exp\left(-\frac{|\theta_i - \theta_j|}{\lambda_i} \right)$$

Finally,

$$\begin{aligned} \lim_{\alpha \uparrow 1} D_\alpha(f_i||f_j) &= \ln\frac{\lambda_j}{\lambda_i} - 2 \\ &\quad + \lambda_i \left[\frac{1}{\lambda_i} + \frac{|\theta_i - \theta_j|}{\lambda_i\lambda_j} + \frac{1}{\lambda_j} \exp\left(-\frac{|\theta_i - \theta_j|}{\lambda_i} \right) \right] \\ &= \ln\frac{\lambda_j}{\lambda_i} + \frac{|\theta_i - \theta_j|}{\lambda_j} + \frac{\lambda_i}{\lambda_j} \exp\left(-\frac{|\theta_i - \theta_j|}{\lambda_i} \right) - 1 \end{aligned}$$

as given by Proposition 2.3.3. □

Proposition 2.3.6. *$D_\alpha(f_i||f_j)$ is continuous at $\alpha = \lambda_i/(\lambda_i + \lambda_j)$.*

Proof. Let $\alpha_0 = \lambda_i/(\lambda_i + \lambda_j)$. Since

$$\frac{\alpha_0}{\lambda_i} = \frac{1-\alpha_0}{\lambda_j} = \frac{1}{\lambda_i + \lambda_j} ,$$

we see that both the terms

$$\frac{\alpha_0}{\lambda_i}\exp\left(-(1-\alpha_0)\frac{|\theta_i-\theta_j|}{\lambda_j}\right)-\frac{(1-\alpha_0)}{\lambda_j}\exp\left(-\alpha_0\frac{|\theta_i-\theta_j|}{\lambda_i}\right)$$

and $\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2$ approach 0 as $\alpha\to\alpha_0$, the limit

$$\begin{aligned}\lim_{\alpha\to\alpha_0}\Bigg(&\frac{\lambda_i\lambda_j^2}{\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2}\\ &\cdot\left[\frac{\alpha}{\lambda_i}\exp\left(-(1-\alpha)\frac{|\theta_i-\theta_j|}{\lambda_j}\right)-\frac{(1-\alpha)}{\lambda_j}\exp\left(-\alpha\frac{|\theta_i-\theta_j|}{\lambda_i}\right)\right]\Bigg)\end{aligned}$$

has an indeterminate form. Applying l'Hospital's this limit becomes

$$\lim_{\alpha\to\alpha_0}\left[\frac{\lambda_i\lambda_j^2 g'(\alpha)}{2\alpha\lambda_j^2+2(1-\alpha)\lambda_i^2}\right],$$

where

$$\begin{aligned}g'(\alpha)=&\exp\left(-(1-\alpha)\frac{|\theta_i-\theta_j|}{\lambda_j}\right)\left[\frac{1}{\lambda_i}+\frac{\alpha|\theta_i-\theta_j|}{\lambda_i\lambda_j}\right]\\ &+\exp\left(-\alpha\frac{|\theta_i-\theta_j|}{\lambda_i}\right)\left[\frac{1}{\lambda_j}+\frac{(1-\alpha)|\theta_i-\theta_j|}{\lambda_i\lambda_j}\right],\end{aligned}$$

(defining $g(\alpha)$ as in the proof of Remark 2.3.5). But

$$\begin{aligned}\lim_{\alpha\to\alpha_0}g'(\alpha)&=\exp\left(-\frac{|\theta_i-\theta_j|}{\lambda_i+\lambda_j}\right)\left(\frac{1}{\lambda_i}+\frac{1}{\lambda_j}\right)\left[1+\frac{|\theta_i-\theta_j|}{\lambda_i+\lambda_j}\right]\\ &=\frac{1}{\lambda_i\lambda_j}\exp\left(\frac{-|\theta_i-\theta_j|}{\lambda_i+\lambda_j}\right)\left[\lambda_i+\lambda_j+|\theta_i-\theta_j|\right],\end{aligned}$$

and

$$\lim_{\alpha\to\alpha_0}\alpha\lambda_j^2+(1-\alpha)\lambda_i^2=2\frac{\lambda_i\lambda_j^2+\lambda_i^2\lambda_j}{\lambda_i+\lambda_j}=2\lambda_i\lambda_j\,.$$

Thus,

$$\lim_{\alpha\to\alpha_0}\left(\frac{\lambda_i\lambda_j^2}{\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2}\right.$$
$$\left.\cdot\left[\frac{\alpha}{\lambda_i}\exp\left(-(1-\alpha)\frac{|\theta_i-\theta_j|}{\lambda_j}\right)-\frac{(1-\alpha)}{\lambda_j}\exp\left(-\alpha\frac{|\theta_i-\theta_j|}{\lambda_i}\right)\right]\right)$$
$$=\frac{\lambda_i\lambda_j^2}{2\lambda_i\lambda_j}\frac{1}{\lambda_i\lambda_j}\exp\left(-\frac{|\theta_i-\theta_j|}{\lambda_i+\lambda_j}\right)\left[\lambda_i+\lambda_j+|\theta_i-\theta_j|\right]$$
$$=\frac{1}{2\lambda_i}\exp\left(-\frac{|\theta_i-\theta_j|}{\lambda_i+\lambda_j}\right)\left[\lambda_i+\lambda_j+|\theta_i-\theta_j|\right].$$

Finally, by the continuity of the logarithm on $(0,\infty)$,

$$\lim_{\alpha\to\alpha_0}D_\alpha(f_i||f_j)$$
$$=\ln\frac{\lambda_j}{\lambda_i}+\frac{1}{\alpha_0-1}\ln\left(\frac{1}{2\lambda_i}\exp\left(-\frac{|\theta_i-\theta_j|}{\lambda_i+\lambda_j}\right)\left[\lambda_i+\lambda_j+|\theta_i-\theta_j|\right]\right)$$
$$=\ln\frac{\lambda_j}{\lambda_i}-\frac{\lambda_i+\lambda_j}{\lambda_j}\ln\left(\frac{1}{2\lambda_i}\exp\left(-\frac{|\theta_i-\theta_j|}{\lambda_i+\lambda_j}\right)\left[\lambda_i+\lambda_j+|\theta_i-\theta_j|\right]\right)$$
$$=\ln\frac{\lambda_j}{\lambda_i}+\frac{|\theta_i-\theta_j|}{\lambda_j}+\frac{\lambda_i+\lambda_j}{\lambda_j}\ln\left(\frac{2\lambda_i}{\lambda_i+\lambda_j+|\theta_i-\theta_j|}\right),$$

which was indeed the value we obtained for $D_\alpha(f_i||f_j)$ when $\alpha=\lambda_i/(\lambda_i+\lambda_j)$ in Case 1 of Proposition 2.3.4, as expected from the continuity of D_α. □

Remark 2.3.7. If we set $\theta_i=\theta_j=0$ in Proposition 2.3.4 we obtain

$$D_\alpha(f_i||f_j)=\ln\frac{\lambda_j}{\lambda_i}+\frac{1}{\alpha-1}\ln\left(\frac{\lambda_i\lambda_j^2}{\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2}\right)+\frac{1}{\alpha-1}\ln\left(\frac{\alpha}{\lambda_i}-\frac{1-\alpha}{\lambda_j}\right)$$
$$=\ln\frac{\lambda_j}{\lambda_i}+\frac{1}{\alpha-1}\ln\left(\frac{\lambda_i\lambda_j^2}{\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2}\frac{\alpha\lambda_j-(1-\alpha)\lambda_i}{\lambda_j\lambda_i}\right)$$
$$=\ln\frac{\lambda_j}{\lambda_i}+\frac{1}{\alpha-1}\ln\left(\frac{\lambda_j}{\alpha\lambda_j+(1-\alpha)\lambda_i}\right)$$
$$=\ln\frac{\lambda_j}{\lambda_i}+\frac{1}{\alpha-1}\ln\frac{\lambda_j}{\lambda_\alpha^*},$$

since $\lambda_\alpha^* := \alpha\lambda_j + (1-\alpha)\lambda_i$ (see Section 2.1), which is consistent with the result derived in Section 2.2.6 using the expression for exponential Families.

2.3.2 Rényi and Kullback Divergence for Cramér Distributions

We consider here the distributions identified by Song [57] as Cramér[6]. Let f_i and f_j be two Cramér densities:

$$f_i(x) = \frac{\theta_i}{2(1+\theta_i|x|)^2} \quad \theta_i > 0;\ x \in \mathbb{R}\ .$$

Proposition 2.3.8.

$$E_{f_i}[\ln f_j] = \begin{cases} \ln \dfrac{\theta_j}{2} - \dfrac{2\theta_j}{\theta_j - \theta_i} \ln \dfrac{\theta_j}{\theta_i} & \text{if } \theta_i \neq \theta_j \\ \ln \dfrac{\theta}{2} - 2 & \text{if } \theta_i = \theta_j = \theta \end{cases}$$

Proof.

$$E_{f_i}[\ln f_j] = E_{f_i}\left[\ln \frac{\theta_j}{2} - 2\ln(1+\theta_j|X|)\right] = \ln\frac{\theta_j}{2} - 2E_{f_i}\left[\ln(1+\theta_j|X|)\right]\ ,$$

where

$$E_{f_i}\left[\ln(1+\theta_j|X|)\right] = \int_{\mathbb{R}} \frac{\theta_i \ln(1+\theta_j|x|)}{2(1+\theta_i|x|)^2}\,dx = \int_{\mathbb{R}^+} \frac{\theta_i \ln(1+\theta_j x)}{(1+\theta_i x)^2}\,dx\ .$$

If $\theta_i = \theta_j = \theta$ then

$$\int_{\mathbb{R}^+} \frac{\theta_i \ln(1+\theta_j x)}{(1+\theta_i x)^2}\,dx = \int_{\mathbb{R}^+} w e^{-w}\,dw = 1\ ,$$

we $w = \ln(1+\theta x)$, and the above corresponds to the integration of an exponenetial pdf over its support. Now let $\theta_i \neq \theta_j$. Then

$$\begin{aligned}\int_{\mathbb{R}^+} \frac{\theta_i \ln(1+\theta_j x)}{(1+\theta_i x)^2}\,dx &= \int_1^\infty \frac{\ln(a(u-1)+1)}{u^2}\,du\ ,\ a = \theta_j/\theta_i\ ,\ u = 1+\theta_i x \\ &= \left[-\frac{1}{u}\ln(au-a+1) + a\int \frac{1}{u(au-a+1)}du\right]_1^\infty\ ,\end{aligned}$$

[6] See footnote 1 in Chapter 2

where we have used integration by parts. The first term vanishes at both limits and for the second term we can use the partial fraction decomposition

$$\frac{1}{u(au-a+1)} = \frac{1}{1-a}\left(\frac{1}{u} - \frac{a}{au-a+1}\right) .$$

Then

$$\begin{aligned}\int_{\mathbb{R}^+} \frac{\theta_i \ln(1+\theta_j x)}{(1+\theta_i x)^2} &= \frac{a}{1-a}\int_1^{\infty}\left(\frac{1}{u} - \frac{a}{au-a+1}\right) du \\ &= \frac{a}{1-a}\ln\left(\frac{u}{au-a+1}\right)_1^{\infty} \\ &= \frac{a}{1-a}\left[\lim_{u\to\infty}\ln\left(\frac{u}{au-a+1}\right) - 0\right] \\ &= \frac{a}{1-a}\ln\frac{1}{a} \\ &= \frac{\theta_j}{\theta_i}\frac{\theta_i}{\theta_i-\theta_j}\ln\frac{\theta_i}{\theta_j} \\ &= \frac{\theta_j}{\theta_i-\theta_j}\ln\frac{\theta_i}{\theta_j} ,\end{aligned}$$

where l'Hospital's rule has been used to evaluate the limit. Thus

$$\begin{aligned} E_{f_i}[\ln f_j] &= \ln\frac{\theta_j}{2} - 2E_{f_i}\left[\ln(1+\theta_j|X|)\right] \\ &= \begin{cases} \ln\frac{\theta_j}{2} - \frac{2\theta_j}{\theta_i-\theta_j}\ln\frac{\theta_i}{\theta_j} & \text{if } \theta_i \neq \theta_j \\ \ln\frac{\theta}{2} - 2 & \text{if } \theta_i = \theta_j = \theta \ . \end{cases}\end{aligned}$$

□

Corollary 2.3.9. *The differential entropy of f_i is*

$$h(f_i) = 2 - \ln\frac{\theta_i}{2} .$$

Proof.

$$h(f_i) = -E_{f_i}(\ln f_i) = -\left(\ln\frac{\theta_i}{2} - 2\right) = 2 - \ln\frac{\theta_i}{2} .$$

□

Proposition 2.3.10. *The Kullback-Liebler divergence from f_i to f_j is*

$$D(f_i||f_j) = \frac{\theta_i + \theta_j}{\theta_i - \theta_j} \ln \frac{\theta_i}{\theta_j} - 2 \, .$$

Proof. Using Proposition 2.3.8 and Remark 1.2.4 we have

$$\begin{aligned} D(f_i||f_j) &= E_{f_i}[\ln f_i] - E_{f_i}[\ln f_j] \\ &= \ln \frac{\theta_i}{2} - 2 - \left[\ln \frac{\theta_j}{2} - \frac{2\theta_j}{\theta_i - \theta_j} \ln \frac{\theta_i}{\theta_j} \right] \\ &= \left(1 + \frac{2\theta_j}{\theta_i - \theta_j} \right) \ln \frac{\theta_i}{2} - 2 \\ &= \frac{\theta_i + \theta_j}{\theta_i - \theta_j} \ln \frac{\theta_i}{\theta_j} - 2 \, . \end{aligned}$$

□

Remark 2.3.11. Note that

$$\lim_{\theta_i \to \theta_j} D(f_i||f_j) = \lim_{\theta_i \to \theta_j} \left[\frac{(\theta_i + \theta_j) \ln \frac{\theta_i}{\theta_j}}{\theta_i - \theta_j} - 2 \right] = \lim_{\theta_i \to \theta_j} \left[\frac{\theta_i + \theta_j}{\theta_i} + \ln \frac{\theta_i}{\theta_j} \right] - 2 = 0 \, ,$$

as expected.

Proposition 2.3.12. *Let $\alpha \in \mathbb{R}^+ \setminus \{1\}$. Then the Rényi divergence between f_i and f_j is given by the following cases*

1. *If $\alpha = 1/2$ then*

$$D_{1/2}(f_i||f_j) = \ln \frac{\theta_i}{\theta_j} + 2 \ln \left(\frac{\theta_i}{\theta_i - \theta_j} \ln \frac{\theta_i}{\theta_j} \right) \, .$$

2. *If $\alpha \neq 1/2$ then*

$$D_\alpha(f_i||f_j) = \ln \frac{\theta_i}{\theta_j} + \frac{1}{\alpha - 1} \ln \left(\frac{\theta_i}{(2\alpha - 1)(\theta_i - \theta_j)} \left[1 - \left(\frac{\theta_j}{\theta_i} \right)^{2\alpha - 1} \right] \right) \, .$$

Proof.

$$
\begin{aligned}
f_i^\alpha f_j^{1-\alpha} &= \left[\frac{\theta_i}{2(1+\theta_i|x|)^2}\right]^\alpha \left[\frac{\theta_j}{2(1+\theta_j|x|)^2}\right]^{1-\alpha} \\
&= \theta_i^\alpha \theta_j^{1-\alpha} \frac{1}{2(1+\theta_i|x|)^{2\alpha}} \frac{1}{(1+\theta_j|x|)^{2-2\alpha}} \\
&= \left(\frac{\theta_i}{\theta_j}\right)^{\alpha-1} \frac{\theta_i}{2(1+\theta_i|x|)^{2\alpha}(1+\theta_j|x|)^{2-2\alpha}} \ .
\end{aligned}
$$

Now

$$
\begin{aligned}
&\int_{\mathbb{R}} \frac{\theta_i}{2(1+\theta_i|x|)^{2\alpha}(1+\theta_j|x|)^{2-2\alpha}}\, dx \\
&\quad = 2\int_{\mathbb{R}^+} \frac{\theta_i}{2(1+\theta_i x)^{2\alpha}(1+\theta_j x)^{2-2\alpha}}\, dx \\
&\quad = \int_1^\infty \left(\frac{1}{u}\right)^{2\alpha} \left(\frac{1}{1+a(u-1)}\right)^{2-2\alpha} du\ , \quad u = 1+\theta_i x\ ,\ a = \theta_j/\theta_i \\
&\quad = \int_1^0 v^{2\alpha} \left(\frac{v}{v(1-a)+a}\right)^{2-2\alpha} v^{-2}\, dv\ , \quad v = \frac{1}{u} \\
&\quad = \int_0^1 [v(1-a)+a]^{2\alpha-2}\, dv\ .
\end{aligned}
$$

Since we exclude the case $\alpha = 1$ the exponent in the integrand above is nonzero. Suppose $\alpha = 1/2$. Then

$$
\begin{aligned}
\int_0^1 [v(1-a)+a]^{2\alpha-2}\, dv &= \int_0^1 \frac{1}{v(1-a)+a}\, dv \\
&= \frac{1}{1-a} \ln(v(1-a)+a)\Big|_0^1 \\
&= \frac{\ln a}{a-1} \\
&= \frac{\theta_i}{\theta_j-\theta_i} \ln\frac{\theta_j}{\theta_i} \\
&= \frac{\theta_i}{\theta_i-\theta_j} \ln\frac{\theta_i}{\theta_j}\ .
\end{aligned}
$$

Thus

$$\begin{aligned}
&D_{1/2}(f_i||f_j)\\
&\quad = \frac{1}{\alpha-1}\ln\left(\left(\frac{\theta_i}{\theta_j}\right)^{\alpha-1}\int_{\mathbb{R}}\frac{\theta_i}{2(1+\theta_i|x|)^{2\alpha}(1+\theta_j|x|)^{2-2\alpha}}\,dx\right)\Bigg|_{\alpha=1/2}\\
&\quad = \ln\frac{\theta_i}{\theta_j}+2\ln\left(\frac{\theta_i}{\theta_i-\theta_j}\ln\frac{\theta_i}{\theta_j}\right)\ .
\end{aligned}$$

If $\alpha \neq 1/2$

$$\begin{aligned}
\int_0^1 [v(1-a)+a]^{2\alpha-2}\,dv &= \frac{1}{(1-a)(2\alpha-1)}(v(1-a)+a)^{2\alpha-1}\Bigg|_0^1\\
&= \frac{1}{(1-a)(2\alpha-1)}\left[1-a^{2\alpha-1}\right]\\
&= \frac{\theta_i}{(2\alpha-1)(\theta_i-\theta_j)}\left[1-\left(\frac{\theta_j}{\theta_i}\right)^{2\alpha-1}\right]\ ,
\end{aligned}$$

and

$$D_\alpha(f_i||f_j) = \ln\frac{\theta_i}{\theta_j}+\frac{1}{\alpha-1}\ln\left(\frac{\theta_i}{(2\alpha-1)(\theta_i-\theta_j)}\left[1-\left(\frac{\theta_j}{\theta_i}\right)^{2\alpha-1}\right]\right)\ .$$

□

Remark 2.3.13.

$$\lim_{\alpha\uparrow 1} D_\alpha(f_i||f_j) = D(f_i||f_j)\ .$$

Proof. Since the term

$$\frac{\theta_i}{(2\alpha-1)(\theta_i-\theta_j)}\left[1-\left(\frac{\theta_j}{\theta_i}\right)^{2\alpha-1}\right]$$

approaches 1 as $\alpha \to 1$, we see that the second term in $D_\alpha(f_i||f_j)$ is of indeterminate form. Applying l'Hospitals rule

$$\lim_{\alpha\uparrow 1} D_\alpha(f_i||f_j) = \ln\frac{\theta_i}{\theta_j}+\lim_{\alpha\uparrow 1}\left[-\frac{2}{2\alpha-1}+\frac{g'(\alpha)}{g(\alpha)}\right]\ ,$$

where

$$g(\alpha) := 1 - \left(\frac{\theta_j}{\theta_i}\right)^{2\alpha-1}\;, \text{ and so } \; g'(\alpha) = -2\left(\frac{\theta_j}{\theta_i}\right)^{2\alpha-1} \ln\frac{\theta_j}{\theta_i}\;.$$

Then

$$\lim_{\alpha\uparrow 1}\frac{g'(\alpha)}{g(\alpha)} = -2\frac{\theta_i}{\theta_i-\theta_j}\frac{\theta_j}{\theta_i}\ln\frac{\theta_j}{\theta_i} = 2\frac{\theta_j}{\theta_i-\theta_j}\ln\frac{\theta_i}{\theta_j}\;,$$

and

$$\begin{aligned}\lim_{\alpha\uparrow 1} D_\alpha(f_i||f_j) &= \ln\frac{\theta_i}{\theta_j} + 2\frac{\theta_j}{\theta_i-\theta_j}\ln\frac{\theta_i}{\theta_j} - 2\\ &= \ln\frac{\theta_i}{\theta_j}\left(\frac{\theta_i-\theta_j+2\theta_j}{\theta_i-\theta_j}\right) - 2\\ &= \frac{\theta_i+\theta_j}{\theta_i-\theta_j}\ln\frac{\theta_i}{\theta_j} - 2\;,\end{aligned}$$

which is the expression for $D(f_i||f_j)$ obtained in Proposition 2.3.10, as expected. □

Remark 2.3.14. The Rényi divergence is continuous at $\alpha = 1/2$.

Proof. For $\alpha = 1/2$ the term

$$\frac{\theta_i}{(2\alpha-1)(\theta_i-\theta_j)}\left[1-\left(\frac{\theta_j}{\theta_i}\right)^{2\alpha-1}\right]$$

is of indeterminate form. We proceed to evaluate the limit with l'Hospital's rule:

$$\begin{aligned}&\lim_{\alpha\to 1/2}\left(\frac{\theta_i}{(2\alpha-1)(\theta_i-\theta_j)}\left[1-\left(\frac{\theta_j}{\theta_i}\right)^{2\alpha-1}\right]\right)\\ &= \frac{\theta_i}{\theta_i-\theta_j}\lim_{\alpha\to 1/2}\left[\frac{g'(\alpha)}{2}\right]\\ &= \frac{\theta_i}{\theta_i-\theta_j}\left[-\frac{2}{2}\ln\frac{\theta_j}{\theta_i}\right]\\ &= \frac{\theta_i}{\theta_i-\theta_j}\ln\frac{\theta_i}{\theta_j}\;,\end{aligned}$$

with $g(\alpha)$ defined as in Remark 2.3.13. Thus by the continuity of the logarithm function

$$\lim_{\alpha\to 1/2} D_\alpha(f_i||f_j) = \ln\frac{\theta_i}{\theta_j} + 2\ln\left(\frac{\theta_i}{\theta_i-\theta_j}\ln\frac{\theta_i}{\theta_j}\right) ,$$

which is indeed the value of $D_{1/2}(f_i||f_j)$ given in Proposition 2.3.12. □

2.3.3 Rényi and Kullback Divergence for General Univariate Pareto Distributions

In this section we take f_i and f_j to be two Pareto densities with generally different supports:

$$f_i(x) = a_i m_i^{a_i} x^{-(a_i+1)} , \quad a_i, m_i > 0 \,; x > m_i \,.$$

Proposition 2.3.15. *The Kullback-Leibler Divergence berween f_i and f_j is*

$$\ln\left(\frac{m_i}{m_j}\right)^{a_j} + \ln\frac{a_i}{a_j} + \frac{a_j-a_i}{a_i}$$

if $m_i \geq m_j$, and ∞ otherwise.

Proof. If $m_i < m_j$ then $D(f_i||f_j) = \infty$ by the definition of the KLD. Suppose from now on that $m_i \geq m_j$. We have

$$E_{f_i}\left[\ln f_j\right] = E_{f_i}\left[\ln\left(a_j m_j^{a_j}\right) - (a_j+1)\ln X\right] = \ln\left(a_j m_j^{a_j}\right) - (a_j+1)E_{f_i}\left[\ln X\right] \,.$$

Now

$$
\begin{aligned}
E_{f_i}[\ln X] &= \int_{m_i}^{\infty} a_i m_i^{a_i} x^{-(a_i+1)} \ln x \, dx \\
&= a_i m_i^{a_i} \left[-\frac{1}{a_i} x^{-a_i} \ln x \Big|_{m_i}^{\infty} + \frac{1}{a_i} \int_{m_i}^{\infty} x^{-(a_i+1)} dx \right] \\
&= \frac{a_i m_i^{a_i} m_i^{-a_i} \ln m_i}{a_i} + \frac{1}{a_i} \int_{m_i}^{\infty} a_i m_i^{a_i} x^{-(a_i+1)} dx \\
&= \ln m_i + \frac{1}{a_i} \ .
\end{aligned}
$$

Thus

$$
\begin{aligned}
E_{f_i}\left[\ln f_j\right] &= \ln\left(a_j m_j^{a_j}\right) - (a_j+1)\left[\ln m_i + \frac{1}{a_i}\right] \\
&= \ln\left(\frac{m_j}{m_i}\right)^{a_j} + \ln\frac{a_j}{m_i} - \frac{a_j+1}{a_i} \ ,
\end{aligned}
$$

and

$$
\begin{aligned}
D(f_i||f_j) &= E_{f_i}[\ln f_i] - E_{f_i}[\ln f_j] \\
&= -\left[\ln\frac{m_i}{a_i} + \frac{(a_i+1)}{a_i}\right] - \left[\ln\left(\frac{m_j}{m_i}\right)^{a_j} + \ln\frac{a_j}{m_i} - \frac{a_j+1}{a_i}\right] \\
&= \ln\left(\frac{m_i}{m_j}\right)^{a_j} + \ln\frac{a_i}{a_j} + \frac{a_j - a_i}{a_i} \ .
\end{aligned}
$$

□

Next consider we the Rényi divergence.

Proposition 2.3.16. *Let $a_\alpha = \alpha a_i + (1-\alpha)a_j$ and $M = \max\{m_i, m_j\}$. The Rényi divergence between f_i and f_j is*

$$D_\alpha(f_i||f_j) = \begin{cases} \ln \dfrac{m_i^{a_i}}{m_j^{a_j}} + \ln \dfrac{a_i}{a_j} + \dfrac{1}{\alpha-1} \ln \dfrac{a_i m_i^{a_i}}{a_\alpha M^{a_\alpha}} & \text{for } \alpha \in (0,1) \\ \ln \left(\dfrac{m_i}{m_j}\right)^{a_j} + \ln \dfrac{a_i}{a_j} + \dfrac{1}{\alpha-1} \ln \dfrac{a_i}{a_\alpha} & \alpha > 1, m_i \geq m_j, \text{ and } a_\alpha > 0 \\ \infty & \text{otherwise} \end{cases}$$

Proof. Note that if $\alpha > 1$ then the integral

$$\int f_i^\alpha f_j^{1-\alpha} dx$$

is ∞ (and also $D_\alpha(f_i||f_j)$) for $m_i < m_j$. Also, for $\alpha < 1$ the integrand is nonzero only if $x > \max\left\{m_i, m_j\right\} := M$. Let's suppose that $\alpha > 1$ and $m_i \geq m_j$. Now

$$f_i^\alpha f_j^{1-\alpha} = \left[a_i m_i^{a_i} x^{-(a_i+1)}\right]^\alpha \left[a_j m_j^{a_j} x^{-(a_j+1)}\right]^{1-\alpha} = \left(\frac{a_i}{a_j}\right)^{\alpha-1} a_i m_i^{\alpha a_i} m_j^{(1-\alpha)a_j} x^{a_\alpha - 1} \ , \ x > m_i$$

where $a_\alpha = \alpha a_i + (1-\alpha)a_j$. If $a_\alpha \leq 0$ then

$$\int_{m_i}^\infty f_i^\alpha f_j^{1-\alpha} \, dx = A \int_{m_i}^\infty x^{a_0 - 1} dx = \infty \ , \quad (A > 0)$$

and since nonpositive a_α only occurs for $\alpha > 1$ we have $D_\alpha(f_i||f_j) = \infty$ as well. Now, if

$a_\alpha > 0$ then

$$\begin{aligned}
\int f_i^\alpha f_j^{1-\alpha} dx &= \left(\frac{a_i}{a_j}\right)^{\alpha-1} a_i m_i^{\alpha a_i} m_j^{(1-\alpha)a_j} \int_{m_i}^{\infty} x^{a_\alpha - 1} dx \\
&= \left(\frac{a_i}{a_j}\right)^{\alpha-1} m_i^{\alpha a_i} m_j^{(1-\alpha)a_j} \frac{a_i}{a_\alpha m_i^{a_\alpha}} \int_{m_i}^{\infty} a_\alpha m_i^{a_\alpha} x^{a_\alpha - 1} dx \\
&= \left(\frac{a_i}{a_j}\right)^{\alpha-1} m_i^{\alpha a_i - a_\alpha} m_j^{(1-\alpha)a_j} \frac{a_i}{a_\alpha} \\
&= \left(\frac{a_i}{a_j}\right)^{\alpha-1} m_i^{(\alpha-1)a_j} m_j^{(1-\alpha)a_j} \frac{a_i}{a_\alpha} \\
&= \left(\frac{a_i}{a_j}\right)^{\alpha-1} \left(\frac{m_i}{m_j}\right)^{(\alpha-1)a_j} \frac{a_i}{a_\alpha} .
\end{aligned}$$

Then

$$D_\alpha(f_i || f_j) = \ln \left(\frac{m_i}{m_j}\right)^{a_j} + \ln \frac{a_i}{a_j} + \frac{1}{\alpha - 1} \ln \frac{a_i}{a_\alpha} .$$

Note that setting $m_i = m_j$ we get the earlier result for equal supports given in Section 2.2.7. Lastly, consider the case $\alpha \in (0,1)$. In this case a_α is automatically positive, and we have

$$\begin{aligned}
\int f_i^\alpha f_j^{1-\alpha} dx &= \left(\frac{a_i}{a_j}\right)^{\alpha-1} a_i m_i^{\alpha a_i} m_j^{(1-\alpha)a_j} \int_{M}^{\infty} x^{a_\alpha - 1} dx \\
&= \left(\frac{a_i}{a_j}\right)^{\alpha-1} m_i^{\alpha a_i} m_j^{(1-\alpha)a_j} \frac{a_i}{a_\alpha M^{a_\alpha}} \\
&= \left(\frac{a_i}{a_j} \frac{m_i^{a_i}}{m_j^{a_j}}\right)^{\alpha-1} \frac{a_i m_i^{a_i}}{a_\alpha M^{a_\alpha}} ,
\end{aligned}$$

where $M = \max\{m_i, m_j\}$. Hence

$$D_\alpha(f_i || f_j) = \ln \frac{m_i^{a_i}}{m_j^{a_j}} + \ln \frac{a_i}{a_j} + \frac{1}{\alpha - 1} \ln \frac{a_i m_i^{a_i}}{a_\alpha M^{a_\alpha}} ,$$

which agrees with the above result for $m_i \geq m_j$. In summary we have

$$D_\alpha(f_i||f_j) = \begin{cases} \ln \dfrac{m_i^{a_i}}{m_j^{a_j}} + \ln \dfrac{a_i}{a_j} + \dfrac{1}{\alpha - 1} \ln \dfrac{a_i m_i^{a_i}}{a_\alpha M^{a_\alpha}}, & \alpha \in (0,1) \\ M = \max\{m_i, m_j\} & \\ & \\ \ln \left(\dfrac{m_i}{m_j}\right)^{a_j} + \ln \dfrac{a_i}{a_j} + \dfrac{1}{\alpha - 1} \ln \dfrac{a_i}{a_\alpha} & \alpha > 1, m_i \geq m_j, \text{ and } a_\alpha > 0 \\ & \\ \infty & \text{otherwise} \end{cases}$$

Remark 2.3.17. To verify that we do obtain $D(f_i||f_j)$ as $\alpha \uparrow 1$ observe that for $M = m_i$ the expression for $D_\alpha(f_i||f_j)$ is the same for all $\alpha > 0$ ($\alpha \neq 1$). Then, by l'Hospitals rule,

$$\lim_{\alpha \uparrow 1} \frac{1}{\alpha - 1} \ln \frac{a_i}{a_\alpha} = \frac{a_j - a_i}{a_i} \; .$$

Moreover, note that for $m_i < m_j$,

$$\lim_{\alpha \uparrow 1} \frac{a_i m_i^{a_i}}{a_\alpha m_j^{a_\alpha}} = \left(\frac{m_i}{m_j}\right)^{a_i} < 1 \; ,$$

hence

$$\lim_{\alpha \uparrow 1} \frac{1}{\alpha - 1} \ln \frac{a_i m_i^{a_i}}{a_\alpha m_j^{a_\alpha}} = \infty \; .$$

□

2.3.4 Rényi and Kullback Divergence for Uniform Distributions

We consider two uniform densities f_i and f_j

$$f_i = \frac{1}{b_i - a_i} \; , \; a_i < x < b_i \; .$$

Proposition 2.3.18. *The Kullback-Leibler divergence between f_i and f_j is*

$$D(f_i||f_j) = \ln \frac{b_j - a_j}{b_i - a_i} \ .$$

for $(a_i, b_i) \subseteq (a_j, b_j)$, and ∞ otherwise.

Proof. Assume that $(a_i, b_i) \subseteq (a_j, b_j)$ since, by definition, the KLD is ∞ otherwise. Then

$$D(f_i||f_j) = E_{f_i}\left[\ln(f_i/f_j)\right] = \ln \frac{b_j - a_j}{b_i - a_i} \ .$$

□

Proposition 2.3.19. *Let $b_m = \min\{b_i, b_j\}$ and $a_M = \max\{a_i, a_j\}$. Then the Rényi divergnce between f_i and f_j is*

$$D_\alpha(f_i||f_j) = \begin{cases} \ln \dfrac{b_j - a_j}{b_i - a_i} + \dfrac{1}{\alpha - 1} \ln \dfrac{b_m - a_M}{b_i - a_i} \ , & \alpha \in (0,1) \ , \ b_m > a_M \\ \\ \ln \dfrac{b_j - a_j}{b_i - a_i} & \alpha > 1 \ , \ (a_i, b_i) \subset (a_j, b_j). \\ \\ \infty & \textit{otherwise} \end{cases}$$

Proof. We calculate the Rényi divergence using the same line of argument as in the Pareto case. For $\alpha \in (0,1)$ we need to look at two cases. If $b_m = \min\{b_i, b_j\} \leq a_M = \max\{a_i, a_j\}$, then

$$\int f_i^\alpha f_j^{1-\alpha} dx = 0 \ ,$$

hence

$$D_\alpha(f_i||f_j) = \frac{1}{\alpha - 1} \ln \int f_i^\alpha f_j^{1-\alpha} dx = \infty \ .$$

Suppose then that $b_m > a_M$. In this case,

$$\int f_i^{\alpha} f_j^{1-\alpha} dx = \int_{a_M}^{b_m} \left(\frac{1}{b_i - a_i}\right)^{\alpha} \left(\frac{1}{b_j - a_j}\right)^{1-\alpha} dx$$
$$= \left(\frac{b_j - a_j}{b_i - a_i}\right)^{\alpha-1} \frac{b_m - a_M}{b_i - a_i} ,$$

For $\alpha > 1$ the integral above is finite only when $(a_i, b_i) \subset (a_j, b_j)$[7]. In this case

$$\int f_i^{\alpha} f_j^{1-\alpha} dx = \int_{a_i}^{b_i} \left(\frac{1}{b_i - a_i}\right)^{\alpha} \left(\frac{1}{b_j - a_j}\right)^{1-\alpha} dx$$
$$= \left(\frac{b_j - a_j}{b_i - a_i}\right)^{\alpha-1} .$$

Thus,

$$D_\alpha(f_i||f_j) = \begin{cases} \ln \dfrac{b_j - a_j}{b_i - a_i} + \dfrac{1}{\alpha - 1} \ln \dfrac{b_m - a_M}{b_i - a_i} , & \alpha \in (0,1) \, , \, b_m > a_M \\ b_m = \min\{b_i, b_j\} \, , \; a_M = \max\{a_i, a_j\} \, , & \\ & \\ \ln \dfrac{b_j - a_j}{b_i - a_i} & \alpha > 1 \, , \, (a_i, b_i) \subset (a_j, b_j). \\ & \\ \infty & \text{otherwise} \end{cases}$$

□

Remark 2.3.20. Note that $D_\alpha(f_i||f_j) = D(f_i||f_j)$ for $\alpha > 1$. To verify that $D_\alpha(f_i||f_j)$ approaches $D(f_i||f_j)$ as $\alpha \uparrow 1$, note first that $(a_i, b_i) \subset (a_j, b_j) \Longleftrightarrow a_M = a_i$ and $b_m = b_i$, in which case the expression for $\alpha \in (0,1)$ is the same as $D(f_i||f_j)$. Suppose then that $(a_i, b_i) \neq (a_M, b_m)$. If $b_m > a_M$ then

$$0 < \frac{b_m - a_M}{b_i - a_i} < \frac{b_i - a_i}{b_i - a_i} = 1 ,$$

[7] Inclusion here is strict since it is tacitly assumed that $(a_i, b_i) \neq (a_j, b_j)$

hence

$$\lim_{\alpha\uparrow 1} D_\alpha(f_i||f_j) = \infty ;$$

and if $b_m \le a_M$ then $D_\alpha(f_i||f_j) = \infty$ for all $\alpha \in (0,1)$. Thus the limit is verified.

2.3.5 Kullback-Liebler Divergence for General Univariate Gumbel Distributions

It is possible to derive the Kullback-Leibler divergence (Rényi divergence for $\alpha = 1$) without the assumption of a fixed β value made in Section 2.2.5. Throughout this section let f_i and f_j be two Gumbel densities

$$\begin{aligned} f_i(x) &= \beta_i^{-1} e^{-(x-\mu_i)/\beta_i} \exp\left(-e^{-(x-\mu_i)/\beta_i}\right) \\ &= \beta_i^{-1} w_i e^{-w_i} \;, \quad w_i = e^{-(x-\mu_i)/\beta_i} \quad \mu_i \in \mathbb{R},\ \beta_i > 0 \;; x \in \mathbb{R} \;. \end{aligned}$$

Proposition 2.3.21.

$$E_{f_i}\left[\ln f_j\right] = -\ln \beta_j + \frac{\mu_j - \mu_i}{\beta_j} - \frac{\beta_i}{\beta_j}\gamma - e^{(\mu_j-\mu_i)/\beta_j}\Gamma\left(\frac{\beta_i}{\beta_j} + 1\right) .$$

Proof.

$$\begin{aligned} E_{f_i}\left[\ln f_j\right] &= E_{f_i}\left[-\ln \beta_j + \ln W_j - W_j\right] \\ &= -\ln \beta_j + E_{f_i}\left[\ln W_j\right] - E_{f_i}\left[W_j\right] . \end{aligned}$$

Let $r > -\beta_j/\beta_i$. Then

$$E_{f_i}\left[W_j^r\right] = \int_{\mathbb{R}} \beta_i^{-1} w_i(x) e^{-w_i(x)} w_j(x)^r \, dx .$$

Note that

$$w_i(x) = e^{-(x-\mu)/\beta_i} \Rightarrow x = -\beta_i \ln w_i + \mu_i \Rightarrow dx = -\frac{\beta_i}{w_i}\, dw_i ,$$

and

$$\begin{aligned}
w_j &= e^{-(x-\mu_j)/\beta_j} \\
&= \exp\left(-\frac{\left[-\beta_i \ln w_i + \mu_i - \mu_j\right]}{\beta_j}\right) \\
&= \exp\left(\frac{\beta_i}{\beta_j}\ln w_i + \frac{\mu_j - \mu_i}{\beta_j}\right) \\
&= w_i^{(\beta_i/\beta_j)}\ e^{(\mu_j-\mu_i)/\beta_j}\ .
\end{aligned}$$

Also, $x \to \infty \Rightarrow w_i \to 0$ and $x \to -\infty \Rightarrow w_i \to \infty$ since $\beta_i > 0$. Thus

$$\begin{aligned}
E_{f_i}\left[W_j^r\right] &= \int_\infty^0 \beta_i^{-1} w_i e^{-w_i}\left(w_i^{(\beta_i/\beta_j)} e^{(\mu_j-\mu_i)/\beta_j}\right)^r \left(-\frac{\beta_i}{w_i}\right)\ dw_i \\
&= e^{r(\mu_j-\mu_i)/\beta_j}\int_{\mathbb{R}^+} e^{-w_i} w_i^{(r\beta_i/\beta_j)}\ dw_i \\
&= e^{r(\mu_j-\mu_i)/\beta_j}\Gamma\left(\frac{r\beta_i}{\beta_j}+1\right)\ .
\end{aligned}$$

Then,

$$E_{f_i}\left[W_j\right] = e^{(\mu_j-\mu_i)/\beta_j}\Gamma\left(\frac{\beta_i}{\beta_j}+1\right)\ .$$

Also,

$$\begin{aligned}
E_{f_i}\left[\ln W_j\right] &= \frac{d}{dr}E_{f_i}\left[W_j^r\right]\Big|_{r=0} \\
&= \frac{d}{dr}\left[e^{r(\mu_j-\mu_i)/\beta_j}\Gamma\left(\frac{r\beta_i}{\beta_j}+1\right)\right]_{r=0} \\
&= \left[e^{r(\mu_j-\mu_i)/\beta_j}\left(\frac{\mu_j-\mu_i}{\beta_j}\Gamma\left(\frac{r\beta_i}{\beta_j}+1\right)+\Gamma'\left(\frac{r\beta_i}{\beta_j}+1\right)\frac{\beta_i}{\beta_j}\right)\right]_{r=0} \\
&= \frac{\mu_j-\mu_i}{\beta_j} + \frac{\beta_i}{\beta_j}\Gamma'(1) \\
&= \frac{\mu_j-\mu_i}{\beta_j} - \frac{\beta_i}{\beta_j}\gamma\ ,
\end{aligned}$$

where γ is the Euler-Mascheroni constant introduced in Section A.3.1. Finally,

$$\begin{aligned} E_{f_i}\left[\ln f_j\right] &= -\ln \beta_j + E_{f_i}\left[\ln W_j\right] - E_{f_i}\left[W_j\right] \\ &= -\ln \beta_j + \frac{\mu_j - \mu_i}{\beta_j} - \frac{\beta_i}{\beta_j}\gamma - e^{(\mu_j-\mu_i)/\beta_j}\Gamma\left(\frac{\beta_i}{\beta_j}+1\right) . \end{aligned}$$

□

Corollary 2.3.22. *The differential entropy of f_i is*

$$h(f_i) = \ln \beta_i + \gamma + 1.$$

Proof. Setting $i = j$ in Proposition 2.3.21 we have

$$\begin{aligned} h(f_i) = -E_{f_i}\left[\ln f_i\right] &= -\left[-\ln \beta_i + \frac{\mu_i - \mu_i}{\beta_i} - \frac{\beta_i}{\beta_i}\gamma - e^{(\mu_i-\mu_i)/\beta_i}\Gamma\left(\frac{\beta_i}{\beta_i}+1\right)\right] \\ &= \ln \beta_i + \gamma + \Gamma(2) \\ &= \ln \beta_i + \gamma + 1 . \end{aligned}$$

□

Proposition 2.3.23. *The Kullback-Liebler divergence between f_i and f_j is*

$$D(f_i||f_j) = \ln \frac{\beta_j}{\beta_i} + \gamma\left(\frac{\beta_i}{\beta_j} - 1\right) + e^{(\mu_j-\mu_i)/\beta_j}\Gamma\left(\frac{\beta_i}{\beta_j}+1\right) - 1 .$$

Proof. Using Proposition 2.3.21 and Remark 1.2.4 we have

$$\begin{aligned} D(f_i||f_j) &= E_{f_i}[\ln f_i] - E_{f_i}[\ln f_j] \\ &= -\left[\ln \beta_i + \gamma + 1\right] \\ &\quad - \left[-\ln \beta_j + \frac{\mu_j - \mu_i}{\beta_j} - \frac{\beta_i}{\beta_j}\gamma - e^{(\mu_j-\mu_i)/\beta_j}\Gamma\left(\frac{\beta_i}{\beta_j}+1\right)\right] \\ &= \ln \frac{\beta_j}{\beta_i} + \gamma\left(\frac{\beta_i}{\beta_j} - 1\right) + e^{(\mu_j-\mu_i)/\beta_j}\Gamma\left(\frac{\beta_i}{\beta_j}+1\right) - 1 . \end{aligned}$$

□

Remark 2.3.24. If we consider the expression for $D_\alpha(f_i||f_j)$ in the case $\beta_i = \beta_j = \beta$, which we derived in Section 2.2.5, then we find

$$\lim_{\alpha\to 1} D_\alpha(f_i||f_j) = -\lim_{\alpha\to 1} \frac{e^{\mu_i/\beta} - e^{\mu_j/\beta}}{\alpha e^{\mu_i/\beta} + (1-\alpha)e^{\mu_j/\beta}} = \frac{e^{\mu_j/\beta} - e^{\mu_i/\beta}}{e^{\mu_i/\beta}} = e^{(\mu_j-\mu_i)/\beta} - 1 ,$$

where we have used l'Hospital's rule to evaluate the indeterminate limit. As expected, this is also the expression obtained by setting $\beta_i = \beta_j = \beta$ in the Proposition 2.3.23.

2.3.6 Kullback-Liebler Divergence for General Univariate Weibull Distributions

It is possible to derive the Kullback-Leibler divergence (Rényi divergence for $\alpha = 1$) without the assumption of a fixed k value made in Section 2.2.8. Throughout this section let f_i and f_j be two univariate Weibull densities

$$f_i(x) = k_i \lambda_i^{-k_i} x^{k_i-1} e^{-(x/\lambda_i)^{k_i}} \ , \quad k_i, \lambda_i > 0; x \in \mathbb{R}^+ \ .$$

Proposition 2.3.25.

$$E_{f_i}\left[\ln f_j\right] = (k_j - 1)\ln\frac{\lambda_i}{\lambda_j} + \ln\frac{k_j}{\lambda_j} - \left(k_j - 1\right)\frac{\gamma}{k_i} - \left(\frac{\lambda_i}{\lambda_j}\right)^{k_j}\Gamma\left(1 + \frac{k_j}{k_i}\right) \ .$$

Proof.

$$\begin{aligned} E_{f_i}\left[\ln f_j\right] &= E_{f_i}\left[\ln\left(k_j\lambda_j^{-k_j}\right) + \left(k_j - 1\right)\ln X - \left(\frac{X}{\lambda_j}\right)^{k_j}\right] \\ &= \ln\left(k_j\lambda_j^{-k_j}\right) + \left(k_j - 1\right)E_{f_i}\left[\ln X\right] - \lambda_j^{-k_j}E_{f_i}\left[X^{k_j}\right] \ . \end{aligned}$$

Let $r \geq -k_i$. Then

$$E_{f_i}\left[X^r\right] = \int_{\mathbb{R}^+} x^r\, k_i \lambda_i^{-k_i} x^{k_i-1} e^{-(x/\lambda_i)^{k_i}}\, dx$$
$$= \lambda_i^r \int_{\mathbb{R}^+} y^{r/k_i} e^{-y}\, dy\,, \quad y = \left(\frac{x}{\lambda_i}\right)^{k_i} = \lambda_i^r\, \Gamma\left(1 + \frac{r}{k_i}\right)\,.$$

Also

$$\frac{d}{dr} E_{f_i}\left[X^r\right] = E\left[\frac{d}{dr} X^r\right] = E\left[X^r \ln X\right]\,.$$

Thus

$$E\left[X \ln X\right] = \frac{d}{dr} E_{f_i}\left[X^r\right]\bigg|_{r=0}$$
$$= \left[\lambda_i^r \ln \lambda_i\, \Gamma\left(1 + \frac{r}{k_i}\right) + \lambda_i^r \frac{d}{dr} \Gamma\left(1 + \frac{r}{k_i}\right)\right]\bigg|_{r=0}$$
$$= \Gamma(1)\, \ln \lambda_i + \frac{\Gamma'(1)}{k_i}$$
$$= \ln \lambda_i - \frac{\gamma}{k_i}\,,$$

where γ is the Euler-Mascheroni constant introduced in Definition A.3.3 and by Proposition A.3.4, $\Gamma'(1) = -\gamma$. Thus,

$$E_{f_i}\left[\ln f_j\right] = \ln\left(k_j \lambda_j^{-k_j}\right) + \left(k_j - 1\right) E_{f_i}\left[\ln X\right] - \lambda_j^{-k_j} E_{f_i}\left[X^{k_j}\right]$$
$$= \ln\left(k_j \lambda_j^{-k_j}\right) + \left(k_j - 1\right)\left(\ln \lambda_i - \frac{\gamma}{k_i}\right) - \lambda_j^{-k_j} \lambda_i^{k_j}\, \Gamma\left(1 + \frac{k_j}{k_i}\right)$$
$$= \ln k_j - k_j \ln \lambda_j + \left(k_j - 1\right) \ln \lambda_i - \left(k_j - 1\right) \frac{\gamma}{k_i} - \left(\frac{\lambda_i}{\lambda_j}\right)^{k_j} \Gamma\left(1 + \frac{k_j}{k_i}\right)$$
$$= (k_j - 1) \ln \frac{\lambda_i}{\lambda_j} + \ln \frac{k_j}{\lambda_j} - \left(k_j - 1\right) \frac{\gamma}{k_i} - \left(\frac{\lambda_i}{\lambda_j}\right)^{k_j} \Gamma\left(1 + \frac{k_j}{k_i}\right)\,.$$

□

Corollary 2.3.26. *The differential entropy of f_i is*

$$h(f_i) = \ln \frac{\lambda_i}{k_i} + \left(1 - \frac{1}{k_i}\right)\gamma + 1 \,.$$

Proof. Setting $i = j$ in Proposition 2.3.25 we have

$$\begin{aligned}
h(f_i) &= -E_{f_i}\left[\ln f_i\right] \\
&= -\left[(k_i - 1)\ln \frac{\lambda_i}{\lambda_i} + \ln \frac{k_i}{\lambda_i} - (k_i - 1)\frac{\gamma}{k_i} - \left(\frac{\lambda_i}{\lambda_i}\right)^{k_i}\Gamma\left(1 + \frac{k_i}{k_i}\right)\right] \\
&= -\left[\ln \frac{k_i}{\lambda_i} - (k_i - 1)\frac{\gamma}{k_i} - \Gamma(2)\right] \\
&= \ln \frac{\lambda_i}{k_i} + \left(1 - \frac{1}{k_i}\right)\gamma + 1 \,.
\end{aligned}$$

□

Proposition 2.3.27. *The Kullback-Liebler divergence between f_i and f_j is*

$$D(f_i||f_j) = \ln\left(\frac{k_i}{k_j}\left[\frac{\lambda_j}{\lambda_i}\right]^{k_j}\right) + \gamma\frac{k_j - k_i}{k_i} + \left(\frac{\lambda_i}{\lambda_j}\right)^{k_j}\Gamma\left(1 + \frac{k_j}{k_i}\right) - 1 \,.$$

Proof. Using Proposition 2.3.25 and Remark 1.2.4 we have

$$\begin{aligned}
D(f_i||f_j) &= E_{f_i}[\ln f_i] - E_{f_i}[\ln f_j] \\
&= \ln \frac{k_i}{\lambda_i} - (k_i - 1)\frac{\gamma}{k_i} - 1 \\
&\quad - \left[(k_j - 1)\ln \frac{\lambda_i}{\lambda_j} + \ln \frac{k_j}{\lambda_j} - (k_j - 1)\frac{\gamma}{k_i} - \left(\frac{\lambda_i}{\lambda_j}\right)^{k_j}\Gamma\left(1 + \frac{k_j}{k_i}\right)\right] \\
&= \ln\left(\frac{k_i\lambda_j}{k_j\lambda_i}\right) + \frac{\gamma}{k_i}(k_j - 1 - k_i + 1) + (k_j - 1)\ln \frac{\lambda_j}{\lambda_i} + \left(\frac{\lambda_i}{\lambda_j}\right)^{k_j}\Gamma\left(1 + \frac{k_j}{k_i}\right) - 1 \\
&= \ln\left(\frac{k_i}{k_j}\left[\frac{\lambda_j}{\lambda_i}\right]^{k_j}\right) + \gamma\frac{k_j - k_i}{k_i} + \left(\frac{\lambda_i}{\lambda_j}\right)^{k_j}\Gamma\left(1 + \frac{k_j}{k_i}\right) - 1 \,.
\end{aligned}$$

□

2.4 Tables for Continuous Rényi and Kullback Divergences

We summarize the results of this chapter in Table 2.2 and Table 2.3, where we present the expressions for Rényi and Kullback divergences, respectively. The densities associated with the distributions are given in Table 2.1. The table of Rényi divergences includes a finiteness constraint for which the given expression is valid. For all other cases (and $\alpha > 0$), $D_\alpha(f_i||f_j) = \infty$. In the cases where the closed-form expression is a piece-wise function the conditions for each case are presented alongside the corresponding formula, and it is implied that for all other cases $D_\alpha(f_i||f_j) = \infty$. The expressions for the Rényi divergence of Laplace and Cramer distributions are still continuous at $\alpha = \lambda_i/(\lambda_i + \lambda_j)$ and $\alpha = 1/2$, respectively (as shown in the corresponding sections of this work).

One important property of Rényi divergence is that $D_\alpha(T(X)||T(Y)) = D_\alpha(X||Y)$ for any invertible transformation T. This follows from the more general data process inequality (see [60]). For example, the Rényi divergence between two lognormal densities is the same as that between two normal densities, hence the absence of the former in the tables.

Table 2.1: Continuous Distributions

Name	**Density**	**Restrictions**
Beta	$\dfrac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$	$a, b > 0;\ x \in (0,1)$
Chi	$\dfrac{2^{1-k/2}x^{k-1}e^{-x^2/2\sigma^2}}{\sigma^k\Gamma\left(\frac{k}{2}\right)}$	$\sigma > 0,\ k \in \mathbb{N};\ x > 0$

Name	Density	Restrictions
$\boldsymbol{\chi^2}$	$\frac{x^{d/2-1}e^{-x/2}}{2^{d/2}\Gamma(d/2)}$	$d \in \mathbb{N};\ x > 0$
Cramér	$\frac{\theta}{2(1+\theta\|x\|)^2}$	$\theta > 0;\ x \in \mathbb{R}$
Dirichlet	$\frac{1}{B(\boldsymbol{a})}\prod_{k=1}^{d} x_k^{a_k-1}$	$\boldsymbol{a} \in \mathbb{R}^d,\ a_k > 0, d \geq 2;$ $\boldsymbol{x} \in \mathbb{R}^d, \sum x_k = 1$
Exponential	$\lambda e^{-\lambda x}$	$\lambda > 0;\ x > 0$
Gamma	$\frac{x^{k-1}e^{-x/\theta}}{\theta^k\Gamma(k)}$	$\sigma > 0,\ k > 0;\ x > 0$
Multivariate Gaussian	$\frac{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}}{(2\pi)^{n/2}\|\Sigma\|^{1/2}}$	$\boldsymbol{\mu} \in \mathbb{R}^n; \boldsymbol{x} \in \mathbb{R}^n$ Σ symmetric positive definite
Univariate Gaussian	$\frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$	$\sigma > 0, \mu \in \mathbb{R}; x \in \mathbb{R}$
Special Bivariate Gaussian	$\frac{e^{-\frac{1}{2}\boldsymbol{x}'\Phi^{-1}\boldsymbol{x}}}{2\pi(1-\rho^2)^{1/2}}$	$\rho \in (-1,1), \Phi = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix};$ $\boldsymbol{x} \in \mathbb{R}^2$
Gumbel	$\frac{e^{-(x-\mu)/\beta}e^{-e^{-(x-\mu)/\beta}}}{\beta}$	$\mu \in \mathbb{R},\ \beta > 0; x \in \mathbb{R}$
Half-Normal	$\sqrt{\frac{2}{\pi\sigma^2}}\, e^{-x^2/(2\sigma^2)}$	$\sigma > 0;\ x > 0$
Laplace	$\frac{1}{2\lambda}e^{-\|x-\theta\|/\lambda}$	$\lambda > 0, \theta \in \mathbb{R};\ x \in \mathbb{R}$
Maxwell-Boltzmann	$\sqrt{\frac{2}{\pi}}\frac{x^2 e^{-\frac{x^2}{2\sigma^2}}}{\sigma^3}$	$\sigma > 0;\ x > 0$

Name	Density	Restrictions
Pareto	$am^a x^{-(a+1)}$	$a\,, m > 0\,; x > m$
Rayleigh	$\frac{x}{\sigma^2}\, e^{-x^2/(2\sigma^2)}$	$\sigma > 0\,;\; x > 0$
Uniform	$\frac{1}{b-a}$	$a < x < b$
Weibull	$k\lambda^{-k} x^{k-1} e^{-(x/\lambda)^k}$	$k\,, \lambda > 0\,;\; x \in \mathbb{R}^+$

Table 2.2: Rényi Divergences for Continuous Distributions

Name	$\boldsymbol{D_\alpha(f_i \| f_j)}$	Finiteness Condition
Beta	$\ln \frac{B(a_j, b_j)}{B(a_i, b_i)} + \frac{1}{\alpha - 1} \ln \frac{B(a_\alpha, b_\alpha)}{B(a_i, b_i)}$ $a_\alpha = \alpha a_i + (1-\alpha) a_j,\, b_\alpha = \alpha b_i + (1-\alpha) b_j$	$a_\alpha, b_\alpha \geq 0$
Chi	$\ln \left(\frac{\sigma_j^{k_j} \Gamma(k_j/2)}{\sigma_i^{k_i} \Gamma(k_i/2)} \right)$ $+ \frac{1}{\alpha - 1} \ln \left(\frac{\Gamma(k_\alpha/2)}{\sigma_i^{k_i} \Gamma(k_i/2)} \left(\frac{\sigma_i^2 \sigma_j^2}{(\sigma^2)^*_\alpha} \right)^{k_\alpha/2} \right)$ $(\sigma^2)^*_\alpha = \alpha \sigma_j^2 + (1-\alpha)\sigma_i^2,\;\; k_\alpha = \alpha k_i + (1-\alpha) k_j$	$(\sigma^2)^*_\alpha > 0,\, k_\alpha > 0$
$\boldsymbol{\chi^2}$	$\ln \left(\frac{\Gamma(d_j/2)}{\Gamma(d_i/2)} \right) + \frac{1}{\alpha - 1} \ln \left(\frac{\Gamma(d_\alpha/2)}{\Gamma(d_i/2)} \right)$ $d_\alpha = \alpha d_i + (1-\alpha) d_j$	$d_\alpha > 0$

Name	$\boldsymbol{D_\alpha(f_i\|f_j)}$	**Finiteness Condition**
Cramér	*For* $\alpha = 1/2$ $\ln\frac{\theta_i}{\theta_j} + 2\ln\left(\frac{\theta_i}{\theta_i-\theta_j}\ln\frac{\theta_i}{\theta_j}\right)$ *For* $\alpha \neq 1/2$ $\ln\frac{\theta_i}{\theta_j} + \frac{1}{\alpha-1}\ln\left(\frac{\theta_i\left[1-\left(\theta_j/\theta_i\right)^{2\alpha-1}\right]}{(\theta_i-\theta_j)(2\alpha-1)}\right)$	
Dirichlet	$\ln\frac{B(\boldsymbol{a}_j)}{B(\boldsymbol{a}_i)} + \frac{1}{\alpha-1}\ln\left(\frac{B(\boldsymbol{a}_\alpha)}{B(\boldsymbol{a}_i)}\right)$ $\boldsymbol{a}_\alpha = \alpha\boldsymbol{a}_i + (1-a)\boldsymbol{a}_j$	$a_{\alpha_k} > 0\ \forall k$
Exponential	$\ln\frac{\lambda_i}{\lambda_j} + \frac{1}{\alpha-1}\ln\frac{\lambda_i}{\lambda_\alpha}$ $\lambda_\alpha = \alpha\lambda_i + (1-\alpha)\lambda_j$	$\lambda_\alpha > 0$
Gamma	$\ln\left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}}\right)$ $+\frac{1}{\alpha-1}\ln\left(\frac{\Gamma(k_\alpha)}{\theta_i^{k_i}\Gamma(k_i)}\left(\frac{\theta_i\theta_j}{\theta_\alpha^*}\right)^{k_\alpha}\right)$ $\theta_\alpha^* = \alpha\theta_j + (1-a)\theta_i,\ k_\alpha = \alpha k_i + (1-\alpha)k_j$	$\theta_\alpha^* > 0$ and $k_\alpha > 0$
Multivariate Gaussian	$\frac{\alpha}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)'(\Sigma_\alpha)^*(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$ $-\frac{1}{2(\alpha-1)}\ln\frac{\lvert(\Sigma_\alpha)^*\rvert}{\lvert\Sigma_i\rvert^{1-\alpha}\lvert\Sigma_j\rvert^\alpha}$ $(\Sigma_\alpha)^* = \alpha\Sigma_j + (1-\alpha)\Sigma_i$	$\alpha\Sigma_i^{-1} + (1-\alpha)\Sigma_j^{-1}$ positive definite

Name	$\boldsymbol{D_\alpha(f_i\|\|f_j)}$	**Finiteness Condition**
Univariate Gaussian	$\ln\frac{\sigma_j}{\sigma_i}+\frac{1}{2(\alpha-1)}\ln\left(\frac{\sigma_j^2}{(\sigma^2)^*_\alpha}\right)+\frac{1}{2}\frac{\alpha(\mu_i-\mu_j)^2}{(\sigma^2)^*_\alpha}$ $(\sigma^2)^*_\alpha=\alpha\sigma_j^2+(1-\alpha)\sigma_i^2$	$(\sigma^2)^*_\alpha>0$
Special Bivariate Gaussian	$\frac{1}{2}\ln\left(\frac{1-\rho_j^2}{1-\rho_i^2}\right)-\frac{1}{2(\alpha-1)}\ln\left(\frac{1-(\rho^*_\alpha)^2}{(1-\rho_j^2)}\right)$ $\rho^*_\alpha=\alpha\rho_j+(1-\alpha)\rho_i$	$\alpha\Phi_i^{-1}+(1-\alpha)\Phi_j^{-1}$ positive definite
Gumbel Fixed Scale $\boldsymbol{(\beta_i=\beta_j)}$	$\frac{\mu_i-\mu_j}{\beta}+\frac{1}{\alpha-1}\ln\frac{e^{\mu_i/\beta}}{\left(e^{\mu_i/\beta}\right)_\alpha}$ $\left(e^{\mu_i/\beta}\right)_\alpha=\alpha e^{\mu_i/\beta}+(1-\alpha)e^{\mu_j/\beta}$	$\left(e^{\mu_i/\beta}\right)_\alpha>0$
Half-Normal	$\ln\frac{\sigma_j}{\sigma_i}+\frac{1}{\alpha-1}\ln\left(\frac{\sigma_j^2}{(\sigma^2)^*_\alpha}\right)^{1/2}$ $(\sigma^2)^*_\alpha=\alpha\sigma_j^2+(1-\alpha)\sigma_i^2$	$(\sigma^2)^*_\alpha>0$
Laplace	*For* $\alpha=\lambda_i/(\lambda_i+\lambda_j)$ $\ln\frac{\lambda_j}{\lambda_i}+\frac{\|\theta_i-\theta_j\|}{\lambda_j}+\frac{\lambda_i+\lambda_j}{\lambda_j}\ln\left(\frac{2\lambda_i}{\lambda_i+\lambda_j+\|\theta_i-\theta_j\|}\right)$ *For* $\alpha\neq\lambda_i/(\lambda_i+\lambda_j)$ and $\alpha\lambda_j+(1-\alpha)\lambda_i>0$ $\ln\frac{\lambda_j}{\lambda_i}+\frac{1}{\alpha-1}\ln\left(\frac{\lambda_i\lambda_j^2\,g(\alpha)}{\alpha^2\lambda_j^2-(1-\alpha)^2\lambda_i^2}\right)$ where $g(\alpha)=\frac{\alpha}{\lambda_i}\exp\left(-\frac{(1-\alpha)\|\theta_i-\theta_j\|}{\lambda_j}\right)-\frac{1-\alpha}{\lambda_j}\exp\left(\frac{-\alpha\|\theta_i-\theta_j\|}{\lambda_i}\right)$	

Name	$D_\alpha(f_i \| f_j)$	**Finiteness Condition**
Maxwell Boltzmann	$3\ln\frac{\sigma_j}{\sigma_i} + \frac{1}{\alpha-1}\ln\left(\frac{\sigma_j^2}{(\sigma^2)^*_\alpha}\right)^{3/2}$ $(\sigma^2)^*_\alpha = \alpha\sigma_j^2 + (1-\alpha)\sigma_i^2$	$(\sigma^2)^*_\alpha > 0$
Pareto	*For* $\alpha \in (0,1)$ $\ln\frac{m_i^{a_i}}{m_j^{a_j}} + \ln\frac{a_i}{a_j} + \frac{1}{\alpha-1}\ln\frac{a_i m_i^{a_i}}{a_\alpha M^{a_\alpha}}$, $M = \max\{m_i, m_j\}$ *For* $\alpha > 1, m_i \geq m_j,$ and $a_\alpha = \alpha a_i + (1-\alpha)a_j > 0$ $\ln\left(\frac{m_i}{m_j}\right)^{a_j} + \ln\frac{a_i}{a_j} + \frac{1}{\alpha-1}\ln\frac{a_i}{a_\alpha}$	
Rayleigh	$2\ln\frac{\sigma_j}{\sigma_i} + \frac{1}{\alpha-1}\ln\left(\frac{\sigma_j^2}{(\sigma^2)^*_\alpha}\right)$ $(\sigma^2)^*_\alpha = \alpha\sigma_j^2 + (1-\alpha)\sigma_i^2$	$(\sigma^2)^*_\alpha > 0$
Uniform	*For* $\alpha \in (0,1)$ and $b_m = \min\{b_i, b_j\} > a_M = \max\{a_i, a_j\}$ $\ln\frac{b_j - a_j}{b_i - a_i} + \frac{1}{\alpha-1}\ln\frac{b_m - a_M}{b_i - a_i}$, *For* $\alpha > 1$, $(a_i, b_i) \subset (a_j, b_j)$ $\ln\frac{b_j - a_j}{b_i - a_i}$	
Weibull Fixed Shape $(k_i = k_j)$	$\ln\left(\frac{\lambda_j}{\lambda_i}\right)^k + \frac{1}{\alpha-1}\ln\frac{\lambda_j^k}{(\lambda^k)^*_\alpha}$ $(\lambda^k)^*_\alpha = \alpha\lambda_j^k + (1-\alpha)\lambda_i^k$	$(\lambda^k)^*_\alpha > 0$

Table 2.3: Kullback Divergences for Continuous Distributions

Name	$\boldsymbol{D(f_i \vert\vert f_j)}$
Beta	$\ln \frac{B(a_j, b_j)}{B(a_i, b_i)} + \psi(a_i)(a_i - a_j) + \psi(b_i)(b_i - b_j) + [a_j + b_j - (a_i + b_i)]\psi(a_i + b_i)$
Chi	$\frac{1}{2}\psi(k_i/2)\left(k_i - k_j\right) + \ln\left[\left(\frac{\sigma_j}{\sigma_i}\right)^{k_j} \frac{\Gamma(k_j/2)}{\Gamma(k_i/2)}\right] + \frac{k_i}{2\sigma_j^2}\left(\sigma_i^2 - \sigma_j^2\right)$
$\boldsymbol{\chi^2}$	$\ln \frac{\Gamma(d_j/2)}{\Gamma(d_i/2)} + \frac{d_i - d_j}{2}\psi(d_i/2)$
Cramér	$\frac{\theta_i + \theta_j}{\theta_i - \theta_j} \ln \frac{\theta_i}{\theta_j} - 2$
Dirichlet	$\log \frac{B(\boldsymbol{a_j})}{B(\boldsymbol{a_i})} + \sum_{k=1}^{d}\left[a_{i_k} - a_{j_k}\right]\left[\psi(a_{i_k}) - \psi\left(\sum_{k=1}^{d} a_{i_k}\right)\right]$
Exponential	$\ln \frac{\lambda_i}{\lambda_j} + \frac{\lambda_j - \lambda_i}{\lambda_i}$
Gamma	$\left(\frac{\theta_i - \theta_j}{\theta_j}\right) k_i + \ln\left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}}\right) + \left(k_i - k_j\right)\left(\ln\theta_i + \psi(k_i)\right)$
Multivariate Gaussian	$\frac{1}{2}\left(\ln \frac{\vert\Sigma_j\vert}{\vert\Sigma_i\vert} + \mathrm{tr}\left(\Sigma_j^{-1}\Sigma_i\right)\right) + \frac{1}{2}\left[\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)' \Sigma_j^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) - n\right]$
Univariate Gaussian	$\frac{1}{2\sigma_j^2}\left[(\mu_i - \mu_j)^2 + \sigma_i^2 - \sigma_j^2\right] + \ln \frac{\sigma_j}{\sigma_i}$
Special Bivariate Gaussian	$\frac{1}{2}\ln\left(\frac{1 - \rho_j^2}{1 - \rho_i^2}\right) + \frac{\rho_j^2 - \rho_j\rho_i}{1 - \rho_j^2}$
General Gumbel	$\ln \frac{\beta_j}{\beta_i} + \gamma\left(\frac{\beta_i}{\beta_j} - 1\right) + e^{(\mu_j - \mu_i)/\beta_j}\Gamma\left(\frac{\beta_i}{\beta_j} + 1\right) - 1$

Name	$D(f_i \| f_j)$
Half-Normal	$\ln\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{\sigma_i^2 - \sigma_j^2}{2\sigma_j^2}$
Laplace	$\ln \frac{\lambda_j}{\lambda_i} + \frac{\lvert\theta_i - \theta_j\rvert}{\lambda_j} + \frac{\lambda_i}{\lambda_j} \exp\left(-\lvert\theta_i - \theta_j\rvert/\lambda_i\right) - 1$
Maxwell Boltzmann	$3\ln\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{3(\sigma_i^2 - \sigma_j^2)}{2\sigma_j^2}$
Pareto	$\ln\left(\frac{m_i}{m_j}\right)^{a_j} + \ln \frac{a_i}{a_j} + \frac{a_j - a_i}{a_i}$, for $m_i \geq m_j$ and ∞ otherwise.
Rayleigh	$2\ln\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{\sigma_i^2 - \sigma_j^2}{\sigma_j^2}$
Uniform	$\ln \frac{b_j - a_j}{b_i - a_i}$ for $(a_i, b_i) \subseteq (a_j, b_j)$ and ∞ otherwise.
General Weibull	$\ln\left(\frac{k_i}{k_j}\left[\frac{\lambda_j}{\lambda_i}\right]^{k_j}\right) + \gamma \frac{k_j - k_i}{k_i} + \left(\frac{\lambda_i}{\lambda_j}\right)^{k_j} \Gamma\left(1 + \frac{k_j}{k_i}\right) - 1$

Chapter 3

Rényi Divergence and the Log-likelihood Ratio

3.1 Rényi entropy and the log-likelihood function

In his 2001 paper [57] Song established the following connection between the variance of the log-likelihood function and the differential Rényi entropy of order α, h_α, which we present in the proposition below. We provide a proof with additional steps and more detail than as it was originally presented by Song.

Proposition 3.1.1. *Let f be a probability density, then*

$$\lim_{\alpha \to 1} \frac{d}{d\alpha} h_\alpha(f) = -\frac{1}{2} Var(\ln f(X)) \ ,$$

assuming the integrals involved are well-defined and differentiation operations are legitimate.

Proof. For $\alpha \in \mathbb{R}^+ \setminus \{1\}$, let

$$F(\alpha) := \int f(x)^\alpha \, dx \ .$$

Sufficient conditions to exchange differentiation and integration in this context are provided for example by Theorem A.1.4. We suppose that the differentiability assumptions carry over up to the nth derivative for some $n \geq 2$. Then

$$\frac{d^n}{d\alpha^n} F(\alpha) = \int \frac{d^n}{d\alpha^n} f(x)^\alpha \, dx = \int f^\alpha(x) \left(\ln f(x)\right)^n \, dx \ .$$

Also,

$$\lim_{\alpha \to 1} F(\alpha) = \int f(x) \, dx = 1 \ , \text{ and}$$

$$\lim_{\alpha \to 1} \frac{d^n}{d\alpha^n} F(\alpha) = \int f(x) \left(\ln f(x)\right)^n \, dx = E_f \left[\left(\ln f(X)\right)^n \right] ,$$

where the continuity follows from the stronger assumptions on differentiability. For $\alpha \in \mathbb{R}^+ \setminus \{1\}$,

$$\frac{d}{d\alpha} h_\alpha(f) = \frac{d}{d\alpha} \left[\frac{1}{1-\alpha} \ln F(\alpha) \right] = \frac{1}{(1-\alpha)^2} \left[(1-\alpha) F(\alpha)^{-1} \frac{dF}{d\alpha} + \ln F(\alpha) \right] \ ,$$

and as $\alpha \to 1$ this becomes an indeterminate limit. Note that

$$\begin{aligned}
&\frac{d}{d\alpha} \left[(1-\alpha) F(\alpha)^{-1} \frac{d\alpha}{dF} + \ln F(\alpha) \right] \\
&\quad = (1-\alpha) \left[(-1) F(\alpha)^{-2} \left(\frac{dF}{d\alpha} \right)^2 + F(\alpha)^{-1} \frac{d^2 F}{d\alpha^2} \right] - F(\alpha)^{-1} \frac{dF}{d\alpha} + F^{-1}(\alpha) \frac{dF}{d\alpha} \\
&\quad = (1-\alpha) \left[(-1) F(\alpha)^{-2} \left(\frac{dF}{d\alpha} \right)^2 + F(\alpha)^{-1} \frac{d^2 F}{d\alpha^2} \right]
\end{aligned}$$

Hence, by l'Hospital's rule,

$$\begin{aligned}
\lim_{\alpha \to 1} \frac{d}{d\alpha} h_\alpha(f) &= \lim_{\alpha \to 1} \frac{(1-\alpha)}{-2(1-\alpha)} \left[(-1)F(\alpha)^{-2} \left(\frac{dF}{d\alpha} \right)^2 + F(\alpha)^{-1} \frac{d^2F}{d\alpha^2} \right] \\
&= -\frac{1}{2} \lim_{\alpha \to 1} \left[(-1)F(\alpha)^{-2} \left(\frac{dF}{d\alpha} \right)^2 + F(\alpha)^{-1} \frac{d^2F}{d\alpha^2} \right] \\
&= -\frac{1}{2} \left(-E_f [\ln f(X)]^2 + E[(\ln f(X))^2] \right) \\
&= -\frac{1}{2} \mathrm{Var} \left(\ln f(X) \right) \ .
\end{aligned}$$

□

When taking h_α as a function of α, Song [57] calls this function the *spectrum of Rényi information*.

3.2 Rényi divergence and the log-likelihood ratio

Motivated by the result above, we derive a similar expression involving the Rényi divergence and the log-likelihood ratio between two densities f_i and f_j.

Proposition 3.2.1. *Let f_i and f_j be two probability densities such that the integral definition of $D_\alpha(f_i \,||\, f_j)$ can be differentiated n times with respect to α ($n \geq 2$) by interchanging differentiation and integration, then*

$$\lim_{\alpha \to 1} \frac{d}{d\alpha} D_\alpha(f_i || f_j) = \frac{1}{2} Var_{f_i} \left(\ln \frac{f_i(X)}{f_j(X)} \right) \ .$$

Proof. The proof follows the same approach as above. If one considers the integral

$$G(\alpha) := \int f_i(x)^\alpha f_j^{1-\alpha}(x)\, dx$$

for $\alpha \in \mathbb{R}^+ \setminus \{1\}$, then under the differentiability assumptions

$$\frac{d}{d\alpha} G(\alpha) = \int f_i(x)^\alpha f_j^{1-\alpha} \left[\ln f_i(x) - \ln f_j(x)\right] dx = \int f_i(x)^\alpha f_j^{1-\alpha} \ln \frac{f_i(x)}{f_j(x)} dx ,$$

and similarly

$$\frac{d^n}{d\alpha^n} G(\alpha) = \int f_i(x)^\alpha f_j^{1-\alpha} \ln \left(\frac{f_i(x)}{f_j(x)}\right)^n dx ,$$

hence

$$\lim_{\alpha \to 1} \frac{d^n}{d\alpha^n} G(\alpha) = E_{f_i}\left[\left(\ln \frac{f_i(X)}{f_j(X)}\right)^n\right] .$$

The expression

$$\frac{d}{d\alpha} D_\alpha(f_i || f_j) = \frac{d}{d\alpha} \frac{1}{\alpha - 1} \ln G(\alpha) = \frac{1}{(\alpha-1)^2}\left[(\alpha-1)G(\alpha)^{-1}\frac{dG(\alpha)}{d\alpha} - \ln G(\alpha)\right]$$

becomes an indeterminate limit as $\alpha \to 1$, since $G(\alpha) \to 1$. Evaluating it with l'Hospital's rule yields

$$\begin{aligned}
\lim_{\alpha \to 1} \frac{d}{d\alpha} D_\alpha(f_i || f_j) &= \frac{1}{2} \lim_{\alpha \to 1} \left[(-1)G(\alpha)^{-2}\left(\frac{dG(\alpha)}{d\alpha}\right)^2 + G(\alpha)^{-1}\frac{d^2}{d\alpha^2} G(\alpha)\right] \\
&= \frac{1}{2}\left(-E_{f_i}\left[\left(\ln \frac{f_i(X)}{f_j(X)}\right)\right]^2 + E_{f_i}\left[\left(\ln \frac{f_i(X)}{f_j(X)}\right)^2\right]\right) \\
&= \frac{1}{2} \mathrm{Var}_{f_i}\left(\ln \frac{f_i(X)}{f_j(X)}\right) .
\end{aligned}$$

□

As an example we consider the case where f_i and f_j are two univariate Gaussian densities. From Proposition B.4.4 we see that

$$2\frac{d}{d\alpha} D_\alpha(f_i || f_j) = \frac{d}{d\alpha}\left[\frac{1}{(\alpha-1)} \ln\left(\frac{\sigma_j^2}{\sigma_0}\right) + \frac{\alpha \mu_0}{\sigma_0}\right]$$

where $\sigma_0 = \alpha\sigma_j^2 + (1-\alpha)\sigma_i^2$ and $\mu_0 = (\mu_i - \mu_j)^2$, hence

$$\frac{d\sigma_0}{d\alpha} = \sigma_j^2 - \sigma_i^2\,, \ \lim_{\alpha\to 1}\sigma_0 = \sigma_j^2\,, \ \text{ and } \lim_{\alpha\to 1}\frac{d}{d\alpha}\left(\frac{\alpha\mu_0}{\sigma_0}\right) = \frac{(\mu_i-\mu_j)^2\sigma_i^2}{\sigma_j^4}\,.$$

Thus

$$\begin{aligned}
\lim_{\alpha\to 1}\frac{d}{d\alpha}\frac{1}{(\alpha-1)}\ln\left(\frac{\sigma_j^2}{\sigma_0}\right) &= \lim_{\alpha\to 1}\left[\frac{-\frac{1}{\sigma_0}\frac{d\sigma_0}{d\alpha}(\alpha-1) - \ln\left(\frac{\sigma_j^2}{\sigma_0}\right)}{(\alpha-1)^2}\right] \\
&= \left[\lim_{\alpha\to 1}\frac{(\alpha-1)\left(\frac{1}{\sigma_0}\frac{d\sigma_0}{d\alpha}\right)^2 - \frac{1}{\sigma_0}\frac{d\sigma_0}{d\alpha} + \frac{1}{\sigma_0}\frac{d\sigma_0}{d\alpha}}{2(\alpha-1)}\right] \\
&= \frac{1}{2}\left(\frac{\sigma_j^2-\sigma_i^2}{\sigma_j^2}\right)^2,
\end{aligned}$$

where we have used l'Hospital's rule to evaluate the limit and the fact that $\dfrac{d^2\sigma_0}{d\alpha^2} = 0$. Thus, by Proposition 3.2.1

$$\begin{aligned}
\mathrm{Var}_{f_i}\left(\ln\frac{f_i(X)}{f_j(X)}\right) &= 2\lim_{\alpha\to 1}\frac{d}{d\alpha}D_\alpha(f_i||f_j) \\
&= \lim_{\alpha\to 1}\frac{d}{d\alpha}\left[\frac{1}{(\alpha-1)}\ln\left(\frac{\sigma_j^2}{\sigma_0}\right) + \frac{\alpha\mu_0}{\sigma_0}\right] \\
&= \frac{1}{2}\left(\frac{\sigma_j^2-\sigma_i^2}{\sigma_j^2}\right)^2 + \frac{(\mu_i-\mu_j)^2\sigma_i^2}{\sigma_j^4}\,.
\end{aligned}$$

Chapter 4

Rényi Divergence Rate for Stationary Gaussian Processes

In this chapter we consider information measure rates for Stationary Gaussian processes, in particular differential entropy rate, Rényi entropy rate, Kullback divergence rate, and Rényi divergence rate.

4.1 Toeplitz Matrices and Toeplitz Forms

We first introduce some results from the theory of Toeplitz matrices, which constitutes the main tool used in the calculation.

A Toeplitz matrix is an $n \times n$ matrix $T_n = \left[T_{kj}\right]$ s.t. $T_{kj} = T_{k-1\,j-1} = t_{k-j}$, i.e., :

$$\begin{pmatrix} t_0 & t_{-1} & t_{-2} & \dots & t_{-(n-1)} \\ t_1 & t_0 & t_{-1} & & \\ t_2 & t_1 & t_0 & & \vdots \\ \vdots & & & \ddots & \\ t_{n-1} & & & \cdots & t_0 \end{pmatrix}$$

A lot of applications in signal analysis and information theory assume the covariance matrix of the given process is Toeplitz, and that it has a constant mean function [28]. These processes are called weakly stationary. For Toeplitz covariance matrices, the autocorrelation function satisfies $K_X(k,j) = K_X(k-j)$.

Let $f(x)$ be a real-valued function in $L_1(-\pi,\pi)$ with Fourier series

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{inx} \ ,$$

where

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-inx} f(x)\, dx \ .$$

Then the Hermitian form

$$T_n = \sum_{i,j=0}^{n} c_{i-j} u_i \bar{u}_j \ ,$$

is called the (finite) Toeplitz form associated with $f(x)$. The asymptotic properties of the eigenvalues of Hermitian Toeplitz forms have been studied by Grenander and Szegö [29].

Note that

$$\overline{c_k} = \frac{1}{2\pi}\int_0^{2\pi} \overline{f(\lambda)e^{-ik\lambda}}\ dx = \frac{1}{2\pi}\int_0^{2\pi} f(\lambda)e^{ik\lambda}\, dx = c_{-k}$$

since f is real. Thus the Toeplitz form above can be represented by an $(n+1)\times(n+1)$ Toeplitz matrix with coefficients c_{i-j} via $T(u)=\bar{u}'Au$. To emphasize the relationship to the function f, T_n is denoted as $T_n(f)$.

A very important property of Toeplitz forms is the following theorem regarding the asymptotic distribution of eigenvalues, which can be found in Chapter 5, p.65 of [29]. Denote the eigenvalues of $T_n(f)$, by $\tau_1^{(n)}, \tau_2^{(n)}, ..., \tau_{n+1}^{(n)}$. Then the following holds

Theorem 4.1.1. *Let $f(\lambda)$ be a real-valued function in $L_1(-\pi,\pi)$, and denote by m and M the essential lower and upper bound of f, respectively, and assume that m and M are finite. If $F(\tau)$ is any continuous function defined on $[m,M]$, we have*

$$\lim_{n\to\infty} \frac{F\left(\tau_1^{(n)}\right)+F\left(\tau_2^{(n)}\right)+...+F\left(\tau_{n+1}^{(n)}\right)}{n+1} = \frac{1}{2\pi}\int_{-\pi}^{\pi} F[f(\lambda)]\,d\lambda\,.$$

4.2 Differential Entropy Rate for Gaussian Processes

The results of [29] can be used to evaluate limits of Toeplitz covariance matrix determinants, which becomes specially useful in the context of information rates for stationary Gaussian processes. The following specialized version is given by Gray [28]:

Theorem 4.2.1. *Let $T_n(f)$ be a sequence of Hermitian Toeplitz matrices with absolutely summable entries such that $\ln f(\lambda)$ is Riemann integrable and $f(\lambda)\geq m_f>0$. Then*

$$\lim_{n\to\infty} \left(\det\left(T_n(f)\right)\right)^{1/n} = \exp\left(\frac{1}{2\pi}\int_0^{2\pi} \ln f(\lambda)\,d\lambda\right)\,.$$

which is equivalent to

$$\lim_{n\to\infty}\frac{1}{n}\ln\left(\det\left(T_n(f)\right)\right) = \frac{1}{2\pi}\int_0^{2\pi} \ln f(\lambda)\,d\lambda\,.$$

While a complete development of these results is beyond the scope of this work, we note that the main idea linking the two results is that for Hermitian matrices T_n we can always express the determinant as the product of the eigenvalues τ_i of T_n, and so

$$\ln|T_n| = \ln\left(\prod_{i=1}^{n} \tau_i\right) = \sum_{i=1}^{n} \ln \tau_i \ .$$

Using the above result, Gray shows how to arrive at the differential entropy rate for Gaussian processes, a result originally obtained by Kolmogorov [37]. Consider a stationary zero mean Gaussian process $\{X^n\}$ determined by its mean autocorrelation function $\sigma_{k,j} = \sigma_{k-j} = E\left[X_k X_j\right]$, that is

$$f(\lambda) = \sum_{k=\infty}^{\infty} \sigma_k e^{ik\lambda} \ , \quad \sigma_l = \frac{1}{2\pi}\int_0^{2\pi} f(\lambda) e^{-i\lambda l}\, d\lambda \ .$$

For a fixed n the pdf of $X^n = (X_1, .., X_n)$ is

$$p_{X^n}(\boldsymbol{x}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{x}'\Sigma_n^{-1}\boldsymbol{x}\right)}{(2\pi)^{n/2}|\Sigma_n|^{1/2}} \ ,$$

where Σ_n is the $n \times n$ covariance matrix with entries σ_{k-j}. Since the process is stationary, the determinant $|\Sigma_n|$ is Toeplitz, and so

$$\lim_{n\to\infty} \frac{1}{n} \ln|\Sigma_n| = \frac{1}{2\pi}\int_0^{2\pi} \ln f(\lambda)\, d\lambda \ .$$

Using the the expression for differential entropy derived in Corollary B.4.7 we then have

$$\lim_{n\to\infty} \frac{1}{n} h(X^n) = \lim_{n\to\infty} \frac{1}{n}\left[\frac{1}{2}\ln(2\pi e)^n + \frac{1}{2}\ln|\Sigma_n|\right] = \frac{1}{2}\ln(2\pi e) + \frac{1}{4\pi}\int_0^{2\pi} \ln f(\lambda)\, d\lambda \ .$$

This expression was originally derive by Kolmogorov [37] and can also be found in p. 417 of [15], and in p. 76 of [34].

4.3 Divergence Rate for Stationary Gaussian Processes

A similar expression for the Kullback-Leibler divergence between two stationary Gaussian processes can be derived. We present below the expression given in p. 81 of [34]. The problem has also been considered in [61].

Theorem 4.3.1. *Let $X = \{X_n : n \in \mathbb{Z}\}$ and $Y = \{Y_n : n \in \mathbb{Z}\}$ be purely nondeterministic stationary Gaussian processes with spectral densities f and g, respectively. Then the relative entropy rate (Kullback divergence rate) is given by*

$$\overline{H}(f\,; g) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\frac{f(\lambda)}{g(\lambda)} - 1 - \ln\frac{f(\lambda)}{g(\lambda)}\right) d\lambda\ ,$$

provided that at least one of the following conditions is satisfied:

a) $f(\lambda)/g(\lambda)$ *is bounded.*

b) $g(\lambda) > a > 0\ , \forall\lambda \in [-\pi, \pi]$, *and* $f \in L^2[-\pi, \pi]$.

Proof. See [34]. □

4.4 Rényi Entropy Rate for Stationary Gaussian Processes

In [25], Golshani and Pasha derive the entropy rate for stationary Gaussian processes, starting from the following definition of conditional Rényi entropy between two continuous random variables X and Y having joint density $f(x, y)$:

$$h_\alpha(Y|X) = \frac{1}{1-\alpha}\ln\frac{\int_{\mathbb{R}^2} f^\alpha(x,y)\,dx\,dy}{\int_{\mathbb{R}} f(x)^\alpha dx}\ ,\ \ \alpha > 0\ ,\ \alpha \neq 1\ .$$

The above definition, based on the axioms of Jizba and Arimitsu [35], is studied by Golshani et. al in [26] and it is shown to be more suitable than the definition of conditional Rényi entropy found in [11]. Considering a stationary process $X = \{X_n\}_{n\in\mathbb{Z}}$, they show that the Rényi entropy rate, $\overline{h}_\alpha(X)$, exists and can be found via

$$\overline{h}_\alpha(X) = \lim_{n\to\infty} h_\alpha(X_n|X_{n-1},...,X_1) \ ,$$

which is used to arrive at the following theorem:

Theorem 4.4.1. *For a stationary Gaussian process, the rate of Rényi entropy is equal to*

$$\overline{h}_\alpha(X) = \frac{1}{2}\ln 2\pi\alpha^{\frac{1}{\alpha-1}} + \frac{1}{4\pi}\int_{-\pi}^{\pi} \ln 2\pi f(\lambda)\, d\lambda \ .$$

It is worth noting that while the Rényi information measures for distributions approach their Shannon counterparts as $\alpha \uparrow 1$, this does not in general hold for information rates. The work [3] provides some counterexamples to this. However, in this particular case, we can see that since

$$\lim_{\alpha\to 1}\frac{\ln\alpha}{\alpha-1} = 1 \ ,$$

the Rényi entropy rate approaches the differential entropy rate as $\alpha \to 1$. As shown below, this convergence is also seen for Rényi divergences.

4.5 Rényi Divergence Rate for Stationary Gaussian Processes

We now consider the Rényi divergence rate between two zero-mean stationary Gaussian processes $X = \{X_k : k \in \mathbb{N}\}$ and $Y = \{Y_k : k \in \mathbb{N}\}$, so that for a given n the vectors

$X^n = (X_1, ..., X_n)$ and $Y^n = (Y_1, ..., Y_n)$ have multivariate normal densities with covariance matrices Σ_{X^n} and Σ_{Y^n}. Assume X and Y have power spectral densities $f(\lambda)$ and $g(\lambda)$, respectively; i.e.,

$$f(\lambda) = \sum_{n=-\infty}^{\infty} r_n e^{in\lambda} \ , \quad r_k = \frac{1}{2\pi} \int_0^{2\pi} f(\lambda) e^{-i\lambda k} \, d\lambda \ ,$$

and,

$$g(\lambda) = \sum_{n=-\infty}^{\infty} l_n e^{in\lambda} \ , \quad l_k = \frac{1}{2\pi} \int_0^{2\pi} g(\lambda) e^{-i\lambda k} \, d\lambda \ ,$$

where $f(\lambda)$ and $g(\lambda)$ are assumed to have finite essential lower and upper bounds. For $\alpha \in (0,1)$ define $h(\lambda) := \alpha g(\lambda) + (1-\alpha) f(\lambda)$ and $s_k := \alpha l_k + (1-\alpha) r_k$. Then the Fourier series

$$\sum_{n=-N}^{N} s_n e^{in\lambda} = \alpha \sum_{n=-N}^{N} l_n e^{in\lambda} + (1-\alpha) \sum_{n=-N}^{N} r_n e^{in\lambda}$$

converges to $h(\lambda)$ as $N \to \infty$, where the kind of convergence is inherited via the triangle inequality from the convergence of the individual series involving r_n and l_n, whether it is meant as point-wise, uniform, in L^p, or in any other metric. Thus,

$$h(\lambda) = \sum_{n=-\infty}^{\infty} s_n e^{in\lambda} \ , \quad s_k = \frac{1}{2\pi} \int_0^{2\pi} h(\lambda) e^{-i\lambda k} \, d\lambda \ .$$

Note also that since Σ_{X^n} and Σ_{Y^n} are Toeplitz matrices, so is the matrix defined as $S_n := \alpha \Sigma_{Y^n} + (1-\alpha) \Sigma_{X^n}$. We are interested in finding the limit[1]

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} D_\alpha(X^n || Y^n) &= -\lim_{n\to\infty} \frac{1}{2(\alpha-1)} \frac{1}{n} \ln \frac{|\alpha \Sigma_{Y^n} + (1-\alpha) \Sigma_{X^n}|}{|\Sigma_{X^n}|^{1-\alpha} |\Sigma_{Y^n}|^{\alpha}} \\ &= \frac{1}{2(1-\alpha)} \lim_{n\to\infty} \left[\frac{1}{n} \left(\ln |S| - (1-\alpha) \ln |\Sigma_{X^n}| - \alpha \ln |\Sigma_{Y^n}| \right) \right] \ , \end{aligned}$$

[1] We perform the calculation using the significantly simpler expression for the Rényi divergence between two multivariate Gaussian distributions given in Section 2.2.4, as opposed to the originally derived (and equivalent) expression from Section B.4. Also, since we are considering zero-mean processes, one of the terms vanishes and we are left with the expression above.

where $\alpha\Sigma_{X^n}^{-1}+(1-\alpha)\Sigma_{Y^n}^{-1}$ is a positive-definite matrix so that the expression for $D_\alpha(X^n||Y^n)$ remains valid. Note also that all the determinants in the right hand side are Toeplitz, and $h(\lambda)$ satisfies the required assumptions of Theorem 4.1.1. Hence the limit in question can be computed as follows:

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}D_\alpha(X^n||Y^n)\\
&\qquad=\frac{1}{2(1-\alpha)}\left[\lim_{n\to\infty}\frac{1}{n}\ln|S|-(1-\alpha)\lim_{n\to\infty}\frac{1}{n}\ln|\Sigma_{X^n}|-\alpha\lim_{n\to\infty}\frac{1}{n}\ln|\Sigma_{Y^n}|\right]\\
&\qquad=\frac{1}{2(1-\alpha)}\left[\frac{1}{2\pi}\int_0^{2\pi}\ln h(\lambda)d\lambda-\frac{(1-\alpha)}{2\pi}\int\ln f(\lambda)d\lambda-\frac{\alpha}{2\pi}\int\ln g(\lambda)d\lambda\right]\\
&\qquad=\frac{1}{4\pi(1-\alpha)}\int_0^{2\pi}\ln\left(\frac{h(\lambda)}{f(\lambda)^{1-\alpha}g(\lambda)^\alpha}\right)d\lambda\\
&\qquad=\frac{1}{4\pi(1-\alpha)}\int_0^{2\pi}\ln\left(\frac{\alpha g(\lambda)+(1-\alpha)f(\lambda)}{f(\lambda)^{1-\alpha}g(\lambda)^\alpha}\right)d\lambda\,.
\end{aligned}$$

We can rearrange this result as

$$\begin{aligned}
D_\alpha(X||Y)=&\frac{1}{2(1-\alpha)}\left[\frac{1}{2\pi}\int_0^{2\pi}\ln\left(\alpha g(\lambda)+(1-\alpha)f(\lambda)\right)d\lambda\right.\\
&\left.-\frac{(1-\alpha)}{2\pi}\int_0^{2\pi}\ln f(\lambda)d\lambda-\frac{\alpha}{2\pi}\int_0^{2\pi}\ln g(\lambda)d\lambda\right]\\
=&-\frac{1}{2(\alpha-1)}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left(\left[1-\alpha+\alpha\frac{g(\lambda)}{f(\lambda)}\right]f(\lambda)\right)\right.\\
&\left.-\frac{(1-\alpha)}{2\pi}\int_{-\pi}^{\pi}\ln f(\lambda)d\lambda-\frac{\alpha}{2\pi}\int_{-\pi}^{\pi}\ln g(\lambda)d\lambda\right]\\
=&-\frac{1}{2(\alpha-1)}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left(1-\alpha+\alpha\frac{g(\lambda)}{f(\lambda)}\right)-\frac{\alpha}{2\pi}\int_{-\pi}^{\pi}\ln\frac{g(\lambda)}{f(\lambda)}d\lambda\right].
\end{aligned}$$

Proposition 8.29 in Vajda's book [59] asserts that if f and g are bounded above by a positive constant, then there exists $\epsilon>0$ such that for every $-\epsilon<\alpha<1+\epsilon$, the Rényi divergence rate is given by the above expression[2]. The parameter α in their case

[2]Strictly speaking, their expression is in terms of R_α so there is the α factor discrepancy we have

is allowed to take on negative values, and the constraint above ensures that positive-definiteness is maintained for the original expression of $D(X^n||Y^n)$ to hold [59]. The proof is also based on Theorem 4.1.1 from [29].

As a very special case we may consider zero-mean stationary Gauss-Markov processes X and Y with equal time constant β^{-1} and power spectral densities

$$f_X(\lambda) = \frac{2\sigma_X^2\beta}{\beta^2-\lambda^2}, \text{ and } f_Y(\lambda) = \frac{2\sigma_Y^2\beta}{\beta^2-\lambda^2}\ ,$$

where $\beta > \pi$. Then

$$\begin{aligned}\frac{\alpha f_Y(\lambda)+(1-\alpha)f_X(\lambda)}{f_X(\lambda)^{1-\alpha}f_Y(\lambda)^{\alpha}} &= \frac{2\beta\left(\alpha\sigma_Y^2+(1-\alpha)\sigma_X^2\right)}{\beta^2-\lambda^2}\left(\frac{\beta^2-\lambda^2}{2\sigma_X^2\beta}\right)^{1-\alpha}\left(\frac{\beta^2-\lambda^2}{2\sigma_Y^2\beta}\right)^{\alpha} \\ &= \frac{\alpha\sigma_Y^2+(1-\alpha)\sigma_X^2}{\left(\sigma_X^2\right)^{1-\alpha}\left(\sigma_Y^2\right)^{\alpha}}\end{aligned}$$

and

$$\begin{aligned}D_\alpha(X||Y) &= \frac{1}{4\pi(1-\alpha)}\int_0^{2\pi}\ln\left(\frac{\alpha g(\lambda)+(1-\alpha)f(\lambda)}{f(\lambda)^{1-\alpha}g(\lambda)^{\alpha}}\right)d\lambda \\ &= \frac{1}{2(1-\alpha)}\ln\frac{\alpha\sigma_Y^2+(1-\alpha)\sigma_X^2}{\left(\sigma_X^2\right)^{1-\alpha}\left(\sigma_Y^2\right)^{\alpha}}\ .\end{aligned}$$

---

mentioned before.

As a final remark, note that if the integral expression above is continuously differentiable in α by exchanging integration and differentiation, then

$$
\begin{aligned}
&\lim_{\alpha\to 1} D_\alpha(X^n||Y^n) \\
&= \lim_{\alpha\to 1}\frac{1}{4\pi(1-\alpha)}\int_0^{2\pi}\ln\left(\frac{\alpha g(\lambda)+(1-\alpha)f(\lambda)}{f(\lambda)^{1-\alpha}g(\lambda)^{\alpha}}\right)d\lambda \\
&= -\frac{1}{4\pi}\int_0^{2\pi}\lim_{\alpha\to 1}\frac{d}{d\alpha}\left[\ln\left(\alpha g(\lambda)+(1-\alpha)f(\lambda)\right)+(\alpha-1)\ln f(\lambda)-\alpha\ln g(\lambda)\right]d\lambda \\
&= -\frac{1}{4\pi}\int_0^{2\pi}\lim_{\alpha\to 1}\left[\frac{g(\lambda)-f(\lambda)}{\alpha g(\lambda)+(1-\alpha)f(\lambda)}+\ln f(\lambda)-\ln g(\lambda)\right]d\lambda \\
&= -\frac{1}{4\pi}\int_0^{2\pi}\frac{g(\lambda)-f(\lambda)}{g(\lambda)}+\ln\frac{f(\lambda)}{g(\lambda)}d\lambda \\
&= \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\frac{f(\lambda)}{g(\lambda)}-1-\ln\frac{f(\lambda)}{g(\lambda)}\right)d\lambda\,,
\end{aligned}
$$

where the limit was evaluated using l'Hospital's rule. This last expression corresponds to the Kullback divergence rate given in Theorem 4.3.1. In Table 4.1 we present the information rate expressions for stationary processes considered in this chapter.

Table 4.1: Information Rates for Stationary Gaussian Processes

Information Measure	**Rate**
Differential Entropy	$\frac{1}{2}\ln(2\pi e)+\frac{1}{4\pi}\int_0^{2\pi}\ln f(\lambda)\,d\lambda$
Rényi Entropy	$\frac{1}{2}\ln 2\pi\alpha^{\frac{1}{\alpha-1}}+\frac{1}{4\pi}\int_{-\pi}^{\pi}\ln 2\pi f(\lambda)\,d\lambda$
Kullback Divergence	$\frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\frac{f(\lambda)}{g(\lambda)}-1-\ln\frac{f(\lambda)}{g(\lambda)}\right)d\lambda$
Rényi Divergence	$\frac{1}{4\pi(1-\alpha)}\int_0^{2\pi}\ln\left(\frac{\alpha g(\lambda)+(1-\alpha)f(\lambda)}{f(\lambda)^{1-\alpha}g(\lambda)^{\alpha}}\right)d\lambda$

Chapter 5

Conclusion

In this thesis we derived closed-form expressions for Rényi and Kullback-Leibler divergences for several commonly used continuous distributions, and presented these results in a summarized form in Table 2.2 and Table 2.3. We demonstrated that the expressions corresponding to Exponential Families are in agreement with the results obtained by Liese and Vajda [40]. This compilation constitutes a useful addition to the literature, given that these measures are widely used in statistical and information theoretical applications, possess operational definitions in the sense of [30], and are related in simple manner to other probabilistic distances like the Chernoff and Hellinger divergences[1]

We also established a connection between the log-likelihood ratio between two distributions and their Rényi divergence, extending the work of Song [57], who consider the log-likelihood function and its relation to Rényi entopy. Given the compilation for Rényi divergence expressions we have provided, this result also becomes practically relevant.

[1] As given by Liese and Vajda [41] and introduced in Chapter 1.

Lastly, we investigated information rates for stationary Gaussian Sources, and derived an expression for the Rényi divergence rate using the asymptotic theory of Toeplitz matrices presented in [29] and [28]. This result was also shown to be in agreement with a later discovered expression presented in [59].

Natural extensions of this work are the expansion of the compilation with additional univariate and multivariate distributions, as well as considering information rates for more general Gaussian processes, beginning with the consideration on nonzero mean stationary Gaussian processes.

Appendices

Appendix A

Miscellaneous Background Results

A.1 Some Integration Results

Standard references for this material include for example [22] and [55].

Theorem A.1.1. *Dominated Convergence Theorem: Let* $\{f_n\}$ *be a sequence in* $L^1(\mu)$ *such that* $f_n \to f$ *and there exists a nonnegative* $g \in L^1(\mu)$ *such that* $|f_n| \leq g$ *for all* n*. Then* $f \in L^1(\mu)$ *and*

$$\int f \, d\mu(x) = \lim_{n\to\infty} \int f_n \, d\mu(x) \, .$$

Theorem A.1.2. *Let* f *be a bounded real-valued function on* $[a, b]$.

1. *If* f *is Riemann integrable, then* f *is Lebesgue integrable on* $[a, b]$ *and*

$$\int_a^b f(x)\, dx = \int_{[a,b]} f(x)\, d\mu(x) \, .$$

2. f *is Riemann integrable iff* $\{x \in [a.b] : f(x)$ *is discontinuous at* $x\}$ *has Lebesgue measure zero.*

Remark A.1.3. (See page 56 in [22]) If f is Riemann integrable on $[0, b]$ for all $b > 0$ and Lebesgue integrable on $[0, \infty)$ then

$$\int_{[0,\infty)} f \, d\mu = \lim_{b\to\infty} \int_0^b f(x)\, dx$$

by the dominated convergence theorem. In particular, this observation allows us to use standard Riemann improper integration tests for Lebesgue integrals of positive, a.s. Riemann integrable functions.

The following theorem (2.27 in [22]) provides conditions to exchange the order of integration and differentiation, a method we employ in several instances throughout the calculations of this work.

Theorem A.1.4. *Suppose $f : X \times [a, b] \to \mathbb{C}$ ($-\infty < a < b < \infty$) and $f(\cdot, t) : X \to \mathbb{C}$ is integrable for each $t \in [a, b]$. Let*

$$F(t) = \int_X f(x, t)\, d\mu(x)\ .$$

1. *Suppose there exists $g \in L^1(\mu)$ such that $|f(x, t)| \leq g(x)$ for all x, t. If $\lim_{t\to t_0} f(x, t) = f(x, t_0)$ for every x, then $\lim_{t\to t_0} F(t) = F(t_0)$; in particular if $f(\cdot, x)$ is continuous for each x, then F is continuous.*

2. *Let $\epsilon > 0$ and $V = (t_0 - \epsilon, t_0 + \epsilon) \subseteq [a, b]$[1]. Suppose $\dfrac{\partial f}{\partial t_0}$ exists and there is a $g \in L^1(\mu)$ such that $\left|\dfrac{\partial f}{\partial t}(x, t)\right| \leq g(x)$ for all $x \in X$ and $t \in V$. Then F is differentiable at t_0 and*

$$F'(t_0) = \int \frac{\partial f}{\partial t}(x, t_0) d\mu(x)\ .$$

[1] *This part was adapted to obtain diffetentiability at a particular point t_0 as opposed to the whole interval.*

A.2 Exponential Families

This material is based mainly on [39]. Let X be an n-dimensional Euclidean space and let $\mathscr{B}$ be the Borel algebra on X.

Definition A.2.1. Let μ be a $\sigma-$finite measure on $\mathscr{B}$. A family of distributions $P_{\boldsymbol{\theta}}$ is said to be an exponential family if the corresponding probability densities $p_{\boldsymbol{\theta}}$ with respect to μ are of the form

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = K(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{k} Q_j(\boldsymbol{\theta}) T_j(\boldsymbol{x})\right] h(\boldsymbol{x}) ,$$

where $\boldsymbol{\theta}$ is parameter vector and $T_j(\boldsymbol{x})$ and $Q(\boldsymbol{\theta})$ are real-valued measurable functions.

As seen above, in an exponential family the probability densities are of the form

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = g_{\boldsymbol{\theta}}(T(\boldsymbol{x}))h(\boldsymbol{x}) ,$$

hence by the factorization criterion[2] the vector $\boldsymbol{T}(\boldsymbol{x})$ is a $k-$dimensional sufficient statistic for a sample $(X_1, ..., X_n)$ drawn from the corresponding distribution.

The parametrization in Definition A.2.1 can be turned into the *natural parametrization* by replacing the original measure μ by a new measure ν, so as to include the factor of $h(\boldsymbol{x})$ and taking the functions $Q_j(\boldsymbol{\theta}) \to \tau_j$ as the new parameters; i.e., $h(\boldsymbol{x})$ is the density of ν with respect to μ,

$$\frac{d\nu}{d\mu} = h(\boldsymbol{x}) ,$$

and

$$\begin{aligned} \int_A K(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{k} Q_j(\boldsymbol{\theta}) T_j(\boldsymbol{x})\right] h(\boldsymbol{x}) d\mu(\boldsymbol{x}) \\ = \int_A \frac{1}{C(\boldsymbol{\tau})} \exp\left[\sum_{j=1}^{k} \tau_j T_j(\boldsymbol{x})\right] d\nu(\boldsymbol{x}) , \end{aligned}$$

[2] For a discussion of statistical sufficiency see for example [39, 13].

where $\boldsymbol{\tau} = (\tau_1, ..., \tau_k)$ and $C(\boldsymbol{\tau})$[3] is a normalization factor. Hence the definition below

Definition A.2.2. A family of distributions P_τ is a natural exponential family if it is an exponential family (as defined in Definition A.2.1) with respect to a measure ν on X where the corresponding densities have the form

$$\begin{aligned} \frac{dP_\tau}{d\nu} = p_\tau &= \frac{1}{C(\boldsymbol{\tau})} \exp\left[\sum_{j=1}^{k} \tau_j T_j(\boldsymbol{x})\right] \\ &= \frac{1}{C(\boldsymbol{\tau})} \exp\langle \boldsymbol{\tau}, \boldsymbol{T}(x)\rangle , \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^k$.

Definition A.2.3. The set $\Theta = \{\boldsymbol{\tau} \in \mathbb{R}^k : C(\boldsymbol{\tau}) < \infty\}$ is called the *natural parameter space*.

A.3 Special Functions

These results can be found in standard Mathematical Methods and Special Functions literature, such as [1, 54, 51]. For brevity we include here only the definitions and results that are immediately relevant to this work and leave out many of the interesting properties of these functions.

A.3.1 Gamma Function

There are several equivalent definitions of the Gamma function, but we consider only the integral representation.

[3] The density above written in this form to be consistent with the notation of the rest of this work.

Definition A.3.1. The Gamma Function is defined as the integral:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \ , \quad x > 0 \ .$$

Proposition A.3.2.

- *For* $r > 0$*,* $\Gamma(r+1) = r\Gamma(r)$*.*
- *Special Values:*

  (1) $\Gamma(1) = 1$.

  (2) $\Gamma(n) = (n-1)!$ *for* $n \in \mathbb{N}$*;*

  (3) $\Gamma(1/2) = \sqrt{\pi}$.

Definition A.3.3. The Euler-Mascheroni constant, γ is defined as

$$\gamma = \lim_{n\to\infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \log(n) \right) \approx 0.5772 \ .$$

Proposition A.3.4. $\Gamma'(1) = -\gamma$.

A.3.2 The Digamma Function

Definition A.3.5. The Digamma function $\psi(z)$ is defined as

$$\psi(z) = \frac{d}{dz} \ln \Gamma(z) \ .$$

Remark A.3.6.

$$\psi(x) = \frac{d}{dx} \ln \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)} = \frac{1}{\Gamma(x)} \int_0^\infty t^{x-1} e^{-t} \ln t \ dt \ .$$

A.3.3 The Beta Function

Definition A.3.7. The Beta function is defined by

$$B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}\,dt\ ,\quad x,y>0\ .$$

The general integral is sometimes referred to as the Beta integral.

Remark A.3.8. For real $x,y\leq 0$, we can see the Beta integral becomes $+\infty$, by the limit comparison test for integrals (see for example [21]).

An important and useful identity relating the Gamma and Beta functions is the following result:

Proposition A.3.9. *For $x,y>0$,*

$$B(x,y)=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\ .$$

Definition A.3.10. The above result can be used to define the Beta function for a vector argument $\boldsymbol{x}=(x_1,...x_n)$

$$B(\boldsymbol{x})=\frac{\prod_{k=1}^{n}\Gamma(x_k)}{\Gamma\left(\sum_{k=1}^{n}x_k\right)}\ ,\quad x_i>0\ ,\ i=1,..n\ .$$

Remark A.3.11. Using Proposition A.3.9 we can express the partial derivatives of $B(x,y)$ in terms of the Digamma function and the Beta function itself:

$$\begin{aligned}\frac{\partial}{\partial x}B(x,y)&=\frac{\partial}{\partial x}\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\\&=\Gamma(y)\left[\frac{\Gamma'(x)\Gamma(x+y)-\Gamma(x)\Gamma'(x+y)}{\Gamma^2(x+y)}\right]\\&=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\left[\frac{\Gamma'(x)}{\Gamma(x)}-\frac{\Gamma'(x+y)}{\Gamma(x+y)}\right]\\&=B(x,y)\left[\psi(x)-\psi(x+y)\right]\ ,\end{aligned}$$

and by symmetry,

$$\frac{\partial}{\partial y} B(x,y) = \left[\psi(y) - \psi(x+y)\right] \, .$$

A.3.4 Signum Function

Definition A.3.12. The signum function, denoted, $\operatorname{sgn}(x)$, is defined by

$$\operatorname{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

Remark A.3.13. For nonzero x, $\operatorname{sgn}(x) = \dfrac{x}{|x|}$.

A.4 Some Results from Matrix Algebra and Matrix Calculus

Many of results presented here can be found in standard linear algebra references such as [23], so we omit the proofs for such standard results here. The material on matrix derivatives can be found for example in [43, 44, 27].

A.4.1 Matrix Algebra

Definition A.4.1. An $n \times n$ real symmetric matrix A is said to be positive definite if $\boldsymbol{x}'A\boldsymbol{x} > 0$ for all nonzero vectors $\boldsymbol{x} \in \mathbb{R}^n$, where $(.)'$ denotes transposition.

Proposition A.4.2. *Two important properties of positive-definite matrices:*

- *A matrix A is positive-definite iff all of its eigenvalues are positive.*
- *A positive-definite matrix is invertible and its inverse is also positive definite.*

Proposition A.4.3. *Let A and B be two invertible $n \times n$ matrices and define a matrix C as*

$$C = \alpha A + (1 - \alpha)B \ , \quad \alpha \in \mathbb{R} \ .$$

Then

(1) If A and B are symmetric so is C.

(2) If A and B are positive-definite and $\alpha \in (0, 1)$, then C is also positive-definite and hence invertible as well.

Proof. The first claim is obvious. Let $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \neq 0$, $\alpha \in (0, 1)$ and A and B be positive-definite. Then

$$\boldsymbol{x}'C\boldsymbol{x} = \alpha \boldsymbol{x}'A\boldsymbol{x} + (1 - \alpha)\boldsymbol{x}'B\boldsymbol{x} > 0 \ ,$$

since all the terms are positive; which is also clear geometrically since this is just the convex combination of the two positive numbers $\boldsymbol{x}'A\boldsymbol{x}$ and $\boldsymbol{x}'B\boldsymbol{x}$. □

Proposition A.4.4. *Let $\boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^n$ be n-dimensional column vectors, c be a scalar, and let A be an invertible, symmetric, $n \times n$ matrix. Then*

$$\boldsymbol{x}'A\boldsymbol{x} - 2\boldsymbol{x}'\boldsymbol{b} + c = (\boldsymbol{x} - v)'A(\boldsymbol{x} - v) + d \ ,$$

where

$$v = A^{-1}\boldsymbol{b} \quad \textit{and} \quad d = c - \boldsymbol{b}A^{-1}\boldsymbol{b} \ .$$

Proof. Since A is invertible, we can write

$$\boldsymbol{x}'A\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{b} - \boldsymbol{b}'\boldsymbol{x} + c = \left(\boldsymbol{x}' - \boldsymbol{b}'A^{-1}\right)A\left(\boldsymbol{x} - A^{-1}\boldsymbol{b}\right) - \boldsymbol{b}'A^{-1}\boldsymbol{b} + c\ .$$

By the symmetry of A we have

$$\left(A^{-1}\boldsymbol{b}\right)' = \boldsymbol{b}'\left(A^{-1}\right)' = \boldsymbol{b}'A^{-1}\ ,$$

and so

$$\left(\boldsymbol{x}' - \boldsymbol{b}'A^{-1}\right) = \left(\boldsymbol{x}' - \left(A^{-1}\boldsymbol{b}\right)'\right) = \left(\boldsymbol{x} - A^{-1}\boldsymbol{b}\right)'\ .$$

Thus

$$\boldsymbol{x}'A\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{b} - \boldsymbol{b}'\boldsymbol{x} + c = \left(\boldsymbol{x} - A^{-1}\boldsymbol{b}\right)'A\left(\boldsymbol{x} - A^{-1}\boldsymbol{b}\right) - \boldsymbol{b}'A^{-1}\boldsymbol{b} + c.$$

□

Remark A.4.5. For $n = 1$ this is just the method of ‘completing the square’.

Definition A.4.6. A matrix A is orthogonally equivalent to another matrix B if there exists an orthogonal matrix P s.t. $B = PAP'$

Theorem A.4.7. *Let A be a real $n \times n$ matrix. Then A is symmetric iff A is orthogonally equivalent to a real diagonal matrix.*

Remark A.4.8. Given a real symmetric matrix, the diagonal matrix above consists of the eigenvalues of A and the orthogonal matrix is constructed from the eigenvectors of A.

A.4.2 Matrix Calculus

Throughout this section, all matrices considered, unless otherwise specified, are assumed to have entries which are differentiable functions of a parameter α.

Definition A.4.9. Let M be the matrix $[M_{ij}] = [M(\alpha)_{ij}]$. Then the derivative of M with respect to α is the matrix given by

$$\left[\frac{dM}{d\alpha}\right]_{ij} := \frac{dM(\alpha)_{ij}}{d\alpha} \; .$$

Proposition A.4.10. *Some differentiation results:*

(1) Given two constant matrices A and B, let $C = f(\alpha)A + g(\alpha)B$, where $f(\alpha)$ and $g(\alpha)$ are differentiable. Then

$$\frac{dC}{d\alpha} = \frac{df}{d\alpha}A + \frac{dg}{d\alpha}B \; .$$

(2) If A and B are of conforming dimensions then

$$\frac{d}{d\alpha}(AB) = \frac{dA}{d\alpha}B + A\frac{dB}{d\alpha} \; .$$

(3) If $B'AB$ is a well defined product, then

$$\frac{d}{d\alpha}\left(B'AB\right) = \frac{dB'}{d\alpha}AB + B'\frac{dA}{d\alpha}B + B'A\frac{dB}{d\alpha}$$

(4) If A is invertible, then

$$\frac{dA^{-1}}{d\alpha} = -A^{-1}\frac{dA}{d\alpha}A^{-1} \; .$$

Proof. (1) is obvious; for (2) observe that

$$\left[\frac{d}{d\alpha}(AB)\right]_{ij} = \frac{d}{d\alpha}\left[\sum_k A_{ik}B_{kj}\right] = \sum_k \frac{dA_{ik}}{d\alpha}B_{kj} + \sum_k A_{ik}\frac{dB_{kj}}{d\alpha} \; ;$$

(3) follows by applying the product rule twice; finally, we see that the (4) holds since

$$0 = \frac{dI}{d\alpha} = \frac{dAA^{-1}}{d\alpha} = \frac{dA}{d\alpha}A^{-1} + A\frac{dA^{-1}}{d\alpha} \Rightarrow \frac{dA^{-1}}{d\alpha} = -A^{-1}\frac{dA}{d\alpha}A^{-1} \; .$$

□

Proposition A.4.11. *If A is an invertible matrix, then*

$$\frac{d|A|}{d\alpha} = |A| tr \left(A^{-1} \frac{dA}{d\alpha} \right) ,$$

Proof. See [43]. □

Appendix B

Original Derivations for Exponential Families

This chapter contains the original calculations for the Rényi divergence expressions presented in Chapter 2, which are there derived in the framework of the results from [40]. The integration calculations presented in this chapter make use of the following techniques:

1. Single-variable integration techniques such as substitution and the method of integration by parts.

2. Reparametrization of some of the integrals so as to express the integrand as a known probability distribution scaled by some factor. Thus, if $f(x)$ can be written as $f(x) = K(\boldsymbol{\theta})g(x)$ where $g(x)$ is a pdf over a suport $\mathcal{X} \subseteq \mathbb{R}^n$ and $\boldsymbol{\theta}$ is a parameter vector, then

$$\int_{\mathcal{X}} f(x)dx = K(\boldsymbol{\theta}) \; .$$

3. Applying integral representations of special functions, in particular the Gamma and related functions.

As mentioned in Section 1.2 we follow the convention $0 \ln 0 = 0$, which is justified by continuity. Lastly, observe that these original calculations were not originally performed using the $(\theta)_\alpha$ notation we introduced in Chapter 2, and instead there appear parameters denoted as θ_0 which sometimes are equal to θ_α but sometimes are equal to θ^*_α. However, as we show in the each section of Chapter 2, the parameters are in agreement in all the derivations.

B.1 Gamma Distributions

Throughout this section let f_i and f_j be two univariate Gamma densities

$$f_i(x) = \frac{x^{k_i-1} e^{-x/\theta_i}}{\theta_i^{k_i}\Gamma(k_i)} \quad k_i, \theta_i > 0;\ x \in \mathbb{R}^+ .$$

where $\Gamma(x)$ is the Gamma function.

Proposition B.1.1.

$$E_{f_i}\left[\ln f_j\right] = -\ln\left(\theta_j^{k_j}\Gamma(k_j)\right) + \left(k_j - 1\right)\left[\ln\theta_i + \psi(k_i)\right] - \frac{\theta_i k_i}{\theta_j} ,$$

where $\psi(x)$ is the Digamma function.

Proof.

$$\begin{aligned} E_{f_i}\left[\ln f_j\right] &= E_{f_i}\left[-\ln\left(\theta_j^{k_j}\Gamma(k_j)\right) + \left(k_j-1\right)\ln X - \frac{X}{\theta_j}\right] \\ &= -\ln\left(\theta_j^{k_j}\Gamma(k_j)\right) + \left(k_j-1\right)E_{f_i}\left[\ln X\right] - \frac{1}{\theta_j}E_{f_i}\left[X\right] . \end{aligned}$$

For $r \geq -k_i$

$$
\begin{aligned}
E_{f_i}[X^r] &= \int_{\mathbb{R}^+} x^r \, \frac{x^{k_i-1} e^{-x/\theta_i}}{\theta_i^{k_i}\Gamma(k_i)} \, dx \\
&= \theta_i^r \frac{\Gamma(k_i+r)}{\Gamma(k_i)} \int_{\mathbb{R}^+} \frac{x^{k_i+r-1} e^{-x/\theta_i}}{\theta_i^{k_i+r}\Gamma(k_i+r)} \, dx \\
&= \theta_i^r \frac{\Gamma(k_i+r)}{\Gamma(k_i)} \, ,
\end{aligned}
$$

since the integrand corresponds to a reparametrized Gamma density with $k_i \mapsto k_i + r$. Hence

$$
E_{f_i}[X] = \theta_i \frac{\Gamma(k_i+1)}{\Gamma(k_i)} = \theta_i \frac{k_i \Gamma(k_i)}{\Gamma(k_i)} = \theta_i k_i \, ,
$$

where we have used the recursive relation for the Gamma function given in Proposition A.3.2. Consider now $E_{f_i}[\ln X]$:

$$
\begin{aligned}
E_{f_i}[\ln X] &= \int_{\mathbb{R}^+} \ln x \frac{x^{k_i-1} e^{-x/\theta_i}}{\theta_i^{k_i}\Gamma(k_i)} \, dx \\
&= \int_{\mathbb{R}^+} \ln(\theta_i y) \frac{(\theta_i y)^{k_i-1} e^{-y}}{\theta_i^{k_i}\Gamma(k_i)} \theta_i \, dy \, , \\
&\quad (\text{ where } y = x/\theta_i, \text{ and } \theta_i > 0 \Rightarrow y \in \mathbb{R}^+) \\
&= \frac{1}{\Gamma(k_i)} \int_{\mathbb{R}^+} \left[\ln \theta_i + \ln y\right] y^{k_i-1} e^{-y} \, dy \\
&= \frac{1}{\Gamma(k_i)} \ln \theta_i \Gamma(k_i) + \frac{1}{\Gamma(k_i)} \int_{\mathbb{R}^+} \ln y \, y^{k_i-1} e^{-y} \, dy \\
&= \ln \theta_i + \frac{\Gamma'(k_i)}{\Gamma(k_i)} \\
&= \ln \theta_i + \psi(k_i) \, ,
\end{aligned}
$$

where we have used Remark A.3.6. Then

$$\begin{aligned}
E_{f_i}\left[\ln f_j\right] &= -\ln\left(\theta_j^{k_j}\Gamma(k_j)\right) + \left(k_j - 1\right) E_{f_i}\left[\ln X\right] - \frac{1}{\theta_j} E_{f_i}\left[X\right] \\
&= -\ln\left(\theta_j^{k_j}\Gamma(k_j)\right) + \left(k_j - 1\right)\left[\ln\theta_i + \psi(k_i)\right] - \frac{\theta_i k_i}{\theta_j} .
\end{aligned}$$

□

Corollary B.1.2. *The differential entropy of f_i is*

$$h(f_i) = \ln\theta_i + \ln\Gamma(k_i) + \left(1 - k_i\right)\psi(k_i) + k_i .$$

Proof. Setting $i = j$ in Proposition B.1.1 we have

$$\begin{aligned}
h(f_i) &= -E_{f_i}\left[\ln f_i\right] \\
&= -\left[-\ln\left(\theta_i^{k_i}\Gamma(k_i)\right) + \left(k_i - 1\right)\left[\ln\theta_i + \psi(k_i)\right] - \frac{\theta_i k_i}{\theta_i}\right] \\
&= -\left[-k_i\ln\theta_i - \ln\Gamma(k_i) + \left(k_i - 1\right)\ln\theta_i + \left(k_i - 1\right)\psi(k_i) - k_i\right] \\
&= \ln\theta_i + \ln\Gamma(k_i) + \left(1 - k_i\right)\psi(k_i) + k_i .
\end{aligned}$$

□

Proposition B.1.3. *The Kullback-Liebler divergence between f_i and f_j is*

$$D(f_i||f_j) = \left(\frac{\theta_i - \theta_j}{\theta_j}\right) k_i + \ln\left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}}\right) + \left(k_i - k_j\right)\left(\ln\theta_i + \psi(k_i)\right) .$$

Proof. Using Proposition B.1.1 and Remark 1.2.4 we have

$$\begin{aligned}
D(f_i||f_j) &= E_{f_i}[\ln f_i] - E_{f_i}[\ln f_j] \\
&= -\left[\ln\theta_i + \ln\Gamma(k_i) + (1-k_i)\,\psi(k_i) + k_i\right] \\
&\quad - \left[-\ln\left(\theta_j^{k_j}\Gamma(k_j)\right) + \left(k_j - 1\right)\left[\ln\theta_i + \psi(k_i)\right] - \frac{\theta_i k_i}{\theta_j}\right] \\
&= k_i\left[\frac{\theta_i}{\theta_j} - 1\right] + \ln\theta_i\left[1 - k_j - 1\right] + \ln\frac{\Gamma(k_j)}{\Gamma(k_i)} \\
&\quad + \psi(k_i)\left[1 - k_j + k_i - 1\right] + k_j\ln\theta_j \\
&= \left(\frac{\theta_i - \theta_j}{\theta_j}\right)k_i - k_j\ln\theta_i + \ln\frac{\Gamma(k_j)}{\Gamma(k_i)} + \psi(k_i)\left[k_i - k_j\right] + k_j\ln\theta_j \\
&= \left(\frac{\theta_i - \theta_j}{\theta_j}\right)k_i + k_j\ln\theta_j + (k_i\ln\theta_i - k_i\ln\theta_i) - k_j\ln\theta_i \\
&\quad + \ln\frac{\Gamma(k_j)}{\Gamma(k_i)} + \psi(k_i)\left[k_i - k_j\right] \\
&= \left(\frac{\theta_i - \theta_j}{\theta_j}\right)k_i + \ln\left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}}\right) + \left(k_i - k_j\right)\left(\ln\theta_i + \psi(k_i)\right) .
\end{aligned}$$

□

Corollary B.1.4. *Let g_i and g_j be two exponential densities*

$$f_i = \lambda_i e^{-\lambda_i x}\ , \quad \lambda_i > 0;\ x > 0\ ,$$

then the Kullback-Leibler divergence between g_i and g_j is

$$D(g_i||g_j) = \ln\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j - \lambda_i}{\lambda_i}\ .$$

Proof. Note that the exponential densities g_i and g_j are obtained from the Gamma densities by setting $\theta_n = 1/\lambda_n$ and $k_n = 1$, $n = i, j$. Then

$$D(g_i||g_j) = \left(\frac{\frac{1}{\lambda_i} - \frac{1}{\lambda_j}}{\frac{1}{\lambda_j}}\right) + \ln\left(\frac{\Gamma(1)\lambda_i}{\Gamma(1)\lambda_j}\right) = \frac{\lambda_j - \lambda_i}{\lambda_i} + \ln\frac{\lambda_i}{\lambda_j}\ .$$

□

Corollary B.1.5. *Let h_i and h_j be two χ^2 densities*

$$h_i = \frac{x^{d_i/2-1}e^{-x/2}}{2^{d_i/2}\Gamma(d_i/2)}\ , \quad d_i \in \mathbb{N};\ x \in \mathbb{R}^+\ .$$

Then the Kullback-Leibler divergence between g_i and g_j is

$$D(h_i||h_j) = \ln\frac{\Gamma(d_j/2)}{\Gamma(d_i/2)} + \frac{d_i - d_j}{2}\psi(d_i/2)\ .$$

Proof. Note that the χ^2 densities g_i and g_j are obtained from the Gamma densities by setting $\theta_n = 2$ and $k_n = d_n/2$, $n = i, j$. Then

$$\begin{aligned} D(h_i||h_j) &= \ln\left(\frac{\Gamma(d_j/2)}{\Gamma(d_i/2)}2^{(d_j-d_i)/2}\right) + \left(\frac{d_i - d_j}{2}\right)(\ln 2 + \psi(d_i/2)) \\ &= \ln\frac{\Gamma(d_j/2)}{\Gamma(d_i/2)} + \frac{d_i - d_j}{2}\psi(d_i/2)\ . \end{aligned}$$

□

Proposition B.1.6. *For $\alpha \in \mathbb{R}^+ \setminus \{1\}$ let $k_0 = \alpha k_i + (1-\alpha)k_j$ and $\theta_0 = \alpha\theta_j + (1-a)\theta_i$. Then the Rényi divergence between f_i and f_j is given by*

$$D_\alpha(f_i||f_j) = \ln\left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}}\right) + \frac{1}{\alpha-1}\ln\left(\frac{\Gamma(k_0)}{\theta_i^{k_i}\Gamma(k_i)}\left(\frac{\theta_i\theta_j}{\theta_0}\right)^{k_0}\right)$$

for $k_0 > 0, \theta_0 > 0$, and

$$D_\alpha(f_i||f_j) = +\infty$$

otherwise.

Proof. We have

$$
\begin{aligned}
f_i^{\alpha} f_j^{1-\alpha} &= \left[\frac{x^{k_i - 1} e^{-x/\theta_i}}{\theta_i^{k_i} \Gamma(k_i)} \right]^{\alpha} \left[\frac{x^{k_j - 1} e^{-x/\theta_j}}{\theta_j^{k_j} \Gamma(k_j)} \right]^{1-\alpha} \\
&= x^{k_0 - 1} e^{-\frac{x}{\xi}} \frac{\theta_i^{-\alpha k_i} \theta_j^{k_j(\alpha-1)}}{\Gamma(k_i)^{\alpha} \Gamma(k_j)^{1-\alpha}} \\
&= x^{k_0 - 1} e^{-\frac{x}{\xi}} \frac{\theta_i^{k_i(1-\alpha)} \theta_j^{k_j(\alpha-1)} \theta_i^{-k_i}}{\Gamma(k_i)^{\alpha-1} \Gamma(k_j)^{1-\alpha} \Gamma(k_i)} \\
&= x^{k_0 - 1} e^{-\frac{x}{\xi}} \left(\frac{\Gamma(k_j) \theta_j^{k_j}}{\Gamma(k_i) \theta_i^{k_i}} \right)^{\alpha-1} \frac{1}{\theta_i^{k_i} \Gamma(k_i)} ,
\end{aligned}
$$

where

$$
k_0 = \alpha k_i + (1-\alpha) k_j \ , \ \frac{1}{\xi} = \frac{\alpha \theta_j + (1-a)\theta_i}{\theta_i \theta_j} = \frac{\theta_0}{\theta_i \theta_j} \ .
$$

- If $k_0 > 0$ and $\theta_0 > 0$, then

$$
\begin{aligned}
&\int_{\mathbb{R}^+} f_i^{\alpha} f_j^{1-\alpha} \, dx \\
&\quad = \int_{\mathbb{R}^+} \left[x^{k_0 - 1} e^{-\frac{x}{\xi}} \left(\frac{\Gamma(k_j) \theta_j^{k_j}}{\Gamma(k_i) \theta_i^{k_i}} \right)^{\alpha-1} \frac{1}{\theta_i^{k_i} \Gamma(k_i)} \right] dx \\
&\quad = \left(\frac{\Gamma(k_j) \theta_j^{k_j}}{\Gamma(k_i) \theta_i^{k_i}} \right)^{\alpha-1} \frac{\Gamma(k_0) \xi^{k_0}}{\theta_i^{k_i} \Gamma(k_i)} \int_{\mathbb{R}^+} \frac{x^{k_0 - 1} e^{-\frac{x}{\xi}}}{\xi^{k_0} \Gamma(k_0)} \, dx \\
&\quad = \left(\frac{\Gamma(k_j) \theta_j^{k_j}}{\Gamma(k_i) \theta_i^{k_i}} \right)^{\alpha-1} \frac{\Gamma(k_0) \xi^{k_0}}{\theta_i^{k_i} \Gamma(k_i)} ,
\end{aligned}
$$

since the integrand is Gamma density with parameters k_0 and ξ. Then

$$\begin{aligned}
D_\alpha(f_i||f_j) &= \frac{1}{\alpha-1}\ln\int_{\mathbb{R}^+} f_i^{\alpha} f_j^{1-\alpha}\,dx \\
&= \frac{1}{\alpha-1}\ln\left(\left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}}\right)^{\alpha-1}\frac{\Gamma(k_0)\xi^{k_0}}{\theta_i^{k_i}\Gamma(k_i)}\right) \\
&= \ln\left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}}\right) + \frac{1}{\alpha-1}\ln\left(\frac{\Gamma(k_0)\xi^{k_0}}{\theta_i^{k_i}\Gamma(k_i)}\right) \\
&= \ln\left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}}\right) + \frac{1}{\alpha-1}\ln\left(\frac{\Gamma(k_0)}{\theta_i^{k_i}\Gamma(k_i)}\left(\frac{\theta_i\theta_j}{\theta_0}\right)^{k_0}\right) .
\end{aligned}$$

Note that for $\alpha \in (0,1)$ we always have $k_0 > 0$ and $\theta_0 > 0$ given the positivity of k_i, k_j, θ_i and θ_j.

- If $\theta_0 \le 0$ then $(1/\xi) \le 0$ [1] Then

$$\begin{aligned}
\int_{R^+} f_i^{\alpha} f_j^{1-a}\,dx &= A\int_{\mathbb{R}^+} x^{k_0-1}e^{Kx}dx\ , \quad K, A \ge 0 \\
&\ge A\int_{\mathbb{R}^+} x^{k_0-1} = \infty,
\end{aligned}$$

for all real values of k_0.

- If $\theta_0 > 0$ but $k_0 < 0$ then

$$\begin{aligned}
\int_{\mathbb{R}^+} f_i^{\alpha} f_j^{1-\alpha}dx &= A_1\int_{\mathbb{R}^+} x^{k_0-1}e^{-|\theta_0|x}\,dx\ , \quad A_1 > 0 \\
&= A_2\int_{\mathbb{R}^+} y^{k_0-1}e^{-y}dy, \quad A_2 > 0 \\
&> A_2\int_0^1 y^{k_0-1}e^{-y}dy\ .
\end{aligned}$$

[1] Here we use the parentheses to emphasize that $1/\xi$ is just the symbol defined above as opposed to the reciprocal of some $\xi \in \mathbb{R}$, so that it can in fact equal 0. When $(1/\xi) \neq 0$ then ξ, defined as its reciprocal, is indeed a real number.

Since $e^{-y} \to 1$ as $y \to 0$, then

$$\int_0^1 y^{k_0-1} e^{-y} dy = \infty \text{ , since } \int_0^1 y^p dy = \infty$$

for $p < -1$. Finally, since nonpositive k_0 and θ_0 only occur for $\alpha > 1$ we have

$$D_\alpha(f_i || f_j) = \frac{1}{\alpha - 1} \ln \int_{\mathbb{R}^+} f_i^{\alpha} f_j^{1-\alpha} dx = +\infty$$

for these cases.

□

Corollary B.1.7. *Let g_i and g_j be two exponential densities*

$$g_i = \lambda_i e^{-\lambda_i x} \text{ ,} \quad \lambda_i > 0; \; x \in \mathbb{R}^+ \text{ .}$$

For $\alpha \in \mathbb{R}^+ \setminus \{1\}$, let $\lambda_0 = \alpha\lambda_i + (1-\alpha)\lambda_j$. Then the Rényi divergence between g_i and g_j is given by

$$D_\alpha(g_i || g_j) = \ln \frac{\lambda_i}{\lambda_j} + \frac{1}{\alpha - 1} \ln \frac{\lambda_i}{\lambda_0}$$

for $\lambda_0 > 0$, and

$$D_\alpha(g_i || g_j) = +\infty$$

otherwise.

Proof. Setting $\theta_n = 1/\lambda_n$ and $k_n = 1$, $n = i, j$ we have

$$k_0 = \alpha k_i + (1-\alpha) k_j = 1 \text{ ,} \quad \lambda_0 = \alpha \lambda_i + (1-\alpha)\lambda_j = \frac{\theta_0}{\theta_i \theta_j}$$

so that $k_0 > 0$, and $\theta_0 > 0 \Leftrightarrow \lambda_0 > 0$. Then it follows from Proposition B.1.6 that

$$\begin{aligned} D_\alpha(g_i || g_j) &= \ln \left(\frac{\Gamma(1)\lambda_i}{\Gamma(1)\lambda_j} \right) + \frac{1}{\alpha - 1} \ln \left(\frac{\lambda_i \Gamma(1)}{\Gamma(1)} \frac{1}{\lambda_0} \right) \\ &= \ln \frac{\lambda_i}{\lambda_j} + \frac{1}{\alpha - 1} \ln \frac{\lambda_i}{\lambda_0} \text{ .} \end{aligned}$$

□

Corollary B.1.8. *Let h_i and h_j be two χ^2 densities*

$$h_i = \frac{x^{d_i/2-1}e^{-x/2}}{2^{d_i/2}\Gamma(d_i/2)} \ , \quad d_i \in \mathbb{N};\ x \in \mathbb{R}^+ \ .$$

For $\alpha \in \mathbb{R}^+ \setminus \{1\}$, let $d_0 = \alpha d_i + (1-\alpha)d_j$. Then the Rényi divergence between h_i and h_j is given by

$$D_\alpha(h_i||h_j) = \ln\left(\frac{\Gamma(d_j/2)}{\Gamma(d_i/2)}\right) + \frac{1}{\alpha-1}\ln\left(\frac{\Gamma(d_0/2)}{\Gamma(d_i/2)}\right) \ .$$

for $d_0 > 0$ and

$$D_\alpha(h_i||h_j) = +\infty$$

otherwise.

Proof. Setting $\theta_n = 2$ and $k_n = d_i/2$, $n = i, j$, we have

$$k_0 = \alpha k_i + (1-\alpha)k_j = \frac{d_0}{2} \ , \quad \theta_0 = \alpha\, 2 + (1-\alpha)2 = 2 \ ,$$

so that $k_0 > 0 \Leftrightarrow d_0 > 0$, and $\theta_0 > 0$. Then by Proposition B.1.6

$$\begin{aligned} D_\alpha(h_i||h_j) &= \ln\left(\frac{\Gamma(d_j/2)2^{d_j/2}}{\Gamma(d_i/2)2^{d_i/2}}\right) + \frac{1}{\alpha-1}\ln\left(\frac{\Gamma(d_0/2)\,2^{d_0/2}}{2^{d_i/2}\Gamma(d_i/2)}\right) \\ &= \ln\left(\frac{\Gamma(d_j/2)}{\Gamma(d_i/2)}\right) + \frac{1}{\alpha-1}\ln\left(\frac{\Gamma(d_0/2)}{\Gamma(d_i/2)}\right) \\ &\quad + \frac{\ln 2}{2}\left(d_j - d_i + \frac{d_0 - d_i}{\alpha-1}\right) \ . \end{aligned}$$

But

$$d_j - d_i + \frac{d_0 - d_i}{\alpha - 1} = \frac{1}{\alpha-1}\left[(\alpha-1)(d_j - d_i) + (\alpha d_i + (1-\alpha)d_j) - d_i\right] = 0 \ .$$

Thus,

$$D_\alpha(h_i||h_j) = \ln\left(\frac{\Gamma(d_j/2)}{\Gamma(d_i/2)}\right) + \frac{1}{\alpha-1}\ln\left(\frac{\Gamma(d_0/2)}{\Gamma(d_i/2)}\right) \ .$$

□

Remark B.1.9.

$$\lim_{\alpha \uparrow 1} D_\alpha(f_i || f_j) = D(f_i || f_j) \ .$$

Proof.

$$\lim_{\alpha \uparrow 1} D_\alpha(f_i || f_j) = \ln \left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}} \right) + \lim_{\alpha \uparrow 1} \frac{1}{\alpha - 1} \ln \left(\frac{\Gamma(k_0)}{\theta_i^{k_i}\Gamma(k_i)} \left(\frac{\theta_i \theta_j}{\theta_0} \right)^{k_0} \right)$$

Note that

$$\begin{aligned}
\lim_{\alpha \to 1} k_0 &= \lim_{\alpha \to 1} \left(\alpha k_i + (1-\alpha) k_j \right) = k_i \ , \quad \text{and} \\
\lim_{\alpha \to 1} \theta_0 &= \lim_{\alpha \to 1} \left(\alpha \theta_j + (1-\alpha) \theta_i \right) = \theta_j
\end{aligned}$$

so that the second limit is of indeterminate form. Applying l'Hospital's rule

$$\begin{aligned}
&\lim_{\alpha \uparrow 1} \left[\frac{1}{\alpha - 1} \ln \left(\frac{\Gamma(k_0)}{\theta_i^{k_i}\Gamma(k_i)} \left(\frac{\theta_i \theta_j}{\theta_0} \right)^{k_0} \right) \right] \\
&= \lim_{\alpha \uparrow 1} \frac{d}{d\alpha} \left[\ln \Gamma(k_0) + k_0 \ln(\theta_i \theta_j) - k_0 \ln \theta_0 \right] \\
&= \lim_{\alpha \uparrow 1} \left[\frac{dk_0}{d\alpha} \left(\frac{d}{dk_0} \ln \Gamma(k_0) + \ln(\theta_i \theta_j) - \ln \theta_0 \right) - \frac{k_0}{\theta_0} \frac{d\theta_0}{d\alpha} \right] \\
&= \lim_{\alpha \uparrow 1} \left[(k_i - k_j) \left(\psi(k_0) + \ln \frac{\theta_i \theta_j}{\theta_0} \right) - \frac{k_0}{\theta_0} (\theta_j - \theta_i) \right] \\
&= (k_i - k_j) \left(\psi(k_i) + \ln \theta_i \right) - \frac{k_i}{\theta_j} (\theta_j - \theta_i) \\
&= \left(\frac{\theta_i - \theta_j}{\theta_j} \right) k_i + (k_i - k_j) \left(\psi(k_i) + \ln \theta_i \right) \ .
\end{aligned}$$

Hence,

$$\lim_{\alpha \uparrow 1} D_\alpha(f_i || f_j) = \ln \left(\frac{\Gamma(k_j)\theta_j^{k_j}}{\Gamma(k_i)\theta_i^{k_i}} \right) + \left(\frac{\theta_i - \theta_j}{\theta_j} \right) k_i + (k_i - k_j) \left(\psi(k_i) + \ln \theta_i \right)$$

which equals the expression for $D(f_i || f_j)$ given in Proposition B.1.3, as expected. □

B.2 Chi Distributions

Throughout this section let f_i and f_j be two Chi densities

$$f_i(x) = \frac{2^{1-k_i/2} x^{k_i-1} e^{-x^2/2\sigma_i^2}}{\sigma_i^{k_i} \Gamma\left(\frac{k_i}{2}\right)} \ , \sigma_i > 0 \ , \ k_i \in \mathbb{N} \ ; \ x \in \mathbb{R}^+ \ .$$

Proposition B.2.1.

$$E_{f_i}\left[\ln f_j\right] = \frac{1}{2}(k_j - 1)\psi(k_i/2) + \ln\left[\frac{\sqrt{2}\sigma_i^{k_j-1}}{\sigma_j^{k_j}\Gamma\left(\frac{k_j}{2}\right)}\right] - \frac{\sigma_i^2}{2\sigma_j^2}k_i \ .$$

Proof.

$$\begin{aligned} E_{f_i}\left[\ln f_j\right] &= E_{f_i}\left[(1-k_j/2)\ln 2 + (k_j - 1)\ln X - \frac{X^2}{2\sigma_j^2} - \ln\left(\sigma_j^{k_j}\Gamma\left(\frac{k_j}{2}\right)\right)\right] \\ &= (1-k_j/2)\ln 2 - \ln\left(\sigma_j^{k_j}\Gamma\left(\frac{k_j}{2}\right)\right) + (k_j-1)E_{f_i}\left[\ln X\right] - \frac{1}{2\sigma_j^2}E_{f_i}\left[X^2\right] \end{aligned}$$

Let $r > -k_i$. Note that

$$\begin{aligned} E_{f_i}\left[X^r\right] &= \int_{\mathbb{R}^+} x^r \frac{2^{1-k_i/2} x^{k_i-1} e^{-x^2/2\sigma_i^2}}{\sigma_i^{k_i}\Gamma\left(\frac{k_i}{2}\right)} dx \\ &= 2^{r/2} \ \sigma_i^r \frac{\Gamma\left(\frac{k_i+r}{2}\right)}{\Gamma\left(\frac{k_i}{2}\right)} \int_{\mathbb{R}^+} \frac{2^{1-(k_i+r)/2} x^{k_i+r-1} e^{-x^2/2\sigma_i^2}}{\sigma_i^{k_i+r}\Gamma\left(\frac{k_i+r}{2}\right)} dx \\ &= \left(2^{1/2}\sigma_i\right)^r \frac{\Gamma\left(\frac{k_i+r}{2}\right)}{\Gamma\left(\frac{k_i}{2}\right)} \ , \end{aligned}$$

since the last integrand corresponds to a reparametrized Chi density with $k_i \mapsto k_i + r > 0$. Then

$$E_{f_i}\left[X^2\right] = 2\sigma_i^2 \frac{\Gamma\left(\frac{k_i}{2}+1\right)}{\Gamma\left(\frac{k_i}{2}\right)} = \sigma_i^2 k_i \ ,$$

where we have used the recursion relation for the Gamma function (Proposition A.3.2). Also,

$$\begin{aligned}
E_{f_i}[\ln X] &= \frac{d}{dr} E_{f_i}[X^r]\bigg|_{r=0} = \frac{d}{dr}\left[\left(2^{1/2}\sigma_i\right)^r \frac{\Gamma\left(\frac{k_i+r}{2}\right)}{\Gamma\left(\frac{k_i}{2}\right)}\right]_{r=0} \\
&= \frac{\left(2^{1/2}\sigma_i\right)^r \ln\left(2^{1/2}\sigma_i\right)\,\Gamma\left(\frac{k_i+r}{2}\right) + \left(2^{1/2}\sigma_i\right)^r\,\Gamma'\left(\frac{k_i+r}{2}\right)\frac{1}{2}}{\Gamma\left(\frac{k_i}{2}\right)}\Bigg|_{r=0} \\
&= \frac{1}{\Gamma\left(\frac{k_i}{2}\right)}\left[\ln\left(2^{1/2}\sigma_i\right)\,\Gamma\left(\frac{k_i}{2}\right) + \Gamma'\left(\frac{k_i}{2}\right)\frac{1}{2}\right] \\
&= \frac{1}{2}\left[\ln\left(2\sigma_i^2\right) + \psi(k_i/2)\right] .
\end{aligned}$$

Finally,

$$\begin{aligned}
E_{f_i}\left[\ln f_j\right] &= (1-k_j/2)\ln 2 - \ln\left(\sigma_j^{k_j}\Gamma\left(\frac{k_j}{2}\right)\right) \\
&\qquad + (k_j-1)E_{f_i}[\ln X] - \frac{1}{2\sigma_j^2}E_{f_i}\left[X^2\right] \\
&= (1-k_j/2)\ln 2 - \ln\left(\sigma_j^{k_j}\Gamma\left(\frac{k_j}{2}\right)\right) \\
&\qquad + (k_j-1)\frac{1}{2}\left[\ln\left(2\sigma_i^2\right) + \psi(k_i/2)\right] - \frac{\sigma_i^2}{2\sigma_j^2}k_i \\
&= \frac{1}{2}(k_j-1)\psi(k_i/2) + (k_j-1)\ln\sigma_i \\
&\quad + \ln 2\left[\frac{1}{2}(k_j-1) + (1-k_j/2)\right] - \frac{\sigma_i^2}{2\sigma_j^2}k_i - \ln\left(\sigma_j^{k_j}\Gamma\left(\frac{k_j}{2}\right)\right) \\
&= \frac{1}{2}(k_j-1)\psi(k_i/2) + \ln\left(\sigma_i^{k_j-1}\right) \\
&\quad + \frac{1}{2}\ln 2 - \frac{\sigma_i^2}{2\sigma_j^2}k_i - \ln\left(\sigma_j^{k_j}\Gamma\left(\frac{k_j}{2}\right)\right) \\
&= \frac{1}{2}(k_j-1)\psi(k_i/2) + \ln\left[\frac{\sqrt{2}\sigma_i^{k_j-1}}{\sigma_j^{k_j}\Gamma\left(\frac{k_j}{2}\right)}\right] - \frac{\sigma_i^2}{2\sigma_j^2}k_i .
\end{aligned}$$

□

Corollary B.2.2. *The differential entropy of f_i is*

$$h(f_i) = \frac{1}{2}(1-k_i)\psi\left(k_i/2\right) + \ln\left[\frac{\sigma_i \Gamma\left(\frac{k_i}{2}\right)}{\sqrt{2}}\right] + \frac{k_i}{2}.$$

Proof. Setting $i = j$ in Proposition B.2.1 we have

$$\begin{aligned} h(f_i) = -E_{f_i}\left[\ln f_i\right] &= -\left[\frac{1}{2}(k_i-1)\psi(k_i/2) + \ln\left[\frac{\sqrt{2}\sigma_i^{k_i-1}}{\sigma_i^{k_i}\Gamma\left(\frac{k_i}{2}\right)}\right] - \frac{\sigma_i^2}{2\sigma_i^2}k_i\right] \\ &= \frac{1}{2}(1-k_i)\psi\left(k_i/2\right) + \ln\left[\frac{\sigma_i \Gamma\left(\frac{k_i}{2}\right)}{\sqrt{2}}\right] + \frac{k_i}{2}. \end{aligned}$$

□

Proposition B.2.3. *The Kullback-Liebler divergence between f_i and f_j is*

$$D(f_i||f_j) = \frac{1}{2}\psi(k_i/2)\left(k_i - k_j\right) + \ln\left[\left(\frac{\sigma_j}{\sigma_i}\right)^{k_j}\frac{\Gamma(k_j/2)}{\Gamma(k_i/2)}\right] + \frac{k_i}{2\sigma_j^2}\left(\sigma_i^2 - \sigma_j^2\right).$$

Proof. Using Proposition B.2.1 and Remark 1.2.4 we have

$$\begin{aligned} D(f_i||f_j) &= E_{f_i}[\ln f_i] - E_{f_i}[\ln f_j] \\ &= -\left[\frac{1}{2}(1-k_i)\psi\left(k_i/2\right) + \ln\left[\frac{\sigma_i\Gamma\left(\frac{k_i}{2}\right)}{\sqrt{2}}\right] + \frac{k_i}{2}\right] \\ &\quad - \left[\frac{1}{2}(k_j-1)\psi(k_i/2) + \ln\left[\frac{\sqrt{2}\sigma_i^{k_j-1}}{\sigma_j^{k_j}\Gamma\left(\frac{k_j}{2}\right)}\right] - \frac{\sigma_i^2}{2\sigma_j^2}k_i\right] \\ &= \frac{1}{2}\psi(k_i/2)\left(k_i - 1 + 1 - k_j\right) - \ln\left[\left(\frac{\sigma_i}{\sigma_j}\right)^{k_j}\frac{\Gamma(k_i/2)}{\Gamma(k_j/2)}\right] + \frac{k_i}{2}\left(\frac{\sigma_i^2}{\sigma_j^2} - 1\right) \\ &= \frac{1}{2}\psi(k_i/2)\left(k_i - k_j\right) + \ln\left[\left(\frac{\sigma_j}{\sigma_i}\right)^{k_j}\frac{\Gamma(k_j/2)}{\Gamma(k_i/2)}\right] + \frac{k_i}{2\sigma_j^2}\left(\sigma_i^2 - \sigma_j^2\right). \end{aligned}$$

□

Remark B.2.4. For $k_i = k_n = k$, we have

$$\begin{aligned} & D(f_i(x;k,\sigma_i)||f_j(x;k,\sigma_j)) \\ & = \frac{1}{2}\psi(k/2)(k-k) + \ln\left[\left(\frac{\sigma_j}{\sigma_i}\right)^k \frac{\Gamma(k/2)}{\Gamma(k/2)}\right] + \frac{k}{2\sigma_j^2}\left(\sigma_i^2 - \sigma_j^2\right) \\ & = k\left[\ln\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{\sigma_i^2 - \sigma_j^2}{2\sigma_j^2}\right] . \end{aligned}$$

Corollary B.2.5. *Special Cases:*

1. *Let h_i and h_j be two half-normal densities*

$$h_i = \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{x^2}{2\sigma_i^2}}}{\sigma_i} \quad , \quad \sigma_i > 0;\ x \in \mathbb{R}^+ .$$

Then the Kullback-Leibler divergence between h_i and h_j is

$$D(h_i||h_j) = \ln\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{\sigma_i^2 - \sigma_j^2}{2\sigma_j^2} .$$

2. *Let r_i and r_j be two Rayleigh densities*

$$r_i = \frac{x}{\sigma_i^2} e^{-\frac{x^2}{2\sigma_i^2}} \quad , \quad \sigma_i > 0;\ x \in \mathbb{R}^+ .$$

Then the Kullback-Leibler divergence between r_i and r_j is

$$D(r_i||r_j) = 2\ln\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{\sigma_i^2 - \sigma_j^2}{\sigma_j^2} .$$

3. *Let m_i and m_j be two Maxwell-Boltzmann densities*

$$m_i = \sqrt{\frac{2}{\pi}} \frac{x^2 e^{-\frac{x^2}{2\sigma_i^2}}}{\sigma_i^3}, \quad \sigma_i > 0;\ x \in \mathbb{R}^+ .$$

Then the Kullback-Leibler divergence between m_i and m_j is

$$D(m_i||m_j) = 3\ln\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{3(\sigma_i^2 - \sigma_j^2)}{2\sigma_j^2} .$$

Proof.

1. The half-normal densities h_i and h_j are obtained from the Chi densities by setting $k_n = 1$, $n = i, j$:

$$f_i(x; k_i = 1) = \frac{2^{1-1/2} x^{1-1} e^{-\frac{x^2}{2\sigma_i^2}}}{\sigma_i^1 \Gamma\left(\frac{1}{2}\right)} = \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{x^2}{2\sigma_i^2}}}{\sigma_i} ,$$

   where $\Gamma(1/2) = \sqrt{\pi}$.

2. The Rayleigh densities r_i and r_j are obtained from the Chi densities by setting $k_n = 2$, $n = i, j$:

$$f_i(x; k_i = 1) = \frac{2^{1-2/2} x^{2-1} e^{-\frac{x^2}{2\sigma_i^2}}}{\sigma_i^2 \Gamma\left(\frac{2}{2}\right)} = \frac{x}{\sigma_i^2} e^{-\frac{x^2}{2\sigma_i^2}} .$$

3. The Maxwell-Boltzmann densities m_i and m_j are obtained from the Chi densities by setting $k_n = 3$, $n = i, j$:

$$f_i(x; k_i = 1) = \frac{2^{1-3/2} x^{3-1} e^{-\frac{x^2}{2\sigma_i^2}}}{\sigma_i^3 \Gamma\left(\frac{3}{2}\right)} = \sqrt{\frac{2}{\pi}} \frac{x^2 \, e^{-\frac{x^2}{2\sigma_i^2}}}{\sigma_i^3} ,$$

   where $\Gamma(3/2) = \Gamma(1 + 1/2) = \frac{1}{2}\Gamma(1/2) = \frac{1}{2}\sqrt{\pi}$.

The corresponding expressions follow from substituting the values $k = 1, 2, 3$ in the expression given in Remark B.2.4. $\square$

Proposition B.2.6. *For $\alpha \in \mathbb{R}^+ \setminus \{1\}$ let $\sigma_0 = \alpha\sigma_j^2 + (1-\alpha)\sigma_i^2$ and $k_0 = \alpha k_i + (1-\alpha)k_j$. Then the Rényi divergence between f_i and f_j is given by*

$$D_\alpha(f_i || f_j) = \ln\left(\frac{\sigma_j^{k_j}\Gamma(k_j/2)}{\sigma_i^{k_i}\Gamma(k_i/2)}\right) + \frac{1}{\alpha-1}\ln\left(\left(\frac{\sigma_i^2\sigma_j^2}{\sigma_0}\right)^{k_0/2} \frac{\Gamma(k_0/2)}{\sigma_i^{k_i}\Gamma(k_i/2)}\right)$$

for $k_0 > 0, \sigma_0 > 0$ *and*

$$D_\alpha(f_i || f_j) = +\infty$$

otherwise.

Proof.

$$\begin{aligned}
f_i^\alpha f_j^{1-\alpha} &= \left[\frac{2^{1-k_i/2} x^{k_i-1} e^{-x^2/2\sigma_i^2}}{\sigma_i^{k_i} \Gamma\left(\frac{k_i}{2}\right)} \right]^\alpha \left[\frac{2^{1-k_j/2} x^{k_j-1} e^{-x^2/2\sigma_j^2}}{\sigma_j^{k_j} \Gamma\left(\frac{k_j}{2}\right)} \right]^{1-\alpha} \\
&= \left[\frac{\sigma_j^{k_j} \Gamma(k_j/2)}{\sigma_i^{k_i} \Gamma(k_i/2)} \right]^{\alpha-1} \frac{2^{1-k_0/2} x^{k_0-1} e^{-x^2/2\xi}}{\sigma_i^{k_i} \Gamma(k_i/2)} ,
\end{aligned}$$

where

$$k_0 = \alpha k_i + (1-\alpha) k_j ,$$

and

$$\frac{1}{\xi} = \frac{\alpha}{\sigma_i^2} + \frac{(1-\alpha)}{\sigma_j^2} = \frac{\alpha \sigma_j^2 + (1-\alpha)\sigma_i^2}{\sigma_i^2 \sigma_j^2} = \frac{\sigma_0}{\sigma_i^2 \sigma_j^2} .$$

- If $k_0 > 0$ and $\sigma_0 > 0$, then

$$\begin{aligned}
&\int_{\mathbb{R}^+} f_i^\alpha f_j^{1-\alpha} \, dx \\
&\quad = \left(\frac{\sigma_j^{k_j} \Gamma(k_j/2)}{\sigma_i^{k_i} \Gamma(k_i/2)} \right)^{\alpha-1} \frac{\xi^{k_0/2} \Gamma(k_0/2)}{\sigma_i^{k_i} \Gamma(k_i/2)} \int_{\mathbb{R}^+} \frac{2^{1-k_0/2} x^{k_0-1} e^{-x^2/2\xi}}{\xi^{k_0/2} \Gamma(k_0/2)} \, dx \\
&\quad = \left(\frac{\sigma_j^{k_j} \Gamma(k_j/2)}{\sigma_i^{k_i} \Gamma(k_i/2)} \right)^{\alpha-1} \frac{\xi^{k_0/2} \Gamma(k_0/2)}{\sigma_i^{k_i} \Gamma(k_i/2)} ,
\end{aligned}$$

since the integrand corresponds to a Chi density ($k = k_0$ and $\sigma^2 = \xi$). Then

$$
\begin{aligned}
D_\alpha(f_i||f_j) &= \frac{1}{\alpha-1}\ln\int_{\mathbb{R}^+} f_i^\alpha f_j^{1-\alpha}\,dx \\
&= \frac{1}{\alpha-1}\ln\left[\left(\frac{\sigma_j^{k_j}\Gamma(k_j/2)}{\sigma_i^{k_i}\Gamma(k_i/2)}\right)^{\alpha-1}\frac{\xi^{k_0/2}\Gamma(k_0/2)}{\sigma_i^{k_i}\Gamma(k_i/2)}\right] \\
&= \ln\left(\frac{\sigma_j^{k_j}\Gamma(k_j/2)}{\sigma_i^{k_i}\Gamma(k_i/2)}\right) + \frac{1}{\alpha-1}\ln\left(\frac{\xi^{k_0/2}\Gamma(k_0/2)}{\sigma_i^{k_i}\Gamma(k_i/2)}\right) \\
&= \ln\left(\frac{\sigma_j^{k_j}\Gamma(k_j/2)}{\sigma_i^{k_i}\Gamma(k_i/2)}\right) + \frac{1}{\alpha-1}\ln\left(\left(\frac{\sigma_i^2\sigma_j^2}{\sigma_0}\right)^{k_0/2}\frac{\Gamma(k_0/2)}{\sigma_i^{k_i}\Gamma(k_i/2)}\right) .
\end{aligned}
$$

Note that for $\alpha \in (0,1)$ we always have $k_0 > 0$ and $\sigma_0 > 0$ given the positivity of k_i, k_j, σ_i^2 and σ_j^2.

- If $\sigma_0 \leq 0$ then $(1/\xi) \leq 0$[2] and

$$
\begin{aligned}
\int_{R^+} f_i^\alpha f_j^{1-\alpha}\,dx &= A\int_{\mathbb{R}^+} x^{k_0-1}e^{Kx^2}\,dx\ , \quad K,A \geq 0 \\
&\geq A\int_{\mathbb{R}^+} x^{k_0-1}\,dx = \infty
\end{aligned}
$$

  for all real values of k_0.

- If $\sigma_0 > 0$ but $k_0 < 0$ then

$$
\begin{aligned}
\int_{\mathbb{R}^+} f_i^\alpha f_j^{1-\alpha}dx &= A_1\int_{\mathbb{R}^+} x^{k_0-1}e^{-\frac{x^2}{2|\xi|}}\,dx\ , \quad A_1 > 0 \\
&= A_2\int_{\mathbb{R}^+} y^{k_0/2-1}e^{-y}dy, \quad A_2 > 0 \\
&> A_2\int_0^1 y^{k_0/2-1}e^{-y}dy\ .
\end{aligned}
$$

[2] see footnote 1

Since $e^{-y} \to 1$ as $y \to 0$, then

$$\int_0^1 y^{k_0/2-1} e^{-y} dy = \infty \ , \ \ \text{since} \ \int_0^1 y^p dy = \infty$$

for $p < -1$. Finally, since nonpositive k_0 and σ_0 only occur for $\alpha > 1$ we have

$$D_\alpha(f_i||f_j) = \frac{1}{\alpha-1} \ln \int_{\mathbb{R}^+} f_i^\alpha f_j^{1-\alpha} \, dx = \infty$$

for these cases.

$\square$

Remark B.2.7. For $k_i = k_j = k$ we have $k_0 = k$ and

$$\begin{aligned} &D_\alpha(f_i(x;k,\sigma_i^2)||f_j(x;k,\sigma_j^2)) \\ &= \ln\left(\frac{\sigma_j^k \Gamma(k/2)}{\sigma_i^k \Gamma(k/2)}\right) + \frac{1}{\alpha-1}\ln\left(\left(\frac{\sigma_i^2\sigma_j^2}{\sigma_0}\right)^{k/2} \frac{\Gamma(k/2)}{\sigma_i^k\Gamma(k/2)}\right) \\ &= k \ln\frac{\sigma_j}{\sigma_i} + \frac{1}{\alpha-1}\ln\left(\frac{\sigma_j^2}{\sigma_0}\right)^{k/2} . \end{aligned}$$

Corollary B.2.8. *Let $\alpha \in \mathbb{R}^+ \setminus \{1\}$ and $\sigma_0 = \alpha\sigma_i^2 + (1-\alpha)\sigma_j^2$. If $\sigma_0 > 0$, then*

1. *If h_i and h_j are two half-normal densities*

$$h_n = \sqrt{\frac{2}{\pi}} \frac{1}{\sigma_n} \, e^{-\frac{x^2}{2\sigma_n^2}} \ , \quad \sigma_n > 0; \ x > 0 \ , \ n = i, j \ ,$$

the Rényi divergence between h_i and h_j is

$$D_\alpha(h_i||h_j) = \ln\frac{\sigma_j}{\sigma_i} + \frac{1}{\alpha-1}\ln\left(\frac{\sigma_j^2}{\sigma_0}\right)^{1/2}$$

2. *If r_i and r_j are two Rayleigh densities*

$$r_n = \frac{x}{\sigma_n^2} e^{-\frac{x^2}{2\sigma_n^2}} \ , \quad \sigma_n > 0; \ x > 0 \ , \ n = i, j \ ,$$

the Rényi divergence between r_i and r_j is

$$D_\alpha(r_i || r_j) = 2 \ln \frac{\sigma_j}{\sigma_i} + \frac{1}{\alpha - 1} \ln \left(\frac{\sigma_j^2}{\sigma_0} \right) \ .$$

3. *If m_i and m_j are two Maxwell-Boltzmann densities*

$$m_n = \sqrt{\frac{2}{\pi}} \frac{x^2 e^{-\frac{x^2}{2\sigma_n^2}}}{\sigma_n^3}, \quad \sigma_n > 0; \ x > 0 \ , \ n = i, j \ ,$$

the Rényi divergence between m_i and m_j is

$$D_\alpha(m_i || m_j) = 3 \ln \frac{\sigma_j}{\sigma_i} + \frac{1}{\alpha - 1} \ln \left(\frac{\sigma_j^2}{\sigma_0} \right)^{3/2} \ .$$

For all cases above, if $\sigma_0 \leq 0$ then $D_\alpha(\cdot||\cdot) = \infty$.

Proof. Just as in Corollary B.2.5, the expressions follow from setting $k = 1, 2, 3$ in the expression given in Remark B.2.7. □

Remark B.2.9.

$$\lim_{\alpha \uparrow 1} D_\alpha(f_i || f_j) = D(f_i || f_j) \ .$$

Proof. Since

$$\lim_{\alpha \to 1} k_0 = \lim_{\alpha \to 1} \alpha k_i + (1 - \alpha) k_j = k_i \ , \quad \text{and}$$
$$\lim_{\alpha \to 1} \sigma_0 = \lim_{\alpha \to 1} \alpha \sigma_j^2 + (1 - \alpha) \sigma_i^2 = \sigma_j^2 \ ,$$

the limit of the second term in the Rényi divergence (Proposition B.2.6) is of indeterminate form. Applying l'Hospital's rule we have

$$
\begin{aligned}
&\lim_{\alpha\uparrow 1}\left[\frac{1}{\alpha-1}\ln\left(\left(\frac{\sigma_i^2\sigma_j^2}{\sigma_0}\right)^{k_0/2}\frac{\Gamma(k_0/2)}{\sigma_i^{k_i}\Gamma(k_i/2)}\right)\right]\\
&=\lim_{\alpha\uparrow 1}\frac{d}{d\alpha}\left[k_0\ln(\sigma_i\sigma_j)+\ln\Gamma(k_0/2)-\frac{k_0}{2}\ln\sigma_0\right]\\
&=\lim_{\alpha\uparrow 1}\left[\frac{dk_0}{d\alpha}\left(\ln(\sigma_i\sigma_j)+\frac{1}{2}\psi(k_0/2)-\frac{1}{2}\ln\sigma_0\right)-\frac{k_0}{2\sigma_0}\frac{d\sigma_0}{d\alpha}\right]\\
&=\lim_{\alpha\uparrow 1}\left[(k_i-k_j)\left(\ln(\sigma_i\sigma_j)+\frac{1}{2}\psi(k_0/2)-\frac{1}{2}\ln\sigma_0\right)-\frac{k_0}{2\sigma_0}(\sigma_j^2-\sigma_i^2)\right]\\
&=(k_i-k_j)\left(\ln(\sigma_i\sigma_j)+\frac{1}{2}\psi(k_i/2)-\frac{1}{2}\ln\sigma_j^2\right)+\frac{k_i}{2\sigma_j^2}(\sigma_i^2-\sigma_j^2)\\
&=\frac{1}{2}\psi(k_i/2)(k_i-k_j)+\ln\sigma_i^{(k_i-k_j)}+\frac{k_i}{2\sigma_j^2}(\sigma_i^2-\sigma_j^2)
\end{aligned}
$$

and so

$$
\begin{aligned}
&\lim_{\alpha\uparrow 1}D_\alpha(f_i\|f_j)\\
&\quad=\ln\left(\frac{\sigma_j^{k_j}\Gamma(k_j/2)}{\sigma_i^{k_i}\Gamma(k_i/2)}\right)+\frac{1}{2}\psi(k_i/2)(k_i-k_j)+\ln\sigma_i^{(k_i-k_j)}+\frac{k_i}{2\sigma_j^2}(\sigma_i^2-\sigma_j^2)\\
&\quad=\frac{1}{2}\psi(k_i/2)(k_i-k_j)+\ln\left[\left(\frac{\sigma_j}{\sigma_i}\right)^{k_j}\frac{\Gamma(k_j/2)}{\Gamma(k_i/2)}\right]+\frac{k_i}{2\sigma_j^2}(\sigma_i^2-\sigma_j^2)\,.
\end{aligned}
$$

which was the expression obtained in Proposition B.2.3, as expected. □

B.3 Beta and Dirichlet Distributions

B.3.1 Beta distributions

Throughout this section let f_i and f_j be two Beta densities

$$f_i(x)=\frac{x^{a_i-1}(1-x)^{b_i-1}}{B(a_i,b_i)}\ ,\quad a_i,b_i>0;x\in(0,1)\ ,$$

where $B(x,y)$ is the Beta function introduced in Section A.3.3.

Proposition B.3.1.

$$\begin{aligned}E_{f_i}\left[\ln f_j\right] &= -\ln B(a_j, b_j) + (a_j - 1)\psi(a_i) + (b_j - 1)\psi(b_i) \\ &\quad + (2 - a_j - b_j)\psi(a_i + b_i)\ .\end{aligned}$$

Proof.

$$\begin{aligned}E_{f_i}\left[\ln f_j\right] &= E_{f_i}\left[-\ln B(a_j, b_j) + (a_j - 1)\ln X + (b_j - 1)\ln(1 - X)\right] \\ &= -\ln B(a_j, b_j) + (a_j - 1)E_{f_i}\left[\ln X\right] + (b_j - 1)E_{f_i}\left[\ln(1 - X)\right]\ .\end{aligned}$$

Let $r > -a_i$. Then

$$\begin{aligned}E_{f_i}\left[X^r\right] &= \int_0^1 x^r \frac{x^{a_i-1}(1-x)^{b_i-1}}{B(a_i, b_i)}\,dx \\ &= \int_0^1 \frac{x^{a_i+r-1}(1-x)^{b_i-1}}{B(a_i, b_i)}\,dx \\ &= \frac{B(a_i + r, b_i)}{B(a_i, b_i)} \int_0^1 \frac{x^{a_i+r-1}(1-x)^{b_i-1}}{B(a_i + r, b_i)} \\ &= \frac{B(a_i + r, b_i)}{B(a_i, b_i)}\ ,\end{aligned}$$

since the last integrand corresponds to a reparametrized Beta distribution with $a_i \mapsto a_i + r > 0$. Then

$$\begin{aligned}E_{f_i}\left[\ln X\right] &= \frac{d}{dr} E_{f_i}\left[X^r\right]\bigg|_{r=0} \\ &= \frac{d}{dr}\frac{B(a_i + r, b_i)}{B(a_i, b_i)}\bigg|_{r=0} \\ &= \frac{B(a_i + r, b_i)\left[\psi(a_i + r) - \psi(a_i + b_i + r)\right]}{B(a_i, b_i)}\bigg|_{r=0} \\ &= \psi(a_i) - \psi(a_i + b_i)\ ,\end{aligned}$$

where we have used the expression given in Remark A.3.11 for the partial derivatives of the Beta function. Similarly,

$$E_{f_i}\left[(1-X)^r\right] = \frac{B(a_i, b_i + r)}{B(a_i, b_i)} \ ,$$

and

$$E_{f_i}\left[\ln(1-X)\right] = \frac{d}{dr} E_{f_i}\left[(1-X)^r\right]\Big|_{r=0} = \psi(b_i) - \psi(a_i + b_i) \ .$$

Finally,

$$\begin{aligned} E_{f_i}\left[\ln f_j\right] &= -\ln B(a_j, b_j) + (a_j - 1)E_{f_i}\left[\ln X\right] + (b_j - 1)E_{f_i}\left[\ln(1-X)\right] \\ &= -\ln B(a_j, b_j) + (a_j - 1)\left[\psi(a_i) - \psi(a_i + b_i)\right] \\ &\quad + (b_j - 1)\left[\psi(b_i) - \psi(a_i + b_i)\right] \\ &= -\ln B(a_j, b_j) + (a_j - 1)\psi(a_i) + (b_j - 1)\psi(b_i) \\ &\quad + (2 - a_j - b_j)\psi(a_i + b_i) \ . \end{aligned}$$

□

Corollary B.3.2. *The differential entropy of f_i is*

$$\begin{aligned} h(f_i) = \ln B(a_i, b_i) &+ (1 - a_i)\psi(a_i) + (1 - b_i)\psi(b_i) \\ &+ (a_i + b_i - 2)\psi(a_i + b_i) \ . \end{aligned}$$

Proof. Setting $i = j$ in Proposition B.3.1 we have

$$\begin{aligned} h(f_i) = -E_{f_i}\left[\ln f_i\right] = -\big[&-\ln B(a_i, b_i) + (a_i - 1)\psi(a_i) + (b_i - 1)\psi(b_i) \\ &+ (2 - a_i - b_i)\psi(a_i + b_i) \ .\big] \\ = \ln B&(a_i, b_i) + (1 - a_i)\psi(a_i) + (1 - b_i)\psi(b_i) \\ &+ (a_i + b_i - 2)\psi(a_i + b_i) \ . \end{aligned}$$

□

Proposition B.3.3. *The Kullback-Liebler divergence between f_i and f_j is*

$$D(f_i||f_j) = \ln \frac{B(a_j, b_j)}{B(a_i, b_i)} + \psi(a_i)(a_i - a_j) + \psi(b_i)(b_i - b_j)$$
$$+ \psi(a_i + b_i)(a_j + b_j - (a_i + b_i)) \, .$$

Proof. Using Proposition B.3.1 and Remark 1.2.4 we have

$$\begin{aligned}
&D(f_i||f_j)\\
&\quad = E_{f_i}[\ln f_i] - E_{f_i}[\ln f_j]\\
&\quad = -\left[\ln B(a_i, b_i) + (1 - a_i)\psi(a_i) + (1 - b_i)\psi(b_i)\right.\\
&\qquad \left. + (a_i + b_i - 2)\psi(a_i + b_i)\right]\\
&\quad - \left[-\ln B(a_j, b_j) + (a_j - 1)\psi(a_i) + (b_j - 1)\psi(b_i)\right.\\
&\qquad \left. +(2 - a_j - b_j)\psi(a_i + b_i) \, .\right]\\
&\quad = \ln B(a_j, b_j) - \ln B(a_i, b_i) + \psi(a_i)(a_i - 1 + 1 - a_j) + \psi(b_i)(b_i - 1 + 1 - b_j)\\
&\qquad \psi(a_i + b_i)(2 - a_i - b_i + a_j + b_j - 2)\\
&\quad = \ln \frac{B(a_j, b_j)}{B(a_i, b_i)} + \psi(a_i)(a_i - a_j) + \psi(b_i)(b_i - b_j)\\
&\qquad + \psi(a_i + b_i)(a_j + b_j - (a_i + b_i)) \, .
\end{aligned}$$

□

Proposition B.3.4. *For $\alpha \in \mathbb{R}^+ \setminus \{1\}$ let $a_0 = \alpha a_i + (1 - \alpha)a_j$ and $b_0 = \alpha b_i + (1 - \alpha)b_j$. Then the Rényi divergence between f_i and f_j is given by*

$$D_\alpha(f_i||f_j) = \ln \frac{B(a_j, b_j)}{B(a_i, b_i)} + \frac{1}{\alpha - 1} \ln \frac{B(a_0, b_0)}{B(a_i, b_i)}$$

for $a_0, b_0 \geq 0$, and

$$D_\alpha(f_i||f_j) = +\infty$$

otherwise.

Proof.

$$
\begin{aligned}
f_i^{\alpha} f_j^{1-\alpha} &= \left[\frac{x^{a_i-1}(1-x)^{b_i-1}}{B(a_i,b_i)}\right]^{\alpha}\left[\frac{x^{a_j-1}(1-x)^{b_j-1}}{B(a_j,b_j)}\right]^{1-\alpha} \\
&= \left[B(a_i,b_i)\right]^{-\alpha}\left[B(a_j,b_j)\right]^{\alpha-1} x^{a_0-1}(1-x)^{b_0-1} \\
&= \left(\frac{B(a_j,b_j)}{B(a_i,b_i)}\right)^{\alpha-1}\frac{1}{B(a_i,b_i)} x^{a_0-1}(1-x)^{b_0-1} \ ,
\end{aligned}
$$

where

$$
a_0 = \alpha a_i + (1-\alpha) a_j \ , \text{ and } \ b_0 = \alpha b_i + (1-\alpha) b_j \ .
$$

- If $a_0, b_0 > 0$, then $B(a_0, b_0)$ is defined and

$$
\begin{aligned}
\int_0^1 f_i^{\alpha} f_j^{1-\alpha}\, dx &= \left(\frac{B(a_j,b_j)}{B(a_i,b_i)}\right)^{\alpha-1}\frac{1}{B(a_i,b_i)}\int_0^1 x^{a_0-1}(1-x)^{b_0-1}\, dx \\
&= \left(\frac{B(a_j,b_j)}{B(a_i,b_i)}\right)^{\alpha-1}\frac{B(a_0,b_0)}{B(a_i,b_i)}\int_0^1 \frac{x^{a_0-1}(1-x)^{b_0-1}}{B(a_0,b_0)}\, dx \\
&= \left(\frac{B(a_j,b_j)}{B(a_i,b_i)}\right)^{\alpha-1}\frac{B(a_0,b_0)}{B(a_i,b_i)} \ ,
\end{aligned}
$$

since the integrand is a Beta distribution with parameters $a_0 > 0$ and $b_0 > 0$. Then,

$$
\begin{aligned}
D_{\alpha}(f_i||f_j) &= \frac{1}{\alpha-1}\ln\int_0^1 f_i^{\alpha} f_j^{1-\alpha}\, dx \\
&= \frac{1}{\alpha-1}\ln\left[\left(\frac{B(a_j,b_j)}{B(a_i,b_i)}\right)^{\alpha-1}\frac{B(a_0,b_0)}{B(a_i,b_i)}\right] \\
&= \ln\frac{B(a_j,b_j)}{B(a_i,b_i)} + \frac{1}{\alpha-1}\ln\frac{B(a_0,b_0)}{B(a_i,b_i)} \ .
\end{aligned}
$$

Note that for $\alpha \in (0,1)$ we always have $a_0 > 0$ and $b_0 > 0$ given the positivity of a_i, b_i, a_j and b_j.

- If $a_0 \leq 0$ or $b_0 \leq 0$ then

$$\int_0^1 x^{a_0-1}(1-x)^{b_0-1}\,dx = \infty$$

as pointed out in Remark A.3.8. Since this can only happen for $\alpha > 1$, then

$$D_\alpha(f_i||f_j) = \frac{1}{\alpha-1}\ln\int_0^1 f_i^{\alpha} f_j^{1-\alpha}\,dx \to +\infty\ .$$

□

Remark B.3.5.

$$\lim_{\alpha\uparrow 1} D_\alpha(f_i||f_j) = D(f_i||f_j)\ .$$

Proof. Comparing the expressions for the Rényi and Kullback divergence (Proposition B.3.4 and Proposition B.3.3, respectively), we need to show that

$$\lim_{\alpha\uparrow 1}\frac{1}{\alpha-1}\ln\frac{B(a_0,b_0)}{B(a_i,b_i)}$$
$$= \psi(a_i)(a_i-a_j) + \psi(b_i)(b_i-b_j) + \psi(a_i+b_i)(a_j+b_j-(a_i+b_i))\ .$$

Note that

$$\lim_{\alpha\to 1} a_0 = \lim_{\alpha\to 1}\alpha a_i + (1-\alpha)a_j = a_i\ ,\ \lim_{\alpha\to 1} b_0 = \lim_{\alpha\to 1}\alpha b_i + (1-\alpha)b_j = b_i\ ,$$

so the limit in question is of indeterminate form. Applying l'Hospital's rule we have

$$\begin{aligned}
&\lim_{\alpha\uparrow 1} \frac{1}{\alpha-1} \ln \frac{B(a_0,b_0)}{B(a_i,b_i)} \\
&= \lim_{\alpha\uparrow 1} \frac{1}{B(a_0,b_0)} \frac{d}{d\alpha} B(a_0,b_0) \\
&= \lim_{\alpha\uparrow 1} \frac{1}{B(a_0,b_0)} \left[\frac{\partial}{\partial a_0} B(a_0,b_0) \frac{da_0}{d\alpha} + \frac{\partial}{\partial b_0} B(a_0,b_0) \frac{db_0}{d\alpha} \right] \\
&= \lim_{\alpha\uparrow 1} \frac{1}{B(a_0,b_0)} \Big\{ B(a_0,b_0) \left[\psi(a_0) - \psi(a_0+b_0) \right] (a_i - a_j) \\
&\quad + B(a_0,b_0) \left[\psi(b_0) - \psi(a_0+b_0) \right] (b_i - b_j) \Big\} \; . \\
&= \psi(a_i+b_i)(a_j+b_j-(a_i+b_i)) + (a_i-a_j)\psi(a_i) \\
&\quad + (b_i-b_j)\psi(b_i) \; ,
\end{aligned}$$

where we have used Proposition B.3.4 to express the partial derivatives of $B(x,y)$ in terms of the Digamma function $\psi(\ldots)$. $\square$

B.3.2 Dirichlet Distributions

The expression for Rényi divergence given in Proposition B.3.4 can be readily generalized to Dirichlet distributions by the same reparametrization argument. The corresponding expression for the finite case was also derived in in [52] in the form of the Chernoff distance of order $\lambda \in (0,1)$.

Let f_i and f_j be two Dirichlet densities of order d:

$$f_i(\boldsymbol{x},\boldsymbol{a_i}) = \frac{1}{B(\boldsymbol{a_i})} \prod_{k=1}^{d} x_k^{a_{i_k}-1} \quad ; \; \boldsymbol{a_i} \in \mathbb{R}^d \; , \; ; \; \boldsymbol{x} \in \mathbb{R}^d \; , d \geq 2, \; ,$$

where $\boldsymbol{x} = (x_1,\ldots,x_d)$ satisfies $\sum_{k=1}^{d} x_k = 1$, $\boldsymbol{a_i} = (a_{i_1},\ldots,a_{i_d})$, $a_k > 0$, and $B(\boldsymbol{y})$ is the beta function of vector argument defined in Definition A.3.10.

Proposition B.3.6. *For $\alpha \in \mathbb{R}^+ \setminus \{1\}$ let $\boldsymbol{a}_0 = \alpha \boldsymbol{a}_i + (1-\alpha)\boldsymbol{a}_j$. Then the Rényi divergence between f_i and f_j is given by*

$$D_\alpha(f_i||f_j) = \ln \frac{B(\boldsymbol{a}_j)}{B(\boldsymbol{a}_i)} + \frac{1}{\alpha - 1} \ln \left(\frac{B(\alpha \boldsymbol{a}_i + (1-\alpha)\boldsymbol{a}_j)}{B(\boldsymbol{a}_i)} \right)$$

if $\forall k$, a_{0_k}, and

$$D_\alpha(f_i||f_j) = +\infty$$

otherwise.

Proof. The proof for the finite case follows the same reparametrization argument as in Proposition B.3.4, so we omit it here. If some a_i fails to be positive, then

$$\int_0^1 x_i^{a_i - 1}\, dx_i = \infty \ ,$$

and since $f_k = x^{a_k - 1}$ are positive functions on $[0,1]$, Fubini's theorem applies (see p. 164 in [55]) and we have

$$\int_{[0,1]^d} \prod_{k=1}^{d} x^{a_k - 1} d\boldsymbol{x} = \int_{[0,1]} x^{a_i - 1} dx_i \cdot \int_{[0,1]^{d-1}} \prod_{k \neq i}^{d} x^{a_k - 1} d\boldsymbol{x} = \infty \ .$$

□

The Chernoff distance between f_i and f_j is given in [52] as

$$D_C(f_i||f_j;\alpha) = -\ln \left(\frac{B(\alpha \boldsymbol{a}_i + (1-\alpha)\boldsymbol{a}_j)}{[B(\boldsymbol{a}_i)]^\alpha [B(\boldsymbol{a}_j)]^{1-\alpha}} \right) \ .$$

We mentioned in Section 1.1 that the Rényi divergence and the Chernoff distance are related via

$$D_\alpha(f_i||f_j) = -\frac{1}{(\alpha - 1)} D_C(f_i||f_j;\alpha) \ .$$

Thus we can verify the consistency between the two expressions above by noting that

$$
\begin{aligned}
-\frac{1}{(\alpha-1)} J_C(f_i, f_j) &= \frac{1}{\alpha-1}\left[\ln\left(\frac{B(\alpha \boldsymbol{a_i} + (1-\alpha)\boldsymbol{a_j})}{[B(\boldsymbol{a_i})]^{\alpha}[B(\boldsymbol{a_j})]^{1-\alpha}}\right)\right] \\
&= \frac{1}{\alpha-1}\ln\left[\frac{B(\alpha \boldsymbol{a_i} + (1-\alpha)\boldsymbol{a_j})}{B(\boldsymbol{a_i})}\left(\frac{[B(\boldsymbol{a_j})]}{[B(\boldsymbol{a_i})]}\right)^{\alpha-1}\right] \\
&= \ln\frac{B(\boldsymbol{a_j})}{B(\boldsymbol{a_i})} + \frac{1}{\alpha-1}\ln\left(\frac{B(\alpha \boldsymbol{a_i} + (1-\alpha)\boldsymbol{a_j})}{B(\boldsymbol{a_i})}\right) .
\end{aligned}
$$

Also, the case $\alpha \in (0,1)$ (assumed for the Chernoff distance expression) is a subset of the finiteness constraint given in terms of $\boldsymbol{a_0}, \boldsymbol{b_0}$ above.

B.4 Gaussian Distributions

B.4.1 Univariate Gaussian Distributions

Throughout this section let f_i and f_j be two univariate normal densities

$$
f_i(x) = \left(\frac{1}{2\pi\sigma_i^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_i^2}(x-\mu_i)^2\right) , \quad \sigma_i > 0 , \mu_i \in \mathbb{R} ; x \in \mathbb{R} .
$$

Proposition B.4.1.

$$
E_{f_i}[\ln f_j] = -\frac{1}{2}\ln\left(2\pi\sigma_j^2\right) - \frac{1}{2\sigma_j^2}\left[\sigma_i^2 + (\mu_i-\mu_j)^2\right] .
$$

Proof.

$$
\begin{aligned}
E_{f_i}[\ln f_j] &= E_{f_i}\left[-\frac{1}{2}\ln\left(2\pi\sigma_j^2\right) - \frac{1}{2\sigma_j^2}(X-\mu_j)^2\right] \\
&= -\frac{1}{2}\ln\left(2\pi\sigma_j^2\right) - \frac{1}{2\sigma_j^2}E_{f_i}[(X-\mu_j)^2] ,
\end{aligned}
$$

where

$$
\begin{aligned}
E_{f_i}[(X-\mu_j)^2] &= E_{f_i}\left[\left((X-\mu_i)+(\mu_i-\mu_j)\right)^2\right] \\
&= E_{f_i}[(X-\mu_i)^2] + 2(\mu_i-\mu_j)E_{f_i}[X-\mu_i] + (\mu_i-\mu_j)^2 \\
&= \sigma_i^2 + (\mu_i-\mu_j)^2 \ .
\end{aligned}
$$

Thus

$$
E_{f_i}[\ln f_j] = -\frac{1}{2}\ln\left(2\pi\sigma_j^2\right) - \frac{1}{2\sigma_j^2}\left[\sigma_i^2+(\mu_i-\mu_j)^2\right] \ .
$$

□

Corollary B.4.2. *The differential entropy of f_i is*

$$
h(f_i) = \frac{1}{2}\ln\left(2\pi e\sigma_i^2\right) \ .
$$

Proof. Setting $i=j$ in Proposition B.4.1 we have

$$
\begin{aligned}
h(f_i) = -E_{f_i}[\ln f_i] &= -\left(-\frac{1}{2}\ln\left(2\pi\sigma_i^2\right) - \frac{1}{2\sigma_i^2}\left[\sigma_i^2+(\mu_i-\mu_i)^2\right]\right) \\
&= \frac{1}{2}\ln\left(2\pi\sigma_i^2\right) + \frac{1}{2} \\
&= \frac{1}{2}\ln\left(2\pi e\sigma_i^2\right) \ .
\end{aligned}
$$

□

Proposition B.4.3. *The Kullback-Liebler divergence between f_i and f_j is*

$$
D(f_i||f_j) = \frac{1}{2\sigma_j^2}\left[(\mu_i-\mu_j)^2+\sigma_i^2-\sigma_j^2\right] + \ln\frac{\sigma_j}{\sigma_i}.
$$

Proof. Using Proposition B.4.1 and Remark 1.2.4 we have

$$\begin{aligned}
D(f_i||f_j) &= E_{f_i}[\ln f_i] - E_{f_i}[\ln f_j] \\
&= -\frac{1}{2}\ln\left(2\pi e\sigma_i^2\right) + \frac{1}{2}\ln\left(2\pi\sigma_j^2\right) + \frac{1}{2\sigma_j^2}\left[\sigma_i^2 + (\mu_i - \mu_j)^2\right] \\
&= \frac{1}{2}\ln\frac{\sigma_j^2}{\sigma_i^2} - \frac{1}{2}\ln e + \frac{1}{2\sigma_j^2}\left[\sigma_i^2 + (\mu_i - \mu_j)^2\right] \\
&= \ln\frac{\sigma_j}{\sigma_i} - \frac{\sigma_j^2}{2\sigma_j^2} + \frac{1}{2\sigma_j^2}\left[\sigma_i^2 + (\mu_i - \mu_j)^2\right] \\
&= \frac{1}{2\sigma_j^2}\left[(\mu_i - \mu_j)^2 + \sigma_i^2 - \sigma_j^2\right] + \ln\frac{\sigma_j}{\sigma_i}\,.
\end{aligned}$$

□

Proposition B.4.4. *For $\alpha \in \mathbb{R}^+ \setminus \{1\}$ let $\sigma_0 = \alpha\sigma_j^2 + (1-\alpha)\sigma_i^2$. Then the Rényi divergence between f_i and f_j is given by*

$$D_\alpha(f_i||f_j) = \ln\frac{\sigma_j}{\sigma_i} + \frac{1}{2(\alpha-1)}\ln\left(\frac{\sigma_j^2}{\sigma_0}\right) + \frac{1}{2}\frac{\alpha(\mu_i-\mu_j)^2}{\sigma_0}$$

if $\sigma_0 > 0$, and

$$D_\alpha(f_i||f_j) = +\infty$$

otherwise.

Proof.

$$
\begin{aligned}
f_i^{\alpha} f_j^{1-\alpha} &= \left[\left(\frac{1}{2\pi\sigma_i^2} \right)^{1/2} \exp\left(-\frac{1}{2\sigma_i^2}(x-\mu_i)^2 \right) \right]^{\alpha} \\
&\cdot \left[\left(\frac{1}{2\pi\sigma_j^2} \right)^{1/2} \exp\left(-\frac{1}{2\sigma_j^2}(x-\mu_j)^2 \right) \right]^{1-\alpha} \\
&= \left(\frac{1}{2\pi\sigma_i^2} \right)^{(\alpha-1)/2} \left(\frac{1}{2\pi\sigma_i^2} \right)^{1/2} \left(\frac{1}{2\pi\sigma_j^2} \right)^{(1-\alpha)/2} \\
&\cdot \exp\left(-\frac{1}{2} \left[\frac{\alpha}{\sigma_i^2}(x-\mu_i)^2 + \frac{(1-\alpha)}{\sigma_j^2}(x-\mu_j)^2 \right] \right) \\
&= \left(\frac{\sigma_j^2}{\sigma_i^2} \right)^{\frac{\alpha-1}{2}} \left(\frac{1}{2\pi\sigma_i^2} \right)^{1/2} \\
&\cdot \exp\left(-\frac{1}{2\sigma_i^2\sigma_j^2} \left[\alpha\sigma_j^2(x-\mu_i)^2 + (1-\alpha)\sigma_i^2(x-\mu_j)^2 \right] \right) .
\end{aligned}
$$

Consider the argument of the exponential above. Note that

$$
\left[\alpha\sigma_j^2(x-\mu_i)^2 + (1-\alpha)\sigma_i^2(x-\mu_j)^2 \right] = ax^2 - 2bx + c
$$

where

$$
a = \sigma_0 := \alpha\sigma_j^2 + (1-\alpha)\sigma_i^2 \ , \quad b = \alpha\sigma_j^2\mu_i + (1-\alpha)\sigma_i^2\mu_j \ ,
$$

and

$$
c = \alpha\sigma_j^2\mu_i^2 + (1-\alpha)\sigma_i^2\mu_j^2 \ .
$$

- If $a = 0$ then

$$
\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2 = 0 \Leftrightarrow \alpha = \frac{\sigma_i^2}{\sigma_i^2 - \sigma_j^2} \ .
$$

Note that this case is only possible if $\sigma_i^2 > \sigma_j^2$ since by assumption $\alpha > 0$. Furthermore, this also implies that $\alpha > 1$; which can also be seen by noting that for $\alpha \in (0,1)$, $\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2$ is the convex combination of two positive numbers, which is clearly positive (and by assumption $\alpha \neq 1$). So we have

$$f_i^\alpha f_j^{1-\alpha} = \left(\frac{\sigma_j^2}{\sigma_i^2}\right)^{\frac{\alpha-1}{2}} \left(\frac{1}{2\pi\sigma_i^2}\right)^{1/2} \exp\left(-2bx + c\right) ,$$

and the integral

$$\int_{\mathbb{R}} f_i^\alpha f_j^{1-\alpha}\, dx$$

is of the form

$$K \int_{\mathbb{R}} e^{sy}\, dy\ , \quad K > 0$$

which equals $+\infty$ for all real values of s. Hence

$$D_\alpha(f_i||f_j) = \frac{1}{\alpha - 1} \ln \int_{\mathbb{R}} f_i^\alpha f_j^{\alpha-1}\, dx = +\infty ,$$

since $\alpha > 1$ for this case, as noted above.

- If $a \neq 0$, we can write

$$\begin{aligned} ax^2 - 2bx + c &= a\left[x^2 - \frac{2b}{a}x + \left(\frac{b}{a}\right)^2 - \left(\frac{b}{a}\right)^2\right] + c \\ &= a\left(x - \frac{b}{a}\right)^2 + c - \frac{b^2}{a} , \end{aligned}$$

so we can express the exponential above as

$$\begin{aligned} &\exp\left(-\frac{1}{2\sigma_i^2\sigma_j^2}\left[a\left(x - \frac{b}{a}\right)^2 + c - \frac{b^2}{a}\right]\right) \\ &\quad = \exp\left(-\frac{a}{2\sigma_i^2\sigma_j^2}\left(x - \frac{b}{a}\right)^2\right) \cdot \exp\left(-\frac{1}{2\sigma_i^2\sigma_j^2}\left(c - \frac{b^2}{a}\right)\right) . \end{aligned}$$

Thus

$$\int_{\mathbb{R}} f_i^{\alpha} f_j^{1-\alpha}\, dx = \left(\frac{\sigma_j^2}{\sigma_i^2}\right)^{\frac{\alpha-1}{2}} \left(\frac{1}{2\pi\sigma_i^2}\right)^{1/2} \exp\left(-\frac{1}{2\sigma_i^2\sigma_j^2}\left(c - \frac{b^2}{a}\right)\right) \cdot \int_{\mathbb{R}} \exp\left(-\frac{a}{2\sigma_i^2\sigma_j^2}\left(x - \frac{b}{a}\right)^2\right) dx\ .$$

- $a < 0$: In this case the integral above is of the form

  $$K \int_{\mathbb{R}} e^{sy^2}\, dy\ , \quad s, K > 0\ ,$$

  which equals $+\infty$, so that

  $$D_\alpha(f_i || f_j) = \frac{1}{\alpha - 1} \ln \int_{\mathbb{R}} f_i^{\alpha} f_j^{1-\alpha}\, dx \to +\infty\ ;$$

  since, as before, this case must be a subset of the cases $\alpha > 1$.

- $a > 0$:

  $$\begin{aligned} &\int_{\mathbb{R}} f_i^{\alpha} f_j^{1-\alpha}\, dx \\ &= \left(\frac{\sigma_j^2}{\sigma_i^2}\right)^{\frac{\alpha-1}{2}} \left(\frac{1}{\sigma_i^2}\right)^{1/2} \left(\frac{\sigma_i^2\sigma_j^2}{a}\right)^{1/2} \exp\left(-\frac{1}{2\sigma_i^2\sigma_j^2}\left(c - \frac{b^2}{a}\right)\right) \\ &\quad \cdot \int_{\mathbb{R}} \left(\frac{a}{2\pi\sigma_i^2\sigma_j^2}\right)^{1/2} \exp\left(-\frac{a}{2\sigma_i^2\sigma_j^2}\left(x - \frac{b}{a}\right)^2\right) dx \\ &= \left(\frac{\sigma_j^2}{\sigma_i^2}\right)^{\frac{\alpha-1}{2}} \left(\frac{1}{\sigma_i^2}\right)^{1/2} \left(\frac{\sigma_i^2\sigma_j^2}{a}\right)^{1/2} \exp\left(-\frac{1}{2\sigma_i^2\sigma_j^2}\left(c - \frac{b^2}{a}\right)\right)\ , \end{aligned}$$

  since in this case the integrand is proportional to the pdf of a Gaussian distribution with mean b/a and variance $\sigma_i^2\sigma_j^2/a$. Note that

$$\begin{aligned}
ac - b^2 &= \left(\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2\right)\left(\alpha\sigma_j^2\mu_i^2 + (1-\alpha)\sigma_i^2\mu_j^2\right) \\
&\quad - \left(\alpha\sigma_j^2\mu_i + (1-\alpha)\sigma_i^2\mu_j\right)^2 \\
&= \alpha^2\sigma_j^4\mu_i^2 + (1-a)^2\sigma_i^4\mu_j^2 + \alpha(1-\alpha)\sigma_i^2\sigma_j^2\left(\mu_j^2 + \mu_i^2\right) \\
&\quad - \left(\alpha\sigma_j^2\mu_i + (1-\alpha)\sigma_i^2\mu_j\right)^2 \\
&= \alpha(1-\alpha)\sigma_i^2\sigma_j^2\left(\mu_j^2 + \mu_i^2\right) - 2\alpha(1-\alpha)\sigma_i^2\sigma_j^2\mu_i\mu_j \\
&= \alpha(1-\alpha)\sigma_i^2\sigma_j^2(\mu_i - \mu_j)^2 \ ,
\end{aligned}$$

hence

$$-\frac{1}{2\sigma_i^2\sigma_j^2}\left(c - \frac{b^2}{a}\right) = -\frac{1}{2\sigma_i^2\sigma_j^2}\left(\frac{ac-b^2}{a}\right) = -\frac{1}{2}\frac{\alpha(1-\alpha)(\mu_i-\mu_j)^2}{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2} \ ,$$

$$\begin{aligned}
&\int_{\mathbb{R}} f_i^\alpha f_j^{1-\alpha}\,dx \\
&\quad = \left(\frac{\sigma_j}{\sigma_i}\right)^{\alpha-1}\left(\frac{\sigma_j^2}{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2}\right)^{1/2}\exp\left(-\frac{1}{2}\frac{\alpha(1-\alpha)(\mu_i-\mu_j)^2}{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2}\right) \ ,
\end{aligned}$$

and

$$\begin{aligned}
D_\alpha(f_i||f_j) = \ln\frac{\sigma_j}{\sigma_i} + \frac{1}{2(\alpha-1)}\ln\left(\frac{\sigma_j^2}{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2}\right) \\
+ \frac{1}{2}\frac{\alpha(\mu_i-\mu_j)^2}{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2} \ .
\end{aligned}$$

□

Remark B.4.5.

$$\lim_{\alpha \uparrow 1} D_\alpha(f_i || f_j) = D(f_i || f_j) \; .$$

Proof. Note that the limit

$$\lim_{\alpha \uparrow 1} \frac{1}{2(\alpha - 1)} \ln \left(\frac{\sigma_j^2}{\alpha \sigma_j^2 + (1-\alpha)\sigma_i^2} \right) \; ,$$

has an indeterminate form. Applying l'Hospital's rule we find

$$\begin{aligned} &\lim_{\alpha \uparrow 1} \frac{1}{2(\alpha - 1)} \ln \left(\frac{\sigma_j^2}{\alpha \sigma_j^2 + (1-\alpha)\sigma_i^2} \right) \\ &\quad = \lim_{\alpha \uparrow 1} \; -\frac{1}{2} \frac{\sigma_j^2 - \sigma_i^2}{\alpha \sigma_j^2 + (1-\alpha)\sigma_i^2} = \frac{1}{2} \frac{\sigma_i^2 - \sigma_j^2}{\sigma_j^2} \; , \end{aligned}$$

and so

$$\begin{aligned} \lim_{\alpha \uparrow 1} \; D_\alpha(f_i || f_j) &= \ln \frac{\sigma_j}{\sigma_i} + \frac{1}{2} \frac{\sigma_i^2 - \sigma_j^2}{\sigma_j^2} + \frac{1}{2} \frac{(\mu_i - \mu_j)^2}{\sigma_j^2} \\ &= \frac{1}{2\sigma_j^2} \left[(\mu_i - \mu_j)^2 + \sigma_i^2 - \sigma_j^2 \right] + \ln \frac{\sigma_j}{\sigma_i} \; , \end{aligned}$$

as given by Proposition B.4.3. □

B.4.2 Multivariate Gaussian Distributions

Throughout this section let f_i and f_j be two multivariate normal densities over $\mathbb{R}^n$:

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)} \; , \quad \boldsymbol{x} \in \mathbb{R}^n \; ,$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^n$, Σ_i is a symmetric positive-definite matrix, and $(\cdot)'$ denotes transposition.

Proposition B.4.6.

$$E_{f_i} \left[\ln f_j \right] = -\frac{1}{2} \left[\ln \left((2\pi)^n |\Sigma_j| \right) + \operatorname{tr} \left(\Sigma_j^{-1} \Sigma_i \right) + \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j \right)' \Sigma_j^{-1} \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j \right) \right] \; .$$

Proof.

$$\begin{aligned}E_{f_i}\left[\ln f_j\right] &= E_{f_i}\left[-\ln\left((2\pi)^{n/2}|\Sigma_j|^{1/2}\right) - \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_j)\right]\\ &= -\frac{1}{2}\ln\left((2\pi)^n|\Sigma_j|\right) - \frac{1}{2}E_{f_i}\left[(\boldsymbol{X}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_j)\right] \ ,\end{aligned}$$

where

$$E_{f_i}\left[(\boldsymbol{X}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_j)\right] = \sum_{k,l=1}^{n}\Sigma_{j_{kl}}^{-1}E_{f_i}\left[(X_k-\mu_{j_k})(X_l-\mu_{j_l})\right] \ ,$$

and

$$\begin{aligned}&E_{f_i}\left[(X_k-\mu_{j_k})(X_l-\mu_{j_l})\right]\\ &\quad= E_{f_i}\left[\left(X_k-\mu_{i_k}+\mu_{i_k}-\mu_{j_k}\right)\left(X_l-\mu_{i_l}+\mu_{i_l}-\mu_{j_l}\right)\right]\\ &\quad= E_{f_i}\left[\left(X_k-\mu_{i_k}\right)\left(X_l-\mu_{i_l}\right)\right] + E_{f_i}\left[\left(\mu_{i_k}-\mu_{j_k}\right)\left(\mu_{i_l}-\mu_{j_l}\right)\right]\\ &\qquad+ E_{f_i}\left[\left(X_k-\mu_{i_k}\right)\left(\mu_{i_l}-\mu_{j_l}\right)\right] + E_{f_i}\left[\left(X_l-\mu_{i_l}\right)\left(\mu_{i_l}-\mu_{j_l}\right)\right]\\ &\quad= \left(\Sigma_i\right)_{kl} + \left(\mu_{i_k}-\mu_{j_k}\right)\left(\mu_{i_l}-\mu_{j_l}\right) \ ,\end{aligned}$$

since the last two expectations above are 0. Also, since $\left(\Sigma_i\right)_{kl} = \left(\Sigma_i\right)_{lk}$, then

$$\begin{aligned}&\sum_{k,l=1}^{n}\left(\Sigma_j^{-1}\right)_{kl}\left[\left(\Sigma_i\right)_{kl} + \left(\mu_{i_k}-\mu_{j_k}\right)\left(\mu_{i_l}-\mu_{j_l}\right)\right]\\ &\qquad= \sum_{k,l=1}^{n}\left(\Sigma_j^{-1}\right)_{kl}\left(\Sigma_i\right)_{lk} + \sum_{k,l=1}^{n}\left(\Sigma_j^{-1}\right)_{kl}\left(\mu_{i_k}-\mu_{j_k}\right)\left(\mu_{i_l}-\mu_{j_l}\right)\\ &\qquad= \mathrm{tr}\left(\Sigma_j^{-1}\Sigma_i\right) + \left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)'\Sigma_j^{-1}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right) \ .\end{aligned}$$

Thus,

$$E_{f_i}\left[\ln f_j\right] = -\frac{1}{2}\left[\ln\left((2\pi)^n|\Sigma_j|\right) + \mathrm{tr}\left(\Sigma_j^{-1}\Sigma_i\right) + \left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)'\Sigma_j^{-1}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)\right] \ .$$

□

Corollary B.4.7. *The differential entropy of f_i is*

$$h(f_i) = \frac{1}{2}\ln\left((2\pi e)^n|\Sigma_i|\right) \ .$$

Proof. Setting $i = j$ in Proposition B.4.6 we have

$$\begin{aligned}
h(f_i) &= -E_{f_i}\left[\ln f_i\right] \\
&= \frac{1}{2}\left[\ln\left((2\pi)^n|\Sigma_i|\right) + \operatorname{tr}\left(\Sigma_i^{-1}\Sigma_i\right) + \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i\right)'\Sigma_i^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i\right)\right] \\
&= \frac{1}{2}\left[\ln\left((2\pi)^n|\Sigma_i|\right) + \operatorname{tr}(I)\right] \\
&= \frac{1}{2}\ln\left((2\pi e)^n|\Sigma_i|\right) \ .
\end{aligned}$$

□

Proposition B.4.8. *The Kullback-Liebler divergence between f_i and f_j is*

$$D(f_i||f_j) = \frac{1}{2}\left[\ln\frac{|\Sigma_j|}{|\Sigma_i|} + \operatorname{tr}\left(\Sigma_j^{-1}\Sigma_i\right) + \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)'\Sigma_j^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) - n\right] \ .$$

Proof. Using Proposition B.4.6 and Remark 1.2.4 we have

$$\begin{aligned}
D(f_i||f_j) &= E_{f_i}[\ln f_i] - E_{f_i}[\ln f_j] \\
&= -\frac{1}{2}\ln\left((2\pi e)^n|\Sigma_i|\right) \\
&\quad + \frac{1}{2}\left[\ln\left((2\pi)^n|\Sigma_j|\right) + \operatorname{tr}\left(\Sigma_j^{-1}\Sigma_i\right) + \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)'\Sigma_j^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)\right] \\
&= \frac{1}{2}\left[\ln\frac{|\Sigma_j|}{|\Sigma_i|} + \operatorname{tr}\left(\Sigma_j^{-1}\Sigma_i\right) + \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)'\Sigma_j^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) - n\right] \ .
\end{aligned}$$

□

Remark B.4.9. If we set $n = 1$ in Proposition B.4.8 we get

$$\begin{aligned}
D(f_i||f_j) &= \frac{1}{2}\left[\ln\frac{\sigma_j^2}{\sigma_i^2} + \frac{\sigma_i^2}{\sigma_j^2} + \left(\mu_i - \mu_j\right)\frac{1}{\sigma_j^2}\left(\mu_i - \mu_j\right) - 1\right] \\
&= \frac{1}{2\sigma_j^2}\left[(\mu_i - \mu_j)^2 + \sigma_i^2 - \sigma_j^2\right] + \ln\frac{\sigma_j}{\sigma_i} \ ,
\end{aligned}$$

which is the expression for the Kullback divergence between two univariate Gaussian distributions obtained in Proposition B.4.3 (as expected).

Proposition B.4.10. *For $\alpha \in \mathbb{R}^+ \setminus \{1\}$ let A be the matrix $A := \alpha\Sigma_i^{-1} + (1-\alpha)\Sigma_j^{-1}$. Then if A is positive definite the Rényi divergence between f_i and f_j is given by*

$$D_\alpha(f_i||f_j) = \frac{1}{2}\ln\left(\frac{|\Sigma_j|}{|\Sigma_i|}\right) + \frac{1}{2(\alpha-1)}\ln\left(\frac{1}{|A||\Sigma_i|}\right) - \frac{F(\alpha)}{2(\alpha-1)} ,$$

where

$$\begin{aligned} F(\alpha) &:= \left[\alpha\boldsymbol{\mu}_i'\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\boldsymbol{\mu}_j'\Sigma_j^{-1}\boldsymbol{\mu}_j\right] \\ &\quad - \left[\alpha\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\Sigma_j^{-1}\boldsymbol{\mu}_j\right]' A^{-1}\left[\alpha\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\Sigma_j^{-1}\boldsymbol{\mu}_j\right] . \end{aligned}$$

If A is not positive-definite then

$$D_\alpha(f_i||f_j) = \infty .$$

Proof.

$$\begin{aligned} f_i^\alpha f_j^{1-\alpha} &= \left[\frac{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}}{(2\pi)^{n/2}|\Sigma_i|^{1/2}}\right]^\alpha \left[\frac{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)}}{(2\pi)^{n/2}|\Sigma_j|^{1/2}}\right]^{1-\alpha} \\ &= \left[(2\pi)^n|\Sigma_i|\right]^{-\frac{\alpha}{2}}\left[(2\pi)^n|\Sigma_j|\right]^{\frac{\alpha-1}{2}} \\ &\quad \cdot \exp\left(-\frac{1}{2}\left[\alpha(\boldsymbol{x}-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + (1-\alpha)(\boldsymbol{x}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right]\right) . \end{aligned}$$

Consider the argument in the exponential:

$$\begin{aligned}
&\alpha(\boldsymbol{x}-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)+(1-\alpha)(\boldsymbol{x}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\\
&=\alpha\boldsymbol{x}'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)-\alpha\boldsymbol{\mu}_i'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)+(1-\alpha)\boldsymbol{x}'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\\
&\quad-(1-\alpha)\boldsymbol{\mu}_j'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\\
&=\alpha\boldsymbol{x}'\Sigma_i^{-1}\boldsymbol{x}-\alpha\boldsymbol{x}'\Sigma_i^{-1}\boldsymbol{\mu}_i-\alpha\boldsymbol{\mu}_i'\Sigma_i^{-1}\boldsymbol{x}+\alpha\boldsymbol{\mu}_i'\Sigma_i^{-1}\boldsymbol{\mu}_i\\
&\quad+(1-\alpha)\boldsymbol{x}'\Sigma_j^{-1}\boldsymbol{x}-(1-\alpha)\boldsymbol{x}'\Sigma_j^{-1}\boldsymbol{\mu}_j-(1-\alpha)\boldsymbol{\mu}_j'\Sigma_j^{-1}\boldsymbol{x}+(1-\alpha)\boldsymbol{\mu}_j'\Sigma_j^{-1}\boldsymbol{\mu}_j\\
&=\boldsymbol{x}'(\alpha\Sigma_i^{-1}+(1-\alpha)\Sigma_j^{-1})\boldsymbol{x}-\boldsymbol{x}'(\alpha\Sigma_i^{-1}\boldsymbol{\mu}_i+(1-\alpha)\Sigma_j^{-1}\boldsymbol{\mu}_j)\\
&\quad-(\alpha\boldsymbol{\mu}_i'\Sigma_i^{-1}+(1-\alpha)\boldsymbol{\mu}_j'\Sigma_j^{-1})\boldsymbol{x}+[\alpha\boldsymbol{\mu}_i'\Sigma_i^{-1}\boldsymbol{\mu}_i+(1-\alpha)\boldsymbol{\mu}_j'\Sigma_j^{-1}\boldsymbol{\mu}_j]\ .
\end{aligned}$$

Note that

$$\left[\alpha\Sigma_i^{-1}\boldsymbol{\mu}_i+(1-\alpha)\Sigma_j^{-1}\boldsymbol{\mu}_j\right]'=\alpha\boldsymbol{\mu}_i'\Sigma_i^{-1}+(1-\alpha)\boldsymbol{\mu}_j'\Sigma_j^{-1}\ ,$$

since Σ_i and Σ_j are symmetric (hence also their inverses). Then

$$\alpha(\boldsymbol{x}-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)+(1-\alpha)(\boldsymbol{x}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)$$

is of the form $\boldsymbol{x}'A\boldsymbol{x}-2\boldsymbol{x}'b+c$, with

$$\begin{aligned}
&A=\alpha\Sigma_i^{-1}+(1-\alpha)\Sigma_j^{-1}\ ,\quad b=\alpha\Sigma_i^{-1}\boldsymbol{\mu}_i+(1-\alpha)\Sigma_j^{-1}\boldsymbol{\mu}_j\ ,\ \text{and}\\
&\qquad c=\alpha\boldsymbol{\mu}_i'\Sigma_i^{-1}\boldsymbol{\mu}_i+(1-\alpha)\boldsymbol{\mu}_j'\Sigma_j^{-1}\boldsymbol{\mu}_j\ .
\end{aligned}$$

- If A is a symmetric matrix, positive-definite, then it is invertible and applying Proposition A.4.4 we can write

  $$\boldsymbol{x}'A\boldsymbol{x}-2\boldsymbol{x}'b+c=(\boldsymbol{x}-\nu)'A(\boldsymbol{x}-\nu)+d\ ,$$

  where

  $$\nu=A^{-1}b\quad\text{and}\quad d=c-b'A^{-1}b\ .$$

Then

$$
\begin{aligned}
& f_i^{\alpha} f_j^{1-\alpha} \\
&\quad = \left[(2\pi)^n|\Sigma_i|\right]^{-\frac{\alpha}{2}} \left[(2\pi)^n|\Sigma_j|\right]^{\frac{\alpha-1}{2}} \\
&\qquad \cdot \exp\left(-\frac{1}{2}\left[\alpha(\boldsymbol{x}-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)+(1-\alpha)(\boldsymbol{x}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right]\right) \\
&\quad = \left[(2\pi)^n|\Sigma_i|\right]^{-\frac{\alpha}{2}} \left[(2\pi)^n|\Sigma_j|\right]^{\frac{\alpha-1}{2}} \exp\left(-\frac{1}{2}\left[(\boldsymbol{x}-\nu)'A(\boldsymbol{x}-\nu)+d\right]\right) \\
&\quad = \left[(2\pi)^n|\Sigma_i|\right]^{-\frac{\alpha}{2}} \left[(2\pi)^n|\Sigma_j|\right]^{\frac{\alpha-1}{2}} e^{-\frac{d}{2}} \exp\left(-\frac{1}{2}\left[(\boldsymbol{x}-\nu)'A(\boldsymbol{x}-\nu)\right]\right) .
\end{aligned}
$$

Letting $B = A^{-1} \Leftrightarrow A = B^{-1}$ we recognize the above as being proportional to the pdf of a multivariate normal distribution with mean ν and covariance matrix B. As shown in Proposition A.4.3, A will always be symmetric, and it will be positive-definite for any $\alpha \in (0,1)$, given that Σ_i and Σ_j are positive-definite and symmetric by assumption. The invertibility of $B = A^{-1}$ also ensures that $|B| \neq 0$, so we may write

$$
\begin{aligned}
\int_{\mathbb{R}^n} f_i^{\alpha} f_j^{\alpha-1}\, d\boldsymbol{x} &= \left(\frac{|\Sigma_j|}{|\Sigma_i|}\right)^{\frac{\alpha-1}{2}} \left(\frac{|B|}{|\Sigma_i|}\right)^{1/2} e^{-\frac{d}{2}} \\
&\qquad \int_{\mathbb{R}^n} ((2\pi)^n|B|)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left[(\boldsymbol{x}-\nu)'B^{-1}(\boldsymbol{x}-\nu)\right]\right) d\boldsymbol{x} \\
&= \left(\frac{|\Sigma_j|}{|\Sigma_i|}\right)^{\frac{\alpha-1}{2}} \left(\frac{|B|}{|\Sigma_i|}\right)^{1/2} e^{-\frac{d}{2}} ,
\end{aligned}
$$

hence

$$
\begin{aligned}
D_\alpha(f_i||f_j) &= \frac{1}{\alpha-1}\ln\left(\left(\frac{|\Sigma_j|}{|\Sigma_i|}\right)^{\frac{\alpha-1}{2}} \left(\frac{|B|}{|\Sigma_i|}\right)^{1/2} e^{-\frac{d}{2}}\right) \\
&= \frac{1}{2}\ln\left(\frac{|\Sigma_j|}{|\Sigma_i|}\right) + \frac{1}{2(\alpha-1)}\ln\left(\frac{|B|}{|\Sigma_i|}\right) - \frac{d}{2(\alpha-1)} .
\end{aligned}
$$

But

$$\begin{aligned} d &= c - b'A^{-1}b \\ &= \left[\alpha\boldsymbol{\mu}_i'\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\boldsymbol{\mu}_j'\Sigma_j^{-1}\boldsymbol{\mu}_j\right] \\ &- \left[\alpha\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\Sigma_j^{-1}\boldsymbol{\mu}_j\right]' A^{-1} \left[\alpha\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\Sigma_j^{-1}\boldsymbol{\mu}_j\right] , \end{aligned}$$

and

$$|B| = |A^{-1}| = \frac{1}{|A|}.$$

Thus,

$$D_\alpha(f_i||f_j) = \frac{1}{2}\ln\left(\frac{|\Sigma_j|}{|\Sigma_i|}\right) + \frac{1}{2(\alpha-1)}\ln\left(\frac{1}{|A||\Sigma_i|}\right) - \frac{F(\alpha)}{2(\alpha-1)} ,$$

where

$$A = \alpha\Sigma_i^{-1} + (1-\alpha)\Sigma_j^{-1}$$

and

$$\begin{aligned} F(\alpha) := &\left[\alpha\boldsymbol{\mu}_i'\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\boldsymbol{\mu}_j'\Sigma_j^{-1}\boldsymbol{\mu}_j\right] \\ &- \left[\alpha\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\Sigma_j^{-1}\boldsymbol{\mu}_j\right]' A^{-1} \left[\alpha\Sigma_i^{-1}\boldsymbol{\mu}_i + (1-\alpha)\Sigma_j^{-1}\boldsymbol{\mu}_j\right] . \end{aligned}$$

- Suppose now that A is not positive definite. Since A is always symmetric, we can always find a Q and Λ such that $\Lambda = Q'AQ$, where Λ is the diagonal matrix $\Lambda = \text{diag}(\lambda_1, ...\lambda_n)$ with the eigenvalues of A, and Q is an orthogonal matrix of eigenvectors of A. Let $\boldsymbol{u} = Q'\boldsymbol{x}$. Then $\boldsymbol{x} = Q\boldsymbol{u}$ since $Q' = Q^{-1}$ and $\boldsymbol{x}'A\boldsymbol{x} = \boldsymbol{u}'Q'AQ\boldsymbol{u} = \boldsymbol{u}'\Lambda\boldsymbol{u}$. Thus

$$\boldsymbol{x}'A\boldsymbol{x} - 2\boldsymbol{x}'\boldsymbol{b} = \boldsymbol{u}'\Lambda\boldsymbol{u} - 2\boldsymbol{u}'Q'\boldsymbol{b} = \sum_{i=1}^{n}\lambda_i u_i^2 - 2\sum_{i,j=1}^{n} Q_{ji}u_i b_j .$$

Moreover, for some $k \in \{1, ..., n\}$, there exists an eigenvalue $\lambda_k \le 0$. With this in mind, we rewrite the above as

$$\sum_{i=1}^n \lambda_i u_i^2 - 2\sum_{i,j=1}^n Q_{ji} u_i b_j = \left(\lambda_k u_k^2 - 2\sum_{j=1}^n Q_{jk} u_k b_j \right) + \sum_{i=1, i\neq k}^n \lambda_i u_i^2 - 2 \sum_{i,j=1; i\neq k}^n Q_{ji} u_i b_j \ .$$

Observe also that

$$\begin{aligned}
&\int_{\mathbb{R}} \exp\left(-\frac{1}{2} \left[\lambda_k u_k^2 - 2\sum_{j=1}^n Q_{ij} u_k b_j \right] \right) du_k \\
&\quad = \int_{\mathbb{R}} \exp\left(\frac{1}{2} \left[|\lambda_k| u_k^2 + \left(2\sum_{j=1}^n Q_{ij} b_j \right) u_k \right] \right) du_k \\
&\quad = \int_{\mathbb{R}} \exp\left(s y^2 + t y \right) dy \ , \ \ s \ge 0 \ , \ t \in \mathbb{R} \\
&\quad = \infty \ ,
\end{aligned}$$

since the matrix Q and the vector b are fixed. Also, $Q'(\mathbb{R}^n) = \mathbb{R}^n$ and for $\boldsymbol{u}(\boldsymbol{x}) = Q'\boldsymbol{x}$ the Jacobian determinant of the transformation $\boldsymbol{x}(\boldsymbol{u}) = Q\boldsymbol{u}$ is simply $|J| = |Q| = 1$ since Q is orthogonal. Thus,

$$\begin{aligned}
\int_{\mathbb{R}^n} f_i^{\alpha} f_j^{\alpha-1} \, d\boldsymbol{x} &= K \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2} \left[\boldsymbol{x}' A \boldsymbol{x} - 2\boldsymbol{x}'\boldsymbol{b} \right] \right) d\boldsymbol{x} \ , \ K > 0 \ , \\
&= K \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2} \left[\boldsymbol{u}' \Lambda \boldsymbol{u} - 2\boldsymbol{u}' Q' \boldsymbol{b} \right] \right) d\boldsymbol{u} \\
&= K \int_{\mathbb{R}} \exp\left(-\frac{1}{2} \left[\lambda_k u_k^2 - 2\sum_{j=1}^n Q_{ij} u_k b_j \right] \right) du_k \\
&\qquad \cdot \int_{\mathbb{R}^{n-1}} \exp\left(-\frac{1}{2} \left[\sum_{i=1, i\neq k}^n \lambda_i u_i^2 - 2 \sum_{i,j=1; i\neq k}^n Q_{ji} u_i b_j \right] \right) d\boldsymbol{u} \ ,
\end{aligned}$$

where we have used Fubini's Theorem[3] as the integrand is always positive. Hence, by the observation above, the whole expression equals $+\infty$ as a result of the first

[3] See for example [55], p.164

integral. Finally, since the case of A not being positive-definite requires $\alpha > 1$ (see Proposition A.4.3) then

$$D_\alpha(f_i||f_f) = \frac{1}{\alpha - 1}\int_{\mathbb{R}^n} f_i^\alpha f_j^{1-\alpha} = \infty\,.$$

□

Remark B.4.11. If we set $n = 1$ in Proposition B.4.10 Then

$$\begin{aligned}
F(\alpha) &= \frac{\alpha\mu_i^2}{\sigma_i^2} + \frac{(1-a)\mu_j^2}{\sigma_j^2} - \left(\frac{\alpha\mu_i}{\sigma_i^2} + \frac{(1-a)\mu_j}{\sigma_j^2}\right)^2 \left(\frac{\alpha}{\sigma_i^2} + \frac{(1-a)}{\sigma_j^2}\right)^{-1} \\
&= \frac{1}{\sigma_i^2\sigma_j^2}\left(\alpha\sigma_j^2\mu_i^2 + (1-\alpha)\sigma_j^2\mu_j^2\right) \\
&\quad - \frac{1}{\sigma_i^4\sigma_j^4}\left(\alpha\sigma_j^2\mu_i + (1-\alpha)\sigma_i^2\mu_j\right)^2 \frac{\sigma_i^2\sigma_j^2}{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2} \\
&= \frac{1}{\sigma_i^2\sigma_j^2\left(\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2\right)} \\
&\qquad \cdot \Big[\left(\alpha\sigma_j^2\mu_i^2 + (1-\alpha)\sigma_j^2\mu_j^2\right)\left(\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2\right) \\
&\qquad\quad - \left(\alpha\sigma_j^2\mu_i + (1-\alpha)\sigma_i^2\mu_j\right)^2\Big]
\end{aligned}$$

As shown in the proof of Proposition B.4.4,

$$\begin{aligned}
&\left(\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2\right)\left(\alpha\sigma_j^2\mu_i^2 + (1-\alpha)\sigma_i^2\mu_j^2\right) - \left(\alpha\sigma_j^2\mu_i + (1-\alpha)\sigma_i^2\mu_j\right)^2 \\
&= \alpha(1-\alpha)\sigma_i^2\sigma_j^2(\mu_i - \mu_j)^2\,,
\end{aligned}$$

So

$$F(\alpha) = \frac{\alpha(1-\alpha)(\mu_i - \mu_j)^2}{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2}\,.$$

Also,

$$|A||\Sigma_i| = \left(\frac{\alpha}{\sigma_i^2} + \frac{(1-\alpha)}{\sigma_j^2}\right)\sigma_i^2 = \frac{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2}{\sigma_j^2}\,.$$

Thus,

$$
\begin{aligned}
D_\alpha(f_i||f_j) &= \frac{1}{2}\ln\left(\frac{\sigma_j^2}{\sigma_i^2}\right) + \frac{1}{2(\alpha-1)}\ln\left(\frac{\sigma_j^2}{\alpha\sigma_j^2+(1-\alpha)\sigma_i^2}\right) \\
&\quad - \frac{\alpha(1-\alpha)(\mu_i-\mu_j)^2}{\alpha\sigma_j^2+(1-\alpha)\sigma_i^2}\frac{1}{2(\alpha-1)} \\
&= \ln\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{1}{2(\alpha-1)}\ln\left(\frac{\sigma_j^2}{\alpha\sigma_j^2+(1-\alpha)\sigma_i^2}\right) \\
&\quad + \frac{1}{2}\frac{\alpha(\mu_i-\mu_j)^2}{\alpha\sigma_j^2+(1-\alpha)\sigma_i^2}\,,
\end{aligned}
$$

which is the expression for the Rényi divergence between two univariate Gaussian distributions obtained in Proposition B.4.4 (for $\alpha\sigma_j^2+(1-\alpha)\sigma_i^2>0$), as expected. Moreover note that for $n=1$ the positive-definiteness constraint in Proposition B.4.10 in this case becomes

$$
\left(\frac{\alpha}{\sigma_i^2}+\frac{(1-\alpha)}{\sigma_j^2}\right)x^2>0\,,\ \forall x\in\mathbb{R}\setminus\{0\}\ \Leftrightarrow \alpha\sigma_j^2+(1-\alpha)\sigma_i^2>0\,,
$$

which was the corresponding constraint in Proposition B.4.4.

Remark B.4.12.

$$
\lim_{\alpha\uparrow 1} D_\alpha(f_i||f_j) = D(f_i||f_j)\,.
$$

Proof. First observe the following:

$$
\lim_{\alpha\uparrow 1} A = \lim_{\alpha\uparrow 1}\left[\alpha\Sigma_i^{-1}+(1-\alpha)\Sigma_j^{-1}\right] = \Sigma_i^{-1} \Rightarrow \lim_{\alpha\uparrow 1}|A||\Sigma_i| = |\Sigma_i^{-1}\Sigma_i| = |I| = 1\,.
$$

Also,

$$\begin{aligned}
\lim_{\alpha\uparrow 1} F(\alpha) &= \lim_{\alpha\uparrow 1}\left(\left[\alpha\mu_i'\Sigma_i^{-1}\mu_i + (1-\alpha)\mu_j'\Sigma_j^{-1}\mu_j\right]\right. \\
&\quad \left. - \left[\alpha\Sigma_i^{-1}\mu_i + (1-\alpha)\Sigma_j^{-1}\mu_j\right]' A^{-1}\left[\alpha\Sigma_i^{-1}\mu_i + (1-\alpha)\Sigma_j^{-1}\mu_j\right]\right) \\
&= \mu'\Sigma_i^{-1}\mu_i - \left(\Sigma_i^{-1}\mu_i\right)'\Sigma_i\Sigma_i^{-1}\mu_i \\
&= \mu'\Sigma_i^{-1}\mu_i - \mu_i'\Sigma_i^{-1}\Sigma_i\Sigma_i^{-1}\mu \\
&= 0\,.
\end{aligned}$$

Then we see that

$$\lim_{\alpha\uparrow 1}\left[\frac{1}{2(\alpha-1)}\ln\left(\frac{1}{|A||\Sigma_i|}\right) - \frac{F(\alpha)}{2(\alpha-1)}\right]$$

has an indeterminate form. From Proposition A.4.11,

$$\frac{d|A|}{d\alpha} = |A|\operatorname{tr}\left(A^{-1}\frac{dA}{d\alpha}\right)\,,$$

and so

$$\begin{aligned}
\lim_{\alpha\uparrow 1}\frac{1}{|A|}\frac{d|A|}{d\alpha} &= \operatorname{tr}\left(\lim_{\alpha\uparrow 1}\left[A^{-1}\frac{dA}{d\alpha}\right]\right) \\
&= \operatorname{tr}\left(\left(\Sigma_i^{-1}\right)^{-1}\left(\Sigma_i^{-1} - \Sigma_j^{-1}\right)\right) \\
&= n - \operatorname{tr}\left(\Sigma_j^{-1}\Sigma_i\right)\,,
\end{aligned}$$

where we have used Proposition A.4.10, as well as the linearity, symmetry, and continuity of the trace operator. Thus, applying l'Hospital's rule,

$$\lim_{\alpha\uparrow 1}\frac{1}{2(\alpha-1)}\ln\left(\frac{1}{|A||\Sigma_i|}\right) = -\frac{1}{2}\lim_{\alpha\uparrow 1}\frac{1}{|A|}\frac{d|A|}{d\alpha} = \frac{1}{2}\left[\operatorname{tr}\left(\Sigma_j^{-1}\Sigma_i\right) - n\right]\,.$$

Now, also from Proposition A.4.10

$$\frac{d}{d\alpha}\left(x'Ax\right) = \frac{dx'}{d\alpha}(Ax) + x'\frac{dA}{d\alpha}x + x'A\frac{dx}{d\alpha}\,,$$

and so

$$\begin{aligned}
\frac{d}{d\alpha}F(\alpha) = \frac{d}{d\alpha}\, &\Big(\Big[\alpha\mu_i'\Sigma_i^{-1}\mu_i + (1-\alpha)\mu_j'\Sigma_j^{-1}\mu_j\Big] \\
&- \Big[\alpha\Sigma_i^{-1}\mu_i + (1-\alpha)\Sigma_j^{-1}\mu_j\Big]' A^{-1}\Big[\alpha\Sigma_i^{-1}\mu_i + (1-\alpha)\Sigma_j^{-1}\mu_j\Big]\Big) \\
= \mu_i'\Sigma_i^{-1}\mu_i &- \mu_j'\Sigma_j^{-1}\mu_j \\
&- \Big[\Sigma_i^{-1}\mu_i - \Sigma_j^{-1}\mu_j\Big]'\Big(A^{-1}\Big[\alpha\Sigma_i^{-1}\mu_i + (1-\alpha)\Sigma_j^{-1}\mu_j\Big]\Big) \\
&- \Big[\alpha\Sigma_i^{-1}\mu_i + (1-\alpha)\Sigma_j^{-1}\mu_j\Big]'\frac{dA^{-1}}{d\alpha}\Big[\alpha\Sigma_i^{-1}\mu_i + (1-\alpha)\Sigma_j^{-1}\mu_j\Big] \\
&- \Big[\alpha\Sigma_i^{-1}\mu_i + (1-\alpha)\Sigma_j^{-1}\mu_j\Big]' A^{-1}\Big[\Sigma_i^{-1}\mu_i - \Sigma_j^{-1}\mu_j\Big] ,
\end{aligned}$$

and also

$$\frac{dA^{-1}}{d\alpha} = -A^{-1}\frac{dA}{d\alpha}A^{-1} .$$

Since $\lim_{\alpha\uparrow 1} A = \Sigma_i^{-1}$, then $\lim_{\alpha\uparrow 1} A^{-1} = \Sigma_i$ and

$$\lim_{\alpha\uparrow 1}\frac{dA^{-1}}{d\alpha} = -\Sigma_i\Big[\Sigma_i^{-1} - \Sigma_j^{-1}\Big]\Sigma_i = -\Sigma_i + \Sigma_i\Sigma_j^{-1}\Sigma_i .$$

Hence

$$\begin{aligned}
\lim_{\alpha\uparrow 1}\frac{d}{d\alpha}F(\alpha) &= \mu_i'\Sigma_i^{-1}\mu_i - \mu_j'\Sigma_j^{-1}\mu_j - \Big[\Sigma_i^{-1}\mu_i - \Sigma_j^{-1}\mu_j\Big]'\Sigma_i\Sigma_i^{-1}\mu_i \\
&\quad - \Big[\Sigma_i^{-1}\mu_i\Big]'\Big(-\Sigma_i + \Sigma_i\Sigma_j^{-1}\Sigma_i\Big)\Big[\Sigma_i^{-1}\mu_i\Big] \\
&\quad - \Big[\Sigma_i^{-1}\mu_i\Big]'\Sigma_i\Big[\Sigma_i^{-1}\mu_i - \Sigma_j^{-1}\mu_j\Big] \\
&= \Big[\mu_i'\Sigma_i^{-1}\mu_i - \mu_j'\Sigma_j^{-1}\mu_j\Big] - \mu_i'\Sigma_i^{-1}\mu_i + \mu_j'\Sigma_j^{-1}\mu_i \\
&\quad - \mu_i'\Sigma_i^{-1}\Big(\Sigma_i\Sigma_j^{-1}\mu_i - I\mu_i\Big) - \mu_i'\Sigma_i^{-1}\mu_i + \mu_i'\Sigma_j^{-1}\mu_j \\
&= -\mu_j'\Sigma_j^{-1}\mu_j + \mu_j'\Sigma_j^{-1}\mu_i - \mu_i'\Sigma_j^{-1}\mu_i + \mu_i'\Sigma_j^{-1}\mu_j \\
&= -\Big[\mu_i'\Sigma_j^{-1}\mu_i - \mu_i'\Sigma_j^{-1}\mu_j - \mu_j'\Sigma_j^{-1}\mu_i + \mu_j'\Sigma_j^{-1}\mu_j\Big] \\
&= -(\mu_i - \mu_j)'\Sigma_j^{-1}(\mu_i - \mu_j) ,
\end{aligned}$$

and from l'Hospital's rule

$$\lim_{\alpha\uparrow 1}\left[-\frac{F(\alpha)}{2(\alpha-1)}\right]=\frac{1}{2}(\mu_i-\mu_j)'\Sigma_j^{-1}(\mu_i-\mu_j)\,.$$

Finally,

$$\begin{aligned}\lim_{\alpha\uparrow 1}D_\alpha(f_i||f_j)&=\lim_{\alpha\uparrow 1}\left[\frac{1}{2}\ln\left(\frac{|\Sigma_j|}{|\Sigma_i|}\right)+\frac{1}{2(\alpha-1)}\ln\left(\frac{1}{|A||\Sigma_i|}\right)-\frac{F(\alpha)}{2(\alpha-1)}\right]\\&=\frac{1}{2}\left[\ln\left(\frac{|\Sigma_j|}{|\Sigma_i|}\right)+\mathrm{tr}\left(\Sigma_j^{-1}\Sigma_i\right)+(\mu_i-\mu_j)'\Sigma_j^{-1}(\mu_i-\mu_j)-n\right]\,,\end{aligned}$$

which is the expression for the Kullback divergence between two multivariate Gaussian distributions obtained in Proposition B.4.8, as expected. □

B.4.3 A Special Bivariate Case

Consider the expression for $D_\alpha(f_i||f_j)$ for the zero-mean, unit-variance, bivariate case:

$$f_i(\boldsymbol{x})=\frac{e^{-\frac{1}{2}\boldsymbol{x}'\Phi_i^{-1}\boldsymbol{x}}}{2\pi(1-\rho_i^2)^{1/2}}\,,x\in\mathbb{R}^2\,,$$

where

$$\Sigma_k=\begin{pmatrix}1&\rho_k\\\rho_k&1\end{pmatrix}\,,\quad k=i,j\,.$$

We have

$$\Sigma_k^{-1}=\frac{1}{1-\rho_k^2}\begin{pmatrix}1&-\rho_k\\-\rho_k&1\end{pmatrix}\,,\quad k=i,j\,,$$

and

$$\begin{aligned}A&=\alpha\Sigma_i^{-1}+(1-\alpha)\Sigma_j^{-1}\\&=\frac{\alpha}{1-\rho_i^2}\begin{pmatrix}1&-\rho_i\\-\rho_i&1\end{pmatrix}+\frac{(1-\alpha)}{1-\rho_j^2}\begin{pmatrix}1&-\rho_j\\-\rho_j&1\end{pmatrix}\,.\end{aligned}$$

Writing the above as a single matrix and taking the determinant we find

$$
\begin{aligned}
|A| &= \left[\frac{1}{(1-\rho_i^2)(1-\rho_j^2)}\right]^2 \left(\left[\alpha(1-\rho_j^2)+(1-\alpha)(1-\rho_i^2)\right]^2\right.\\
&\qquad\left. - \left[\alpha(1-\rho_j^2)\rho_i+(1-\alpha)(1-\rho_i^2)\rho_j\right]^2\right)\\
&= \left[\frac{1}{(1-\rho_i^2)(1-\rho_j^2)}\right]^2 \left(\alpha^2(1-\rho_j^2)^2(1-\rho_i^2)\right.\\
&\quad \left.+(1-\alpha)^2(1-\rho_i^2)^2(1-\rho_j^2)+2\alpha(1-\alpha)(1-\rho_i^2)(1-\rho_j^2)(1-\rho_i\rho_j)\right)\\
&= \frac{\alpha^2(1-\rho_j^2)+(1-\alpha)^2(1-\rho_i^2)+2\alpha(1-\alpha)(1-\rho_i\rho_j)}{(1-\rho_i^2)(1-\rho_j^2)}\\
&= \frac{(\alpha^2+(1-\alpha)^2+2\alpha(1-\alpha))-\alpha^2\rho_j^2-(1-\alpha)^2\rho_i^2-2\alpha(1-\alpha)\rho_i\rho_j}{(1-\rho_i^2)(1-\rho_j^2)}\\
&= \frac{1-(\alpha\rho_j+(1-\alpha)\rho_i)^2}{(1-\rho_i^2)(1-\rho_j^2)}\ .
\end{aligned}
$$

Thus

$$
|A||\Sigma_i| = \frac{1-(\alpha\rho_j+(1-\alpha)\rho_i)^2}{(1-\rho_j^2)}\ .
$$

Also, since $\boldsymbol{\mu_i} = \boldsymbol{\mu_j} = (0,0)'$ then $F(\alpha)=0$. Thus when A is positive definite (e.g. when $\alpha \in (0,1)$) we have

$$
D_\alpha(f_i||f_j) = \frac{1}{2}\ln\left(\frac{1-\rho_j^2}{1-\rho_i^2}\right) - \frac{1}{2(\alpha-1)}\ln\left(\frac{1-(\alpha\rho_j+(1-\alpha)\rho_i)^2}{(1-\rho_j^2)}\right)\ .
$$

Now consider the KLD case. The multivariate Gaussian KLD

$$
D(f_i||f_j) = \frac{1}{2}\left(\ln\frac{|\Sigma_j|}{|\Sigma_i|} + \mathrm{tr}\left(\Sigma_j^{-1}\Sigma_i\right)\right) + \frac{1}{2}\left[\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)'\Sigma_j^{-1}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)-n\right]
$$

becomes

$$
D(f_i||f_j) = \frac{1}{2}\left(\ln\frac{1-\rho_j^2}{1-\rho_i^2} + \mathrm{tr}\left(\Sigma_j^{-1}\Sigma_i\right) - 2\right)\ .
$$

We have

$$\Sigma_j^{-1}\Sigma_i = \frac{1}{1-\rho_j^2}\begin{pmatrix} 1 & -\rho_j \\ -\rho_j & 1 \end{pmatrix}\begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix} ,$$

hence

$$\mathrm{tr}(\Sigma_j^{-1}\Sigma_i) = \frac{2(1-\rho_j\rho_i)}{1-\rho_j^2} ,$$

and

$$\begin{aligned} D(f_i||f_j) &= \frac{1}{2}\ln\left(\frac{1-\rho_j^2}{1-\rho_i^2}\right) + \frac{1-\rho_j\rho_i}{1-\rho_j^2} - 1 \\ &= \frac{1}{2}\ln\left(\frac{1-\rho_j^2}{1-\rho_i^2}\right) + \frac{\rho_j^2-\rho_j\rho_i}{1-\rho_j^2} . \end{aligned}$$

We can see that taking the limit $\alpha \to 1$ of the second term for the Rényi expression above we have

$$-\frac{1}{2}\lim_{\alpha\to 1}\frac{-2\rho_\alpha^*(\rho_j-\rho_i)}{1-(\rho_\alpha^*)^2} = \frac{\rho_j^2-\rho_j\rho_i}{1-\rho_j^2} ,$$

so that the expressions agree in the limit $\alpha \to 1$, as expected.

B.5 Pareto Distributions

Throughout this section let f_i and f_j be two Pareto densities (over the same support)

$$f_i(x) = a_i m^{a_i} x^{-(a_i+1)} \ , \quad a_i, m > 0 \ ; x > m \ .$$

Proposition B.5.1.

$$E_{f_i}\left[\ln f_j\right] = \ln \frac{a_j}{m} - \frac{(a_j+1)}{a_i} \ .$$

Proof.

$$\begin{aligned} E_{f_i}\left[\ln f_j\right] &= E_{f_i}\left[\ln\left(a_j m^{a_j}\right) - (a_j+1)\ln X\right] \\ &= \ln\left(a_j m^{a_j}\right) - (a_j+1) E_{f_i}\left[\ln X\right] \ . \end{aligned}$$

Now

$$\begin{aligned} E_{f_i}[\ln X] &= \int_m^\infty a_i m^{a_i} x^{-(a_i+1)} \ln x \, dx \\ &= a_i m^{a_i} \left[-\frac{1}{a_i} x^{-a_i} \ln x \Big|_m^\infty + \frac{1}{a_i} \int_m^\infty x^{-(a_i+1)} dx \right] \\ &= \frac{a_i m^{a_i} m^{-a_i} \ln m}{a_i} + \frac{1}{a_i} \int_m^\infty a_i m^{a_i} x^{-(a_i+1)} dx \\ &= \ln m + \frac{1}{a_i} \ , \end{aligned}$$

where we have used integration by parts, and the last term integrates to 1 since it corresponds to integrating the original density over its support. Thus

$$E_{f_i}\left[\ln f_j\right] = \ln\left(a_j m^{a_j}\right) - (a_j+1)\left[\ln m + \frac{1}{a_i}\right] = \ln \frac{a_j}{m} - \frac{(a_j+1)}{a_i} \ .$$

□

Corollary B.5.2. *The differential entropy of f_i is*

$$h(f_i) = \ln \frac{m}{a_i} + \frac{(a_i + 1)}{a_i} .$$

Proof. Setting $i = j$ in Proposition B.5.1 we have

$$h(f_i) = -E_{f_i}\left[\ln f_i\right] = -\left[\ln \frac{a_i}{m} - \frac{(a_i+1)}{a_i}\right] = \ln \frac{m}{a_i} + \frac{(a_i+1)}{a_i} .$$

□

Proposition B.5.3. *The Kullback-Liebler divergence between f_i and f_j is*

$$D(f_i || f_j) = \ln \frac{a_i}{a_j} + \frac{a_j - a_i}{a_i} .$$

Proof. Using Proposition B.5.1 and Remark 1.2.4 we have

$$\begin{aligned} D(f_i || f_j) &= E_{f_i}[\ln f_i] - E_{f_i}[\ln f_j] \\ &= -\left[\ln \frac{m}{a_i} + \frac{(a_i+1)}{a_i}\right] - \left[\ln \frac{a_j}{m} - \frac{(a_j+1)}{a_i}\right] \\ &= \ln \frac{a_i}{a_j} + \frac{a_j - a_i}{a_i} . \end{aligned}$$

□

Proposition B.5.4. *For $\alpha \in \mathbb{R}^+ \setminus \{1\}$ let $a_0 = \alpha a_i + (1-\alpha) a_j$. Then the Rényi divergence between f_i and f_j is given by*

$$D_\alpha(f_i || f_j) = \ln \frac{a_i}{a_j} + \frac{1}{\alpha - 1} \ln \frac{a_i}{a_0}$$

for $a_0 > 0$ and

$$D_\alpha(f_i || f_j) = +\infty$$

otherwise.

Proof.

$$f_i^{\alpha} f_j^{1-\alpha} = \left[a_i m^{a_i} x^{-(a_i+1)}\right]^{\alpha} \left[a_j m^{a_j} x^{-(a_j+1)}\right]^{1-\alpha} = \left(\frac{a_i}{a_j}\right)^{\alpha-1} a_i m^{a_0} x^{a_0-1} ,$$

where $a_0 = \alpha a_i + (1-\alpha)a_j$.

- If $a_0 > 0$ then

  $$\int_m^{\infty} f_i^{\alpha} f_j^{1-\alpha}\, dx = \left(\frac{a_i}{a_j}\right)^{\alpha-1} \frac{a_i}{a_0} \int_m^{\infty} a_0 m^{a_0} x^{a_0-1}\, dx = \left(\frac{a_i}{a_j}\right)^{\alpha-1} \frac{a_i}{a_0} ,$$

  since the integrand is Pareto density with parameters m and a_0. Then

  $$D_{\alpha}(f_i||f_j) = \frac{1}{\alpha-1} \ln \left[\left(\frac{a_i}{a_j}\right)^{\alpha-1} \frac{a_i}{a_0}\right] = \ln \frac{a_i}{a_j} + \frac{1}{\alpha-1} \ln \frac{a_i}{a_0} .$$

  Note that for $\alpha \in (0,1)$ we always have $a_0 > 0$ given the positivity of a_i and a_j.

- If $a_0 \leq 0$ then

  $$\begin{aligned} \int_m^{\infty} f_i^{\alpha} f_j^{1-\alpha}\, dx &= A \int_m^{\infty} x^{a_0-1} dx , \quad A \geq 0 \\ &= \infty . \end{aligned}$$

  Finally, since nonpositive a_0 only occurs for $\alpha > 1$ we have

  $$D_{\alpha}(f_i||f_j) = \frac{1}{\alpha-1} \ln \int_m^{\infty} f_i^{\alpha} f_j^{1-\alpha}\, dx = \infty$$

  for these cases.

□

Remark B.5.5.

$$\lim_{\alpha \uparrow 1} D_{\alpha}(f_i||f_j) = D(f_i||f_j) .$$

Proof.

$$\lim_{\alpha\uparrow 1} D_\alpha(f_i||f_j) = \ln\frac{a_i}{a_j} + \lim_{\alpha\uparrow 1}\left[\frac{1}{\alpha-1}\ln\frac{a_i}{a_0}\right]$$

Since

$$\lim_{\alpha\uparrow 1} a_0 = \lim_{\alpha\uparrow 1}\left(\alpha a_i + (1-\alpha)a_j\right) = a_i,$$

the limit of the second term is of indeterminate form. Applying l'Hospital's rule

$$\lim_{\alpha\uparrow 1}\left[\frac{1}{\alpha-1}\ln\frac{a_i}{a_0}\right] = -\lim_{\alpha\uparrow 1}\frac{a_i-a_j}{a_0} = \frac{a_j-a_i}{a_i}\ .$$

Therefore,

$$\lim_{\alpha\uparrow 1} D_\alpha(f_i||f_j) = \ln\frac{a_i}{a_j} + \frac{a_j-a_i}{a_i} = D(f_i||f_j)\ ,$$

as given by Proposition B.5.3. □

Information measures for univariate Pareto distributions are also considered in [7]. Introduced as "The Pareto distribution with survival function

$$\bar{F}_\beta(x) = (x+1)^{-\beta},\ x>0,\ \beta>0\ \text{''}\ ,$$

denoted by $\mathscr{P}_\beta$. The authors denote the Shannon entropy, Rényi entropy, Kullback-Leibler divergence, and Rényi divergence by $H(\mathscr{P}_i)$, $H_\alpha(\mathscr{P}_i)$, $K(\mathscr{P}_i : \mathscr{P}_j)$, and $K_\alpha(\mathscr{P}_i : \mathscr{P}_j)$, respectively. Thus, they present

$$\begin{aligned}
K(\mathscr{P}_{\beta_1} : \mathscr{P}_{\beta_2}) &= \rho - \log\rho - 1\ ,\\
H(\mathscr{P}_\beta) &= 1 + \frac{1}{\beta} - \log\beta\ ,\\
K_\alpha(\mathscr{P}_{\beta_1} : \mathscr{P}_{\beta_2}) &= \frac{1}{1-\alpha}\log\left(\alpha\rho^{\alpha-1} + (1-\alpha)\rho^\alpha\right)\ , \quad \alpha + (1-\alpha)\rho > 0 \quad \text{and}\\
H_\alpha(\mathscr{P}_b) &= \frac{1}{1-\alpha}\log\frac{\beta^\alpha}{\alpha(\beta+1)-1}\ , \quad \alpha > \frac{1}{\beta+1}\ ,
\end{aligned}$$

where $\rho = \beta_2/\beta_1$. Since the survival function is defined as $1 - F(x)$, where $F(x)$ is the distribution function of X, then the corresponding density is

$$f(x) = \beta(x+1)^{-(\beta+1)}\ , \quad x > 0 \quad \equiv \quad \beta y^{-(\beta+1)}\ , \quad y > 1\ .$$

In our notation this corresponds to the case $a = \beta$ and $m = 1$. Substituting these values into Corollary B.5.2, Proposition B.5.3 and Proposition B.5.4 we obtain

$$\begin{aligned}
h(f) &= \ln\frac{1}{\beta} + \frac{\beta+1}{\beta} = 1 + \frac{1}{\beta} - \log\beta\ , \\
D(f_i||f_j) &= \ln\frac{\beta_i}{\beta_j} + \frac{\beta_j - \beta_i}{\beta_i} = \rho - \log\rho - 1\ , \text{ and} \\
D_\alpha(f_i||f_j) &= \ln\frac{\beta_i}{\beta_j} + \frac{1}{\alpha-1}\ln\frac{\beta_i}{\beta_0} \\
&= \ln\frac{\beta_i}{\beta_j} + \frac{1}{\alpha-1}\ln\frac{\beta_i}{\alpha\beta_i + (1-\alpha)\beta_j} \\
&= -\ln\rho + \frac{1}{1-\alpha}\ln\frac{\alpha\beta_i + (1-\alpha)\beta_j}{\beta_i} \\
&= \frac{1}{1-\alpha}\ln\rho^{\alpha-1} + \frac{1}{1-\alpha}\ln(\alpha + (1-\alpha)\rho) \\
&= \frac{1}{1-\alpha}\ln(\alpha\rho^{\alpha-1} + (1-\alpha)\rho^{\alpha})\ ,
\end{aligned}$$

where $\rho = \beta_j/\beta_i$. Moreover,

$$\begin{aligned}
\beta_0 > 0 &\Leftrightarrow \alpha\beta_i + (1-\alpha)\beta_j > 0 \\
&\Leftrightarrow \alpha + (1-\alpha)\frac{\beta_j}{\beta_i} > 0 \\
&\Leftrightarrow \alpha + (1-\alpha)\rho > 0\ ,
\end{aligned}$$

and we see the two sets of expressions are in agreement.

B.6 Weibull Distributions

Throughout this section let f_i and f_j be two univariate Weibull densities

$$f_i(x) = k_i \lambda_i^{-k_i} x^{k_i - 1} e^{-(x/\lambda_i)^{k_i}} \ , \quad k_i, \lambda_i > 0; x \in \mathbb{R}^+ \quad .$$

Proposition B.6.1. *For $\alpha \in \mathbb{R}^+ \setminus \{1\}$ let $k_i = k_j = k$ and $\lambda_0 = \alpha \lambda_j^k + (1-\alpha)\lambda_i^k$. Then the Rényi divergence between f_i and f_j is given by*

$$D_\alpha(f_i||f_j) = \ln \left(\frac{\lambda_j}{\lambda_i}\right)^k + \frac{1}{\alpha - 1} \ln \frac{\lambda_j^k}{\lambda_0} \ ,$$

for $\lambda_0 > 0$ and

$$D_\alpha(f_i||f_j) = +\infty$$

otherwise.

Proof.

$$\begin{aligned}
f_i^\alpha f_j^{1-\alpha} &= \left[k_i \lambda_i^{-k_i} x^{k_i-1} e^{-(x/\lambda_i)^{k_i}}\right]^\alpha \left[k_j \lambda_j^{-k_j} x^{k_j-1} e^{-(x/\lambda_j)^{k_j}}\right]^{1-\alpha} \\
&= \left[k \lambda_i^{-k} x^{k-1} e^{-(x/\lambda_i)^{k}}\right]^\alpha \left[k \lambda_j^{-k} x^{k-1} e^{-(x/\lambda_j)^{k}}\right]^{1-\alpha} \\
&= k \lambda_i^{-\alpha k} \lambda_j^{-(1-\alpha)k} x^{k-1} \exp\left(-\xi x^k\right) \\
&= \left(\frac{\lambda_j}{\lambda_i}\right)^{k(\alpha-1)} \lambda_i^{-k} k x^{k-1} \exp\left(-\xi x^k\right) \ ,
\end{aligned}$$

where

$$\xi = \frac{\alpha}{\lambda_i^k} + \frac{1-\alpha}{\lambda_j^k} = \frac{\alpha \lambda_j^k + (1-\alpha)\lambda_i^k}{\left(\lambda_i \lambda_j\right)^k} = \frac{\lambda_0}{\left(\lambda_i \lambda_j\right)^k} \ .$$

- If $\lambda_0 > 0$ then $\xi > 0$ and

$$\begin{aligned}
\int_{\mathbb{R}^+} f_i^{\alpha} f_j^{1-\alpha}\, dx &= \left(\frac{\lambda_j}{\lambda_i}\right)^{k(\alpha-1)} \lambda_i^{-k} \int_{\mathbb{R}^+} k x^{k-1} \exp\left(-\xi x^k\right)\, dx \\
&= \left(\frac{\lambda_j}{\lambda_i}\right)^{k(\alpha-1)} \frac{\lambda_i^{-k}}{\xi} \int_{\mathbb{R}^+} e^{-y}\, dy\ , \quad y = \xi x^k \\
&= \left(\frac{\lambda_j}{\lambda_i}\right)^{k(\alpha-1)} \frac{\lambda_i^{-k}}{\xi} \\
&= \left(\frac{\lambda_j}{\lambda_i}\right)^{k(\alpha-1)} \lambda_i^{-k} \frac{\left(\lambda_i \lambda_j\right)^k}{\lambda_0}.
\end{aligned}$$

Then,

$$D_\alpha(f_i||f_j) = \frac{1}{\alpha-1} \ln\left[\left(\frac{\lambda_j}{\lambda_i}\right)^{k(\alpha-1)} \frac{\lambda_j^k}{\lambda_0}\right] = \ln\left(\frac{\lambda_j}{\lambda_i}\right)^k + \frac{1}{\alpha-1} \ln\frac{\lambda_j^k}{\lambda_0}\ .$$

Note that for $\alpha \in (0,1)$ we always have $\lambda_0 > 0$ given the positivity of λ_i and λ_j.

- If $\lambda_0 \leq 0$ then $\xi \leq 0$ and

$$\begin{aligned}
\int_{\mathbb{R}^+} f_i^{\alpha} f_j^{1-\alpha} dx &= A_1 \int_{\mathbb{R}^+} x^{k-1} e^{|\xi| x^k}\, dx\ , \quad A_1 > 0 \\
&> A_1 \int_{\mathbb{R}^+} x^{k-1} dx = \infty\ ,
\end{aligned}$$

for all values of k. Finally, since nonpositive λ_0 only occurs for $\alpha > 1$ we have

$$D_\alpha(f_i||f_j) = \frac{1}{\alpha-1} \ln \int_{\mathbb{R}^+} f_i^{\alpha} f_j^{1-\alpha}\, dx = +\infty$$

for this case.

□

Remark B.6.2. For $k_i = k_j = k$,

$$\lim_{\alpha \uparrow 1} D_\alpha(f_i||f_j) = D(f_i||f_j)\ .$$

Proof. Note that setting $k_i = k_j = k$ in the expression for the Kullback divergence, $D(f_i||f_j)$ (Proposition 2.3.27), we obtain

$$D(f_i||f_j) = \ln\left(\frac{\lambda_j}{\lambda_i}\right)^k + \left(\frac{\lambda_i}{\lambda_j}\right)^k - 1\,.$$

Comparing this to the expression for the corresponding Rényi divergence (Proposition B.6.1), it remains to show that

$$\lim_{\alpha\uparrow 1}\frac{1}{\alpha-1}\ln\left(\frac{\lambda_j^k}{\lambda_0}\right) = \left(\frac{\lambda_i}{\lambda_j}\right)^k - 1\,.$$

Since

$$\lim_{\alpha\uparrow 1}\lambda_0 = \lim_{\alpha\uparrow 1}\alpha\lambda_j^k + (1-\alpha)\lambda_i^k = \lambda_j^k$$

we see the limit in question is of indeterminate form. Applying l'Hospital's rule,

$$\lim_{\alpha\uparrow 1}\frac{1}{\alpha-1}\ln\left(\frac{\lambda_j^k}{\lambda_0}\right) = -\lim_{\alpha\uparrow 1}\frac{1}{\lambda_0}\left(\lambda_j^k - \lambda_i^k\right) = \frac{\lambda_i^k - \lambda_j^k}{\lambda_j^k} = \left(\frac{\lambda_i}{\lambda_i}\right)^k - 1\,.$$

□

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition, 1964.

[2] J. Aczél and Z. Daróczy. *On Measures of Information and their Characterization*. Academic Press, 1975.

[3] F. Alajaji, P. N. Chen, and Z. Rached. Csiszár's cutoff rates for the general hypothesis testing problem. *IEEE Transactions on Information Theory*, 50(4):663 – 678, 2004.

[4] M. Ali, L. Rueda, and M. Herrera. On the performance of Chernoff-distance-based linear dimensionality reduction techniques. In Luc Lamontagne and Marion Marchand, editors, *Advances in Artificial Intelligence*, volume 4013 of *Lecture Notes in Computer Science*, pages 467–478. Springer Berlin / Heidelberg, 2006.

[5] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):pp. 131–142, 1966.

[6] C. Arndt. *Information Measures: Information and its Description in Science and Engineering*. Signals and Communication Technology. Springer, 2004.

[7] M. Asadi, N. Ebrahimi, G. G. Hamedani, and E. S. Soofi. Information measures for Pareto distributions and order statistics. In *Advances in Distribution Theory, Order Statistics, and Inference*, pages 207–223. Springer, 2006.

[8] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349 – 369, 1989.

[9] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

[10] J. Burbea. The convexity with respect to Gaussian distributions of divergences of order α. *Utilitas Mathematica*, 26:171–192, 1984.

[11] C. Cachin. *Entropy Measures and Unconditional Security in Cryptography*. PhD thesis, Swiss Federal Institute of Technology Zürich, 1997.

[12] L.L. Campbell. A coding theorem and Rényi's entropy. *Information and Control*, 8(4):423 – 429, 1965.

[13] G. Casella and R. Berger. *Statistical Inference*. Duxbury Resource Center, 2001.

[14] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):pp. 493–507, 1952.

[15] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, 2nd edition, 2006.

[16] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, (8):85–108, 1963.

[17] I. Csiszar. Generalized cutoff rates and Renyi's information measures. *IEEE Transactions on Information Theory*, 41(1):26 –34, 1995.

[18] I. Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10:261–273, September 2008.

[19] G. A. Darbellay and I. Vajda. Entropy expressions for multivariate continuous distributions. *IEEE Transactions on Information Theory*, pages 709–712, 2000.

[20] P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice - Hall International, 1982.

[21] R. Estrada and R. P. Kanwal. *Singular Integral Equations*. Birkäuser, 2000.

[22] G. B. Folland. *Real Analysis - Modern Techniques and Applications*. John Wiley and Sons, 1984.

[23] S. H. Friedberg, A. J. Insel, and L. E. Spence. *Linear Algebra*. Prentice Hall, fourth edition, 2003.

[24] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3):419–435, 2002.

[25] L. Golshani and E. Pasha. Rényi entropy rate for Gaussian processes. *Information Sciences*, 180(8):1486–1491, 2010.

[26] L. Golshani, E. Pasha, and G. Yari. Some properties of Rényi entropy and Rényi entropy rate. *Information Sciences.*, 179(14):2426–2433, 2009.

[27] A. Graham. *Kronecker Products and Matrix Calculus*. Ellis Horwood, 1981.

[28] R. M. Gray. *Toeplitz and Circulant Matrices: A Review*, volume 2 of *Foundations and Trends in Communications and Information Theory*. NOW, 2005.

[29] U. Grenander and G. Szegö. *Toeplitz forms and their applications*. California Monographs in Mathematical Sciences. University of California Press, Berkeley, 1958.

[30] P. Harremoës. Interpretations of Rényi entropies and divergences. *Physica A: Statistical Mechanics and its Applications*, 365(1):57 – 62, 2006.

[31] E. Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.

[32] A. O. Hero, B. Ma, O. Michel, and J. Gorman. Alpha-divergence for classification, indexing and retrieval. Technical report, University of Michigan, 2001.

[33] T. Hobza, D. Morales, and L. Pardo. Rényi statistics for testing equality of autocorrelation coefficients. *Statistical Methodology*, 6(4):424 – 436, 2009.

[34] S. Ihara. *Information Theory for Continuous Systems*. World Scientific, 1993.

[35] P. Jizba and T. Arimitsu. The world according to Renyi: Thermodynamics of multifractal systems. *Annals of Physics*, 312:17, 2004.

[36] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. John Wiley & Sons, 1994.

[37] A. N. Kolmogorov. A new metric invariant of transient dynamical systems and automorphisms. In *in Lebesgue Spaces.Ť Dokl. Akad. Nauk. SSSR 119*, page 861, 1958.

[38] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[39] E. L. Lehmann. *Testing Statistical Hypotheses*. John Wiley and Sons, second edition, 1986.

[40] F. Liese and I. Vajda. *Convex Statistical Distances*. Treubner, 1987.

[41] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394 –4412, 2006.

[42] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37:145–151, 1991.

[43] J. R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2nd edition, 1999.

[44] A. D. Michal. *Matrix and Tensor Calculus*. John Willey & Sons, 1947.

[45] S. Nadarajah and K. Zografos. Formulas for Rényi information and related measures for univariate distributions. *Information Sciences*, 155(1-2):119 – 138, 2003.

[46] M. C. Pardo and J. A. Pardo. Use of Rényi's divergence to test for the equality of the coefficients of variation. *Journal of Computational and Applied Mathematics*, 116(1):93 – 104, 2000.

[47] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Phil. Mag*, 5(50):157–175, 1900.

[48] W. D. Penny. Kl-divergences of normal, gamma, dirichlet and wishart densities. Technical report, Wellcome Department of Cognitive Neurology, 2001.

[49] W. D. Penny and S. Roberts. Bayesian methods for autoregressive models. In *Proceedings of Neural Networks for Signal Processing, Technical report PARG-00-3*, 2000.

[50] Z. Rached, F. Alajaji, and L. L. Campbell. Renyi's divergence and entropy rates for finite alphabet Markov sources. *IEEE Transactions on Information Theory*, 47(4):1553 –1561, 2001.

[51] E. D. Rainville. *Special Functions*. The Macmillan Company, 1960.

[52] T. W. Rauber, T. Braun, and K. Berns. Probabilistic distance measures of the Dirichlet and beta distributions. *Pattern Recognition*, 41(2):637 – 645, 2008.

[53] A. Rényi . On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561, 1961.

[54] K. F. Riley, M. P. Hobson, and S. J. Bence. *Mathematical Methods for Physics and Engineering*. Cambridge University Press, third edition, 2006.

[55] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, 1967.

[56] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.

[57] K. Song. Rényi information, loglikelihood and an intrinsic distribution measure. *Journal of Statistical Planning and Inference*, 93(1-2):51 – 69, 2001.

[58] K. Strimmer. Statistical Thinking: Introduction to Probabilistic Data Analysis. Website, 2010.
http://www.uni-leipzig.de/ strimmer/lab/statisticalthinking/pdf/c4.pdf.

[59] I. Vajda. *Theory of Statistical Inference and Information*. Kluwer, Dordrecht, 1989.

[60] T. van Erven and P. Harremoës. Rényi divergence and majorization. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1335 – 1339, june 2010.

[61] J.H. van Schuppen and A. A. Stoorvogel. System identification with information theoretic criteria. Technical report, In: Identification, adaptation, learning (S. Bittanti and Picci G.), 1996.

[62] K. Zografos and S. Nadarajah. Expressions for Rényi and Shannon entropies for multivariate distributions. *Statistics & Probability Letters*, 71(1):71 – 84, 2005.