A precise bare simulation approach to the minimization of some distances, 1. Foundations

Michel Broniatowski, Wolfgang Stummer

A precise bare simulation approach to the minimization of some distances. Foundations

Michel Broniatowski and Wolfgang Stummer

Abstract

In information theory — as well as in the adjacent fields of statistics, machine learning, artificial intelligence, signal processing and pattern recognition — many *flexibilizations* of the omnipresent Kullback-Leibler information distance (relative entropy) and of the closely related Shannon entropy have become frequently used tools. To tackle corresponding constrained minimization (respectively maximization) problems by a newly developed *dimension-free bare (pure) simulation* method, is the main goal of this paper. Almost no assumptions (like convexity) on the set of constraints are needed, within our discrete setup of arbitrary dimension, and our method is precise (i.e., converges in the limit). As a side effect, we also derive an innovative way of constructing new useful distances/divergences. To illustrate the core of our approach, we present numerous examples. The potential for wide-spread applicability is indicated, too; in particular, we deliver many recent references for uses of the involved distances/divergences and entropies in various different research fields (which may also serve as an interdisciplinary interface).

Index Terms

f-divergences of Csiszar-Ali-Silvey-Morimoto type, power divergences, Kullback-Leibler information distance, relative entropy, Renyi divergences, Bhattacharyya distance, Jensen-Shannon divergence/distance, alpha-divergences, Shannon entropy, Renyi entropies, Bhattacharyya coefficient, Tsallis (cross) entropies, Cressie-Read measures, Hellinger distance, Euclidean norms, generalized maximum entropy method, importance sampling, fuzzy divergences.

July 04, 2021

I. Introduction

DIRECTED (i.e., not necessarily symmetric) distances $D(\mathbf{P},\mathbf{Q})$ between two finite discrete[1] (probability) distributions $\mathbf{P},\mathbf{Q}$ or between two general Euclidean vectors $\mathbf{P},\mathbf{Q}$ are known as *divergences*; they serve as important (dis)similarity measures, proximity measures and discrepancy measures in various different research areas such as information theory, statistics, artificial intelligence, machine learning, signal processing, pattern recognition, physics, finance, etc.[2]. Besides Bregman distances/divergences (with which we do not deal here), another major class are the $\varphi-$*divergences* $D_{\varphi}(\mathbf{P},\mathbf{Q})$ *of Csiszar-Ali-Silvey-Morimoto* (in short CASM $\varphi-$divergences, cf. [94], [11], [266]). The latter covers — with corresponding choices of φ — e.g. the omnipresent *Kullback-Leibler information distance/divergence* [200] (also known as relative entropy), the *Jensen-Shannon distance/divergence*, as well as the *power divergences* (also known as alpha-divergences, Cressie-Read measures, and Tsallis cross-entropies). For some comprehensive overviews on CASM $\varphi-$divergences, the reader is referred to the insightful books of e.g. Liese & Vajda [217], Read & Cressie [303], Vajda [371], Csiszar & Shields [99], Stummer [344], Pardo [282], Liese & Miescke [216], the survey articles of e.g. Liese & Vajda [218], Vajda & van der Meulen [374], Reid & Williamson [304], Basseville [34], and the references therein; an imbedding of CASM $\varphi-$divergences to more general frameworks can be found e.g. in Stummer & Vajda [350], Broniatowski & Vajda [65], Stummer & Kißlinger [346] and Broniatowski & Stummer [64].

Frequently used special cases of the above-mentioned power divergences are e.g. the (squared) *Hellinger distance*, the *Pearson chi-square divergence*, and the *Neyman chi-square divergence*. Moreover, several deterministic transformations of power divergences are also prominently used in research, most notably the *Bhattacharyya distance* [48] and the more general *Renyi divergences* [309] (also known as Renyi cross-entropies); a comprehensive exposition of the latter is given e.g. in van Erven & Harremoes [380]. Some other important deterministic transformations of power divergences include the *Bhattacharyya coefficient* (cf. [48],[49],[50]) — which is also called *affinity* (cf. Matusita [256]) and *fidelity similarity* (cf. e.g. Deza & Deza [113]) — as well as the *Bhattacharyya arccos distance* (cf. [50]) and the *Fisher distance* (also known as *Rao distance*, *geodesic distance*, cf. e.g. Deza & Deza [113]). As shown below, by further explicit transformations we can also recover Sundaresan's divergence [352] [353].

M. Broniatowski is with the LPSM, Sorbonne Université, 4 place Jussieu, 75252 Paris, France. ORCID 0000-0001-6301-5531.

W. Stummer is with the Department of Mathematics, University of Erlangen–Nürnberg, Cauerstrasse 11, 91058 Erlangen, Germany; e-mail: stummer@math.fau.de . ORCID 0000-0002-7831-4558. Corresponding author.

[1] for reasons of technicality, in this paper we only deal with such kind of distributions; for instance, these can be also achieved from more involved systems by quantizations of observations represented by finite partitions of the observation/data space, or by making use of the dual representation for CASM $\varphi-$divergences (cf. Liese & Vajda [218], Broniatowski & Keziou [61]).

[2] since there exists a vast literature on divergences and connected entropies in these fields, for the sake of brevity we will give in this introduction only some basic references; many corresponding concrete applications will be mentioned in the following sections, in the course of the method-illuminating examples.

The *minimization* $\inf_{\mathbf{Q}\in\mathbf{\Omega}} D(\mathbf{Q},\mathbf{P})$ of divergences from one distribution (respectively, its equivalent vector of frequencies) $\mathbf{P}$ to an appropriate set $\mathbf{\Omega}$ of distributions (frequency vectors) appears in a natural way in various different contexts, as indicated in the following. For instance, let $\mathbf{P} = \mathbf{P}_{true}$ be the *true* distribution of a mechanism which generates non-deterministic data and $\mathbf{\Omega}$ is a pregiven *model* in the sense of a (parametric or non-parametric) family of distributions which serves as an "approximation" (in fact, a collection of approximations) of the "truth" $\mathbf{P}_{true}$. If $\mathbf{P}_{true} \notin \mathbf{\Omega}$ — e.g. since $\mathbf{\Omega}$ reflects some simplifications of $\mathbf{P}_{true}$ which is in line with the general scientific procedure — then the positive quantity $\Phi_{\mathbf{P}_{true}}(\mathbf{\Omega}) := \inf_{\mathbf{Q}\in\mathbf{\Omega}} D(\mathbf{Q},\mathbf{P}_{true})$ can be used as an *index of model adequacy* in the sense of a degree of departure between the model and the truth (cf. Lindsay [222], see also e.g. Lindsay et al. [223], Markatou & Sofikitou [248], Markatou & Chen [249]); small index values should indicate high adequacy. If $\mathbf{P}_{true} \in \mathbf{\Omega}$, then $\Phi_{\mathbf{P}_{true}}(\mathbf{\Omega}) = 0$ which corresponds to full adequacy. This index of model adequacy $\Phi_{\mathbf{P}_{true}}(\mathbf{\Omega})$ can also be seen as *index of goodness/quality of approximation to the truth* or as *model misspecification error*, and it can be used for model assessment as well as for model search (model selection, model hunting) by comparing the indices $\Phi_{\mathbf{P}_{true}}(\mathbf{\Omega_1}), \Phi_{\mathbf{P}_{true}}(\mathbf{\Omega_2}), \ldots$ of competing models $\mathbf{\Omega_1}, \mathbf{\Omega_2}, \ldots$ and choosing the one with the smallest index; this idea can be also used for classification (e.g. analogously to Bilik & Khomchuk [54] who deal with continuous (rather than discrete) distributions) where the $\mathbf{\Omega_i}$ are interpreted as (possibly data-derived but fixed) classes which are disjoint and non-exhaustive.

Typically, in statistical analyses the true distribution $\mathbf{P}_{true}$ is unknown and is either replaced by a hypothesis-distribution $\mathbf{P}_{hyp}$ or by a distribution $\mathbf{P}_{data}$ derived from data (generated by $\mathbf{P}_{true}$) which converges to $\mathbf{P}_{true}$ as the data/sample size tends to infinity (e.g. $\mathbf{P}_{data}$ may be the well-known empirical distribution or a conditional distribution). Correspondingly, $\Phi_{\mathbf{P}_{data}}(\mathbf{\Omega})$ reflects a data-derived approximation (estimate) of the index of model adequacy (resp. of the model misspecification error) from which one can cast corresponding model-adequacy tests and related goodness-of-fit tests. Moreover, for the case of i.i.d. data-generation and $\mathbf{P}_{data}$ to be the corresponding empirical distribution, the (not necessarily existent or unique) best-model-member/element choice $\arg\min_{\mathbf{Q}\in\mathbf{\Omega}} D(\mathbf{Q},\mathbf{P}_{data})$ amounts to the well-known *minimum distance estimator* which for the Kullback-Leibler information divergence D is equal to the omnipresent *maximum likelihood estimator*; for comprehensive surveys on divergence-based statistical testing and estimation, the reader is referred to e.g. the references in the second half of the first paragraph in the current introduction.

Most of the above-mentioned considerations also hold for deterministic (rather than non-deterministic) frameworks where $\mathbf{P}$ is a general Euclidean vector (rather than a probability-distribution describing frequency vector in the probability simplex) and $\mathbf{\Omega}$ is a model in the sense of a family of general Euclidean vectors (which may be encodings of more complicated context descriptions).

Returning to the general context, let us mention that from CASM $\varphi-$divergences one can also derive the widely used *$\varphi-$entropies* $\mathcal{E}_{\varphi}(\mathbf{Q})$ of a distribution $\mathbf{Q}$ (and non-probability versions thereof) in the sense of Burbea & Rao [68] (see also Csiszar [95], Ben-Bassat [38], Ben-Tal & Teboulle [40], Kesavan & Kapur [187], Dacunha-Castelle & Gamboa [102], Teboulle & Vajda [357], Gamboa & Gassiat [132], Vajda & Zvarova [376]); these entropies can e.g. be basically constructed from $D_{\varphi}(\mathbf{Q},\mathbf{P}^{unif})$ where $\mathbf{P}^{unif}$ denotes the uniform distribution. Moreover, by use of certain deterministic transformations h one can also deduce the more general *$(h,\varphi)-$entropies* (and non-probability versions thereof) in the sense of Salicru et al. [314] (see also e.g. Pardo [282]). As will be worked out in detail in Subsection IV-C below, from this one can deduce as special cases a variety of prominently used quantities in research, such as for instance:

- the omnipresent *Shannon entropy* [328], the *$\gamma-$order Renyi entropy* [309], the *$\gamma-$order entropy of Havrda-Charvat* [157] (also called non-additive *$\gamma-$order Tsallis entropy* [363] in statistical physics), the *$\widetilde{\gamma}-$order entropy of Arimoto* [16], Vajda's quadratic entropy [371], *Sharma-Mittal entropies* [329],
- the Euclidean *$\gamma-$norms*, as well as
- measures of diversity, heterogeneity and unevenness, like the *Gini-Simpson diversity index*, the *diversity index of Hill* [160], the *Simpson-Herfindahl index* (which is also known as *index of coincidence*, cf. Harremoes & Topsoe [155] and its generalization in Harremoes & Vajda [156]), the *diversity index of Patil & Taillie* [286], the *$\gamma-$mean heterogeneity index* (see e.g. van der Lubbe [379]); see also Nayak [271] and Jost [176] for some interrelations with the above-mentioned entropies.

Given that the constraint set $\mathbf{\Omega}$ reflects some *incomplete/partial* information about a system (e.g. moment constraints), the maximization over $\mathbf{Q} \in \mathbf{\Omega}$ of the above-mentioned entropies, norms and diversity indices (and the more general *$(h,\varphi)-$entropies*) is important for many research topics, most notably manifested in Jaynes's [165],[166] omnipresent, "universally applicable" *maximum entropy principle* (which employs the Shannon entropy), and its generalizations (see e.g. the books of Kapur [184], Kapur & Kesavan [185], Arndt [17], and Gzyl et al. [150] for comprehensive surveys).

Besides the above-mentioned principal overview, let us now briefly discuss some existing *technical issues* for the minimization of CASM $\varphi-$divergences $\Phi_{\mathbf{P}}(\mathbf{\Omega}) := \inf_{\mathbf{Q}\in\mathbf{\Omega}} D_{\varphi}(\mathbf{Q},\mathbf{P})$. For (not necessarily discrete) probability distributions/measures $\mathbf{P}$ and sets $\mathbf{\Omega}$ of probability distributions/measures satisfying a finite set of linear equality constraints, $\Phi_{\mathbf{P}}(\mathbf{\Omega})$ has been characterized in Csiszar [96] and more recently by Csiszar & Matus [98], Broniatowski & Keziou [60], Leonard [213], and Pelletier [288] among others, in various contexts; those results extend to inequality constraints. Minimizations of $\gamma-$order Renyi divergences on $\gamma-$convex sets $\mathbf{\Omega}$ are studied e.g. in Kumar & Sason [202], whereas Kumar & Sundaresan [203] [204] investigate minimizations of Sundaresan's divergence on certain convex sets $\mathbf{\Omega}$.

To our knowledge, no general representation for $\Phi_{\mathbf{P}}(\mathbf{\Omega})$ for a positive distribution/measure $\mathbf{P}$ (respectively, for an Euclidean vector with positive components) and a general set $\mathbf{\Omega}$ of signed measures (respectively, of Euclidean vector with components of arbitrary sign) exists. At the contrary, many algorithmic approaches for such minimization problems have been proposed; they mostly aim at finding minimizers more than at the evaluation of the minimum divergence itself, which is obtained as a by-product. Moreover, it is well-known that such kind of CASM $\varphi-$divergence minimization problems may be hard to tackle or even intractable via usual methods such as the omnipresent gradient descent method and versions thereof, especially for non-parametric or semi-parametric $\mathbf{\Omega}$ in sufficiently high-dimensional situations. For instance, $\mathbf{\Omega}$ may consist (only) of constraints on moments or on L-moments (see e.g. Broniatowski & Decurninge [59]); alternatively, $\mathbf{\Omega}$ may be e.g. a tubular neighborhood of a parametric model (see e.g. Liu & Lindsay [225], Ghosh & Basu [135]). The same intractability problem holds for the above-mentioned $(h, \varphi)-$entropy maximization problems. In the light of this, the goals of this paper are:

- to solve constrained minimization problems of a large range of CASM $\varphi-$divergences and deterministic transformations thereof (respectively constrained maximization problems of $(h, \varphi)-$entropies including Euclidean norms and diversity indices), by means of a newly developed *dimension-free bare (pure) simulation* method which is precise (i.e., converges in the limit) and which needs almost no assumptions (like convexity) on the set $\mathbf{\Omega}$ of constraints; in doing so, for the sake of brevity we concentrate on finding/computing the minimum divergences themselves rather than the corresponding minimizers (to achieve the latter, e.g. dichotomous search could be used in a subsequent step, however);
- to derive a method of constructing new useful distances/divergences;
- to present numerous examples in order to illuminate our method and its potential for wide-spread applicability; as we go along, we also deliver many recent references for uses of the outcoming distances/divergences and entropies (covering in particular all the above-mentioned ones).

This agenda is achieved in the following way. In the next Section II, we briefly introduce the principal idea of our new bare-simulation optimization paradigm. After manifesting the fundamentally employed class of CASM $\varphi-$divergences in Section III, we give in Section IV the main cornerstones, construction principles and theorems, for deterministic as well as for statistical divergence-minimization problems; the maximization of generalized entropies is addressed, too. Section V deals with the concrete determination of the involved simulation-weights, as well as with the interrelated issue of creating associated CASM $\varphi-$divergences. Some sampling-concerning details for the principal implementation of our bare-simulation optimization approach are worked out in Section VI. The main proofs are presented in the appendices.

A first simulation-based algorithm in vein with the present proposal has been developed by Broniatowski [58], in the restricted setup of risk estimation for power divergences. The present paper extends this considerably by considering general CASM $\varphi-$divergences and related entropies, and by dealing with corresponding general optimization problems, of both deterministic respectively stochastic type.

II. A NEW MINIMIZATION PARADIGM

We concern with minimization problems of the following type, where $\mathcal{M}$ is a topological space and $\mathcal{T}$ is the Borel $\sigma-$field over a given base on $\mathcal{M}$; e.g. take $\mathcal{M} = \mathbb{R}^K$ to be the $K-$dimensional Euclidean space equipped with the Borel $\sigma-$ field $\mathcal{T}$.

Definition 1: A measurable function $\Phi : \mathcal{M} \mapsto [0, \infty]$ and measurable set $\Omega \subset \mathcal{M}$ [3] are called "bare-simulation minimizable" (BS-minimizable) respectively "bare-simulation maximizable" (BS-maximizable) if for

$$\Phi(\Omega) := \inf_{Q \in \Omega} \{\Phi(Q)\} < \infty \qquad \text{respectively} \qquad \Phi(\Omega) := \sup_{Q \in \Omega} \{\Phi(Q)\} < \infty \tag{1}$$

there exists a measurable function $G : [0, \infty[\mapsto [0, \infty[$ as well as a sequence $((\mathfrak{X}_n, \mathcal{A}_n, \Pi_n))_{n \in \mathbb{N}}$ of probability spaces and on them a sequence $(\xi_n)_{n \in \mathbb{N}}$ [4] of $\mathcal{M}-$valued random variables such that

$$G\left(- \lim_{n \to \infty} \frac{1}{n} \log \Pi_n[\xi_n \in \Omega]\right) = \inf_{Q \in \Omega} \Phi(Q) = \Phi(\Omega) \tag{2}$$

respectively

$$G\left(- \lim_{n \to \infty} \frac{1}{n} \log \Pi_n[\xi_n \in \Omega]\right) = \sup_{Q \in \Omega} \Phi(Q) = \Phi(\Omega); \tag{3}$$

[3] i.e. $\Omega \in \mathcal{T}$

[4] in order to emphasize the dependence on Φ, one should use the notations $(\xi_{\Phi,n})_{n \in \mathbb{N}}$, $\Pi_{\Phi,n}$, etc.; this is avoided for the sake of a better readability.

in situations where Φ is fixed and different Ω's are considered, we say that "Φ is bare-simulation minimizable (BS-minimizable) on Ω" respectively "Φ is bare-simulation maximizable (BS-maximizable) on Ω".

Remark 2: (a) Even in situations where one can uniformly choose $(\mathfrak{X}_n, \mathcal{A}_n, \Pi_n) \equiv (\widetilde{\mathfrak{X}}, \widetilde{\mathcal{A}}, \widetilde{\Pi})$, the sequence $(\xi_n)_{n\in\mathbb{N}}$ may be not "independent and identically distributed" .
(b) Throughout the paper, we shall mainly deal with $BS-$minimizability.

The basic idea/incentive of this new approach is: if a minimization problem (1) has no explicit solution and is computationally intractable (or unfeasible) but can be shown to be BS-minimizable with concretely constructable $(\xi_n)_{n\in\mathbb{N}}$ and $(\Pi_n)_{n\in\mathbb{N}}$, then one can basically simulate the log-probabilities $-\frac{1}{n}\Pi_n[\xi_n \in \Omega]$ for large enough integer $n \in \mathbb{N}$ to obtain an approximation of (1) without having to evaluate the corresponding (not necessarily unique) minimizer, where the latter is typically time-costly. Finding minimizers can be performed through dichotomic search, once an algorithm leading to the minimal value of the divergence on adequate families of sets Ω is at hand; for the sake of brevity, this is omitted in the current paper.

For reasons of transparency, we start to demonstrate this approach for the following important/prominent class of minimization problems with the following components:

(i) $\mathcal{M}$ is the $K-$dimensional Euclidean space $\mathbb{R}^K$, i.e. Ω is a set of vectors Q with a number of K components (where K may be huge, as it is e.g. the case in big data contexts);
(ii) $\Phi(\cdot) := \Phi_P(\cdot)$ depends on some known vector P in $\mathbb{R}^K$ with K nonnegative components;
(iii) $\Phi_P(\cdot)$ is a "directed distance" (divergence) from P into Ω in the sense of $\Omega \ni Q \mapsto \Phi_P(Q) := D(Q,P)$, where $D(\cdot,\cdot)$ has the the two properties "$D(Q,P) \geq 0$" and "$D(Q,P) = 0$ if and only if $Q = P$". In particular, $D(\cdot,\cdot)$ needs neither satisfy the symmetry $D(Q,P) = D(P,Q)$ nor the triangular inequality.

In other words, (1) together with (i)-(iii) constitutes a distance/divergence-minimization problem; we design a "universal" method to solve such problems by constructing appropriate (cf.(2)) sequences $(\xi_n)_{n\in\mathbb{N}}$ of $\mathbb{R}^K-$valued random variables, for all directed distances $D(\cdot,\cdot)$ from a large subclass of the important omnipresent Csiszar-Ali-Silvey-Morimoto CASM $\varphi-$divergences (also called $f-$divergences).

As a second demonstration for the workability of our paradigm, we "extend" (i) to (iii) to the setup where P is a *random* element of the simplex $\mathbb{S}^K$ of $K-$component probability (frequency) vectors (cf. the exact definition below) and $\Omega \subset \mathbb{S}^K$; for the sub-setup where P corresponds to a data-observation-dependent probability distribution and Ω corresponds to a pregiven *model* in the sense of a family of probability distributions, the formula (1) amounts to the corresponding (discrete) "minimization-distance estimation problem (MDEP)" of choosing the best model element/member under given data [5]. This is important/prominent in statistics and in the adjacent research fields of artificial intelligence and machine learning; the concrete solving of the MDEP is especially "hard" for nonparametric respectively semiparametric problems, and our BS method is predestined for such kind of contexts.

III. Directed distances

In detail, concerning the above-mentioned point (i) we take the $K-$ dimensional Euclidean space $\mathcal{M} = \mathbb{R}^K$, denote from now on — as usual — its elements (i.e. vectors) in boldface letters, and also employ the subsets

$$\mathbb{R}^K_{\neq 0} := \{\mathbf{Q} := (q_1,\ldots,q_K) \in \mathbb{R}^K : q_i \neq 0 \text{ for all } i = 1,\ldots,K\},$$
$$\mathbb{R}^K_{>0} := \{\mathbf{Q} := (q_1,\ldots,q_K) \in \mathbb{R}^K : q_i > 0 \text{ for all } i = 1,\ldots,K\},$$
$$\mathbb{R}^K_{\geq 0} := \{\mathbf{Q} := (q_1,\ldots,q_K) \in \mathbb{R}^K : q_i \geq 0 \text{ for all } i = 1,\ldots,K\},$$
$$\mathbb{R}^K_{\leq 0} := \{\mathbf{Q} := (q_1,\ldots,q_K) \in \mathbb{R}^K : q_i \leq 0 \text{ for all } i = 1,\ldots,K\},$$
$$\mathbb{S}^K := \{\mathbf{Q} := (q_1,\ldots,q_K) \in \mathbb{R}^K_{\geq 0} : \sum_{i=1}^K q_i = 1\} \quad \text{(simplex of probability vectors)},$$
$$\mathbb{S}^K_{>0} := \{\mathbf{Q} := (q_1,\ldots,q_K) \in \mathbb{R}^K_{>0} : \sum_{i=1}^K q_i = 1\}.$$

Concerning the directed distances $D(\cdot,\cdot)$ in (ii) and (iii), we deal with the important omnipresent Csiszar-Ali-Silvey-Morimoto $\varphi-$ divergences (CASM $\varphi-$divergences) — adapted to our context:

Definition 3:
(a) Let the "divergence-generator" be a lower semicontinuous convex function $\varphi :]-\infty,\infty[\to [0,\infty]$ satisfying $\varphi(1) = 0$. Furthermore, for the effective domain $dom(\varphi) := \{t \in \mathbb{R} : \varphi(t) < \infty\}$ we assume that its interior $int(dom(\varphi))$ is non-empty which implies that $int(dom(\varphi)) =]a,b[$ for some $-\infty \leq a < 1 < b \leq \infty$. Moreover, we suppose that φ is strictly convex in

[5] an alternative naming also used in literature is to call Ω a model class (rather than model), and each $P \in \Omega$ a model (rather than model element)

a non-empty neighborhood $]t_-^{sc}, t_+^{sc}[\subseteq]a,b[$ of one ($t_-^{sc} < 1 < t_+^{sc}$). Also, we set $\varphi(a) := \lim_{t\downarrow a}\varphi(t)$ and $\varphi(b) := \lim_{t\uparrow b}\varphi(t)$ (these limits always exist). The class of all such functions φ will be denoted by $\widetilde{\Upsilon}(]a,b[)$. A frequent choice is e.g. $]a,b[=]0,\infty[$ or $]a,b[=]-\infty,\infty[$.

(b) For $\varphi \in \widetilde{\Upsilon}(]a,b[)$, $\mathbf{P} := (p_1,\ldots,p_K) \in \mathbb{R}_{\geq 0}^K$ and $\mathbf{Q} := (q_1,\ldots,q_K) \in \mathbf{\Omega} \subset \mathbb{R}^K$, we define the Csiszar-Ali-Silvey-Morimoto $\varphi-$divergence

$$\Phi_{\mathbf{P}}(\mathbf{Q}) := D_{\varphi}(\mathbf{Q},\mathbf{P}) := \sum_{k=1}^{K} p_k \cdot \varphi\left(\frac{q_k}{p_k}\right) \geq 0. \tag{4}$$

As usual, in (4) we employ the three conventions that $p\cdot\varphi\left(\frac{0}{p}\right) = p\cdot\varphi(0) > 0$ for all $p>0$, and $0\cdot\varphi\left(\frac{q}{0}\right) = q\cdot\lim_{x\to\infty}\frac{\varphi(x\cdot\mathrm{sgn}(q))}{x\cdot\mathrm{sgn}(q)} > 0$ for $q \neq 0$ (employing the sign of q), and $0\cdot\varphi\left(\frac{0}{0}\right) := 0$. Throughout the paper, we only consider constellations $(\varphi,\mathbf{P},\mathbf{\Omega})$ for which the very mild condition

$$\Phi_{\mathbf{P}}(\Omega) := \inf_{\mathbf{Q}\in\mathbf{\Omega}} D_{\varphi}(\mathbf{Q},\mathbf{P}) \neq \infty$$ [6]

holds.

For probability vectors $\mathbb{P}$ and $\mathbb{Q}$ in $\mathbb{S}^K$, the $\varphi-$divergences $D_{\varphi}(\mathbb{Q},\mathbb{P})$ were introduced by Csiszar [94], Ali & Silvey [11] and Morimoto [266] (where the first two references even deal with more general probability distributions); for some comprehensive overviews — including statistical applications to goodness-of-fit testing and minimum distance estimation — the reader is referred to the insightful books of e.g. Liese & Vajda [217], Read & Cressie [303], Vajda [371], Csiszar & Shields [99], Stummer [344], Pardo [282], Liese & Miescke [216], the survey articles of e.g. Liese & Vajda [218], Vajda & van der Meulen [374], Reid & Williamson [304], Basseville [34], and the references therein. Some exemplary recent studies and applications of CASM $\varphi-$ divergences appear e.g. in Qiao & Minematsu [298] for invariances in speech recognition, Nguyen et al. [273] in connection with empirical risk optimization, Feixas et al. [124] for various different image processing tasks, Luo et al. [232] for video clip segmentation and key frame generation, Kißlinger & Stummer [189] for model preselection (structure detection) in the context of nonlinear recursive models with additional exogenous inputs, Mahboubi & Kochenderfer [241] within a context of traffic-pattern learning from flight tracks, Guo et al. [145] for local contrastive descriptors in image classification through e.g. regional color distributions, Csiszar & Breuer [100] for modelling generalized-ball type constraints in expectation minimization problems, Kißlinger & Stummer [191] for the detection of distributional changes in random data (streams and clouds), Noh et al. [275] within a context of generative local metric learning for nearest neighbor classification, Yu et al. [416] for adversarial learning within oil spill segmentation, Arslan [19] for automated active reconfiguration in mobile sensor networks, Sason [320] in connection with with data-processing and majorization inequalities, Ciftci et al. [89] for the optimization of multienergy microgrids in energy infrastructure systems, and Stummer [345] for solving some new optimal transport (OT) problems which flexibilize some Wasserstein-distance based OTs.

For the setup of $D_{\varphi}(\mathbf{Q},\mathbf{P})$ for vectors $\mathbf{P}$, $\mathbf{Q}$ with non-negative components the reader is referred to e.g. Stummer & Vajda [349] (who deal with even more general nonnegative measures and giving some statistical as well as information-theoretic applications) and Gietl & Reffel [137] (including applications to iterative proportional fitting). The case of $\varphi-$divergences for vectors with arbitrary components can be extracted from e.g. Broniatowski & Keziou [60] who actually deal with finite *signed* measures. For a comprehensive technical treatment, see also Broniatowski & Stummer [64].

Clearly, from (4) it is obvious that in general $D_{\varphi}(\mathbf{Q},\mathbf{P}) \neq D_{\varphi}(\mathbf{P},\mathbf{Q})$ (non-symmetry). Moreover, it is straightforward to deduce that $D_{\varphi}(\mathbf{Q},\mathbf{P}) = 0$ if and only if $\mathbf{Q}=\mathbf{P}$ (reflexivity). Very prominent and important examples of CASM $\varphi-$divergences are the *power divergences* in the scaling of e.g. Liese & Vajda [217] (in other scalings also called Rathie & Kannapan's non-additive directed divergences of order γ [302], Cressie-Read divergences [93] [303], relative Tsallis entropies or Tsallis cross-entropies [364] (see also Shiino [331]), Amari's alpha-divergences [12]) where basically (up to technicalities) $\varphi(t) := \varphi_{\gamma}(t) := \frac{t^{\gamma}-\gamma\cdot t+\gamma-1}{\gamma\cdot(\gamma-1)}$ ($\gamma \in \mathbb{R}\backslash\{0,1\}$), $\varphi(t) := \varphi_0(t) := \lim_{\gamma\to 0}\varphi_{\gamma}(t) = -\log t + t - 1$, $\varphi(t) := \varphi_1(t) := \lim_{\gamma\to 1}\varphi_{\gamma}(t) = t\cdot\log t + 1 - t$. Usually, in the literature one takes $t \in]0,\infty[$ (and the limit as $t \to 0$), except for the case $\gamma = 2$ where one handles $t \in]-\infty,\infty[$; for our purposes, we have to essentially extend these divergence generators φ_{γ} for $t<0$, which will be carried out and discussed in detail below, namely in (43), (44) (see also Table 1), as well as at several other places in this paper. Notice that $D_{\varphi_1}(\mathbf{Q},\mathbf{P})$ basically corresponds to the (extended form of) the omnipresent Kullback-Leibler information resp. relative entropy. Below, we shall also consider the minimization/maximization of *important transforms of power divergences* such as Renyi divergences/entropies, Sundaresan's divergence, etc., which are frequently used in information theory and its applications to e.g. artificial intelligence, machine learning, and physics.

[6] i.e. $dom(\varphi)$ covers (at least) a non-empty part of $\{1\} \cup \mathcal{R}\big(\frac{\mathbf{\Omega}}{\mathbf{P}}\big)$, where $\mathcal{R}\big(\frac{\mathbf{\Omega}}{\mathbf{P}}\big) := \big\{\frac{q_k}{p_k} : k \in \{1,\ldots,K\}, \mathbf{Q} := (q_1,\ldots,q_K) \in \Omega\big\}$ is the range of all possible entry-ratios.

Remark 4: Since, in general, our methods work also for *non*-probability vectors $\mathbf{Q}, \mathbf{P}$, we can also deal with — plain versions and transformations of — *weighted* $\varphi-$*divergences* of the form

$$D_{\varphi}^{wei}(\mathbf{Q}, \mathbf{P}) := \sum_{k=1}^{K} c_k \cdot p_k \cdot \varphi\left(\frac{q_k}{p_k}\right) \geq 0 \tag{5}$$

where $c_k > 0 \quad (k = 1, \ldots, K)$ are weights which not necessarily add up to one. Indeed, by means of (5) we formally end up with

$$\inf_{\mathbf{Q} \in \mathbf{\Omega}} D_{\varphi}^{wei}(\mathbf{Q}, \mathbf{P}) = \inf_{\mathbf{Q}^{wei} \in \mathbf{\Omega}^{wei}} D_{\varphi}(\mathbf{Q}^{wei}, \mathbf{P}^{wei})$$

where $\mathbf{P}^{wei} := (c_1 \cdot p_1, \ldots, c_K \cdot p_K)$, $\mathbf{Q}^{wei} := (c_1 \cdot q_1, \ldots, c_K \cdot q_K)$ and $\mathbf{\Omega}^{wei}$ is the corresponding rescaling of $\mathbf{\Omega}$. Of course, all the necessary technicalities for the $\varphi-$divergences (see below) have to be adapted to the weighted $\varphi-$divergences; for the sake of brevity, this will not be discussed in detail. Notice that $\mathbf{P}^{wei}$, $\mathbf{Q}^{wei}$ are generally not probability vectors anymore, even if $\mathbf{Q}, \mathbf{P}$ are probability vectors. In the latter case, and under the assumption $\sum_{k=1}^{K} c_k = 1$, the divergences (5) coincide with the discrete versions of the *(*$\mathbf{c}-$*)local divergences* of Avlogiaris et al. [22], [23] who also give absolutely-continuous versions and beyond (see also Broniatowski & Stummer [64] for an imbedding in a general divergence framework).

IV. Construction principles, BS-minimizability/amenability

A. The cornerstone

In this Section IV, we show that a number of deterministic optimization problems and and problems in statistical minimum risk based approaches pertaining to non- or semi-parametric contexts are BS-minimizable/amenable in the sense of Definition 1. The below-mentioned Sections V and VI will draw conclusions, proposing effective solutions.

For the construction of the desired sequence $(\boldsymbol{\xi}_n)_{n \in \mathbb{N}}$ of $\mathbb{R}^K-$valued random variables (viz. random vectors) and a corresponding probability distribution $\mathbb{\Pi}$ (which will not depend on n), we will assume that the divergence generator $\varphi \in \widetilde{\Upsilon}(]a, b[)$ has the additional property that it can be represented as

$$\varphi(t) = \sup_{z \in \mathbb{R}} \Big(z \cdot t - \log \int_{\mathbb{R}} e^{z \cdot y} d\mathfrak{C}(y) \Big), \qquad t \in \mathbb{R}, \tag{6}$$

for some probability distribution/measure $\mathfrak{C}$ on the real line such that the function $z \mapsto MGF_{\mathfrak{C}}(z) := \int_{\mathbb{R}} e^{z \cdot y} d\mathfrak{C}(y)$ is finite on some open interval containing zero [7]. From this, we shall construct — basically in Section V below — a sequence $(W_n)_{n \in \mathbb{N}}$ of i.i.d. copies of a random variable W whose distribution (under $\mathbb{\Pi}$) is $\mathfrak{C}$ (i.e. $\mathbb{\Pi}[W \in \cdot\,] = \mathfrak{C}[\,\cdot\,]$), from which the desired $(\boldsymbol{\xi}_n)_{n \in \mathbb{N}}$ will be constructed.

Since φ attains its minimal value at the point 1, fit follows that $\varphi'(1) = 0$. By (6), for all t in $int(dom(\varphi))$, $\varphi'(t)$ is the reciprocal of $\psi(z) := (d/dz) \log MGF(z)$ at point t, whence $\psi(0) = 1$, which is to say that the expectation $E_{\mathbb{\Pi}}[W] = 1$.

The class of functions $\varphi \in \widetilde{\Upsilon}(]a, b[)$ satisfying the representability (6) will be denoted by $\Upsilon(]a, b[)$.

Remark 5: The condition $\varphi \in \Upsilon(]a, b[)$ implies that $\mathfrak{C}$ can not be a one-point distribution (Dirac mass) δ_y at some point y, since for such a situation one can straightforwardly deduce from (6) that $\varphi(y) = 0$ and $\varphi(t) = \infty$ for all $t \neq y$, which leads to $int(dom(\varphi)) = \emptyset$ and thus $\varphi \notin \widetilde{\Upsilon}(]a, b[)$ (in fact, our requirement $\varphi(y) = 0$ would narrow down to $y = 1$ anyway).

Let us remark that the class $\Upsilon(]a, b[)$ contains many divergence generators; this together with $\varphi-$construction principles will be developed at length in Section V below. Also, for the minimization problems considered in Section IV-B hereunder, we mostly modify the generator φ into $\widetilde{c} \cdot \varphi$ for strictly positive scales $\widetilde{c}$. At this point, for the sake of transparency, we only present a summarizing Table 1 of a selection of concrete examples which will be treated in detail below:

[7] in particular, this implies that $\mathfrak{C}$ has light tails.

$]a,b[$	$\varphi(t)$ for $t \in]a,b[$	$\mathbf{P} \in$	$\mathbf{Q} \in$	$D_\varphi(\mathbf{Q},\mathbf{P})$, $P=(p_k)_{k=1,\ldots K}$ $\mathbf{Q}=(q_k)_{k=1,\ldots K}$	$]t_-^{sc}, t_+^{sc}[$	$\varphi(a)$	$\varphi(b)$	$\varphi'(a)$ $=\lambda_-$	$\varphi'(b)$ $=\lambda_+$	$supp(\mathfrak{C})$	$\mathfrak{C} = \mathbb{\Pi}[W \in \cdot]$ $= \mathbb{\Pi}[W \in \cdot]$	Details
$]0,\infty[$	$\tilde{c}\cdot\varphi_\gamma(t) := \tilde{c}\cdot\frac{t^\gamma-\gamma\cdot t+\gamma-1}{\gamma\cdot(\gamma-1)}$ for $\gamma<0$, $\tilde{c}>0$	$\mathbb{R}^K_{\geq 0}$	$\mathbb{R}^K_{>0}$	$\tilde{c}\cdot\Big\{\frac{\sum_{k=1}^K (q_k)^\gamma\cdot(p_k)^{1-\gamma}}{\gamma\cdot(\gamma-1)}$ $-\frac{1}{\gamma-1}\cdot\sum_{k=1}^K q_k + \frac{1}{\gamma}\cdot\sum_{k=1}^K p_k\Big\}$	$]0,\infty[$	∞	∞	$-\infty$	$\frac{\tilde{c}}{1-\gamma}$	$]0,\infty[$	tilted stable distribution on $[0,\infty[$	Prop. 22, Ex. 39(a),48(d)
$]0,\infty[$	$\tilde{c}\cdot\varphi_\gamma(t)$ for $\gamma\in]0,1[$, $\tilde{c}>0$	$\mathbb{R}^K_{\geq 0}$	$\mathbb{R}^K_{\geq 0}$	as above	$]0,\infty[$	$\frac{c}{\gamma}$	∞	$-\infty$	$\frac{c}{1-\gamma}$	$[0,\infty[$	Compound $POI(\frac{c}{\gamma})$ $-\,GAM(\frac{\tilde{c}}{1-\gamma},\frac{\gamma}{1-\gamma})$	Prop. 24, Ex. 39(a),48(b)
$]-\infty,\infty[$	$\tilde{c}\cdot\{\varphi_\gamma(t)\cdot\mathbf{1}_{]0,\infty[}(t)$ $+(\frac{1}{\gamma}-\frac{t}{\gamma-1})\cdot\mathbf{1}_{]-\infty,0]}(t)\}$ for $\gamma\in]2,\infty[$, $\tilde{c}>0$	$\mathbb{R}^K_{>0}$	$\mathbb{R}^K$	$\tilde{c}\cdot\Big\{\sum_{k=1}^K\frac{(q_k)^\gamma\cdot(p_k)^{1-\gamma}}{\gamma\cdot(\gamma-1)}\cdot\mathbf{1}_{[0,\infty[}(q_k)$ $-\frac{1}{\gamma-1}\cdot\sum_{k=1}^K q_k + \frac{1}{\gamma}\cdot\sum_{k=1}^K p_k\Big\}$	$]0,\infty[$	∞	∞	$\frac{\tilde{c}}{1-\gamma}$	∞	$]-\infty,\infty[$	distorted stable distri- bution on $]-\infty,\infty[$	Prop. 27, Ex. 39(a), Ex. 48(e)
$]-\infty,\infty[$	$\tilde{c}\cdot\varphi_2(t)=\tilde{c}\cdot\frac{(t-1)^2}{2}$ for $\tilde{c}>0$	$\mathbb{R}^K_{>0}$	$\mathbb{R}^K$	$\tilde{c}\cdot\sum_{k=1}^K\frac{(q_k-p_k)^2}{2\cdot p_k}$	$]-\infty,\infty[$	∞	∞	$-\infty$	∞	$]-\infty,\infty[$	$NOR(1,\frac{1}{c})$	Prop. 26, Ex. 39(b),48(c)
$]0,\infty[$	$\tilde{c}\cdot\varphi_0(t):=\tilde{c}\cdot\{-\log t$ $+t-1\}$ for $\tilde{c}>0$	$\mathbb{R}^K_{\geq 0}$	$\mathbb{R}^K_{>0}$	$\tilde{c}\cdot\Big\{\sum_{k=1}^K p_k\cdot\log\big(\frac{p_k}{q_k}\big)$ $+\sum_{k=1}^K q_k-\sum_{k=1}^K p_k\Big\}$	$]0,\infty[$	∞	∞	$-\infty$	1	$]0,\infty[$	$GAM(\tilde{c},\tilde{c})$	Prop. 23, Ex. 39(c),48(a)
$]0,\infty[$	$\tilde{c}\cdot\varphi_1(t):=\tilde{c}\cdot\{t\cdot\log t$ $+1-t\}$ for $\tilde{c}>0$	$\mathbb{R}^K_{>0}$	$\mathbb{R}^K_{\geq 0}$	$\tilde{c}\cdot\Big\{\sum_{k=1}^K q_k\cdot\log\big(\frac{q_k}{p_k}\big)$ $-\sum_{k=1}^K q_k+\sum_{k=1}^K p_k\Big\}$	$]0,\infty[$	1	∞	$-\infty$	∞	$\{0,\frac{1}{c},\ldots\}$ $=\frac{1}{c}\cdot\mathbb{N}_0$	$\frac{1}{c}-$fold of $POI(\tilde{c})$	Prop. 25, Ex. 40(a),50(a)
$]0,\infty[$	$\tilde{c}\cdot\{t\cdot\log t$ $+(t+1)\cdot\log(\frac{2}{t+1})\}$, $\tilde{c}>0$	$\mathbb{R}^K_{>0}$	$\mathbb{R}^K_{\geq 0}$	$\tilde{c}\cdot\Big\{\sum_{k=1}^K q_k\cdot\log\big(\frac{2q_k}{q_k+p_k}\big)$ $+\sum_{k=1}^K p_k\cdot\log\big(\frac{2p_k}{q_k+p_k}\big)\Big\}$	$]0,\infty[$	$\tilde{c}\log 2$	∞	$-\infty$	$\tilde{c}\log 2$	$\{0,\frac{1}{c},\ldots\}$ $=\frac{1}{c}\cdot\mathbb{N}_0$	$\frac{1}{c}-$fold of $NB(\tilde{c},\frac{1}{2})$	Ex. 43, Ex. 53, are more general
$]\frac{\beta-1}{\beta},\infty[$ $\beta\in]0,1]$	$\tilde{c}\cdot\frac{(t-1)^2}{2(\beta\cdot t+1-\beta)}$ for $\tilde{c}>0$	$\mathbb{R}^K_{\geq 0}$	$]\frac{\beta-1}{\beta}\mathbf{P},\infty[$ component-wise	$\frac{\tilde{c}}{2}\cdot\sum_{k=1}^K\frac{(q_k-p_k)^2}{\beta\cdot q_k+(1-\beta)\cdot p_k}$	$]\frac{\beta-1}{\beta},\infty[$	∞	∞	$-\infty$	$\frac{\tilde{c}}{2\beta}$	$]\frac{\beta-1}{\beta},\infty[$	modified tilted stable distribution	Ex. 41, Ex. 52
$]z_1,z_2[$ $z_1<1$ $z_2>1$	$\frac{(t-z_1)}{z_2-z_1}\cdot\log\big(\frac{(t-z_1)\cdot p}{(z_2-t)\cdot(1-p)}\big)$ $-\log\big(\frac{(z_2-z_1)\cdot p}{(z_2-t)}\big)$ for $p:=\frac{z_2-1}{z_2-z_1}\in]0,1[$	$\mathbb{R}^K_{>0}$	$]z_1\mathbf{P},z_2\mathbf{P}[$ comp.-wise	$\sum_{k=1}^K\frac{q_k-z_1\cdot p_k}{z_2-z_1}\cdot\log\Big(\frac{p\cdot(q_k-z_1\cdot p_k)}{(1-p)\cdot(z_2\cdot p_k-q_k)}\Big)$ $-\sum_{k=1}^K p_k\cdot\log\Big(\frac{p\cdot(z_2-z_1)\cdot p_k}{z_2\cdot p_k-q_k}\Big)$	$]z_1,z_2[$	$\log\frac{1}{p}$	$\log\frac{1}{1-p}$	$-\infty$	∞	$\{z_1,z_2\}$	$\mathfrak{C}[\{z_1\}]=p$, $\mathfrak{C}[\{z_2\}]=1-p$	Ex. 45, Ex. 54
$]t_-^{sc},t_+^{sc}[$	$\tilde{c}\cdot\varphi_{\alpha,\beta_1,\beta_2}(t):=\mathbf{1}_{]t_-^{sc},\infty[}(t)\cdot\tilde{c}\alpha\cdot\{$ $\frac{\sqrt{4+(\beta_1+\beta_2)^2g^2}-(\beta_1-\beta_2)g-2}{2}$ $+\log\frac{\sqrt{4+(\beta_1+\beta_2)^2g^2}-2}{\beta_1\beta_2g^2}\}$ $+\mathbf{1}_{]-\infty,t_-^{sc}]}(t)\cdot\{\overline{d}-\tilde{c}\beta_1\cdot(t-t_-^{sc})\}$ for $0<\beta_1<\beta_2$, $\alpha,\tilde{c}>0$, $g:=g(t):=\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1}$, $\overline{d}:=\tilde{c}\alpha\cdot\{\frac{3\beta_1-\beta_2}{2(\beta_2-\beta_1)}+\log\frac{2(\beta_2-\beta_1)}{\beta_2}\}$,	$\mathbb{R}^K_{\geq 0}$	$\mathbb{R}^K$	$\tilde{c}\cdot\sum_{k=1}^K p_k\cdot\varphi_{\alpha,\beta_1,\beta_2}\big(\frac{q_k}{p_k}\big)$	$]1-\alpha\cdot\check{d},\infty[$ $\check{d}=\frac{2\beta_1^2-\beta_1\beta_2+\beta_2^2}{2\beta_1\beta_2(\beta_2-\beta_1)}$	∞	∞	$-\tilde{c}\beta_1$	$\tilde{c}\beta_1$	$]-\infty,\infty[$	law of $\check{\theta}+Z_1-Z_2$ $\check{\theta}:=1+\alpha\cdot\big(\frac{1}{\beta_2}-\frac{1}{\beta_1}\big)$ Z_1,Z_2 independent, $Z_1\sim GAM(\tilde{c}\beta_1,\tilde{c}\alpha)$, $Z_2\sim GAM(\tilde{c}\beta_2,\tilde{c}\alpha)$ (generalized asymmetric Laplace distribution)	Ex. 46(a), Ex. 55
$]t_-^{sc},t_+^{sc}[$	$\tilde{c}\cdot\varphi_{\alpha,\beta_1,\beta_2}(t):=\mathbf{1}_{]-\infty,t_+^{sc}[}(t)\cdot\tilde{c}\alpha\cdot\{$ $\frac{\sqrt{4+(\beta_1+\beta_2)^2g^2}-(\beta_1-\beta_2)g-2}{2}$ $+\log\frac{\sqrt{4+(\beta_1+\beta_2)^2g^2}-2}{\beta_1\beta_2g^2}\}$ $+\mathbf{1}_{[t_+^{sc},\infty[}(t)\cdot\{\overline{d}+\tilde{c}\beta_2\cdot(t-t_+^{sc})\}$ for $\beta_1>\beta_2>0$, $\alpha,\tilde{c}>0$, $g:=g(t):=\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1}$, $\overline{d}:=\tilde{c}\alpha\cdot\{\frac{3\beta_2-\beta_1}{2(\beta_1-\beta_2)}+\log\frac{2(\beta_1-\beta_2)}{\beta_1}\}$,	$\mathbb{R}^K_{\geq 0}$	$\mathbb{R}^K$	$\tilde{c}\cdot\sum_{k=1}^K p_k\cdot\varphi_{\alpha,\beta_1,\beta_2}\big(\frac{q_k}{p_k}\big)$	$]-\infty,1+\alpha\cdot\check{d}[$ $\check{d}=\frac{\beta_1^2-\beta_1\beta_2+2\beta_2^2}{2\beta_1\beta_2(\beta_1-\beta_2)}$	∞	∞	$-\tilde{c}\beta_2$	$\tilde{c}\beta_2$	$]-\infty,\infty[$	as above	Ex. 46(b), Ex. 55
$]-\infty,\infty[$	$\tilde{c}\cdot\varphi_{\alpha,\beta}(t):=\tilde{c}\cdot\alpha\cdot\{\sqrt{1+\beta^2g^2}$ $-1+\log\frac{2\cdot(\sqrt{1+\beta^2g^2}-1)}{\beta^2g^2}\}$ for $\beta,\alpha,\tilde{c}>0$ and $g:=g(t):=\frac{1-t}{\alpha}$	$\mathbb{R}^K_{\geq 0}$	$\mathbb{R}^K$	$\tilde{c}\cdot\sum_{k=1}^K p_k\cdot\varphi_{\alpha,\beta}\big(\frac{q_k}{p_k}\big)$	$]-\infty,\infty[$	∞	∞	$-\tilde{c}\beta$	$\tilde{c}\beta$	$]-\infty,\infty[$	as above, but for $\check{\theta}=1$, $Z_1\sim GAM(\tilde{c}\beta,\tilde{c}\alpha)$, $Z_2\sim GAM(\tilde{c}\beta,\tilde{c}\alpha)$	Ex. 46(c), Ex. 55

Table 1. Selection of concrete examples treated in this paper, included with some important features (to be explained in the course of method build-up).

As already explained above, the representability (6) is *the* cornerstone for our approach, and opens the gate to make use of simulation methods in appropriate contexts. We first develop this approach for *deterministic* minimization problems (cf. Subsection IV-B); thereafter, in Subsection IV-C, we "extend" this to the setup where $\mathbf{P}$ is identified with an unknown probability vector in the simplex $\mathbb{S}^K$ which is supposed to be the limit (as n tends to infinity) of the empirical distribution pertaining to a collection of observations $\mathbf{X}_n := (X_1, .., X_n)$; in the classical statistical setting, this amounts to the estimation of $\Phi_{\mathbf{P}}(\Omega)$ based on $\mathbf{X}_n$, leading to the important "minimization-distance estimation problem" in statistics, artificial intelligence and machine learning. Finally, we end up this Section IV by shortly dealing with divergences between fuzzy sets (cf. Subsection IV-D) and basic belief assignments (cf. Subsection IV-E).

B. Deterministic minimization problems

Problem 6: For pregiven $\varphi \in \Upsilon(]a, b[)$, positive-entries vector $\mathbf{P} := (p_1, .., p_K) \in \mathbb{R}^K_{>0}$ (or from some subset thereof), and subset $\mathbf{\Omega} \subset \mathbb{R}^K$ (also denoted in boldface letters, with a slight abuse of notation) with regularity properties

$$cl(\mathbf{\Omega}) = cl\left(int\left(\mathbf{\Omega}\right)\right), \qquad int\left(\mathbf{\Omega}\right) \neq \emptyset, \tag{7}$$

find

$$\Phi_{\mathbf{P}}(\mathbf{\Omega}) := \inf_{\mathbf{Q}\in\mathbf{\Omega}} D_\varphi(\mathbf{Q}, \mathbf{P}), \tag{8}$$

provided that

$$\inf_{\mathbf{Q}\in\mathbf{\Omega}} D_\varphi(\mathbf{Q}, \mathbf{P}) < \infty. \tag{9}$$

An immediate consequence thereof is — for pregiven $\varphi \in \Upsilon(]a, b[)$ — the treatment of the more flexible problem

$$\Phi_{\mathbf{P},h}(\mathbf{\Omega}) := \inf_{\mathbf{Q}\in\mathbf{\Omega}} h\Big(D_\varphi(\mathbf{Q}, \mathbf{P})\Big) = h\Big(\inf_{\mathbf{Q}\in\mathbf{\Omega}} D_\varphi(\mathbf{Q}, \mathbf{P})\Big) \tag{10}$$

for any continuous strictly increasing function $h : \mathcal{H} \mapsto \mathbb{R}$ with $\mathcal{H} := [0, \infty[$ and extension $h(\infty) := \sup_{y\in\mathcal{H}}(y)$ (depending on the problem, a sufficiently large $\mathcal{H} \subset [0, \infty[$ may be enough), respectively of

$$\sup_{\mathbf{Q}\in\mathbf{\Omega}} h\Big(D_\varphi(\mathbf{Q}, \mathbf{P})\Big) = h\Big(\inf_{\mathbf{Q}\in\mathbf{\Omega}} D_\varphi(\mathbf{Q}, \mathbf{P})\Big) \tag{11}$$

for any continuous strictly decreasing function $h : \mathcal{H} \mapsto \mathbb{R}$ and extension $h(\infty) := \inf_{y\in\mathcal{H}}(y)$.

Remark 7:
(a) By the basic properties of φ, it follows that for all $c > 0$ the level sets $\varphi_c := \{x \in \mathbb{R} : \varphi(x) \leq c\}$ are compact and so are the level sets of $\mathbf{Q} \to D_\varphi(\mathbf{Q}, \mathbf{P})$

$$\Gamma_c := \left\{\mathbf{Q} \in \mathbb{R}^K : D_\varphi(\mathbf{Q}, \mathbf{P}) \leq c\right\}$$

for all $c > 0$.
(b) When $\mathbf{\Omega}$ is not closed but merely satisfies (7), then the infimum in (8) may not be reached in $\mathbf{\Omega}$ although being finite; however we aim for finding the *infimum/minimum* in (8). Finding the *minimizers* in (8) is another question. For instance, this can be solved whenever, additionally, $\mathbf{\Omega}$ is a closed set which implies the existence of minimizers in $\mathbf{\Omega}$. In this case, and when the number of such minimizers is finite, those can be approximated by dichotomic search. For the sake of brevity, this will not be addressed in this paper.
(c) The purpose of the condition (7) is to get rid of the $\limsup$ type and $\liminf$ type results in our below-mentioned "bare-simulation" approach and to obtain simple limit-statements which motivate our construction. In practice, it is enough to verify $\mathbf{\Omega} \subseteq cl\left(int\left(\mathbf{\Omega}\right)\right)$, which is equivalent to the left-hand part of (7). Clearly, any open set $\mathbf{\Omega} \subset \mathbb{R}^K$ satisfies the left-hand part of (7). In the subsetup where $\mathbf{\Omega}$ is a closed convex set and $int(\mathbf{\Omega}) \neq \emptyset$, (7) is satisfied and the minimizer $\mathbf{Q}_{min} \in \mathbf{\Omega}$ in (8) is attained and even unique. When $\mathbf{\Omega}$ is open and satisfies (7), then the infimum in (8) exists but is reached at some generalized projection of $\mathbf{P}$ on $\mathbf{\Omega}$ (see Csiszar [97] for the definition in the Kullback-Leibler case of probability measures, which extends to any $\varphi-$divergence in our framework).
(d) Without further mentioning, the regularity condition (7) is supposed to hold in the *full* topology. Of course, $int\left(\mathbb{S}^K\right) = \emptyset$ and thus, for the important probability-vector setup $\mathbf{\Omega} \subset \mathbb{S}^K$ the condition (7) is violated which requires extra refinements (cf. Subsection IV-C below). The same is needed for $\mathbf{\Omega} \subset A \cdot \mathbb{S}^K$ for some $A \neq 1$, since obviously $int\left(A \cdot \mathbb{S}^K\right) = \emptyset$; such a context appears naturally e.g. in connection with mass transportation problems (cf. (102) below) and with distributed energy management (cf. the paragraph after (113)).
(e) Often, $\mathbf{\Omega}$ will present a (discrete) model [8]. Since $\mathbf{\Omega}$ is assumed to have a non-void interior (cf. the right-hand part of (7)), this will exclude (parametric) models $\mathbf{\Omega} := \{\mathbf{Q}_\theta : \theta \in \Theta\}$ for some $\Theta \subset \mathbb{R}^d$ $(d < K - 1)$, for which $\theta \mapsto \mathbf{Q}_\theta$ constitutes a curve/surface in $\mathbb{R}^K$; however, for such a situation, one can employ standard minimization principles. Our approach is

[8] recall that an alternative naming also used in literature is to call Ω a model class (rather than model), and each $P \in \Omega$ a model (rather than model element)

predestined for non- or semiparametric models, instead. For instance, (7) is valid for appropriate *tubular neighborhoods* of parametric models or for more general non-parametric settings such as e.g. shape constraints.

Let us now present our new bare-simulation approach (cf. Definition 1) for solving the distance-optimization Problem 6:

(BS1) Step 1: equivalently rewrite (8) such that the vector $\mathbf{P}$ "turns into" a probability vector $\widetilde{\mathbb{P}}$. More exactly, define $M_{\mathbf{P}} := \sum_{i=1}^{K} p_i > 0$ and let $\widetilde{\mathbb{P}} := \mathbf{P}/M_{\mathbf{P}}$, and for $\mathbf{Q}$ in $\mathbf{\Omega}$, let $\widetilde{\mathbf{Q}} := \mathbf{Q}/M_{\mathbf{P}}$ (notice that $\widetilde{\mathbf{Q}}$ may be a non-probability vector). With the function $\widetilde{\varphi} \in \Upsilon(]a,b[)$ defined through $\widetilde{\varphi} := M_{\mathbf{P}} \cdot \varphi$, we obtain

$$D_{\varphi}(\mathbf{Q},\mathbf{P}) = \sum_{k=1}^{K} p_k \cdot \varphi\left(\frac{q_k}{p_k}\right) = \sum_{k=1}^{K} M_{\mathbf{P}} \cdot \widetilde{p_k} \cdot \frac{\varphi\left(\frac{M_{\mathbf{P}} \cdot \widetilde{q_k}}{M_{\mathbf{P}} \cdot \widetilde{p_k}}\right)}{M_{\mathbf{P}}} = D_{\widetilde{\varphi}}(\widetilde{\mathbf{Q}},\widetilde{\mathbb{P}}). \tag{12}$$

It follows that the solution of (8) coincides with the one of the problem of finding

$$\widetilde{\Phi}_{\widetilde{\mathbb{P}}}(\widetilde{\mathbf{\Omega}}) := \inf_{\widetilde{\mathbf{Q}} \in \widetilde{\mathbf{\Omega}}} D_{\widetilde{\varphi}}(\widetilde{\mathbf{Q}},\widetilde{\mathbb{P}}), \qquad \text{with } \widetilde{\mathbf{\Omega}} := \mathbf{\Omega}/M_{\mathbf{P}}; \tag{13}$$

as a side remark, one can see that in such a situation the rescaling of the divergence generator φ is important, which is one incentive that we incorporate multiples of φ below.

As an important special case we get for the choice $\mathbf{P} := (1,\ldots,1) := \mathbf{1}$ that the "prominent/frequent" separable nonlinear optimization problem of finding the optimal value $\inf_{\mathbf{Q}\in\mathbf{\Omega}} \sum_{k=1}^{K} \varphi(q_k)$ — with objective (e.g. cost, energy, purpose) function $\varphi \in \Upsilon(]a,b[)$ and constraint set (choice set, search space) $\mathbf{\Omega}$ — can be imbedded into our BS-approach by

$$\inf_{\mathbf{Q}\in\mathbf{\Omega}} \sum_{k=1}^{K} \varphi(q_k) = \inf_{\mathbf{Q}\in\mathbf{\Omega}} D_{\varphi}(\mathbf{Q},\mathbf{1}) = \inf_{\widetilde{\mathbf{Q}}\in\mathbf{\Omega}/K} D_{K\cdot\varphi}(\widetilde{\mathbf{Q}},\mathbb{P}^{unif}), \tag{14}$$

with $\mathbb{P}^{unif} := (\frac{1}{K},\ldots,\frac{1}{K})$ being the probability vector of frequencies of the uniform distribution on $\{1,\ldots,K\}$. Notice that with our new BS approach one may even tackle more general optimization problems of the form $\inf_{\breve{Q}\in\breve{\Omega}} \sum_{k=1}^{K} \breve{\varphi}(\breve{q}_k)$ where $\breve{\varphi}$ is some function which is finite and convex in a non-empty neighborhood (say, $]t_0+a-1, t_0+b-1[$ with $a<1<b$) of some point $t_0 \in \mathbb{R}$ as well as strictly convex in a non-empty sub-neighborhood of t_0; for this, the function

$$\varphi(t) := \breve{\varphi}(t+t_0-1) - \breve{\varphi}'(t_0)\cdot\Big((t+t_0-1)-t_0\Big) - \breve{\varphi}(t_0), \qquad t \in]a,b[,$$

(which corresponds to shifting the argument and adding an affine-linear function) should be a member of $\Upsilon(]a,b[)$, and from the corresponding minimization problem

$$\begin{aligned}
&\inf_{\widetilde{\mathbf{Q}}\in\mathbf{\Omega}/K} D_{K\cdot\varphi}(\widetilde{\mathbf{Q}},\mathbb{P}^{unif}) = \inf_{\mathbf{Q}\in\mathbf{\Omega}} \sum_{k=1}^{K} \varphi(q_k) = \inf_{\mathbf{Q}\in\mathbf{\Omega}} \sum_{k=1}^{K} \Big(\breve{\varphi}(q_k+t_0-1) - \breve{\varphi}'(t_0)\cdot((q_k+t_0-1)-t_0) - \breve{\varphi}(t_0)\Big) \\
&= \inf_{\breve{\mathbf{Q}}\in\mathbf{\Omega}+t_0-1} \sum_{k=1}^{K} \Big(\breve{\varphi}(\breve{q}_k) - \breve{\varphi}'(t_0)\cdot(\breve{q}_k-t_0) - \breve{\varphi}(t_0)\Big) \\
&= K\cdot\Big(t_0\cdot\breve{\varphi}'(t_0) - \breve{\varphi}(t_0)\Big) + \inf_{\breve{\mathbf{Q}}\in\breve{\mathbf{\Omega}}} \left(\sum_{k=1}^{K} \breve{\varphi}(\breve{q}_k) - \breve{\varphi}'(t_0)\cdot\sum_{k=1}^{K} \breve{q}_k\right), \qquad \text{with } \breve{\mathbf{\Omega}} := \mathbf{\Omega}+t_0-1,
\end{aligned} \tag{15}$$

the term $\inf_{\breve{Q}\in\breve{\Omega}} \sum_{k=1}^{K} \breve{\varphi}(\breve{q}_k)$ should be recoverable; for instance, later on we shall employ constraints sets $\breve{\mathbf{\Omega}}$ which particularly include $\sum_{k=1}^{K} \breve{q}_k = A > 0$, whereas another possibility would be to use a $\breve{\varphi}$ which satisfies $\breve{\varphi}'(t_0) = 0$. As a different line of flexibilization of (14), we can also deal with the problem $\inf_{\mathbf{Q}\in\mathbf{\Omega}} h\Big(\sum_{k=1}^{K} \varphi(q_k)\Big)$ through

$$\inf_{\mathbf{Q}\in\mathbf{\Omega}} h\Big(\sum_{k=1}^{K} \varphi(q_k)\Big) = h\Big(\inf_{\widetilde{\mathbf{Q}}\in\mathbf{\Omega}/K} D_{K\cdot\varphi}(\widetilde{\mathbf{Q}},\mathbb{P}^{unif})\Big) \tag{16}$$

for any $\varphi \in \Upsilon(]a,b[)$ and any continuous strictly increasing function $h : \mathcal{H} \mapsto \mathbb{R}$ with $\mathcal{H} := [0,\infty[$ (or a sufficiently large subset thereof), and with the problem $\sup_{\mathbf{Q}\in\mathbf{\Omega}} h\Big(\sum_{k=1}^{K} \varphi(q_k)\Big)$ through

$$\sup_{\mathbf{Q}\in\mathbf{\Omega}} h\Big(\sum_{k=1}^{K} \varphi(q_k)\Big) = h\Big(\inf_{\widetilde{\mathbf{Q}}\in\mathbf{\Omega}/K} D_{K\cdot\varphi}(\widetilde{\mathbf{Q}},\mathbb{P}^{unif})\Big) \tag{17}$$

for any $\varphi \in \Upsilon(]a,b[)$ and any continuous strictly decreasing function $h : \mathcal{H} \mapsto \mathbb{R}$. Combining (15) with (16) (respectively, with (17)) leads to a further flexibilization. Of course, we can also apply our BS method to the maximization

$\sup_{\mathbf{Q}\in\mathbf{\Omega}} h\Big(\sum_{k=1}^{K}\zeta(q_k)\Big)$ for any concave function ζ with $-\zeta \in \Upsilon(]a,b[)$ and any continuous strictly increasing function $h : \mathcal{H} \mapsto \mathbb{R}$ with $\mathcal{H} := -[\infty, 0]$ (or a sufficiently large subset thereof), via

$$\sup_{\mathbf{Q}\in\mathbf{\Omega}} h\Big(\sum_{k=1}^{K}\zeta(q_k)\Big) = h\Big(-\inf_{\widetilde{\mathbf{Q}}\in\mathbf{\Omega}/K} D_{-K\cdot\zeta}(\widetilde{\mathbf{Q}}, \mathbb{P}^{unif})\Big), \tag{18}$$

and to $\inf_{\mathbf{Q}\in\mathbf{\Omega}} h\Big(\sum_{k=1}^{K}\zeta(q_k)\Big)$ for any concave function ζ with $-\zeta \in \Upsilon(]a,b[)$ and any continuous strictly decreasing function $h : \mathcal{H} \mapsto \mathbb{R}$, via

$$\inf_{\mathbf{Q}\in\mathbf{\Omega}} h\Big(\sum_{k=1}^{K}\zeta(q_k)\Big) = h\Big(-\inf_{\widetilde{\mathbf{Q}}\in\mathbf{\Omega}/K} D_{-K\cdot\zeta}(\widetilde{\mathbf{Q}}, \mathbb{P}^{unif})\Big). \tag{19}$$

Moreover, we can tackle $\sup_{\breve{Q}\in\breve{\Omega}} \sum_{k=1}^{K} \breve{\zeta}(\breve{q}_k)$ where $\breve{\zeta}$ is some function which is finite and concave in a non-empty neighborhood $]t_0 + a - 1, t_0 + b - 1[$ (with $a < 1 < b$) of some point $t_0 \in \mathbb{R}$ as well as strictly concave in a non-empty sub-neighborhood of t_0; for this, the function

$$-\zeta(t) := -\breve{\zeta}(t + t_0 - 1) + \breve{\zeta}'(t_0) \cdot \Big((t + t_0 - 1) - t_0\Big) + \breve{\zeta}(t_0), \qquad t \in]a, b[,$$

should be a member of $\Upsilon(]a,b[)$, and from the corresponding minimization problem

$$\begin{aligned}
&-\inf_{\widetilde{\mathbf{Q}}\in\mathbf{\Omega}/K} D_{-K\cdot\zeta}(\widetilde{\mathbf{Q}}, \mathbb{P}^{unif}) = \sup_{\mathbf{Q}\in\mathbf{\Omega}} \sum_{k=1}^{K}\zeta(q_k) = \sup_{\mathbf{Q}\in\mathbf{\Omega}} \sum_{k=1}^{K} \Big(\breve{\zeta}(q_k + t_0 - 1) - \breve{\zeta}'(t_0) \cdot ((q_k + t_0 - 1) - t_0) - \breve{\zeta}(t_0)\Big) \\
&= \sup_{\breve{\mathbf{Q}}\in\mathbf{\Omega}+t_0-1} \sum_{k=1}^{K} \Big(\breve{\zeta}(\breve{q}_k) - \breve{\zeta}'(t_0) \cdot (\breve{q}_k - t_0) - \breve{\zeta}(t_0)\Big) \\
&= K \cdot \Big(t_0 \cdot \breve{\zeta}'(t_0) - \breve{\zeta}(t_0)\Big) + \sup_{\breve{\mathbf{Q}}\in\breve{\mathbf{\Omega}}} \left(\sum_{k=1}^{K} \breve{\zeta}(\breve{q}_k) - \breve{\zeta}'(t_0) \cdot \sum_{k=1}^{K} \breve{q}_k\right), \qquad \text{with } \breve{\mathbf{\Omega}} := \mathbf{\Omega} + t_0 - 1,
\end{aligned} \tag{20}$$

the term $\sup_{\breve{Q}\in\breve{\Omega}} \sum_{k=1}^{K} \breve{\zeta}(\breve{q}_k)$ should be recoverable; the left-hand side of (20) corresponds to the special case $h(x) := x$ of the BS-minimizable (18). A combination of (20) with (18) (respectively, with (19)) leads to a further flexibilization.

Remark 8: (a) Since $\mathbf{1}$ can be seen e.g. as a reference vector with (normalized) equal components, the quantity $\inf_{\mathbf{Q}\in\mathbf{\Omega}} D_{\varphi}(\mathbf{Q}, \mathbf{1})$ in (14) can be interpreted as an "index/degree of (in)equality of the set $\mathbf{\Omega}$", respectively as an "index/degree of diversity of the set $\mathbf{\Omega}$".
(b) The quantity $\sum_{k=1}^{K} \varphi(q_k)$ in (14) can be interpreted as (non-probability extension of an) $\varphi-$entropy in the sense of Burbea & Rao [68] (see also Csiszar [95], Ben-Bassat [38], Ben-Tal & Teboulle [40], Kesavan & Kapur [187], Dacunha-Castelle & Gamboa [102], Teboulle & Vajda [357], Gamboa & Gassiat [132], Vajda & Zvarova [376]); for applications to scalar quantization for lossy coding of information sources see e.g. György & Linder [149]. More generally, the quantity $h\Big(\sum_{k=1}^{K} \varphi(q_k)\Big)$ in (16) can be seen as (non-probability extension of an) $(h, \varphi)-$entropy in the sense of Salicru et al. [314] (see also e.g. Pardo [282], Vajda & Vasek [375], as well as e.g. Chen et al. [78] for uses as supervised adaption criterion within stochastic information gradient algorithms and Ren et al. [308] for applications to tracking in networked control systems). Important special cases will be discussed in more detail, below.

Returning to the original distance-minimizing Problem 6, after the first step (12) and (13), we proceed as follows:

(BS2) Step 2: construct an appropriate sequence $(\boldsymbol{\xi}_n)_{n\in\mathbb{N}}$ of $\mathbb{R}^K-$valued random variables/random vectors (cf. (2) in Definition 1):

The following condition transposes the minimization problem (13) into a *BS minimizable/amenable problem* in the sense of Definition 1 and it is required in order that Problem (13) is equivalent to Problem (8). The connection of this condition with (6) will be discussed in Proposition 34 and its surroundings, see Section V.

Condition 9: With $M_{\mathbf{P}} = \sum_{i=1}^{K} p_i > 0$, the divergence generator φ in (8) (cf. also (12)) satisfies $\widetilde{\varphi} := M_{\mathbf{P}} \cdot \varphi \in \Upsilon(]a,b[)$, i.e. $\widetilde{\varphi} \in \widetilde{\Upsilon}(]a,b[)$ (which is equivalent to $\varphi \in \widetilde{\Upsilon}(]a,b[)$) and there holds the representation

$$\widetilde{\varphi}(t) = \sup_{z\in\mathbb{R}} \left(z \cdot t - \log \int_{\mathbb{R}} e^{zy} d\widetilde{\mathfrak{C}}(y)\right), \qquad t \in \mathbb{R}, \tag{21}$$

for some probability measure $\widetilde{\mathfrak{C}}$ on the real line such that the function $z \mapsto MGF_{\widetilde{\mathfrak{C}}}(z) := \int_{\mathbb{R}} e^{zy} d\widetilde{\mathfrak{C}}(y)$ is finite on some open interval containing zero [9].

In the following, let us explain the above-mentioned Step 2 in detail: for any $n \in \mathbb{N}$ and any $k \in \{1, \ldots, K\}$, let $n_k := \lfloor n \cdot \widetilde{p}_k \rfloor$ where $\lfloor x \rfloor$ denotes the integer part of x. We assume $\mathbf{P} \in \mathbb{R}^K_{>0}$, and since thus none of the $\widetilde{p}_k$'s is zero, one has

$$\lim_{n \to \infty} \frac{n_k}{n} = \widetilde{p}_k. \tag{22}$$

Moreover, we assume that $n \in \mathbb{N}$ is large enough, namely $n \geq \max_{k \in \{1,\ldots,K\}} \frac{1}{\widetilde{p}_k}$, and decompose the set $\{1, \ldots, n\}$ of all integers from 1 to n into the following disjoint blocks: $I_1^{(n)} := \{1, \ldots, n_1\}$, $I_2^{(n)} := \{n_1 + 1, \ldots, n_1 + n_2\}$, and so on until the last block $I_K^{(n)} := \{\sum_{k=1}^{K-1} n_k + 1, \ldots, n\}$ which therefore contains all integers from $n_1 + \ldots + n_{K-1} + 1$ to n. Clearly, $I_k^{(n)}$ has $n_k \geq 1$ elements (i.e. $card(I_k^{(n)}) = n_k$ where $card(A)$ denotes the number of elements in a set A) for all $k \in \{1, \ldots, K-1\}$, and the last block $I_K^{(n)}$ has $n - \sum_{k=1}^{K-1} n_k \geq 1$ elements which anyhow satisfies $\lim_{n\to\infty} card(I_K^{(n)})/n = \widetilde{p}_K$ [10]. Furthermore, consider a vector $\widetilde{\mathbf{W}} := \left(\widetilde{W}_1, \ldots, \widetilde{W}_n\right)$ where the $\widetilde{W}_i$'s are i.i.d. copies of the random variable $\widetilde{W}$ whose distribution is associated with the divergence-generator $\widetilde{\varphi} := M_{\mathbf{P}} \cdot \varphi$ through (21), in the sense that $\mathbb{\Pi}[\widetilde{W} \in \cdot\,] = \widetilde{\mathfrak{C}}[\,\cdot\,]$. We group the $\widetilde{W}_i$'s according the above-mentioned blocks and sum them up blockwise, in order to build the following $K-$ component random vector

$$\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} := \Big(\frac{1}{n} \sum_{i \in I_1^{(n)}} \widetilde{W}_i, \ldots, \frac{1}{n} \sum_{i \in I_K^{(n)}} \widetilde{W}_i\Big); \tag{23}$$

notice that the signs of its components may be negative, depending on the nature of the $\widetilde{W}_i$'s; moreover, the expectation of its $k-$th component converges to $\widetilde{p}_k$ as n tends to infinity (since the expectation of $\widetilde{W}_1$ is 1), whereas the $n-$fold of the corresponding variance converges to $\widetilde{p}_k$ times the variance of $\widetilde{W}_1$.

For such a context, we obtain the following assertion on BS-minimizability:

Theorem 10: Let $\mathbf{P} \in \mathbb{R}^K_{>0}$, $M_{\mathbf{P}} := \sum_{i=1}^K p_i > 0$, and suppose that the divergence generator φ satisfies the Condition 9 above, with $\widetilde{\mathfrak{C}}$ (cf. (21)). Additionally, let $\widetilde{W} := (\widetilde{W}_i)_{i \in \mathbb{N}}$ be a sequence of random variables where the $\widetilde{W}_i$'s are i.i.d. copies of the random variable $\widetilde{W}$ whose distribution is $\mathbb{\Pi}[\widetilde{W} \in \cdot\,] = \widetilde{\mathfrak{C}}[\,\cdot\,]$ [11]. Then, in terms of the random vectors

$$\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} = \Big(\frac{1}{n} \sum_{i \in I_1^{(n)}} \widetilde{W}_i, \ldots, \frac{1}{n} \sum_{i \in I_K^{(n)}} \widetilde{W}_i\Big) \qquad \text{(cf. (23))}$$

there holds

$$-\lim_{n \to \infty} \frac{1}{n} \log \mathbb{\Pi}\left[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \boldsymbol{\Omega}/M_{\mathbf{P}}\right] = \inf_{Q \in \boldsymbol{\Omega}} D_{\varphi}(\mathbf{Q}, \mathbf{P}\ ) \tag{24}$$

for any $\boldsymbol{\Omega} \subset \mathbb{R}^K$ with regularity properties (7) and finiteness property (9). In particular, for each $\mathbf{P} \in \mathbb{R}^K_{>0}$ the function $\Phi_{\mathbf{P}}(\cdot) := D_{\varphi}(\cdot, \mathbf{P})$ (cf. (4)) is bare-simulation minimizable (BS-minimizable, cf. (2))) on any such $\boldsymbol{\Omega} \subset \mathbb{R}^K$.

The proof of Theorem 10 will be given in Appendix A.

Remark 11: (i) Whenever $int(\boldsymbol{\Omega}) \neq \emptyset$, it clearly holds that $\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{\Pi}\left[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \boldsymbol{\Omega}/M_{\mathbf{P}}\right] > 0$; see the proof of Theorem 10. Hence, the limit in (24) exists and is finite when $\boldsymbol{\Omega}$ satisfies (7).
(ii) For some contexts, one can *explicitly* give the distribution of each of the independent (non-deterministic parts of the) components $\left(\sum_{i \in I_k^{(n)}} \widetilde{W}_i\right)_{k=1,\ldots,K}$ of the vector $\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}}$; this will ease the corresponding concrete simulations. For instance, we shall give those in the Examples 48, 50, 53, 54 and 55 in Section V below.
(iii) Let us emphasize that we have assumed $\mathbf{P} \in \mathbb{R}^K_{>0}$ in Theorem 10 which excludes $\mathbf{P}$ from having zero components. However, in cases where $\lim_{x\to\infty} \left|\frac{\varphi(x \cdot sgn(q))}{x \cdot sgn(q)}\right| = +\infty$ for $q \neq 0$ then if $p_{k_0} = 0$ for some k_0 it follows that $q_{k_0} = 0$, which proves that $\mathbf{P} \in \mathbb{R}^K_{>0}$ imposes no restriction in Theorem 9, since the projection of $\mathbf{P}$ in $\boldsymbol{\Omega}$ then belongs to the subspace of $\mathbb{R}^K$ generated by the non-null components of $\mathbf{P}$; such a situation appears e.g. for power divergence generators φ_γ with $\gamma > 2$. So there is no loss of generality assuming $\mathbf{P} \in \mathbb{R}^K_{>0}$ in this case.

[9] in particular, this implies that $\int_{\mathbb{R}} y d\widetilde{\mathfrak{C}}(y) = 1$ (cf. (G11i) below) and that $\widetilde{\mathfrak{C}}$ has light tails.

[10] if all $\widetilde{p}_k$ $(k = 1, \ldots, K)$ are rational numbers in $]0, 1[$ with $\sum_{k=1}^K \widetilde{p}_k = 1$ and N is the (always existing) smallest integer such that all $N \cdot \widetilde{p}_k$ $(k = 1, \ldots, K)$ are integers (i.e. $\in \mathbb{N}$), then for any multiple $n = \ell \cdot N$ $(\ell \in \mathbb{N})$ one gets that all $n_k = n \cdot \widetilde{p}_k$ are integers and that $card(I_K^{(n)}) = n_K$.

[11] and thus, $E_{\mathbb{\Pi}}[\widetilde{W}_i] = 1$

As examples for the applicability of Theorem 10, one can e.g. combine *each* of the divergence generators φ of Table 1 (except for the 9th row) with *any* of the optimization problems (8), (10), (11), (14), (16), (17); the needed distributions $\Pi[\widetilde{W} \in \cdot\,] = \widetilde{\zeta}[\,\cdot\,]$ correspond to the entry in the second last column with the choice $\widetilde{c} \cdot M_{\mathbf{P}}$ instead of $\widetilde{c}$. By taking $\zeta := -\varphi$ instead, one can solve the corresponding problems (18) and (19).

Returning to the general context, the limit statement (24) provides the principle for the approximation of the solution of Problem 8. Indeed, by replacing the left-hand side in (24) by its finite counterpart, we deduce for given large n

$$-\frac{1}{n}\log \Pi\left[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \mathbf{\Omega}/M_{\mathbf{P}}\right] \approx \inf_{Q\in\mathbf{\Omega}} D_\varphi(\mathbf{Q},\mathbf{P}); \tag{25}$$

it remains to estimate the left-hand side of (25). The latter can be performed either by a *naive estimator* of the frequency of those replications of $\boldsymbol{\xi}_{n,\widetilde{\mathbf{x}}}^{\widetilde{\mathbf{W}}}$ which hit $\Omega/M_{\mathbf{P}}$, or more efficiently by some improved estimator; this will be discussed in detail in Section VI below.

Remark 12: According to (24) of Theorem 9 as well as (25), we can principally tackle the (approximative) computation of the minimum value

$$\inf_{Q\in\mathbf{\Omega}} D_\varphi(\mathbf{Q},\mathbf{P}\ ) = \inf_{Q\in\mathbf{\Omega}} \sum_{k=1}^{K} p_k \cdot \varphi\left(\frac{q_k}{p_k}\right)$$

and in particular of

$$\inf_{\mathbf{Q}\in\mathbf{\Omega}} \sum_{k=1}^{K} \varphi(q_k) = \inf_{\mathbf{Q}\in\mathbf{\Omega}} D_\varphi(\mathbf{Q},\mathbf{1}) \qquad \text{(cf. (14))}$$

by basically *only employing a fast and accurate — pseudo, true, natural, quantum — random number generator*[12], provided that the constraint set $\mathbf{\Omega}$ satisfies the mild assumptions (7) and (9). Notice that (7) also covers (e.g. high-dimensional) constraint sets $\mathbf{\Omega}$ which are *non-convex* and even *highly disconnected*, and for which other minimization methods (e.g. pure enumeration, gradient or steepest descent methods, etc. [13]) may be problematic or intractable. For instance, (7) covers kind of "$K-$dimensional (not necessarily regular) polka dot (leopard skin) pattern type" relaxations $\mathbf{\Omega} := \dot{\bigcup}_{i=1}^{N} \mathcal{U}_i(Q_i^{dis})$ of finite discrete constraint sets $\mathbf{\Omega}^{dis} := \{Q_1^{dis}, \ldots, Q_N^{dis}\}$ of high cardinality N (e.g. being exponential or factorial in a large K), where each $K-$dimensional vector Q_i^{dis} (e.g. having pure integer components only) is surrounded by some small (in particular, non-overlapping/disjoint) neighborhood $\mathcal{U}_i(Q_i^{dis})$; in such a context, e.g. $\inf_{\mathbf{Q}\in\mathbf{\Omega}} \sum_{k=1}^{K} \varphi(q_k)$ can be regarded as a *"BS-tractable" relaxation* of the nonlinear discrete (e.g. integer, combinatorial [14]) optimization program $\inf_{\mathbf{Q}\in\mathbf{\Omega}^{dis}} \sum_{k=1}^{K} \varphi(q_k)$.

C. Minimum distance/risk estimation

In statistics of discrete data — and in the adjacent research fields of information theory, artificial intelligence and machine learning — one often encounters the following *minimum distance estimation (MDE) problem* which is often also named as *estimation of the empirical risk*:

(MDE1) for index $i \in \mathbb{N}$, let the generation of the $i-$th (uncertainty-prone) data point be represented by the random variable X_i which takes values in the discrete set $\mathcal{Y} := \{d_1, \cdots, d_K\}$ of K distinct values "of any kind"˙. It is assumed that there exists a probability measure $\mathbb{P}[\cdot]$ on $\mathcal{Y}$ which is the a.s. limit of the empirical measures $\mathbb{P}_n^{emp}$ defined by the collection of collected $(X_1, .., X_n)$ as n tends to infinity, in formula

$$\lim_{n\to\infty} \mathbb{P}_n^{emp} := \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} = \mathbb{P} \qquad \text{a.s.} \tag{26}$$

where δ_y denotes the one-point distribution (Dirac mass) at point y [15]. We will assume that none of the entries of $\mathbb{P}$ bears zero mass so that $\mathbb{P}$ is identified with a point in the interior of $\mathbb{S}^K$ (see below). The underlying probability space (say, $(\mathfrak{X}, \mathcal{A}, \Pi)$) where the above a.s. convergence holds, pertains to the random generation of the sequence $(X_n)_{n\geq 1}$, of which we do not need to know but for (26). Examples include the i.i.d. case (where the X_i's are independent and

[12] see e.g. Tucci [366], Teh et al. [358], Aghamohammadi & Crutchfield [6], Herrero-Collantes & Garcia-Escartin [159], Balygin et al. [31], Dang et al. [103], Gong et al. [138], Chandrasekaran et al. [77], Drahi et al. [117], Kollmitzer et al. [194], Liu et al. [228], Fischer & Gauthier [126], Kim et al. [188], Stoller & Campbell [342]

[13] a detailed discussion and comparisons are beyond the scope of this paper, given its current length

[14] see e.g. Schrijver [324], Bertsimas & Weismantel [45], Chen et al. [83], Onn [279], Korte & Vygen [195], Wolsey [393] for comprehensive books on discrete, integer and combinatorial programming and their vast applications

[15] notice that $\mathbb{P}_n^{emp}$ a probability measure on the data space $\mathcal{Y}$, which is random due to its dependence on the X_i's

have common distribution $\mathbb{P}$), ergodic Markov chains on $\mathcal{Y}$ with stationary distribution $\mathbb{P}$, more globally autoregressive chains with stationary measure $\mathbb{P}$, etc.

Let us briefly discuss our assumption (26) (resp. its vector form (30) below) on the limit behavior of the empirical distribution of the observed sample $\mathbf{X}_n := (X_1, .., X_n)$ as n tends to infinity. In the "basic" statistical context, the sample $\mathbf{X}_n$ consists of i.i.d. replications of a generic random variable X with probability distribution $\mathbf{P}$. However, our approach captures many other sampling schemes, where the distribution $\mathbf{P}$ is defined implicitly through (26) for which we aim at some estimate of $\Phi_{\mathbf{P}}(\Omega)$ of a family Ω of probability distributions on $\mathcal{Y}$. Sometimes the sequence of samples $(\mathbf{X}_n)_{n\geq 1}$ may stem from a triangular array so that $\mathbf{X}_n = (X_{1,n}, .., X_{k_n,n})$ with $k_n \to \infty$ and statement (26) is substituted by

$$\lim_{n\to\infty} \frac{1}{k_n} \sum_{i=1}^{k_n} \delta_{X_{i,n}} = \mathbf{P} \quad \text{a.s.}$$

which does not alter the results of this paper by any means.

(MDE2) given a *model* Ω, i.e. a family Ω of probability distributions on $\mathcal{Y}$ each of which serves as a potential description of the underlying (unknown) data-generating mechanism $\mathbb{P}$, one would like to find

$$\Phi_{\mathbb{P}}(\Omega) := \inf_{\mathbb{Q}\in\Omega} D_{\varphi}(\mathbb{Q}, \mathbb{P}) \tag{27}$$

which quantifies the *adequacy* of the model Ω for modeling $\mathbb{P}$, *via* the minimal distance/dissimilarity of Ω to $\mathbb{P}$; a lower $\Phi_{\mathbb{P}}-$value means a better adequacy (in the sense of a lower departure between the model and the truth, cf. Lindsay [222], Lindsay et al. [223], Markatou & Sofikitou [248], Markatou & Chen [249]).
Hence, especially in the context of *model selection* within complex big-data contexts, for the *search of appropriate models* Ω and model elements/members therein, the (fast and efficient) computation of $\Phi_{\mathbb{P}}(\Omega)$ constitutes a decisive first step, since if the latter is "too large" (respectively "much larger than" $\Phi_{\mathbb{P}}(\overline{\Omega})$ for some competing model $\overline{\Omega}$), then the model Ω is "not adequate enough" (respectively "much less adequate than" $\overline{\Omega}$); in such a situation, the effort of computing the (not necessarily unique) best model element/member $\arg\inf_{\mathbb{Q}\in\Omega} D_{\varphi}(\mathbb{Q}, \mathbb{P})$ within the model Ω is "not very useful" and is thus a "waste of computational time".
Because of such considerations, we concentrate ourselves to finding the infimum (27) rather than finding the corresponding minimizer(s). Variants of (27) are of interest, too.

Since $int(\Omega)$ is supposed to be a non-empty set in the space of probability distribution on $\mathcal{Y}$, the present procedure is fitted for semi-parametric models Ω, e.g. such as defined through moment conditions (as extensions of the Empirical Likelihood paradigm, see e.g. Broniatowski & Keziou [62]), or through L-moment conditions (i.e. moment conditions pertaining to quantile measures, see Broniatowski & Decurninge [59]), or even more involved non-parametric models where the geometry of Ω does not allow for ad-hoc procedures.

The measurement or the estimation of $\Phi_{\mathbb{P}}(\Omega)$ is a tool for the choice of pertinent putative models Ω among a class of specifications. The case when $\Phi_{\mathbb{P}}(\Omega) > 0$ is interesting in its own, since it is quite common in engineering modelling to argue in favor of misspecified models (or (non-void) neighborhoods of such models for sake of robustness issues), due to quest for conservatism; the choice between them is a widely open field e.g. in the practice of reliability. This also opens the question of the choice of the divergence generator φ; although this will not be discussed in this paper, as a motivating running example the reader may keep in mind the generator $\varphi_2(x) := (x-1)^2/2$ which induces the divergence $D_{\varphi_2}(\mathbb{Q}, \mathbb{P})$ (see (44) below for details) which quantifies the expected square relative error when substituting the true distribution $\mathbb{P}$ by the model $\mathbb{Q}$.

As examples of sets Ω of probability distributions on $\mathcal{Y}$ which obey (through their $K-$vector of corresponding probability masses/frequencies) the global assumptions (7), one can consider semi-parametric models defined by moment conditions or defined through L-moment constraints (hence on the quantile functions), as well as more involved ones, for which no closed form of the divergence with respect to any probability distribution is available. In the context of model selection, the choice of Ω may be dictated by various considerations, and misspecification may be assumed as a requisite, for example for conservatism in reliability design.

An estimate of $\Phi_{\mathbb{P}}(\Omega)$ can be used as a statistics for some test of fit, and indeed the likelihood ratio test adapted to some semi-parametric models has been generalized to the divergence setting (see Broniatowski & Keziou [62]). The statement of the limit distributions of our estimate, under the model and under misspecification, is postponed to future work.

In the following, we compute/approximate (27) — and some variants thereof — by our *bare simulation (BS)* method, by "mimicking" the deterministic minimization problem (8) respectively (13). Let us first remark that, as usual, each probability distribution (probability measure) $\mathbb{P}$ on $\mathcal{Y} = \{d_1, \ldots, d_K\}$ can be uniquely identified with the (row) vector $\mathbf{P} := (p_1, \ldots, p_K) \in \mathbb{S}^K$ of the corresponding probability masses (frequencies) $p_k = \mathbb{P}[\{d_k\}]$ via $\mathbb{P}[A] = \sum_{k=1}^{K} p_k \cdot 1_A(d_k)$ for each $A \subset \mathcal{Y}$, where $1_A(\cdot)$ denotes the indicator function on the set A. In particular, the probability distribution $\mathbb{P}$ in (MDE1) can be identified with

$(p_1,\ldots,p_K)$ in terms of $p_k = \mathbb{P}[\{d_k\}]$ (which in the i.i.d. case turns into $p_k = \Pi[X_1 = d_k]$). Along this line, the family Ω of probability distributions in (MDE2) can be identified with a subset $\mathbf{\Omega} \subset \mathbb{S}^K$ of probability vectors (viz. of vectors of probability masses).

Analogously, each finite nonnegative measure Q on $\mathcal{Y}$ can be uniquely identified with a vector $\mathbf{Q} := (q_1,\ldots,q_K) \in \mathbb{R}^K_{\geq 0}$, and each finite signed measure Q with a vector $\mathbf{Q} := (q_1,\ldots,q_K) \in \mathbb{R}^K$. The corresponding divergences between distributions/measures are then, as usual, defined through the divergences between their respective masses/frequencies:

$$D_\varphi(Q,\mathbb{P}) := D_\varphi(\mathbf{Q},\mathbf{P}). \tag{28}$$

In particular, $\mathbb{P}_n^{emp}$ can be identified with the vector $\mathbf{P}_n^{emp} := (p_{n,1}^{emp},\ldots,p_{n,K}^{emp})$ where

$$p_{n,k}^{emp} \;:=\; \frac{1}{n}\cdot n_k \;:=\; \frac{1}{n}\cdot card(\{i\in\{1,\ldots,n\}:\; X_i = d_k\}) \;=:\; \frac{1}{n}\cdot card(I_k^{(n)}), \quad k\in\{1,\ldots,K\}, \tag{29}$$

and accordingly the required limit behaviour (26) is equivalent to the vector-convergence

$$\lim_{n\to\infty}\left(\frac{n_1}{n},\ldots,\frac{n_K}{n}\right) = (p_1,\ldots,p_K) \qquad \text{a.s.} \tag{30}$$

Notice that, in contrast to the case handled in the above Subsection IV-B, the sets $I_k^{(n)}$ of indexes introduced in (29) and their numbers $n_k = card(I_k^{(n)})$ of elements are now *random* (due to their dependence on the X_i's) and $M_{\mathbf{P}_n^{emp}} = 1$. In a *batch procedure*, when $D_\varphi(\mathbf{\Omega},\mathbf{P}_n^{emp})$ is estimated once the sample $(X_1,..,X_n)$ is observed, we may reorder this sample by putting the n_1 sample points X_i which are equal to d_1 in the first places, and so on; accordingly one ends up with index sets $I_k^{(n)}$ as defined in Section IV-B. When the *online acquisition* of the data X_i's is required, then we usually do not reorder the sample, and the $I_k^{(n)}$'s do not consist in consecutive indexes, which does not make any change with respect to the resulting construction nor to the estimator.

The above considerations open the gate to our desired "mimicking" of (8) and (13) to achieve (27) (and some variants thereof) by our bare simulation (BS) method. To proceed, we employ a family of random variables $(W_i)_{i\in\mathbb{N}}$ of independent and identically distributed $\mathbb{R}-$valued random variables with probability distribution $\mathbb{Q}[\cdot\,] := \Pi[W_1 \in \cdot\,]$ (being connected with the divergence generator $\varphi \in \Upsilon(]a,b[)$ via the representability (6)), such that $(W_i)_{i\in\mathbb{N}}$ is independent of $(X_i)_{i\in\mathbb{N}}$ [16].

As a next step, notice that the "natural candidate"

$$\xi_{n,\mathbf{X}}^{\mathbf{W}} := \frac{1}{n}\cdot\sum_{k=1}^{K}\left(\sum_{i\in I_k^{(n)}} W_i\right)\cdot\delta_{d_k} \;=\; \frac{1}{n}\sum_{i=1}^{n} W_i\cdot\delta_{X_i}$$

is not a probability measure since its total mass is not 1 in general, since in terms of its equivalent vector version

$$\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}} := \Big(\frac{1}{n}\sum_{i\in I_1^{(n)}} W_i,\ldots,\frac{1}{n}\sum_{i\in I_K^{(n)}} W_i\Big) \tag{31}$$

the sum $\sum_{k=1}^{K}\frac{1}{n}\sum_{i\in I_k^{(n)}} W_i = \frac{1}{n}\sum_{j=1}^{n} W_i$ of the K vector components of (31) is typically not equal to 1; this implies that no limit result of the form (24) with finite limit can hold, since $\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}}$ takes values in $\mathbb{R}^K$ and $\mathbf{\Omega}$ is a subset in the probability simplex $\mathbb{S}^K$ which has *void* interior in $\mathbb{R}^K$ causing a violation of condition (7) (cf. Remark 7(c)); moreover, depending on the concrete form of the generator φ, the corresponding weights may take *negative values*. Therefore, we need some "rescaling". Indeed, let us introduce the *normalized weighted empirical measure*

$$\xi_{n,\mathbf{X}}^{w\mathbf{W}} \quad:=\quad \begin{cases} \frac{1}{\sum_{k=1}^{K}\sum_{i\in I_k^{(n)}} W_i}\cdot\sum_{k=1}^{K}\Big(\sum_{i\in I_k^{(n)}} W_i\Big)\cdot\delta_{d_k} = \sum_{i=1}^{n}\frac{W_i}{\sum_{j=1}^{n} W_j}\cdot\delta_{X_i}, & \text{if } \sum_{j=1}^{n} W_j \neq 0,\\ \infty\cdot\sum_{k=1}^{K}\delta_{d_k} =: \underline{\infty}, & \text{if } \sum_{j=1}^{n} W_j = 0, \end{cases} \tag{32}$$

which will substitute $\xi_{n,\mathbf{X}}^{\mathbf{W}}$ and which may belong to $\mathbf{\Omega}$ with positive probability. The equivalent vector version of $\xi_{n,\mathbf{X}}^{w\mathbf{W}}$ is given by

$$\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \quad:=\quad \begin{cases} \left(\frac{\sum_{i\in I_1^{(n)}} W_i}{\sum_{k=1}^{K}\sum_{i\in I_k^{(n)}} W_i},\ldots,\frac{\sum_{i\in I_K^{(n)}} W_i}{\sum_{k=1}^{K}\sum_{i\in I_k^{(n)}} W_i}\right), & \text{if } \sum_{j=1}^{n} W_j \neq 0,\\ (\infty,\ldots,\infty) =: \boldsymbol{\infty}, & \text{if } \sum_{j=1}^{n} W_j = 0, \end{cases} \tag{33}$$

a point in the linear subset of $\mathbb{R}^K$ spanned by $\mathbb{S}^K$ at infinity.

[16] on the common underlying probability space $(\mathfrak{X},\mathcal{A},\Pi)$

Remark 13:
(i) (Concerning e.g. computer-program command availability) In case of $\sum_{j=1}^{n} W_j = 0$, in (32) we may equivalently assign to $\xi_{n,\mathbf{X}}^{w\mathbf{W}}$ instead of $\underline{\infty}$ any measure (e.g. probability distribution) which does not belong to Ω, respectively, in (33) we may equivalently choose for $\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}}$ any vector outside of $\boldsymbol{\Omega}$ instead of $\boldsymbol{\infty}$.
(ii) By construction, in case of $\sum_{j=1}^{n} W_j \neq 0$, the sum of the random K vector components of (33) is now automatically equal to 1, but — as (depending on φ) the W_i's may take both positive and negative values — these random components may be negative (resp. nonnegative) with probability strictly greater (resp. smaller) than zero (resp. one); in the framework of (32) this means that $\xi_{n,\mathbf{X}}^{w\mathbf{W}}$ is in general a random signed measure with total mass 1, in case of $\sum_{j=1}^{n} W_j \neq 0$. However, $\Pi[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \mathbb{S}_{>0}^{K}] > 0$ since all the (identically distributed) random variables W_i have expectation 1 (as a consequence of the assumed representability (6)); in case of $\Pi[W_1 > 0] = 1$ one has even $\Pi[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \mathbb{S}_{>0}^{K}] = 1$. In the particular context of Example 48(c), one gets $\Pi[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \mathbb{S}_{>0}^{K}] = (\Pi[W_1 > 0])^n = \left(\int_0^\infty \sqrt{\frac{\widetilde{c}}{2\pi}} \cdot \exp(-\frac{\widetilde{c}\cdot(u-1)^2}{2}) du\right)^n \in]0,1[$.
Summing up things, the probability $\Pi[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \boldsymbol{\Omega}]$ is strictly positive and finite at least for large n, whenever $\Phi_{\mathbf{P}}(\boldsymbol{\Omega}) = \inf_{\mathbf{Q} \in \boldsymbol{\Omega}} D_\varphi(\mathbf{Q}, \mathbf{P})$ is finite.
(iii) By generalizing the terminology of e.g. Vajda [372], through the right-hand side of (32) one can interpret (for $\sum_{j=1}^{n} W_j \neq 0$) the normalized weighted empirical measure $\xi_{n,\mathbf{X}}^{w\mathbf{W}}$ as response of an output neuron in a random perceptron consisting of random inputs $\mathbf{X}$, a layer with n units having one-point-distribution-valued responses $\delta_{X_1}, \ldots, \delta_{X_n}$, and independent random synaptic weights $\left(\frac{W_1}{\sum_{j=1}^{n} W_j}, \ldots, \frac{W_n}{\sum_{j=1}^{n} W_j}\right)$.

With the above-mentioned ingredients, we are now in the position to tackle a variant of the distance minimization problem (27), by our bare simulation method through "mimicking" the deterministic minimization problem (8) respectively (13). For this, we also employ the conditional distributions $\Pi_n[\cdot] := \Pi_{X_1^n}[\cdot] := \Pi[\cdot \,|\, X_1, \ldots, X_n]$ and obtain the following

Theorem 14: Suppose that $(X_i)_{i \in \mathbb{N}}$ is a sequence of random variables with values in $\mathcal{Y} := \{d_1, \cdots, d_K\}$ such that (26) holds for some probability measure $\mathbb{P}[\cdot]$ on $\mathcal{Y}$ having no zero-mass frequencies (or equivalently, (30) holds for some probability vector $\mathbf{P} \in \mathbb{S}_{>0}^{K}$). Moreover, let $(W_i)_{i \in \mathbb{N}}$ be a family of independent and identically distributed $\mathbb{R}-$valued random variables with probability distribution $\mathbb{C}[\cdot] := \Pi[W_1 \in \cdot]$ being connected with the divergence generator $\varphi \in \Upsilon(]a,b[)$ via the representability (6), such that $(W_i)_{i \in \mathbb{N}}$ is independent of $(X_i)_{i \in \mathbb{N}}$. Then there holds

$$-\lim_{n\to\infty} \frac{1}{n} \log \Pi_{X_1^n}\left[\xi_{n,\mathbf{X}}^{w\mathbf{W}} \in \Omega\right] = \inf_{\mathbb{Q} \in \Omega} \inf_{m \neq 0} D_\varphi(m \cdot \mathbb{Q}, \mathbb{P}) \tag{34}$$

$$= \inf_{m \neq 0} \inf_{\mathbb{Q} \in \Omega} D_\varphi(m \cdot \mathbb{Q}, \mathbb{P})$$

$$= \inf_{m \neq 0} \inf_{\mathbf{Q} \in \boldsymbol{\Omega}} D_\varphi(m \cdot \mathbf{Q}, \mathbf{P}) \tag{35}$$

$$= \inf_{\mathbf{Q} \in \boldsymbol{\Omega}} \inf_{m \neq 0} D_\varphi(m \cdot \mathbf{Q}, \mathbf{P}) = -\lim_{n\to\infty} \frac{1}{n} \log \Pi_{X_1^n}\left[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \boldsymbol{\Omega}\right] \tag{36}$$

for all sets Ω of probability distributions such that their equivalent probability-vector form $\boldsymbol{\Omega}$ satisfies the regularity properties (7) *in the relative topology* and the finiteness property (9); notice that for the equality (35) we have used the "divergence link" (28). In particular, for each $\mathbf{P} \in \mathbb{S}_{>0}^{K}$ (respectively, its equivalent probability-distribution $\mathbb{P}$) the function $\mathbf{Q} \mapsto \inf_{m \neq 0} D_\varphi(m \cdot \mathbf{Q}, \mathbf{P})$ (respectively, the function $\mathbb{Q} \mapsto \inf_{m \neq 0} D_\varphi(m \cdot \mathbb{Q}, \mathbb{P})$) is BS-minimizable (cf. (2)) on all sets $\boldsymbol{\Omega} \subset \mathbb{S}^K$ satisfying (7) *in the relative topology* and (9) (respectively, on their probability-distribution-equivalent Ω).

The proof of Theorem 14 will be given in Appendix B. Analogously to Remark 11(iii), let us emphasize that we have assumed $\mathbf{P} \in \mathbb{S}_{>0}^{K}$ in Theorem 14. Henceforth, for sets $\boldsymbol{\Omega} \subset \mathbb{S}^K$ of probability vectors we deal with (7) only in the relative topology; thus, the latter will be unmentioned for the sake of brevity. Remark 7(a),(b),(c),(e) applies accordingly.

Remark 15:
(i) In strong contrast to Theorem 10, the above result does not provide a direct tool for the solution of Problem (27) since the limit in (34) bears no *direct* information on the minimum divergence $D_\varphi(\Omega, \mathbb{P}) := \inf_{\mathbb{Q} \in \Omega} D_\varphi(\mathbb{Q}, \mathbb{P})$; the link between the corresponding quantities can be emphasized and exploited e.g. in the case of power type divergences, which leads to explicit minimization procedures as shown in the Subsection IV-C1 below. For general divergences, Theorem 14 allows for the estimation of upper and lower bounds of $D_\varphi(\Omega, \mathbb{P})$, as developed in the Subsection IV-C2 below.
(ii) Notice that $\breve{D}_\varphi(\mathbb{Q}, \mathbb{P}) := \inf_{m \neq 0} D_\varphi(m \cdot \mathbb{Q}, \mathbb{P})$ satisfies the axioms of a divergence, that is, $\breve{D}_\varphi(\mathbb{Q}, \mathbb{P}) \geq 0$, as well as $\breve{D}_\varphi(\mathbb{Q}, \mathbb{P}) = 0$ if and only if $\mathbb{Q} = \mathbb{P}$ (reflexivity). Hence, in Theorem 14 we are still within our framework of bare simulation of a divergence minimum w.r.t. its first component (however, notice the difference to (i)).
(iii) Viewed from a "reverse" angle, Theorem 14 gives a crude approximation for the probability for $\xi_{n,\mathbf{X}}^{w\mathbf{W}}$ to belong to Ω, conditionally upon $\mathbf{X} = (X_1, \ldots, X_n)$.

(iv) In the same spirit as Remark 11(ii), for some contexts one can *explicitly* give the distribution of each of the independent components $\Big(\sum_{i\in I_k^{(n)}} W_i\Big)_{k=1,\ldots,K}$ of the vector $\boldsymbol{\xi}_n^{w\mathbf{W}}$; this will ease the corresponding concrete simulations in a batch procedure. For instance, we shall give those in the Examples 48, 50, 53, 54 and 55 in the Section V below.
(v) Consider the special "degenerate" case where all the data observations are *certain* and thus $(X_i)_{i\in\mathbb{N}}$ is nothing but a *purely deterministic* sequence, say $(\widetilde{x}_i)_{i\in\mathbb{N}}$, of elements $\widetilde{x}_i$ from the arbitrary set $\mathcal{Y} := \{d_1,\ldots,d_K\}$ of K distinct values "of any kind" (e.g., $\mathcal{Y}$ may consist of K distinct numbers); then the corresponding empirical distribution $\mathbb{P}_n^{emp}$ can be identified with the vector $\mathbb{P}_n^{emp} := (p_{n,1}^{emp},\ldots,p_{n,K}^{emp})$ where

$$p_{n,k}^{emp} \;:=\; \frac{1}{n}\cdot n_k \;:=\; \frac{1}{n}\cdot card(\{i\in\{1,\ldots,n\}:\ \widetilde{x}_i = d_k\}) \;=:\; \frac{1}{n}\cdot card(I_k^{(n)}), \quad k\in\{1,\ldots,K\}, \tag{37}$$

and accordingly the required limit behaviour (26) is equivalent to the vector-convergence

$$\lim_{n\to\infty}\Big(\frac{n_1}{n},\ldots,\frac{n_K}{n}\Big) = (p_1,\ldots,p_K) \quad \text{for some } p_1>0,\ \ldots,\ p_K>0 \text{ such that } \sum_{k=1}^{K} p_k = 1. \tag{38}$$

Correspondingly, with the notations $\mathbb{P} := (p_1,\ldots,p_K)$ and $\widetilde{\mathbf{x}} := (\widetilde{x}_1,\ldots,\widetilde{x}_n)$, the vector-form part of the assertion (34) of Theorem 14 becomes

$$-\lim_{n\to\infty}\frac{1}{n}\log \Pi\Big[\boldsymbol{\xi}_{n,\widetilde{\mathbf{x}}}^{w\mathbf{W}} \in \boldsymbol{\Omega}\Big] \;=\; \inf_{\mathbb{Q}\in\boldsymbol{\Omega}}\ \inf_{m\neq 0} D_\varphi(m\cdot\mathbb{Q},\mathbb{P}) \;=\; \inf_{m\neq 0}\ \inf_{\mathbb{Q}\in\boldsymbol{\Omega}} D_\varphi(m\cdot\mathbb{Q},\mathbb{P}) \tag{39}$$

for all subsets $\boldsymbol{\Omega} \subset \mathbb{S}^K$ satisfying the regularity properties (7) and the finiteness property (9); notice that the conditional probability $\Pi_{X_1^n}[\,\cdot\,]$ has degenerated to the ordinary probability $\Pi[\,\cdot\,]$.
(vi) In a similar fashion to the proof of (the special degenerate case (v) of) Theorem 14, one can show

$$-\lim_{n\to\infty}\frac{1}{n}\log \Pi\Big[\boldsymbol{\xi}_{n}^{w\mathbf{W}} \in \boldsymbol{\Omega}\Big] \;=\; \inf_{\mathbb{Q}\in\boldsymbol{\Omega}}\ \inf_{m\neq 0} D_\varphi(m\cdot\mathbb{Q},\mathbb{P}) \;=\; \inf_{m\neq 0}\ \inf_{\mathbb{Q}\in\boldsymbol{\Omega}} D_\varphi(m\cdot\mathbb{Q},\mathbb{P}) \tag{40}$$

for all subsets $\boldsymbol{\Omega} \subset \mathbb{S}^K$ with regularity properties (7) and the finiteness property (9), where

$$\boldsymbol{\xi}_n^{w\mathbf{W}} \;:=\; \begin{cases} \left(\dfrac{\sum_{i\in I_1^{(n)}} W_i}{\sum_{k=1}^K \sum_{i\in I_k^{(n)}} W_i},\ldots,\dfrac{\sum_{i\in I_K^{(n)}} W_i}{\sum_{k=1}^K \sum_{i\in I_k^{(n)}} W_i}\right) = \dfrac{n\cdot\boldsymbol{\xi}_n^{\mathbf{W}}}{\sum_{i=1}^n W_i}, & \text{if } \sum_{j=1}^n W_j \neq 0,\\ (\infty,\ldots,\infty) =: \boldsymbol{\infty}, & \text{if } \sum_{j=1}^n W_j = 0, \end{cases} \tag{41}$$

with $I_1^{(n)} := \{1,\ldots,n_1\}$, $I_2^{(n)} := \{n_1+1,\ldots,n_1+n_2\}$, $\ldots$, $I_K^{(n)} := \{\sum_{k=1}^{K-1} n_k + 1,\ldots,n\}$ and $n_k := \lfloor n\cdot p_k\rfloor$ ($k\in\{1,\ldots,K\}$) for some pregiven known probability vector $\mathbb{P} := (p_1,\ldots,p_K)$. Recall the definition of $\boldsymbol{\xi}_n^{\mathbf{W}}$ in (23) (with $\mathbf{W}$ instead of $\widetilde{\mathbf{W}}$). The limit behaviour (40) contrasts to the one of Theorem 10, where

$$-\lim_{n\to\infty}\frac{1}{n}\log \Pi\Big[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \Omega/M_{\mathbf{P}}\Big] = \inf_{Q\in\boldsymbol{\Omega}} D_\varphi(\mathbf{Q},\mathbf{P}) \qquad \text{(cf. (24))}$$

for any $\boldsymbol{\Omega} \subset \mathbb{R}^K$ with regularity properties (7) and the finiteness property (9); recall that $(\widetilde{W}_i)_{i\in\mathbb{N}}$ are i.i.d. random variables with probability distribution $\widetilde{\zeta}$ (being connected with the divergence generator $\widetilde{\varphi} := M_{\mathbf{P}}\cdot\varphi$ via the representability (21)), whereas $(W_i)_{i\in\mathbb{N}}$ are i.i.d. random variables with probability distribution ζ (being connected with the divergence generator φ via the representability (6)). Indeed, the construction leading to Theorem 10 does not hold any longer when $\boldsymbol{\Omega} \subset \mathbb{S}^K$ is a set of vectors within the probability simplex $\mathbb{S}^K$ and $\mathbf{P} \in \mathbb{S}_{>0}^K$ is a known vector in this simplex with no zero entries. In such a case, one has to use (40) and (41) instead. Notice that for each constant $A>0$, (40) can be rewritten as

$$-\lim_{n\to\infty}\frac{1}{n}\log \Pi\Big[\boldsymbol{\xi}_{n}^{w\mathbf{W}} \in \boldsymbol{\Omega}\Big] = \inf_{\mathbf{Q}\in A\cdot\boldsymbol{\Omega}}\ \inf_{m\neq 0} D_\varphi\Big(\frac{m}{A}\cdot\mathbf{Q},\mathbb{P}\Big) = \inf_{\mathbf{Q}\in A\cdot\boldsymbol{\Omega}}\ \inf_{\widetilde{m}\neq 0} D_\varphi(\widetilde{m}\cdot\mathbf{Q},\mathbb{P}) = \inf_{\widetilde{m}\neq 0}\ \inf_{\mathbf{Q}\in A\cdot\boldsymbol{\Omega}} D_\varphi(\widetilde{m}\cdot\mathbf{Q},\mathbb{P}); \tag{42}$$

therein, the constraint $\mathbf{Q} \in A\cdot\boldsymbol{\Omega}$ means geometrically that the vector $\mathbf{Q}$ lives in a subset of a simplex which is parallel to the simplex $\mathbb{S}^K$ of probability vectors and which is cut off at the edges of the first/positive orthant; in view of Remark 7(d) and (42), we can also handle such a situation. Namely, in the light of the third expression in (42) in combination with (12) to (14) for the special case of $\boldsymbol{\Omega} := \boldsymbol{\Omega}$ lying in the probability simplex, it makes sense to study e.g. functional relationships between $\inf_{\widetilde{m}\neq 0} D_{\widetilde{c}\cdot\varphi}(\widetilde{m}\cdot\mathbf{Q},\mathbb{P})$ and $D_{\widetilde{c}\cdot\varphi}(\mathbf{Q},\mathbb{P})$ $(\widetilde{c}>0)$ for $\mathbf{Q} \in A\cdot\mathbb{S}^K$ with arbitrary $A>0$ not necessarily being equal to 1 (i.e. $\mathbf{Q} = A\cdot\mathbb{Q}$ for some probability distribution $\mathbb{Q}$). Indeed, such a context appears naturally e.g. in connection with mass transportation problems (cf. (102) below) and with distributed energy management (cf. the paragraph after (113)); the special case $A = 1/K$ of (42) will also be used below for the application of our BS method to solving *(generalized) minimum/maximum entropy problems* for probability vectors (and even for sub-/super-probability vectors) $\mathbb{Q}$ with constraints.

Let us proceed with the main context. As indicated in Remark 15(i), in a number of important cases the limit in the above Theorem 14 can be stated in terms of an invertible function G^{-1} (cf. (2)) of $\inf_{\mathbb{Q}\in\boldsymbol{\Omega}} D_\varphi(\mathbb{Q},\mathbb{P})$ by elimination of m. As explained above, for the degenerate case (cf. Remark 15 (v), (vi)) the search for G^{-1} is even interesting for the more general infimum over non-probability vectors. This is the scope of the following development.

1) Construction principle for the estimation of the minimum divergence, the power-type case :

Within the context of Theorem 14 respectively Remark 15 (v) and (vi), we obtain an explicit solution for the inner (i.e. $m-$concerning) minimization in (34) for the important case of power-divergence generators $\varphi_\gamma : \mathbb{R} \mapsto [0,\infty]$ defined by

$$\varphi_\gamma(t) := \begin{cases} \frac{t^\gamma - \gamma\cdot t + \gamma - 1}{\gamma\cdot(\gamma-1)}, & \text{if } \gamma \in]-\infty,0[\text{ and } t \in]0,\infty[, \\ -\log t + t - 1, & \text{if } \gamma = 0 \text{ and } t \in]0,\infty[, \\ \frac{t^\gamma - \gamma\cdot t + \gamma - 1}{\gamma\cdot(\gamma-1)}, & \text{if } \gamma \in]0,1[\text{ and } t \in [0,\infty[, \\ t\cdot\log t + 1 - t, & \text{if } \gamma = 1 \text{ and } t \in [0,\infty[, \\ \frac{t^\gamma - \gamma\cdot t + \gamma - 1}{\gamma\cdot(\gamma-1)} \cdot \mathbb{1}_{]0,\infty[}(t) + (\frac{1}{\gamma} - \frac{t}{\gamma-1}) \cdot \mathbb{1}_{]-\infty,0]}(t), & \text{if } \gamma \in]1,2[\text{ and } t \in]-\infty,\infty[, \\ \frac{(t-1)^2}{2}, & \text{if } \gamma = 2 \text{ and } t \in]-\infty,\infty[, \\ \frac{t^\gamma - \gamma\cdot t + \gamma - 1}{\gamma\cdot(\gamma-1)} \cdot \mathbb{1}_{]0,\infty[}(t) + (\frac{1}{\gamma} - \frac{t}{\gamma-1}) \cdot \mathbb{1}_{]-\infty,0]}(t), & \text{if } \gamma \in]2,\infty[\text{ and } t \in]-\infty,\infty[, \\ \infty, & \text{else,} \end{cases} \tag{43}$$

which for arbitrary multiplier $\widetilde{c} > 0$ generate (the vector-valued form of) the *generalized power divergences* displayed in the first six rows of Table 1 (and beyond), i.e.

$$D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbf{Q},\mathbf{P}) := \begin{cases} \widetilde{c}\cdot\Big\{ \frac{\sum_{k=1}^K (q_k)^\gamma\cdot(p_k)^{1-\gamma}}{\gamma\cdot(\gamma-1)} - \frac{1}{\gamma-1}\cdot\sum_{k=1}^K q_k + \frac{1}{\gamma}\cdot\sum_{k=1}^K p_k \Big\}, & \text{if } \gamma \in]-\infty,0[,\ \mathbf{P} \in \mathbb{R}^K_{\geq 0} \text{ and } \mathbf{Q} \in \mathbb{R}^K_{>0}, \\ \widetilde{c}\cdot\Big\{ \sum_{k=1}^K p_k\cdot\log\Big(\frac{p_k}{q_k}\Big) + \sum_{k=1}^K q_k - \sum_{k=1}^K p_k \Big\}, & \text{if } \gamma = 0,\ \mathbf{P} \in \mathbb{R}^K_{\geq 0} \text{ and } \mathbf{Q} \in \mathbb{R}^K_{>0}, \\ \widetilde{c}\cdot\Big\{ \frac{\sum_{k=1}^K (q_k)^\gamma\cdot(p_k)^{1-\gamma}}{\gamma\cdot(\gamma-1)} - \frac{1}{\gamma-1}\cdot\sum_{k=1}^K q_k + \frac{1}{\gamma}\cdot\sum_{k=1}^K p_k \Big\}, & \text{if } \gamma \in]0,1[,\ \mathbf{P} \in \mathbb{R}^K_{\geq 0} \text{ and } \mathbf{Q} \in \mathbb{R}^K_{\geq 0}, \\ \widetilde{c}\cdot\Big\{ \sum_{k=1}^K q_k\cdot\log\Big(\frac{q_k}{p_k}\Big) - \sum_{k=1}^K q_k + \sum_{k=1}^K p_k \Big\}, & \text{if } \gamma = 1,\ \mathbf{P} \in \mathbb{R}^K_{>0} \text{ and } \mathbf{Q} \in \mathbb{R}^K_{\geq 0}, \\ \widetilde{c}\cdot\Big\{ \sum_{k=1}^K \frac{(q_k)^\gamma\cdot(p_k)^{1-\gamma}}{\gamma\cdot(\gamma-1)}\cdot\mathbb{1}_{[0,\infty[}(q_k) - \frac{1}{\gamma-1}\cdot\sum_{k=1}^K q_k + \frac{1}{\gamma}\cdot\sum_{k=1}^K p_k \Big\}, & \text{if } \gamma \in]1,2[,\ \mathbf{P} \in \mathbb{R}^K_{>0} \text{ and } \mathbf{Q} \in \mathbb{R}^K, \\ \widetilde{c}\cdot\sum_{k=1}^K \frac{(q_k-p_k)^2}{2\cdot p_k}, & \text{if } \gamma = 2,\ \mathbf{P} \in \mathbb{R}^K_{>0} \text{ and } \mathbf{Q} \in \mathbb{R}^K, \\ \widetilde{c}\cdot\Big\{ \sum_{k=1}^K \frac{(q_k)^\gamma\cdot(p_k)^{1-\gamma}}{\gamma\cdot(\gamma-1)}\cdot\mathbb{1}_{[0,\infty[}(q_k) - \frac{1}{\gamma-1}\cdot\sum_{k=1}^K q_k + \frac{1}{\gamma}\cdot\sum_{k=1}^K p_k \Big\}, & \text{if } \gamma \in]2,\infty[,\ \mathbf{P} \in \mathbb{R}^K_{>0} \text{ and } \mathbf{Q} \in \mathbb{R}^K, \\ \infty, & \text{else;} \end{cases} \tag{44}$$

notice that one has the straightforward relationship $D_{\widetilde{c}\cdot\varphi_\gamma}(\cdot,\cdot) = \widetilde{c}\cdot D_{\varphi_\gamma}(\cdot,\cdot)$; however, as a motivation for the introduction of $\widetilde{c} > 0$, we shall show in the Examples 48, 50, 53, 54 in Section V below that the corresponding probability distribution ζ of the W_i's depends on $\widetilde{c}$ in a non-straightforward way (see also Remark 15 (vi) for another motivation for $\widetilde{c}$). In the course of this, it turns out that $\widetilde{c}\cdot\varphi_\gamma \in \Upsilon(]a_\gamma,\infty[)$ with $a_\gamma = 0$ for $\gamma \in]-\infty,1]$ and $a_\gamma = -\infty$ for $\gamma \in [2,\infty[$.

For $\widetilde{c} = 1$ and probability vectors $\mathbb{Q}$, $\mathbb{P}$ in $\mathbb{S}^K$ respectively $\mathbb{S}^K_{>0}$, the divergences (44) simplify considerably, namely to the well-known *power divergences* $D_{\varphi_\gamma}(\mathbb{Q},\mathbb{P})$ in the scaling of e.g. Liese & Vajda [217] (in other scalings they are also called *Rathie & Kannapan's non-additive directed divergences of order* γ [302], *Cressie-Read divergences* [93] [303], *relative Tsallis entropies or Tsallis cross-entropies* [364] (see also Shiino [331]), *Amari's alpha-divergences* [12]); for some comprehensive overviews on power divergences $D_{\varphi_\gamma}(\mathbb{Q},\mathbb{P})$ — including statistical applications to goodness-of-fit testing and minimum distance estimation — the reader is referred to the insightful books of e.g. Liese & Vajda [217], Read & Cressie [303], Vajda [371], Stummer [344], Pardo [282], Liese & Miescke [216], the survey articles of e.g. Liese & Vajda [218], Vajda & van der Meulen [374], and the references therein. Prominent and widely used special cases of $D_{\varphi_\gamma}(\mathbb{Q},\mathbb{P})$ are the omnipresent *Kullback-Leibler information divergence (relative entropy)* where $\gamma = 1$, the equally important *reverse Kullback-Leibler information divergence (reverse relative entropy)* where $\gamma = 0$, the *Pearson chi-square divergence* ($\gamma = 2$), the *Neyman chi-square divergence* ($\gamma = -1$), the *Hellinger divergence* ($\gamma = \frac{1}{2}$, also called squared Hellinger distance, squared Matusita distance [256] or squared Hellinger-Kakutani metric, see e.g. Deza & Deza [113] [17]). Some exemplary (relatively) recent studies and applications of power divergences $D_{\varphi_\gamma}(\mathbb{Q},\mathbb{P})$ — aside from the vast statistical literature (including in particular maximum likelihood estimation and

[17]in some literature, the (square root of the) Hellinger divergence (HD) is misleadingly called Bhattacharyya distance; however, the latter is *basically* some rescaled logarithm of HD, namely $R_{1/2}(\mathbb{Q},\mathbb{P})$ (cf. (69) with $\gamma = 1/2$)

Pearson's chi-square test) — appear e.g. in Matsuyama [253] for flexibilizations of the well-known expectation-maximization (EM) algorithm and their uses for big-data completion (cf. [254]) and data credit computation in blockchain networks (cf. [255]), Ku & Fine [199] in connection with blind source separation, Stummer & Vajda [348] as well as Stummer & Lao [347] for optimal decisions about some alternative financial models, Berend et al. [41] for the derivation of a kind of reverse Pinsker's inequality (with $\gamma = 1$), Verrelst et al. [383] in geoscientific remote sensing via semiautomatic mapping of biophysical parameters from optical earth observations, Salem at al. [315] for automatic alarm-triggering detection of events (e.g. patient health degradations) from collected data by biomedical sensors, Fu et al. [128] for the study of income distributions in China, Ha et al. [151] for $x-$ray spectrum reconstruction in computer tomography (CT) systems (with $\gamma = 1$), Iqbal & Seghouane [161] for robust sequential dictionary learning, Luppino et al. [233] for unsupervised change detection in heterogeneous multi-temporal satellite images (with $\gamma = \frac{1}{2}$), Sason [320] in connection with data-processing and majorization inequalities, Krömer & Stummer [198] for the smoothing and error-correcting of crude mortality rates (where they even employ non-probability-type vectors), Bekhet & Ahmed [37] for effectiveness evaluations in video retrieval (with $\gamma = -1$, $\gamma = \frac{1}{2}$), Cai et al. [71] for the stabilization of trainings of generative adversarial networks (GANs), Fu et al. [129] for automatic molecule optimization, Görtler et al. [139] for dimensionality reduction on uncertain data in visualization and computer graphics (with $\gamma = \frac{1}{2}$), Kammerer & Stummer [179] for optimal decision making in the presence of pandemics (e.g. COVID-19), Kanapram et al. [180] for the development of collective self-awareness in a network of connected and autonomous vehicles through agent-centered detection of abnormal situations (with $\gamma = \frac{1}{2}$), Kumbhakar [206] for modelling the streamwise velocity profile in open-channel flows, Sigmon et al. [335] for the improvement of genetic quality control in mouse research for biomedical applications (with $\gamma = 2$), Zhang et al. [420] for the design of a noise-adaptation adapted generative adversarial network for medical image analysis (with $\gamma = \frac{1}{2}$), Chen et al. [79] for clustering high-dimensional microbial data from RNA sequencing (with $\gamma = \frac{1}{2}$), Dharmawan et al. [114] for the development of improvements in long-term cell observations via semiconductor-chips-based lensless holographic microscopy, Liu & Sun [229] for analyzing approximate inferences in Bayesian neural networks, Rekavandi et al. [307] for detections in functional magnetic resonance imaging (fMRI) as well as hyperspectral and synthetic aperture radar (SAR) data, Seghouane & Shokouhi [325] for adaptive learning within robust radial basis function networks (RBFN), and Wang et al. [388] for recommender-system relevant collaborative filtering in sparse data.

For $\widetilde{c} = 1$ and nonnegative-component vectors $\mathbf{Q}, \mathbf{P}$ in $\mathbb{R}^K_{\geq 0}$ respectively $\mathbb{R}^K_{>0}$, the generalized power divergences $D_{\varphi_\gamma}(\mathbf{Q}, \mathbf{P})$ of (44) also (partially) simplify, and were treated by Stummer & Vajda [349] (for even more general probability measures, deriving e.g. also generalized Pinsker's inequalities); for a more general comprehensive technical treatment see also e.g. Broniatowski & Stummer [64].

Returning to the general context, in Theorem 14 we stated that for each $\mathbb{P} \in \mathbb{S}^K_{>0}$ the function $\mathbb{Q} \mapsto \inf_{m \neq 0} D_\varphi(m \cdot \mathbb{Q}, \mathbb{P})$ is BS-minimizable (cf. (2)) on all sets $\boldsymbol{\Omega} \subset \mathbb{S}^K$ satisfying (7) and (9). The (corresponding subsetup of the) following Lemma 16 is the *cornerstone* leading from this statement to BS-minimizability of the function $\mathbb{Q} \mapsto D_\varphi(\mathbb{Q}, \mathbb{P}))$ on those same sets, for the special divergences in (44). After giving the *fundamental preparatory* Lemma 16, we shall derive from it some BS-minimizability/BS-maximizability results for (extensions of) a variety of important, widely used, closely related divergences respectively entropy/diversity indices. To achieve this in a transparent way, we employ the following three fundamental quantities $H_\gamma(\mathbf{Q}, \mathbb{P})$, $I(\mathbf{Q}, \mathbb{P})$ and $\widetilde{I}(\mathbf{Q}, \mathbb{P})$. To begin with, let $A > 0$ be an arbitrary constant (notice that for the choice $A = 1$, all the following vectors $\mathbf{Q}$ will turn into probability vectors $\mathbb{Q}$). Moreover — for any constellation $(\gamma, \mathbb{P}, \mathbf{Q}) \in \widetilde{\Gamma} \times \widetilde{\mathcal{M}}_1 \times \widetilde{\mathcal{M}}_2$, where $\widetilde{\Gamma} \times \widetilde{\mathcal{M}}_1 \times \widetilde{\mathcal{M}}_2 :=]0,1[\times \mathbb{S}^K \times A \cdot \mathbb{S}^K$ or $\widetilde{\Gamma} \times \widetilde{\mathcal{M}}_1 \times \widetilde{\mathcal{M}}_2 :=]-\infty, 0[\times \mathbb{S}^K \times A \cdot \mathbb{S}^K_{>0}$ or $\widetilde{\Gamma} \times \widetilde{\mathcal{M}}_1 \times \widetilde{\mathcal{M}}_2 :=]1, \infty[\times \mathbb{S}^K_{>0} \times A \cdot \mathbb{S}^K$ — let

$$0 < H_\gamma(\mathbf{Q}, \mathbb{P}) := \sum_{k=1}^{K} (q_k)^\gamma \cdot (p_k)^{1-\gamma} \ = \ 1 + \gamma \cdot (A-1) + \gamma \cdot (\gamma - 1) \cdot D_{\varphi_\gamma}(\mathbf{Q}, \mathbb{P}), \qquad \gamma \in \mathbb{R} \backslash \{0,1\}, \tag{45}$$

be the *modified $\gamma-$order Hellinger integral of* $\mathbf{Q}$ *and* $\mathbb{P}$. Furthermore, for any $\mathbb{P} \in \mathbb{S}^K_{>0}$, $\mathbf{Q} \in A \cdot \mathbb{S}^K$, let

$$-1 \ < \ I(\mathbf{Q}, \mathbb{P}) := \sum_{k=1}^{K} q_k \cdot \log\left(\frac{q_k}{p_k}\right) \ = \ D_{\varphi_1}(\mathbf{Q}, \mathbb{P}) + A - 1, \tag{46}$$

be the *modified Kullback-Leibler information (modified relative entropy)*. Finally, for any $\mathbb{P} \in \mathbb{S}^K$, $\mathbf{Q} \in A \cdot \mathbb{S}^K_{>0}$, let

$$1 - A \ \leq \ \widetilde{I}(\mathbf{Q}, \mathbb{P}) := \sum_{k=1}^{K} p_k \cdot \log\left(\frac{p_k}{q_k}\right) \ = \ D_{\varphi_0}(\mathbf{Q}, \mathbb{P}) + 1 - A, \tag{47}$$

be the *modified reverse Kullback-Leibler information (modified reverse relative entropy)*.

In terms of (45), (46) and (47) we obtain the following

Lemma 16: Let $A > 0$ be an arbitrary constant.
(a) Let $\widetilde{c} > 0$ be arbitrary and $(\gamma, \mathbb{P}, \mathbf{Q}) \in \widetilde{\Gamma} \times \widetilde{\mathcal{M}}_1 \times \widetilde{\mathcal{M}}_2$ as above. Then one has

$$\inf_{m\neq 0} D_{\widetilde{c}\cdot\varphi_\gamma}(m\cdot\mathbf{Q},\mathbb{P}) = \inf_{m>0} D_{\widetilde{c}\cdot\varphi_\gamma}(m\cdot\mathbf{Q},\mathbb{P}) = \frac{\widetilde{c}}{\gamma}\cdot\left[1 - A^{\gamma/(\gamma-1)}\cdot\left[1+\gamma\cdot(A-1)+\frac{\gamma\cdot(\gamma-1)}{\widetilde{c}}\cdot D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbf{Q},\mathbb{P})\right]^{-1/(\gamma-1)}\right] \tag{48}$$

$$= \frac{\widetilde{c}}{\gamma}\cdot\left[1 - A^{\gamma/(\gamma-1)}\cdot H_\gamma(\mathbf{Q},\mathbb{P})^{-1/(\gamma-1)}\right]$$

and consequently for any subset $A\cdot\mathbf{\Omega}\subset\widetilde{\mathcal{M}}_2$

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0} D_{\widetilde{c}\cdot\varphi_\gamma}(m\cdot\mathbf{Q},\mathbb{P}) = \frac{\widetilde{c}}{\gamma}\cdot\left[1 - A^{\gamma/(\gamma-1)}\cdot\left[1+\gamma\cdot(A-1)+\frac{\gamma\cdot(\gamma-1)}{\widetilde{c}}\cdot\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}} D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbf{Q},\mathbb{P})\right]^{-1/(\gamma-1)}\right], \tag{49}$$

$$\arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0}\ D_{\widetilde{c}\cdot\varphi_\gamma}(m\cdot\mathbf{Q},\mathbb{P}) = \arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbf{Q},\mathbb{P}), \tag{50}$$

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0} D_{\varphi_\gamma}(m\cdot\mathbf{Q},\mathbb{P}) = \frac{1}{\gamma}\cdot\left[1 - A^{\gamma/(\gamma-1)}\cdot\left[\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}} H_\gamma(\mathbf{Q},\mathbb{P})\right]^{-1/(\gamma-1)}\right], \qquad \text{for } \gamma<0 \text{ and } \gamma>1, \tag{51}$$

$$\arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0}\ D_{\varphi_\gamma}(m\cdot\mathbf{Q},\mathbb{P}) = \arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ H_\gamma(\mathbf{Q},\mathbb{P}), \qquad \text{for } \gamma<0 \text{ and } \gamma>1, \tag{52}$$

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0} D_{\varphi_\gamma}(m\cdot\mathbf{Q},\mathbb{P}) = \frac{1}{\gamma}\cdot\left[1 - A^{\gamma/(\gamma-1)}\cdot\left[\sup_{\mathbf{Q}\in A\cdot\mathbf{\Omega}} H_\gamma(\mathbf{Q},\mathbb{P})\right]^{-1/(\gamma-1)}\right], \qquad \text{for } \gamma\in]0,1[, \tag{53}$$

$$\arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0}\ D_{\varphi_\gamma}(m\cdot\mathbf{Q},\mathbb{P}) = \arg\sup_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ H_\gamma(\mathbf{Q},\mathbb{P}), \qquad \text{for } \gamma\in]0,1[, \tag{54}$$

provided that the infimum on the right-hand side of (49) exists.
(b) For any $\mathbb{P}\in\mathbb{S}^K_{>0}$, $\mathbf{Q}\in A\cdot\mathbb{S}^K$, $\widetilde{c}>0$ one gets

$$\inf_{m\neq 0} D_{\widetilde{c}\cdot\varphi_1}(m\cdot\mathbf{Q},\mathbb{P}) = \inf_{m>0} D_{\widetilde{c}\cdot\varphi_1}(m\cdot\mathbf{Q},\mathbb{P}) = \widetilde{c}\cdot\left[1 - A\cdot\exp\left(-\frac{1}{A\cdot\widetilde{c}}\cdot D_{\widetilde{c}\cdot\varphi_1}(\mathbf{Q},\mathbb{P}) + \frac{1}{A} - 1\right)\right] \tag{55}$$

$$= \widetilde{c}\cdot\left[1 - A\cdot\exp\left(-\frac{1}{A}\cdot I(\mathbf{Q},\mathbb{P})\right)\right]$$

and consequently for any subset $A\cdot\mathbf{\Omega}\subset A\cdot\mathbb{S}^K$

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0} D_{\widetilde{c}\cdot\varphi_1}(m\cdot\mathbf{Q},\mathbb{P}) = \widetilde{c}\cdot\left[1 - A\cdot\exp\left(-\frac{1}{A\cdot\widetilde{c}}\cdot\inf_{Q\in A\cdot\mathbf{\Omega}} D_{\widetilde{c}\cdot\varphi_1}(\mathbf{Q},\mathbb{P}) + \frac{1}{A} - 1\right)\right], \tag{56}$$

$$\arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0} D_{\widetilde{c}\cdot\varphi_1}(m\cdot\mathbf{Q},\mathbb{P}) = \arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ D_{\widetilde{c}\cdot\varphi_1}(Q,P), \tag{57}$$

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0} D_{\varphi_1}(m\cdot\mathbf{Q},\mathbb{P}) = \left[1 - A\cdot\exp\left(-\frac{1}{A}\cdot\inf_{Q\in A\cdot\mathbf{\Omega}} I(\mathbf{Q},\mathbb{P})\right)\right], \tag{58}$$

$$\arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0} D_{\varphi_1}(m\cdot\mathbf{Q},\mathbb{P}) = \arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ I(\mathbf{Q},\mathbb{P}), \tag{59}$$

provided that the infimum on the right-hand side of (56) exists.
(c) For any $\mathbb{P}\in\mathbb{S}^K$, $\mathbf{Q}\in A\cdot\mathbb{S}^K_{>0}$, $\widetilde{c}>0$ we obtain

$$\inf_{m\neq 0} D_{\widetilde{c}\cdot\varphi_0}(m\cdot\mathbf{Q},\mathbb{P}) = \inf_{m>0} D_{\widetilde{c}\cdot\varphi_0}(m\cdot\mathbf{Q},\mathbb{P}) = D_{\widetilde{c}\cdot\varphi_0}(\mathbf{Q},\mathbb{P}) + \widetilde{c}\cdot(1 - A + \log A) \tag{60}$$

$$= \widetilde{c}\cdot\left(\widetilde{I}(\mathbf{Q},\mathbb{P}) + \log A\right)$$

and consequently for any set subset $A\cdot\mathbf{\Omega}\subset A\cdot\mathbb{S}^K_{>0}$

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0} D_{\widetilde{c}\cdot\varphi_0}(m\cdot\mathbf{Q},\mathbb{P}) = \widetilde{c}\cdot(1 - A + \log A) + \inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ D_{\widetilde{c}\cdot\varphi_0}(\mathbf{Q},\mathbb{P}), \tag{61}$$

$$\arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0} D_{\widetilde{c}\cdot\varphi_0}(m\cdot\mathbf{Q},\mathbb{P}) = \arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ D_{\widetilde{c}\cdot\varphi_0}(\mathbf{Q},\mathbb{P}), \tag{62}$$

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0} D_{\varphi_1}(m\cdot\mathbf{Q},\mathbb{P}) = \log A + \inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \widetilde{I}(\mathbf{Q},\mathbb{P}), \tag{63}$$

$$\arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \inf_{m\neq 0} D_{\varphi_0}(m\cdot\mathbf{Q},\mathbb{P}) = \arg\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\ \widetilde{I}(\mathbf{Q},\mathbb{P}), \tag{64}$$

provided that the infimum on the right-hand side of (61) exists.

The proof of Lemma 16 is given in Appendix C.

Remark 17: Notice that for $\mathbb{P} \in \mathbb{S}^K_{>0}$ and $\mathbf{Q} \in A \cdot \mathbb{S}^K$, the modified Kullback-Leibler information has the property $I(\mathbf{Q}, \mathbb{P}) \geq 0$ if $A \geq 1$ (cf. (46)); otherwise, $I(\mathbf{Q}, \mathbb{P})$ may become negative, as can be easily seen from the case where $\mathbb{P} := \mathbb{P}^{unif} := (\frac{1}{K}, \ldots, \frac{1}{K})$ is the probability vector of frequencies of the uniform distribution on $\{1, \ldots, K\}$, and $\mathbf{Q} := (\frac{1}{K+1}, 0, \ldots, 0)$. Analogously, for $\mathbb{P} \in \mathbb{S}^K$ and $\mathbf{Q} \in A \cdot \mathbb{S}^K_{>0}$ one gets $\widetilde{I}(\mathbf{Q}, \mathbb{P}) \geq 0$ if $A \leq 1$ (cf. (47)); otherwise, $\widetilde{I}(\mathbf{Q}, \mathbb{P})$ may become negative (take e.g. $\mathbf{Q} = (\frac{K+1}{K}, \ldots, \frac{K+1}{K})$ and $\mathbb{P} := (1, 0, \ldots, 0)$).

Remark 18: (a) In the context of Remark 15(vi), according to (42) applied to $\varphi := \widetilde{c} \cdot \varphi_\gamma$, for all cases $\gamma \in]-\infty, 0[\ \cup \ ]0, 1[\ \cup \ [\ 2, \infty[$ the left-hand side of each of (49), (51), (53) is independent of $A > 0$ and equal to $-\lim_{n\to\infty} \frac{1}{n} \log \Pi \Big[\boldsymbol{\xi}_n^{w\mathbf{W}} \in \boldsymbol{\Omega} \Big]$ where — as will be shown below — the corresponding $\mathbf{W}$'s have probability distribution $\mathfrak{C}[\cdot] = \Pi[W_1 \in \cdot\,]$ (cf. (6)) which varies "quite drastically" with γ (and the case $\gamma \in]1, 2[$ has to be even excluded for analytical difficulties [18]). Analogously, each of the left-hand sides of (56), (58), (61), (63) is also independent of $A > 0$ and equal to $-\lim_{n\to\infty} \frac{1}{n} \log \Pi \Big[\boldsymbol{\xi}_n^{w\mathbf{W}} \in \boldsymbol{\Omega} \Big]$ for some $\mathbf{W}$ of respective distribution. Hence, by inversion, all the extremum-describing target quantities $\inf_{\mathbf{Q} \in A \cdot \boldsymbol{\Omega}} D_{\widetilde{c} \cdot \varphi_\gamma}(\mathbf{Q}, \mathbb{P})$ ($\gamma \in \mathbb{R} \backslash]1, 2[$), $\inf_{\mathbf{Q} \in A \cdot \boldsymbol{\Omega}} H_\gamma(\mathbf{Q}, \mathbb{P})$ ($\gamma \in]-\infty, 0[\cup [2, \infty[$), $\sup_{\mathbf{Q} \in A \cdot \boldsymbol{\Omega}} H_\gamma(\mathbf{Q}, \mathbb{P})$ ($\gamma \in]0, 1[$), $\inf_{Q \in A \cdot \boldsymbol{\Omega}} I(\mathbf{Q}, \mathbb{P})$ and $\inf_{Q \in A \cdot \boldsymbol{\Omega}} \widetilde{I}(\mathbf{Q}, \mathbb{P})$ can be expressed as $G\big(-\lim_{n\to\infty} \frac{1}{n} \log \Pi \Big[\boldsymbol{\xi}_n^{w\mathbf{W}} \in \boldsymbol{\Omega} \big) \Big]$ for some explicitly known ($A-$dependent) function G. This means that — in the sense of Definition 1 — all the corresponding four "cornerstone quantities" $D_{\widetilde{c} \cdot \varphi_\gamma}(\mathbf{Q}, \mathbb{P})$, $H_\gamma(\mathbf{Q}, \mathbb{P})$, $I(\mathbf{Q}, \mathbb{P})$, $\widetilde{I}(\mathbf{Q}, \mathbb{P})$ are BS-minimizable, respectively BS-maximizable, on $\Omega = A \cdot \boldsymbol{\Omega}$. The above-mentioned inversions (i.e. constructions of $G(\cdot)$) will be concretely carried out below — namely in the Propositions 22, 23, 24, 25, 26 and 27. In those, we also involve the BS-minimizability/maximizability of several other important closely related divergences and measures of entropy (measures of diversity, measures of heterogeneity/homogeneity, measures of concentration) which (i) are widely used in information theory and its applications to artificial intelligence, machine learning and physics, and which (ii) can be built from the above-mentioned four cornerstone quantities (power divergences, Hellinger integrals, Kullback-Leibler information divergences).
(b) The special case $\varphi := \widetilde{c} \cdot \varphi_\gamma$ ($\gamma \in]-\infty, 0[\cup]0, 1[\cup [\ 2, \infty[$) of Theorem 14 works analogously to (a), with the differences that we employ $A = 1$ (instead of arbitrary $A > 0$), (36) (instead of (42)), $\Pi_{X_1^n}[\cdot]$ (instead of $\Pi[\cdot]$), and $\xi_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\boldsymbol{\xi}_n^{w\mathbf{W}}$).
(c) From the proof of Lemma 16 in Appendix C below, one can see that for the important case $\gamma = 2$ the formulas (48) to (52) also hold for $A < 0$.

In the following, we further elaborate the three points (a),(b) and (c) of Remark 18 "comprehensively and unifyingly", where the expression "BS minimizable/maximizable" always has to be interpreted accordingly in terms of $-\lim_{n\to\infty} \frac{1}{n} \log \Pi[\boldsymbol{\xi}_n^{w\mathbf{W}} \in \cdot\,]$ respectively $-\lim_{n\to\infty} \frac{1}{n} \log \Pi_{X_1^n}[\xi_{n,\mathbf{X}}^{w\mathbf{W}} \in \cdot\,]$ (without explicit mentioning, for the sake of brevity).

Let us fix $\widetilde{c} = 1$ and an arbitrary triple $(\gamma, \mathbb{P}, \mathbf{Q})$ which satisfies the assumptions of Lemma 16(a) with $A := \sum_{k=1}^K q_k > 0$. For such a setup, we have obtained in (45) the $\gamma-$order Hellinger integral (of $\mathbf{Q}$ and $\mathbb{P}$) $H_\gamma(\mathbf{Q}, \mathbb{P}) > 0$, which is not a divergence; as a terminology-concerning side remark, let us mention that $H_\gamma(\mathbf{Q}, \mathbb{P})$ ($\gamma \geq 1$) is called *relative information generating function* in Guiasu & Reischer [144], see e.g. also Clark [90]; moreover, $H_\gamma(\mathbb{Q}, \mathbb{P})$ is sometimes termed *($\gamma-$order) Chernoff coefficient* being a component of the Chernoff distances/informations [85]. Torgersen [361] uses the name *($\gamma-$order) Hellinger transform*. Notice that the special case $\gamma = \frac{1}{2}$ is nothing but (a multiple of) the well-known important *Bhattacharyya coefficient* (cf. [48],[49],[50])

$$BC(\mathbf{Q}, \mathbb{P}) := H_{1/2}(\mathbf{Q}, \mathbb{P}) = \sum_{k=1}^K \sqrt{q_k \cdot p_k} = 1 + \frac{1}{2} \cdot (A - 1) + \frac{1}{2} \cdot (\frac{1}{2} - 1) \cdot D_{\varphi_{\frac{1}{2}}}(\mathbf{Q}, \mathbb{P})$$

which is also known as *affinity* (cf. Matusita [256], see e.g. also Toussaint [362]) and *(classic, non-quantum) fidelity similarity* (cf. e.g. Deza & Deza [113]); for non-probability vectors $\mathbf{P} \in \mathbb{R}^K_{\geq 0}$ one can simply retransform $\mathbb{P} := \frac{\mathbf{P}}{M_\mathbf{P}}$ and thus imbed $BC(\mathbf{Q}, \mathbf{P}) = \sqrt{M_\mathbf{P}} \cdot BC(\mathbf{Q}, \mathbb{P})$ into our BS context. There is a vast literature on very recent applications of the Bhattacharyya coefficient, for instance it appears exemplarily in Peng & Li [289] for object tracking from successive video frames, Ayed et al. [26] for efficient graph cut algorithms, Patra et al. [287] for collaborative filtering in sparse data, El Merabet et al. [119] for region classification in intelligent transport systems in order to compensate the lack of performance of Global Navigation Satellites Systems, Chiu et al. [86] for the design of interactive mobile augmented reality systems, Noh et al. [274] for dimension reduction in interacting fluid flow models, Bai et al. [29] for material defect detection through ultrasonic array imaging, Dixit & Jain [115] for the design of recommender systems on highly sparse context aware datasets, Guan et al. [143] for visible light positioning methods based on image sensors, Lin et al. [220] for probabilistic representation of color image pixels, Chen et al. [80] for distributed compressive video sensing, Jain et al. [162] for the enhancement of multistage user-based collaborative filtering in recommendation systems, Pascuzzo et al. [285] for brain-diffusion-MRI based early diagnosis of the sporadic

[18]because in this case there are some indications that the representation (6) only holds for some *signed* probability distribution $\mathfrak{C}$ (e.g. having a density with positive and negative values).

Creutzfeldt—Jakob disease, Sun et al. [351] for the design of automatic detection methods multitemporal (e.g. landslide) point clouds, Valpione et al. [377] for the investigation of T cell dynamics in immunotherapy, Wang et al. [387] for the tracking and prediction of downbursts from meteorological data, Xu et al. [403] for adaptive distributed compressed video sensing for coal mine monitoring, Zhao et al. [424] for the shared sparse machine learning of the affective content of images, Chen et al. [82] for image segmentation and domain partitioning, De Oliveira et al. [105] for the prediction of cell-penetrating peptides, Eshaghi et al. [122] for the identification of multiple sclerosis subtypes through machine learning of brain MRI scans, Feng et al. [125] for improvements of MRI-based detection of epilepsy-causing cortical malformations, Hanli et al. [153] for designing pilot protection schemes for transmission lines, Jiang et al. [170] for flow-assisted visual tracking through event cameras, Lysiak & Szmajda [235] for comparisons of selected feature quality evaluations, Joel & Sivakumar [172] for the despeckling enhancement of medical ultrasound image quality, Reising et al. [305] for the design of security protection of Internet-of-Things (IoT) devices, Skrbic et al. [338] for the uncovering of interplays between amino acid sequences and local structures in proteins, Tsiapoki et al. [365] for the improvement of the detection performance of structural health monitoring frameworks, van Molle et al. [381] for uncertainty quantification in deep neural networks, Yang et al. [413] for the determination of the onset of transient signals, and Zhou & Yu [427] for the modelling of spatiotemporal human eye movements.

To proceed with the general context, for any $\gamma \in]-\infty,0[\, \cup\,]0,1[\, \cup\, [\,2,\infty[$ let the function $h_\gamma :]0,\infty[\mapsto]-\infty,\infty[$ be such that $x \mapsto h_\gamma(1+\gamma\cdot(A-1)+\gamma\cdot(\gamma-1)\cdot x)$ is continuous and strictly increasing (respectively, strictly decreasing) for all $x \geq 0$ with $1+\gamma\cdot(A-1)+\gamma\cdot(\gamma-1)\cdot x > 0$; since $D_{\varphi_\gamma}(\mathbf{Q},\mathbb{P})$ is BS-minimizable on $\mathbf{\Omega} = A\cdot\mathbf{\mathfrak{Q}}$, then also the — not necessarily nonnegative — quantity $h_\gamma\Big(1+\gamma\cdot(A-1)+\gamma\cdot(\gamma-1)\cdot D_{\varphi_\gamma}(\mathbf{Q},\mathbb{P})\Big) = h_\gamma\Big(H_\gamma(\mathbf{Q},\mathbb{P})\Big)$ is BS-minimizable (respectively, BS-maximizable) on $\mathbf{\Omega} = A\cdot\mathbf{\mathfrak{Q}}$. If h_γ satisfies additionally $h_\gamma(1) = 0$ as well as $h_\gamma(1+\gamma\cdot(A-1)+\gamma\cdot(\gamma-1)\cdot x) \geq 0$ for all $x \geq 0$ with $1+\gamma\cdot(A-1)+\gamma\cdot(\gamma-1)\cdot x > 0$, then $D_{h_\gamma}(\mathbf{Q},\mathbb{P}) := h_\gamma\Big(1+\gamma\cdot(A-1)+\gamma\cdot(\gamma-1)\cdot D_{\varphi_\gamma}(\mathbf{Q},\mathbb{P})\Big) = h_\gamma\Big(H_\gamma(\mathbf{Q},\mathbb{P})\Big)$ constitutes a divergence [19] which is BS-minimizable on $\mathbf{\Omega} = A\cdot\mathbf{\mathfrak{Q}}$ (respectively, BS-maximizable on $\mathbf{\Omega} = A\cdot\mathbf{\mathfrak{Q}}$).

Let us consider some important examples. For the identity mapping $h_\gamma^{Id}(y) := y \quad (y > 0)$ the function $x \mapsto 1+\gamma\cdot(A-1)+\gamma\cdot(\gamma-1)\cdot x$ is strictly increasing for $\gamma < 0$ and $\gamma > 1$ (on the required domain of x), and strictly decreasing for $\gamma \in]0,1[$. Accordingly, $H_\gamma(\mathbf{Q},\mathbb{P})$ is BS-minimizable on $\mathbf{\Omega} = A\cdot\mathbf{\mathfrak{Q}}$ for $\gamma < 0$ and $\gamma \geq 2$ and BS-maximizable on $\mathbf{\Omega} = A\cdot\mathbf{\mathfrak{Q}}$ for $\gamma \in]0,1[$ (this is consistent with (51), (53)); in particular, the Bhattacharyya coefficient $BC(\mathbf{Q},\mathbb{P})$ is BS-maximizable on $\mathbf{\Omega} = A\cdot\mathbf{\mathfrak{Q}}$. Some other important choices are

$$h_\gamma(y) := h_{c_1,c_2,c_3}(y) := c_1\cdot\Big(y^{c_2}-c_3\Big), \qquad y>0,\ c_1,c_2\in\mathbb{R}\backslash\{0\},\ c_3\in\mathbb{R}, \tag{65}$$

$$h_\gamma(y) := h^R_{c_4,f}(y) := \lim_{c_2\to 0} h_{c_4/f(c_2),c_2,1}(y) = \frac{c_4}{f'(0)}\cdot\log(y), \qquad y>0,\ c_4\in\mathbb{R}\backslash\{0\}, \tag{66}$$

$$h_\gamma(y) := h^{GB2}_{c_5,c_6}(y) := c_5\cdot(\arccos(y))^{c_6}, \qquad \gamma\in]0,1[,\ y\in]0,1],\ c_5>0,\ c_6>0, \tag{67}$$

$$h_\gamma(y) := h^{BB}_{\nu,c_7}(y) := c_7\cdot\frac{\log(1-\frac{1-y}{\nu})}{\log(1-\frac{1}{\nu})}, \qquad \gamma\in]0,1[,\ y\in]0,1],\ c_7>0,\ \nu\in]-\infty,0[\cup]1,\infty[, \tag{68}$$

where the constants c_1 to c_7 may depend on γ, and f is some (maybe $\gamma-$dependent) function which is differentiable in a neighborhood of 0 and satisfies $f(0) = 0$, $f'(0) \neq 0$ (e.g. $f(z) = c_8\cdot z$ for some non-zero constant c_8). Clearly, $h_{c_1,c_2,c_3}(\cdot)$ is strictly increasing (respectively, strictly decreasing) if and only if $c_1\cdot c_2 > 0$ (respectively, $c_1\cdot c_2 < 0$). Moreover, $h^R_{c_4,f}(\cdot)$ is strictly increasing (respectively, strictly decreasing) if and only if $\frac{c_4}{f'(0)} > 0$ (respectively, $\frac{c_4}{f'(0)} < 0$). Furthermore, both $h^{GB2}_{c_5,c_6}(\cdot)$ and $h^{GoBa}_{\nu,c_7}(\cdot)$ are strictly decreasing.

For instance, the special case $h_\gamma(y) = h^R_{c_4,Id}(y)$ with $c_4 := \frac{1}{\gamma\cdot(\gamma-1)}$ (recall that $\gamma \in]-\infty,0[\, \cup\,]0,1[\, \cup\, [\,2,\infty[$) and identity function $f := Id$ leads to the quantities

$$\begin{aligned} R_\gamma(\mathbf{Q},\mathbb{P}) &:= D_{h^R_{c_4,Id}}(\mathbf{Q},\mathbb{P}) = \frac{\log\Big(1+\gamma\cdot(A-1)+\gamma\cdot(\gamma-1)\cdot D_{\varphi_\gamma}(\mathbf{Q},\mathbb{P})\Big)}{\gamma\cdot(\gamma-1)} = \frac{\log\Big(H_\gamma(\mathbf{Q},\mathbb{P})\Big)}{\gamma\cdot(\gamma-1)} \\ &= \frac{\log\Big(\sum_{k=1}^K (q_k)^\gamma\cdot(p_k)^{1-\gamma}\Big)}{\gamma\cdot(\gamma-1)}, \qquad \gamma\in]-\infty,0[\,\cup\,]0,1[\,\cup\,[\,2,\infty[, \end{aligned} \tag{69}$$

(provided that all involved power divergences are finite), which are thus BS-minimizable on $\mathbf{\Omega} = A\cdot\mathbf{\mathfrak{Q}}$; notice that $R_\gamma(\mathbf{Q},\mathbb{P}) \geq 0$ if $\gamma \in]0,1[\, \cup\, [\,2,\infty[$ together with $A \in [1,\infty[$, and if $\gamma \in]-\infty,0[$ together with $A \in]0,1]$. The special subcase $A = 1$ in (69) (and thus, $\mathbf{Q}$ is a probability vector $\mathbb{Q}$) corresponds to the prominent *Renyi divergences/distances* [309] (in the scaling of e.g. Liese & Vajda [217] and in probability-vector form), see e.g. van Erven & Harremoes [380] for a comprehensive study of their properties; as a side remark, $\gamma\cdot(\gamma-1)\cdot R_\gamma(\mathbb{Q},\mathbb{P})$ is also employed in the Chernoff distances/informations [85]. The special subcase $R_{1/2}(\mathbb{Q},\mathbb{P})$ (i.e. $\gamma = 1/2$ and $A = 1$ in (69)) corresponds to (a multiple of) the widely used *Bhattacharyya*

[19] in the usual sense that $D_{h_\gamma}(\mathbb{Q},\mathbb{P}) \geq 0$ with equality iff $\mathbb{Q} = \mathbb{P}$.

distance (of type 1) between $\mathbb{Q}$ and $\mathbb{P}$, cf. [48] (see e.g. also Kailath [178]). Sometimes, $\exp(R_\gamma(\mathbf{Q}, \mathbb{P}))$ is also called *Renyi divergence/distance*. Some exemplary (relatively) recent studies and applications of Renyi divergences $R_\gamma(\mathbb{Q}, \mathbb{P})$ (respectively, their multiple or exponential) — aside from the substantial statistical literature — appear e.g. in Zhao et al. [423] for the study of isomeric stability of fullerenes (which are e.g. employed for state-of-the-art organic solar cells), in the papers of Sundaresan [353], Bunte & Lapidoth [67], Sason [318], Kumar et al. [205] for (mismatch-cases of) coding and guessing as well as task partitioning, in the papers of Prest [296], Bai et al. [30] for lattice-based cryptography, in He et al. [158] for robot active olfaction search (by infotaxis) in turbulent flows, in Momeni et al. [263] for the design of reprogrammable encrypted graphene-based coding metasurfaces, in Staszowska et al. [340] for accurate and precise cluster analysis for super-resolution localization microscopy, in Yu & Tan [414] for distributed source simulation problems, in Zhang et al. [419] for sensor control, in Yu & Tan [415] for the so-called random variable simulation problem, in Blanchet et al. [55] for the robust treatment of extreme values in rainfall accumulation data, in Cai et al. [70] for sensor tasking for search and catalog maintenance of geosynchronous space objects, in Gholami & Hodtani [134] for refinements of safety-and-security-targeted location verification systems in wireless communication networks (e.g in Intelligent Transportation Systems (ITSs) and vehicular technology), in Seweryn et al. [326] for the assessment of similarity and diversity of expression profiles in single cell systems, in Zhou [426] for the study of secrecy constraints in key generation problems where side information might be present at untrusted users, in Makkawi et al. [243] for the design of an automated decision-support framework for adaptive diagnosis of fault–tolerant multi–sensor data fusion for vehicle localization, in Mao et al. [245] for privacy-preserving computation offloading for parallel deep neural networks training.

There is vast literature on recent applications of the above-mentioned special case $R_{1/2}(\mathbb{Q}, \mathbb{P})$ — that is, the Bhattacharyya distance (of type 1); for instance, it appears in Tarighati & Jalden [356] for rate balancing in wireless sensor networks, Bi et al. [51], [52] for certain uncertainty quantifications respectively stochastic sensitivity analyses in mechanical systems and signal processing, Fu & He [130] for the design of multibit quantizers for cooperative spectrum sensing in cognitive radio networks, Cohen et al. [91] for adaptive and causal random linear network coding with forward error correction for a point-to-point communication channel with delayed feedback, Xu et al. [401] for cost minimization problems of big data analytics on geo-distributed data centers connected to renewable energy sources with unpredictable capacity, Xu et al. [402] for community identification in networks, Arrigoni & Madsen [18] for automated discovering of low-energy defect configurations in materials, Fan et al. [123] for region-merging-based methods for synthetic aperture radar (SAR) image segmentation, Mahfouz et al. [242] for some refined ensemble classifications in microarray-based automated cancer diagnosis, Matchev & Shyamsundar [252] for some machine-learning based signal discovery in high energy physics (HEP) experiments, Wang et al. [386] for the investigation of intratumoral heterogeneity (ITH) of some gastric cancer, Webster et al. [389] for the characterization, identification, clustering and classification of disease, and Xiahou et al. [396] for the prediction of remaining useful life (RUL) through fusion of expert knowledge and condition monitoring information.

As a further example, consider

$$\begin{aligned}\mathcal{B}_{\gamma,c_5,c_6}(\mathbb{Q}, \mathbb{P}) &:= D_{h^{GB2}_{c_5,c_6}}(\mathbb{Q}, \mathbb{P}) = c_5 \cdot \Big(\arccos\Big(1 + \gamma\cdot(\gamma-1)\cdot D_{\varphi_\gamma}(\mathbb{Q}, \mathbb{P})\Big)\Big)^{c_6} = c_5 \cdot \Big(\arccos\Big(H_{\varphi_\gamma}(\mathbb{Q}, \mathbb{P})\Big)\Big)^{c_6} \\ &= c_5 \cdot \Big(\arccos\Big(\sum_{k=1}^{K} (q_k)^\gamma \cdot (p_k)^{1-\gamma}\Big)\Big)^{c_6} \geq 0\,, \qquad \gamma \in]0,1[,\ c_5 > 0,\ c_6 > 0,\end{aligned}$$

which is BS-maximizable on $\mathbf{\Omega}$. The case $\mathcal{B}_{1/2,1,1}(\mathbb{Q}, \mathbb{P})$ corresponds to the well-known *Bhattacharyya arccos distance (Bhattacharyya distance of type 2)* in [50] (which is also called Wootters distance [395]), and $\mathcal{B}_{1/2,1,2}(\mathbb{Q}, \mathbb{P})$ to its variant in [49]; the case $\mathcal{B}_{1/2,2,1}(\mathbb{Q}, \mathbb{P})$ is known as *Fisher distance* or *Rao distance* or *geodesic distance* (see e.g. Deza & Deza [113]); a nice graphical illustration of the geometric connection between the Fisher distance $\mathcal{B}_{1/2,2,1}(\mathbb{Q}, \mathbb{P})$ and the Hellinger distance/metric $\sqrt{\frac{1}{2}\cdot D_{\varphi_{1/2}}(\mathbb{Q}, \mathbb{P})}$ can be found e.g. on p.35 in Ay et al. [25]. Some exemplary applications of the Bhattacharyya arccos distance $\mathcal{B}_{1/2,1,1}(\mathbb{Q}, \mathbb{P})$ can be found e.g. in Rao [301] and Juhasz [177] for cluster analysis of human populations, in Martin-Fernandez et al. [251] for general hierarchical clustering, Greenacre [141] for metric scaling, and in Chen et al. [79] for clustering high-dimensional microbial data from RNA sequencing.

Let us give another example, namely

$$\begin{aligned}\widetilde{\mathcal{B}}_{\gamma,\nu,c_7}(\mathbb{Q}, \mathbb{P}) &:= D_{h^{BB}_{\nu,c_7}}(\mathbb{Q}, \mathbb{P}) = \frac{c_7}{\log(1-\frac{1}{\nu})}\cdot \log\left(1 - \frac{1-\Big(1+\gamma\cdot(\gamma-1)\cdot D_{\varphi_\gamma}(\mathbb{Q}, \mathbb{P})\Big)}{\nu}\right) \\ &= \frac{c_7}{\log(1-\frac{1}{\nu})}\cdot \log\left(1 - \frac{1-H_{\varphi_\gamma}(\mathbb{Q}, \mathbb{P})}{\nu}\right) = \frac{c_7}{\log(1-\frac{1}{\nu})}\cdot \log\left(1 - \frac{1-\sum_{k=1}^{K}(q_k)^\gamma\cdot(p_k)^{1-\gamma}}{\nu}\right) \in [0, c_7[\,, \\ &\qquad \gamma \in]0,1[,\ \ c_7 > 0,\ \nu \in]-\infty, 0[\cup]1,\infty[,\end{aligned}$$

which is BS-maximizable on $\mathbf{\Omega}$. The case $\widetilde{\mathcal{B}}_{1/2,\nu,1}(\mathbb{Q}, \mathbb{P})$ corresponds to the *Bounded Bhattacharyya Distance Measures* of Jolad et al. [174].

We can also employ divergences of the form $\breve{R}_\gamma(\mathbb{Q},\mathbb{P}) := R_\gamma(T_1(\mathbb{Q}),T_2(\mathbb{P}))$ [20] where $T_1 : \mathcal{D}_1 \mapsto \mathcal{R}_1$, $T_2 : \mathcal{D}_1 \mapsto \mathcal{R}_2$ are (say) invertible functions on appropriately chosen subsets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{R}_1, \mathcal{R}_2$ of the probability-vector simplex $\mathbb{S}^K$. For instance, consider the following special case (with a slight abuse of notation):

$$\breve{R}_\gamma(\mathbb{Q},\mathbb{P}) := R_\gamma(\widetilde{\mathbb{Q}},\widetilde{\mathbb{P}}) = \frac{1}{\gamma\cdot(\gamma-1)}\cdot\log\left(\sum_{k=1}^{K}\left(\frac{(q_k)^{\nu_1}}{\sum_{j=1}^{K}(q_j)^{\nu_1}}\right)^{\gamma}\cdot\left(\frac{(p_k)^{\nu_2}}{\sum_{j=1}^{K}(p_j)^{\nu_2}}\right)^{1-\gamma}\right) \tag{70}$$

where (i) $\widetilde{\mathbb{Q}} := (\widetilde{q}_k)_{k=1}^K$ with $\widetilde{q}_k := \frac{(q_k)^{\nu_1}}{\sum_{j=1}^{K}(q_j)^{\nu_1}}$ is the *escort probability distribution (in vector form) associated with the probability distribution (in vector form)* $\mathbb{Q} := (q_k)_{k=1}^K \in \mathbb{S}_{>0}^K$, and (ii) $\widetilde{\mathbb{P}} := (\widetilde{p}_k)_{k=1}^K$ with $\widetilde{p}_k := \frac{(p_k)^{\nu_2}}{\sum_{j=1}^{K}(p_j)^{\nu_2}}$ is the escort probability distribution associated with the probability distribution $\mathbb{P} := (p_k)_{k=1}^K \in \mathbb{S}_{>0}^K$, in terms of some fixed escort parameters $\nu_1 > 0$, $\nu_2 > 0$.

In particular, for the special choice $\nu_1 = \nu_2 > 0$ and $\gamma := \frac{\nu}{\nu_1}$ with $\nu \in]0,\nu_1[\cup[2\nu_1,\infty[$ we obtain from (70)

$$\begin{aligned} 0 \le \frac{\nu}{\nu_1}\cdot R_{\nu/\nu_1}(\widetilde{\mathbb{Q}},\widetilde{\mathbb{P}}) &= \frac{\log\Big(\sum_{k=1}^{K}(\widetilde{q}_k)^{\nu/\nu_1}\cdot(\widetilde{p}_k)^{1-(\nu/\nu_1)}\Big)}{\frac{\nu}{\nu_1}-1} \\ &= \frac{\nu_1}{\nu-\nu_1}\cdot\log\Big(\sum_{k=1}^{K}(q_k)^{\nu}\cdot(p_k)^{\nu_1-\nu}\Big) - \frac{\nu}{\nu-\nu_1}\cdot\log\Big(\sum_{k=1}^{K}(q_k)^{\nu_1}\Big) + \log\Big(\sum_{k=1}^{K}(p_k)^{\nu_1}\Big) =: \breve{R}_{\nu/\nu_1}(\mathbb{Q},\mathbb{P}) \end{aligned} \tag{71}$$

which is BS-minimizable (in $\widetilde{\mathbb{Q}}$) on $\boldsymbol{\Omega}$. Our divergence $\breve{R}_{\nu/\nu_1}(\mathbb{Q},\mathbb{P})$ in (71) is basically a multiple of a divergence which has been very recently used in Ghosh & Basu [136]. Moreover, $\breve{R}_{1/\nu_1}(\mathbb{Q},\mathbb{P})$ (i.e. the special case $\nu = 1$ in (71)) is equal to *Sundaresan's divergence* [352] [353] (see also Lutwak et al. [234], Kumar & Sundaresan [203], [204], Yagli et al. [408]); for our BS-approach, we need the restriction $\nu_1 \in]0,\frac{1}{2}]\cup]1,\infty[$. Notice that Sundaresan's divergence can be employed in mismatch-cases of (i) Campbell's coding problem, (ii) Arikan's guessing problem, (iii) memoryless guessing, and (iv) task partitioning problems; see e.g. Sundaresan [353], Bunte & Lapidoth [67], Kumar et al. [205].

Returning to the general context, functions of the modified Kullback-Leibler information $I(\mathbf{Q},\mathbb{P})$ and the modified reverse Kullback-Leibler information $\widetilde{I}(\mathbf{Q},\mathbb{P})$ can be treated analogously. For the sake of brevity, we only deal with the former and fix arbitrary $\mathbb{P} \in \mathbb{S}_{>0}^K$ and $\mathbf{Q} \in A\cdot\mathbb{S}^K$ with $A := \sum_{k=1}^{K} q_k > 0$. For this, in (46) we have obtained $I(\mathbf{Q},\mathbb{P})$ which is generally not a divergence (cf. Remark 17). In the following, let the function $h_1 :]-1,\infty[\mapsto]-\infty,\infty[$ be continuous and strictly increasing (respectively, strictly decreasing); since $D_{\varphi_1}(\mathbf{Q},\mathbb{P})$ is BS-minimizable on $\boldsymbol{\Omega} = A\cdot\boldsymbol{\Omega}$, also the quantity $h_1\Big(A-1+D_{\varphi_1}(\mathbf{Q},\mathbb{P})\Big) = h_1\Big(I(\mathbf{Q},\mathbb{P})\Big)$ is BS-minimizable on $\boldsymbol{\Omega} = A\cdot\boldsymbol{\Omega}$ (respectively, BS-maximizable on $\boldsymbol{\Omega} = A\cdot\boldsymbol{\Omega}$). In particular, by using the negative identity mapping $h_\gamma^{-Id}(y) := -y \quad (y > -1)$ we get that $-I(\mathbf{Q},\mathbb{P})$ is BS-maximizable. Another exemplary choice for h_1 is (cf. Sharma & Mittal [330] in the scaling of e.g. Morales et al. [264])

$$h_1(y) := h_s^{SM}(y) := \frac{e^{(s-1)\cdot y}-1}{s-1}, \qquad y\in\mathbb{R},\ s\in]0,1[\cup]1,\infty[, \tag{72}$$

which is strictly increasing; hence, $h_s^{SM}(I(\mathbf{Q},\mathbb{P}))$ (and also $h_s^{SM}(D_{\varphi_1}(\mathbf{Q},\mathbb{P}))$) is BS-minimizable on $\boldsymbol{\Omega} = A\cdot\boldsymbol{\Omega}$.

As another important application line, let us fix any $(\gamma,\mathbf{Q}) \in (\widetilde{\Gamma}\backslash]1,2[)\times\widetilde{\mathcal{M}}_2$ (cf. Lemma 16(a)) with $A := \sum_{k=1}^{K} q_k > 0$. Moreover, we take $\mathbb{P} := \mathbb{P}^{unif} := (\frac{1}{K},\ldots,\frac{1}{K})$ to be the probability vector of frequencies of the uniform distribution on $\{1,\ldots,K\}$. Then, for $\gamma \in]-\infty,0[\ \cup\ ]0,1[\ \cup\ [\ 2,\infty[$ one gets $H_\gamma(\mathbf{Q},\mathbb{P}^{unif}) = K^{\gamma-1}\cdot\sum_{k=1}^{K} q_k^\gamma$. One can rewrite $K^{1-\gamma}\cdot H_\gamma(\mathbf{Q},\mathbb{P}^{unif}) = \sum_{k=1}^{K} q_k^\gamma$; the latter is sometimes called *heterogeneity index of type* γ, see e.g. van der Lubbe [379], with $\gamma = 2$ being the *Simpson-Herfindahl index* which is also known as *index of coincidence* (cf. Harremoes & Topsoe [155] and its generalization in Harremoes & Vajda [156]). Alternatively, $\sum_{k=1}^{K} q_k^\gamma$ is also called Onicescu's information energy in case of $\gamma = 2$ (cf. Onicescu [278], see also Pardo & Taneja [283] for comprehensive investigations) and in general *information energy of order* γ (cf. Theodorescu [359], see also e.g. Pardo [281]); for exemplary applications to electron density functional theory (DFT) for quantum chemical reactivity, the reader may take (discretized versions of) e.g. Liu et al. [226], Lopez-Rosa et al. [231] and Rong et al. [311]. In some other literature (see e.g. Clark [90]), $\sum_{k=1}^{K} q_k^\gamma$ is alternatively called *Golomb's [140] information generating function (of a probability distribution* $\mathbb{Q}$*)*; yet another name is *generalized information potential* and for $\gamma = 2$ *information potential* (cf. e.g. Principe [297], Acu et al. [4]). From the above-mentioned investigations, we obtain

[20] and analogously power divergences $\breve{D}_{\widetilde{c}\cdot\varphi_\gamma}(\mathbb{Q},\mathbb{P}) := D_{\widetilde{c}\cdot\varphi_\gamma}(T_1(\mathbb{Q}),T_2(\mathbb{P}))$ etc.

that $\sum_{k=1}^{K} q_k^{\gamma}$ is BS-minimizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma < 0$ and $\gamma \geq 2$, and BS-maximizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma \in]0,1[$. More generally, by employing (65) and (66), for the class of entropies (diversity indices)

$$\begin{aligned} \mathcal{E}_{\gamma,c_1,c_2,c_3}(\mathbf{Q}) &:= h_{c_1,c_2,c_3}\left(\sum_{k=1}^{K} q_k^{\gamma}\right) = c_1 \cdot \left(\left(\sum_{k=1}^{K} q_k^{\gamma}\right)^{c_2} - c_3\right) \\ &= c_1 \cdot \Big(K^{c_2\cdot(1-\gamma)} \cdot H_{\gamma}(\mathbf{Q}, \mathbb{P}^{unif})^{c_2} - c_3\Big), \qquad c_1, c_2 \in \mathbb{R}\backslash\{0\},\ c_3 \in \mathbb{R}, \qquad (73) \\ \mathcal{E}^{R}_{c_4,f}(\mathbf{Q}) &:= h^{R}_{c_4,f}\left(\sum_{k=1}^{K} q_k^{\gamma}\right) = \frac{c_4}{f'(0)} \cdot \log\left(\sum_{k=1}^{K} q_k^{\gamma}\right), \\ &= \frac{c_4}{f'(0)} \cdot \Big(\log\big(H_{\gamma}(\mathbf{Q}, \mathbb{P}^{unif})\big) + (1-\gamma)\cdot \log(K)\Big), \qquad c_4 \in \mathbb{R}\backslash\{0\}, \qquad (74) \end{aligned}$$

(which is similar to the entropy-class of Morales et al. [265] who use a different, more restrictive parametrization and probability distributions $\mathbb{Q}$), one gets the following extremum-behaviour:

- $\mathcal{E}_{\gamma,c_1,c_2,c_3}(\mathbf{Q})$ is BS-minimziable if $\gamma < 0$ and $c_1 \cdot c_2 > 0$;
- $\mathcal{E}_{\gamma,c_1,c_2,c_3}(\mathbf{Q})$ is BS-minimizable if $\gamma \geq 2$ and $c_1 \cdot c_2 > 0$;
- $\mathcal{E}_{\gamma,c_1,c_2,c_3}(\mathbf{Q})$ is BS-minimizable if $\gamma \in]0,1[$ and $c_1 \cdot c_2 < 0$;
- $\mathcal{E}_{\gamma,c_1,c_2,c_3}(\mathbf{Q})$ is BS-maximizable if $\gamma < 0$ and $c_1 \cdot c_2 < 0$;
- $\mathcal{E}_{\gamma,c_1,c_2,c_3}(\mathbf{Q})$ is BS-maximizable if $\gamma \geq 2$ and $c_1 \cdot c_2 < 0$;
- $\mathcal{E}_{\gamma,c_1,c_2,c_3}(\mathbf{Q})$ is BS-maximizable if $\gamma \in]0,1[$ and $c_1 \cdot c_2 > 0$;
- $\mathcal{E}^{R}_{c_4,f}(\mathbf{Q})$ is BS-minimizable if $\gamma < 0$ and $\frac{c_4}{f'(0)} > 0$;
- $\mathcal{E}^{R}_{c_4,f}(\mathbf{Q})$ is BS-minimizable if $\gamma \geq 2$ and $\frac{c_4}{f'(0)} > 0$;
- $\mathcal{E}^{R}_{c_4,f}(\mathbf{Q})$ is BS-minimizable if $\gamma \in]0,1[$ and $\frac{c_4}{f'(0)} < 0$;
- $\mathcal{E}^{R}_{c_4,f}(\mathbf{Q})$ is BS-maximizable if $\gamma < 0$ and $\frac{c_4}{f'(0)} < 0$;
- $\mathcal{E}^{R}_{c_4,f}(\mathbf{Q})$ is BS-maximizable if $\gamma \geq 2$ and $\frac{c_4}{f'(0)} < 0$;
- $\mathcal{E}^{R}_{c_4,f}(\mathbf{Q})$ is BS-maximizable if $\gamma \in]0,1[$ and $\frac{c_4}{f'(0)} > 0$.

From this, one can deduce that our new BS method works for the constrained minimization/maximization of the following well-known, prominently used measures of entropy respectively measures of diversity, and beyond:

(E1) $c_1 = 1$, $c_2 = \frac{1}{\gamma}$, $c_3 = 0$: the Euclidean $\gamma-$*norm* (also known as $\gamma-$*norm heterogeneity index*, see e.g. van der Lubbe [379]) $||\mathbf{Q}||_{\gamma} := \Big(\sum_{k=1}^{K} q_k^{\gamma}\Big)^{1/\gamma} = K^{(1-\gamma)/\gamma} \cdot \Big(H_{\gamma}(\mathbf{Q}, \mathbb{P}^{unif})\Big)^{1/\gamma}$ is BS-minimizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma \in]0,1[$ and $\gamma \geq 2$, and BS-maximizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma < 0$ (note that $||\mathbf{Q}||_1 = A$) ;
similarly, the $\gamma-$*mean heterogeneity index* (see e.g. [379], as well as Jost [176] for its interpretation as "effective number of species" respectively as "numbers equivalent") given by $\mathcal{E}^{HI}(\mathbf{Q}) := \Big(\sum_{k=1}^{K} q_k^{\gamma}\Big)^{1/(\gamma-1)} = \frac{1}{K} \cdot \Big(H_{\gamma}(\mathbf{Q}, \mathbb{P}^{unif})\Big)^{1/(\gamma-1)}$ is BS-minimizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma \geq 2$, and BS-maximizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma < 0$ and $\gamma \in]0,1[$. Alternatively, $\mathcal{E}^{HI}(\mathbf{Q})$ is also called *($\gamma-$order) Hill diversity index or Hill number* [160], respectively *($\gamma-$order) Hannah-Kay index* [154], respectively *($\gamma-$order) Renyi heterogeneity* (cf. Nunes et al. [276]), respectively *($\gamma-$order) exponential Renyi entropy or exponential entropy* (cf. Campbell [72]) since it is equal to $\exp(\mathcal{E}^{gR}(\mathbf{Q}))$ (cf. (E6) below). The $\gamma-$mean heterogeneity index (under one of the above-mentioned namings) was recently employed e.g. by Greiff et al. [142] for immunodiagnostic design of fingerprints of an individual's ongoing immunological status (e.g., healthy, infected, vaccinated) — culminating in accurate and early detection of disease and infection, by Ma & Li [237] for the quantification of metagenome diversity and similarity, by Jasinska et al. [164] for studying bacterial evolution — in particular evolution under sub-inhibitory antibiotic levels, by Ma et al. [238] for the definition of individual-level genetic diversity and similarity profiles as well as their applications to datasets from the 1000-Genomes Project, and by Lassance & Vrins [209] for some optimal selection procedure of financial-asset portfolios.

(E2) $c_1 = \frac{1}{2^{1-\gamma}-1}$, $c_2 = 1$, $c_3 = 1$: the entropy

$$\mathcal{E}^{gHC}(\mathbf{Q}) := \frac{1}{2^{1-\gamma}-1} \cdot \left(\sum_{k=1}^{K} q_k^{\gamma} - 1\right) = \frac{1}{2^{1-\gamma}-1} \cdot \big(K^{1-\gamma} \cdot H_{\gamma}(\mathbf{Q}, \mathbb{P}^{unif}) - 1\big) \qquad (75)$$

is BS-minimizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma < 0$, and BS-maximizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma \in]0,1[$ and $\gamma \geq 2$; the special subcase $A = 1$ in (75) (and thus, $\mathbf{Q} = \mathbb{Q}$ is a probability vector) corresponds to the $\gamma-$*order entropy of Havrda-Charvat* [157] (also called non-additive $\gamma-$*order Tsallis entropy* [363] in statistical physics) where the special case $\gamma = 2$ is (a multiple of) *Vajda's quadratic entropy* [371] and *Ahlswede's identification entropy* [7] (see also Ahlswede & Cai [8]). Some exemplary (relatively) recent studies and applications of $\mathcal{E}^{gHC}(\mathbb{Q})$ appear e.g. in Peter & Rangarajan [291] for shape matching, in Liu et al. [226] as well as in Rong et al. [311] to electron density functional theory (DFT) for quantum chemical reactivity, in Yalcin & Beck [409] for the investigation of energy spectra of cosmic rays, in Wen & Jiang [390] for the quantification of complexity degrees in complex networks, in Bhandari [46] for fast multilevel thresholding for color image segmentation, in Erguzel et al. [121] for the investigation of Electroencephalography (EEG) signals of subjects suffering from some psychiatric disorders, in Kang & Kim [182] for automatic synthetic aperture radar (SAR) image registration, in Namdari & Li [269] for the modelling of Lithium-Ion battery capacity fade, in Seweryn et al. [326] for the assessment of similarity and diversity of expression profiles in single cell systems, in Zhang et al. [418] for the search of functional relationships between groundwater depth and vegetation distribution, in Kumbhakar et al. [207] for the modelling of streamwise velocity profiles in wide–open channel turbulent flows (e.g. in rivers, streams, canals, ditches), and in Ramezani & Pourdarvish [300] for transfer learning for image classification of gravitational waves.

For the special case $\gamma = 2$, a directly connected quantity is the *measure of concentration* (cf. e.g. De Wet et al. [107]) $\mathcal{E}^{gMC}(\mathbb{Q}) := 1 - \frac{1}{K} - \mathcal{E}^{gHC}(\mathbb{Q}) = \sum_{k=1}^{K} \left(q_k - \frac{1}{K}\right)^2$ which (up to a multiple) was introduced by Brukner & Zeilinger [66] as an appropriate measure of information for quantum experiments.

(E3) $\gamma := \frac{1}{\widetilde{\gamma}}$, $c_1 = \frac{1}{\widetilde{\gamma}-1}$, $c_2 = \widetilde{\gamma}$, $c_3 = 1$: the entropy

$$\mathcal{E}^{gA}(\mathbf{Q}) := \frac{1}{\widetilde{\gamma}-1} \cdot \left(\left(\sum_{k=1}^{K} q_k^{1/\widetilde{\gamma}} \right)^{\widetilde{\gamma}} - 1 \right) = \frac{1}{\widetilde{\gamma}-1} \cdot \left(K^{\widetilde{\gamma}\cdot(1-\gamma)} \cdot H_{1/\widetilde{\gamma}}(\mathbf{Q}, \mathbb{P}^{unif})^{\widetilde{\gamma}} - 1 \right) \tag{76}$$

is BS-minimizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\widetilde{\gamma} < 0$ and $\widetilde{\gamma} \in]0,1[$, and BS-maximizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma \geq 2$; the special subcase $A = 1$ in (76) (and thus, $\mathbf{Q} = \mathbb{Q}$ is a probability vector) corresponds to the *$\widetilde{\gamma}$−order entropy of Arimoto* [16].

(E4) $s \in \mathbb{R}\backslash\{1\}$, $c_1 = \frac{1}{1-s}$, $c_2 = \frac{1-s}{1-\gamma}$, $c_3 = 1$: the entropy

$$\mathcal{E}^{gSM1}(\mathbf{Q}) := \frac{1}{1-s} \cdot \left(\left(\sum_{k=1}^{K} q_k^{\gamma} \right)^{(1-s)/(1-\gamma)} - 1 \right) = \frac{1}{1-s} \cdot \left(K^{1-s} \cdot H_{\gamma}(\mathbf{Q}, \mathbb{P}^{unif})^{(1-s)/(1-\gamma)} - 1 \right) \tag{77}$$

is BS-minimizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma < 0$ and BS-maximizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma \in]0,1[$ and $\gamma \geq 2$; the special subcase $A = 1$ in (77) (and thus, $\mathbf{Q} = \mathbb{Q}$ is a probability vector) corresponds to the *entropy of order γ and degree s of Sharma & Mittal* [329] in the scaling of e.g. Salicru et al. [314].

(E5) $s \in \mathbb{R}\backslash\{0\}$, $\gamma = s+1$, $c_1 = -\frac{1}{s}$, $c_2 = 1$, $c_3 = 1$: the diversity index

$$\mathcal{E}^{gPT}(\mathbf{Q}) := -\frac{1}{s} \cdot \left(\sum_{k=1}^{K} q_k^{s+1} - 1 \right) = -\frac{1}{s} \cdot \left(K^{-s} \cdot H_{s+1}(\mathbf{Q}, \mathbb{P}^{unif}) - 1 \right) \tag{78}$$

is BS-minimizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $s < -1$ and BS-maximizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $s \in]-1,0[$ and $s > 0$; the special subcase $A = 1$ in (78) (and thus, $\mathbf{Q} = \mathbb{Q}$ is a probability vector) corresponds to the *diversity index of degree s of Patil & Taillie* [286]; the case $s = 1$ for probability measures $\mathbf{Q} = \mathbb{Q}$ gives the well-known *Gini-Simpson diversity index*.

(E6) $c_4 = \frac{1}{1-\gamma}$, $f(z) = z$: the entropy

$$\mathcal{E}^{gR}(\mathbf{Q}) := \frac{1}{1-\gamma} \cdot \log \left(\sum_{k=1}^{K} q_k^{\gamma} \right) = \frac{1}{1-\gamma} \cdot \left(\log \left(H_{\gamma}(\mathbf{Q}, \mathbb{P}^{unif}) \right) + (1-\gamma) \cdot \log(K) \right) = \frac{\log 2}{1-\gamma} \cdot \log_2 \left(\sum_{k=1}^{K} q_k^{\gamma} \right) \tag{79}$$

is BS-minimizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma < 0$, and BS-maximizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$ for $\gamma \in]0,1[$ and $\gamma \geq 2$; the special subcase $A = 1$ in (79) (and thus, $\mathbf{Q} = \mathbb{Q}$ is a probability vector) corresponds to the prominent (additive) *γ−order Renyi entropy* [309]. As well known, there is a vast literature on Renyi entropies $\mathcal{E}^{gR}(\mathbb{Q})$. Some exemplary (mostly recent) studies and applications appear e.g. in Nath [270] — as well as in Arikan [15], Sundaresan [353], Bunte & Lapidoth [67], Sason & Verdu [321], Kumar et al. [205] — for coding and guessing, in Bennett et al. [39] in connection with unconditionally secure secret-key agreement protocols and quantum cryptography, in Mayoral [257] for cluster sampling, in Aviyente et al. [21] for information extraction in certain neurophysiological signals (so-called event-related potentials), in Tao et al. [355] as well as in Jiao et al. [171] for early defect/fault detection of rolling element bearings, in Pham et al.: [292] for blind source separation, in Liu et al. [226] as well as in Rong et al. [311] to electron density functional theory (DFT) for quantum chemical reactivity, in Sason [319] for data compression, in Carravilla et al. [73] for the recognition of HIV-1 antibodies through STED microscopy and the corresponding design of therapeutic interventions, in Joshi et al. [175] for the identification and tracking of relevant T cell receptors for adoptive immunotherapy, in Erguzel et al. [121] for the investigation of Electroencephalography (EEG) signals of subjects suffering from some psychiatric disorders, in German–Sallo [133] for fault–characteristics extraction from discrete signals in manufacturing systems, in Schober et al. [323] for investigations of some evolutions of the T cell antigen receptor (TCR) repertoire, in Seweryn et al. [326] for the assessment of similarity and diversity of expression profiles in single cell systems, in Amezquita-Sanchez

[13] for the detection of incipient damage in high-rise buildings subjected to dynamic vibrations, in Barennes et al. [32] for comparing the accuracy of current T cell receptor sequencing methods employed for the understanding of adaptive immune responses, in Kumar et al. [201] for the segmentation of digital images through multilevel iterative variational mode decomposition (VMD), and in Pandey [280] for the quantification of cosmic homogeneity.

Remark 19: (i) For Renyi entropies there are also matrix versions $\mathcal{E}^{gR}(X) := \frac{1}{1-\gamma} \cdot \log\Big(\sum_{i=1}^{K_1} \sum_{j=1}^{K_2} x_{ij}^{\gamma}\Big)$ where $X := (x_{ij})_{i=1,\ldots,K_1}^{j=1,\ldots,K_2}$ is a $K_1 \times K_2-$matrix whose elements x_{ij} are (say) strictly positive and sum up to A. Such a setup with $A = 1$ is e.g. used in *time-frequency analyses of signals* where the i's correspond to discrete time points, the j's to discrete frequencies, and x_{ij} to the probability that (i, j) occurs; see e.g. Popescu & Aiordachioaie [295] for change detection in seismic signals. Another line of application is to use as X the normalized communicability matrix of a directed network (respectively the upper triangular part of X in case of an unweighted and undirected network). Of course, the matrix version $\mathcal{E}^{gR}(X)$ can be easily and equivalently rewritten in our vector version $\mathcal{E}^{gR}(\mathbf{Q})$ by setting $\mathbf{Q} := (q_1, \ldots, q_{K_1 \cdot K_2})$ such that $x_{ij} = q_{(i-1)\cdot K_2 + j}$ ($i = 1, \ldots, K_1$, $j = 1, \ldots, K_2$ and hence $K := K_1 \cdot K_2$; accordingly, we can apply our BS method.
(ii) The latter conversion works analogously also for matrix versions of all the other entropies, divergences, etc. of this paper; more flexible versions where $i \in \{1, \ldots, K_1\}$, $j \in J_i$ for some $J_i \subseteq \{1, \ldots, K_2\}$ as well as multidimensional-array/tensor versions can be transformed in a similar book-keeping manner, too. For instance, within the above-mentioned framework of unweighted and undirected networks, Chen et al. [81] and Shi et al. [332] employ communicability matrix versions of the Shannon entropy and the Jensen-Shannon divergence (JSD), e.g. in order to derive a new complexity measure of such kind of networks; see also Bagrow and Bollt [28] for similar network applications of the JSD. Moreover, Jena et al. [167] use "3D versions" of Tsallis entropies for brain magnetic resonance (MR) image segmentation.

Remark 20: All the above cases which are BS-maximizable can be interpreted as bare-simulation approach to the solution of *generalized maximum entropy problems on* $\mathbf{\Omega} = A \cdot \mathbf{\Omega}$.

Remark 21: (i) If (all) the above- and below-mentioned entropies are used for probability vectors $\mathbb{Q} \in \mathbf{S}^K$ — i.e. one employs $\mathcal{E}(\mathbb{Q})$ — then typically the components q_k of $\mathbb{Q}$ represent a genuine probability mass (frequency) $q_k = \mathbb{\Pi}[\{d_k\}]$ of some data point (state) d_k. However, $\mathbb{Q} \in \mathbf{S}^K$ may alternatively be artificially generated. For instance, for the purpose of fault detections of mechanical drives, Boskoski & Juricic [57] use Renyi entropies where the q_k's are normalized squared energy-describing coefficients of the wavelet packet transform of measured vibration records. Another exemplary "artificial" operation is concatenation, see e.g. Subsection IV-D below.
(ii) An analogous statement holds for the employment of (all) the above- and below-mentioned divergences $D(\mathbb{Q}, \mathbb{P})$ — and their transformations — between genuine respectively artificially generated probability vectors $\mathbb{Q}, \mathbb{P} \in \mathbf{S}^K$.

The remaining parameter cases $\gamma = 0$ and $\gamma = 1$ can be treated analogously. For the sake of brevity, we only deal with the latter. For this, let $\mathbf{Q} \in A \cdot \mathbb{S}^K$ with $A := \sum_{k=1}^K q_k > 0$ and $\mathbb{P} := \mathbb{P}^{unif}$. Clearly, $I(\mathbf{Q}, \mathbb{P}^{unif}) - \frac{\log K}{K} = \sum_{k=1}^K q_k \cdot \log(q_k)$; thus the latter is BS-minimizable on $\mathbf{\Omega} = A \cdot \mathbf{\Omega}$. More generally, for any continuous strictly increasing (respectively strictly decreasing) function $h_1 : [-\frac{K}{e}, 0[\mapsto \mathbb{R}$, the quantity $h_1\Big(\sum_{k=1}^K q_k \cdot \log(q_k)\Big)$ is BS-minimizable on $\mathbf{\Omega} = A \cdot \mathbf{\Omega}$ (respectively BS-maximizable on $\mathbf{\Omega} = A \cdot \mathbf{\Omega}$). Important special cases are:

(E7) $h_1(y) := h_1^{-Id}(y) = -y$: the entropy

$$\mathcal{E}^{Sh}(\mathbf{Q}) := h_1^{-Id}\Big(\sum_{k=1}^{K} q_k \cdot \log(q_k)\Big) = -\sum_{k=1}^{K} q_k \cdot \log(q_k) \tag{80}$$

is BS-maximizable on $\mathbf{\Omega} = A \cdot \mathbf{\Omega}$; the special subcase $A = 1$ in (80) (and thus, $\mathbf{Q} = \mathbb{Q}$ is a probability vector) corresponds to the omnipresent *Shannon entropy*; hence, by our bare-simulation approach we can particularly tackle *maximum entropy problems* on almost arbitrary sets $\mathbf{\Omega}$ of probability vectors. Analogously, we can treat $\frac{1}{\log(K)} \cdot \mathcal{E}^{Sh}(\mathbb{Q})$ which is called *Pielou's evenness index* [293], and $1 - \frac{1}{\log(K)} \cdot \mathcal{E}^{Sh}(\mathbb{Q}) \in [0, 1]$ which is sometimes used as *clonality (clonotype diversity) index* (see e.g. Gabriel et al. [131] for applications to HIV-connected T cell receptor repertoires, and Bashford-Rogers et al. [33] (with supplementary private communication) for its use for comparative analyses of the BCR repertoire in immune-mediated diseases, for the sake of understanding pathological mechanisms and designing treatment strategies). As a further example for Remark 21, Lyubushin [236] uses q_k's which are normalized squared coefficients of an orthogonal wavelet decomposition of some seismic noise, and accordingly, $\frac{1}{\log(K)} \cdot \mathcal{E}^{Sh}(\mathbb{Q})$ can be interpreted as the entropy of the distribution of energy of oscillations at various frequency and time scales.

Some further exemplary studies and applications of the maximization of $\mathcal{E}^{Sh}(\mathbb{Q})$ — aside from the vast physics literature — appear e.g. in De Santis et al. [106] for cryptanalytic guessing problems for breaking ciphertexts with probabilistic brute-force attacks, Johansson & Sternad [173] for tackling certain resource allocation problems under uncertainty, Marano & Franceschetti [246] for ray propagation in percolating lattices, Miao et al. [260] for unsupervised mixed-pixel decomposition in image processing, Rodrigues et al. [310] for modelling biological species geographic distribution, Xiong et al. [400] for capturing desirable phrasal and hierarchical segmentations within a statistical machine translation context, Chan et al. [76] for alignment-free DNA sequence comparison, Mann & Garnett [244] for capturing some collective behaviours of intelligent agents in social interactions, Singh et al. [336] for the study of finite buffer queueing systems, Baddeley [27] for geoscientifical prediction of the occurrence of mineral deposits on regional scales, Einicke et al. [118] for feature selection within change classification during running, and Han et al. [152] for substructure imaging of blood cells by means of maximum entropy tomography (MET).

(E8) $s \in]0,1[\cup]1,\infty[$, $h_1(y) := h_s^{SM2}(y) := \frac{e^{(s-1)\cdot y}-1}{1-s}$ (cf. (72)) with $y \in \mathbb{R}$: the entropy

$$\mathcal{E}^{SM2}(\mathbf{Q}) := h_s^{SM2}\Big(\sum_{k=1}^{K} q_k \cdot \log(q_k)\Big) = \frac{1}{1-s} \cdot \left(\exp\Big\{(s-1)\cdot \sum_{k=1}^{K} q_k \cdot \log(q_k)\Big\} - 1\right) \tag{81}$$

is BS-maximizable on $\mathbf{\Omega} = A \cdot \mathbb{\Omega}$; the special subcase $A = 1$ in (81) (and thus, $\mathbf{Q} = \mathbb{Q}$ is a probability vector) corresponds to the (second type) *entropy of Sharma & Mittal* [329] in the scaling of e.g. Pardo [282] (p.20).

Returning to the general context, we now (as already indicated above) state *explicitly* the corresponding bare-simulation-minimizations (respectively maximizations) of the power divergences $\inf_{\mathbf{Q}\in\mathbf{\Omega}} D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbf{Q},\mathbb{P})$ $(\gamma \in \mathbb{R})$, the Renyi divergences $\inf_{\mathbf{Q}\in\mathbf{\Omega}} R_\gamma(\mathbf{Q},\mathbb{P})$ $(\gamma \in \mathbb{R})$, the Hellinger integrals $\inf_{\mathbf{Q}\in A\cdot\mathbb{\Omega}} H_\gamma(\mathbf{Q},\mathbb{P})$ $(\gamma \in]-\infty,0[\cup]1,\infty[)$, $\sup_{\mathbf{Q}\in A\cdot\mathbb{\Omega}} H_\gamma(\mathbf{Q},\mathbb{P})$ $(\gamma \in]0,1[)$, the modified Kullback-Leibler information $\inf_{Q\in A\cdot\mathbb{\Omega}} I(\mathbf{Q},\mathbb{P})$, the modified reverse Kullback-Leibler information $\inf_{Q\in A\cdot\mathbb{\Omega}} \widetilde{I}(\mathbf{Q},\mathbb{P})$, as well as the above-mentioned measures of entropy (diversity). Since the corresponding probability distribution $\mathbb{C}[\cdot] = \Pi[W_1 \in \cdot\,]$ of the W_i's (cf. the representability (6)) varies "quite drastically" with γ, we split this issue into several pieces.

Proposition 22: (a) Consider the context of Remark 15(vi) for $\varphi := \widetilde{c}\cdot\varphi_\gamma$ with $\gamma < 0$, and let $\mathbb{P} \in \mathbb{S}_{>0}^K$ as well as $\widetilde{c} > 0$ be arbitrary but fixed. Furthermore, let $W := (W_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of non-negative real-valued random variables having density[21]

$$f_{W_1}(y) \;:=\; \frac{\exp\{-\frac{y\cdot\widetilde{c}}{1-\gamma}\}}{\exp\{\widetilde{c}/\gamma\}} \cdot f_Z(y)\cdot \mathbb{1}_{]-\infty,0[}(y), \qquad y \in \mathbb{R}, \tag{82}$$

where f_Z is the density of a random variable Z which has stable law with parameter-quadruple $(\frac{-\gamma}{1-\gamma}, 1, 0, -\frac{\widetilde{c}^{1/(1-\gamma)}\cdot(1-\gamma)^{-\gamma/(1-\gamma)}}{\gamma})$ in terms of "form-B notation" in Zolotarev [428], p.12. Then for all $A > 0$ and all $\mathbb{\Omega} \subset \mathbb{S}_{>0}^K$ with (7) there holds

$$-\lim_{n\to\infty} \frac{1}{n} \log \Pi\Big[\boldsymbol{\xi}_n^{w\mathbf{W}} \in \mathbb{\Omega}\Big] = \inf_{\mathbf{Q}\in A\cdot\mathbb{\Omega}} \frac{\widetilde{c}}{\gamma} \cdot \left[1 - A^{\gamma/(\gamma-1)} \cdot \left[1 + \gamma\cdot(A-1) + \frac{\gamma\cdot(\gamma-1)}{\widetilde{c}} \cdot D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbf{Q},\mathbb{P})\right]^{-1/(\gamma-1)}\right] \tag{83}$$

as well as the BS minimizabilities/maximizabilites (cf. Definition 1)

$$\inf_{\mathbf{Q}\in A\cdot\mathbb{\Omega}} D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbf{Q},\mathbb{P}) = \lim_{n\to\infty} \frac{\widetilde{c}}{\gamma\cdot(\gamma-1)} \cdot \left\{A^\gamma \cdot \left(1 + \frac{\gamma}{\widetilde{c}} \cdot \frac{1}{n} \cdot \log \Pi\Big[\boldsymbol{\xi}_n^{w\mathbf{W}} \in \mathbb{\Omega}\Big]\right)^{1-\gamma} + \gamma\cdot(1-A) - 1\right\}, \tag{84}$$

$$\inf_{\mathbf{Q}\in A\cdot\mathbb{\Omega}} H_\gamma(\mathbf{Q},\mathbb{P}) = \lim_{n\to\infty} A^\gamma \cdot \left(1 + \gamma\cdot\frac{1}{n}\cdot\log \Pi\Big[\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}} \in \mathbb{\Omega}\Big]\right)^{1-\gamma}, \tag{85}$$

$$\inf_{\mathbf{Q}\in A\cdot\mathbb{\Omega}} c_1\cdot\Big(H_\gamma(\mathbf{Q},\mathbb{P})^{c_2} - c_3\Big) = \lim_{n\to\infty} c_1\cdot\left\{A^{c_2\cdot\gamma}\cdot\Big(1 + \frac{\gamma}{n}\cdot\log \Pi\Big[\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}} \in \mathbb{\Omega}\Big]\Big)^{c_2\cdot(1-\gamma)} - c_3\right\}, \quad \text{if } c_1\cdot c_2 > 0,\ c_3 \in \mathbb{R}, \tag{86}$$

$$\sup_{\mathbf{Q}\in A\cdot\mathbb{\Omega}} c_1\cdot\Big(H_\gamma(\mathbf{Q},\mathbb{P})^{c_2} - c_3\Big) = \lim_{n\to\infty} c_1\cdot\left\{A^{c_2\cdot\gamma}\cdot\Big(1 + \frac{\gamma}{n}\cdot\log \Pi\Big[\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}} \in \mathbb{\Omega}\Big]\Big)^{c_2\cdot(1-\gamma)} - c_3\right\}, \quad \text{if } c_1\cdot c_2 < 0,\ c_3 \in \mathbb{R}, \tag{87}$$

$$\inf_{\mathbf{Q}\in A\cdot\mathbb{\Omega}} R_\gamma(\mathbf{Q},\mathbb{P}) = \lim_{n\to\infty} \frac{1}{\gamma\cdot(\gamma-1)} \cdot \log\Big(A^\gamma\cdot\Big(1 + \gamma\cdot\frac{1}{n}\cdot\log \Pi\Big[\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}} \in \mathbb{\Omega}\Big]\Big)^{1-\gamma}\Big), \tag{88}$$

$$\inf_{\mathbf{Q}\in A\cdot\mathbb{\Omega}} c_1\cdot\Big(\Big(\sum_{k=1}^{K} q_k^\gamma\Big)^{c_2} - c_3\Big) = \lim_{n\to\infty} c_1\cdot\left\{K^{c_2\cdot(1-\gamma)}\cdot A^{c_2\cdot\gamma}\cdot\Big(1 + \frac{\gamma}{n}\cdot\log \Pi\Big[\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}} \in \mathbb{\Omega}\Big]\Big)^{c_2\cdot(1-\gamma)} - c_3\right\},$$
$$\text{if } c_1\cdot c_2 > 0,\ c_3 \in \mathbb{R}, \tag{89}$$

[21] in the classical sense, with respect to Lebesgue measure

$$\sup_{\mathbf{Q}\in A\cdot\mathbf{\Omega}} c_1\cdot\Big(\Big(\sum_{k=1}^{K} q_k^{\gamma}\Big)^{c_2}-c_3\Big)=\lim_{n\to\infty} c_1\cdot\Big\{K^{c_2\cdot(1-\gamma)}\cdot A^{c_2\cdot\gamma}\cdot\Big(1+\frac{\gamma}{n}\cdot\log \Pi\Big[\breve{\breve{\boldsymbol{\xi}}}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]\Big)^{c_2\cdot(1-\gamma)}-c_3\Big\}, \quad \text{if } c_1\cdot c_2<0,\ c_3\in\mathbb{R}, \tag{90}$$

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\frac{1}{1-\gamma}\cdot\log\Big(\sum_{k=1}^{K} q_k^{\gamma}\Big)=\lim_{n\to\infty}\frac{1}{\gamma\cdot(\gamma-1)}\cdot\Big[\log\Big(A^{\gamma}\cdot\Big(1+\gamma\cdot\frac{1}{n}\cdot\log\Pi\Big[\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]\Big)^{1-\gamma}\Big)+(1-\gamma)\cdot\log(K)\Big], \tag{91}$$

where $\boldsymbol{\xi}_n^{w\mathbf{W}}$ is the normalized randomly weighted empirical measure given in (41), $\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}}$ is its special case for $\widetilde{c}=1$, and $\breve{\breve{\boldsymbol{\xi}}}_n^{w\mathbf{W}}$ is its special case for $\widetilde{c}=1$ together with $\mathbb{P}=\mathbb{P}^{unif}$ [22]. From this, the BS-minimizability/maximizability of the important norms/entropies/diversity indices (E1) to (E6) follow immediately as special cases.

(b) The special case $\varphi:=\widetilde{c}\cdot\varphi_{\gamma}\quad(\gamma<0)$ of Theorem 14 works analogously to (a), with the differences that we employ (i) additionally a sequence $(X_i)_{i\in\mathbb{N}}$ of random variables being independent of $(W_i)_{i\in\mathbb{N}}$ and satisfying condition (26) (resp. (30)), (ii) $A=1$ (instead of arbitrary $A>0$), (iii) $\Pi_{X_1^n}[\cdot]$ (instead of $\Pi[\cdot]$), (iv) $\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\boldsymbol{\xi}_n^{w\mathbf{W}}$), (v) $\breve{\boldsymbol{\xi}}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}}$), and (vi) $\breve{\breve{\boldsymbol{\xi}}}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\breve{\breve{\boldsymbol{\xi}}}_n^{w\mathbf{W}}$).

The assertions of Proposition 22 can be deduced from Theorem 14, Remark 15(vi), Lemma 16(a), (65), (69), (73), (74) and the below-mentioned $\zeta-$concerning Example 48(d). Employing Lemma 16(c) and Example 48(a) instead, one ends up with the following proposition on the reverse Kullback-Leibler divergence:

Proposition 23: (a) Consider the context of Remark 15(vi) for $\varphi:=\widetilde{c}\cdot\varphi_{\gamma}$ with $\gamma=0$, and let $\mathbb{P}\in\mathbb{S}_{>0}^{K}$ as well as $\widetilde{c}>0$ be arbitrary but fixed. Furthermore, let $W:=(W_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of non-negative real-valued random variables with Gamma distribution $\zeta=GAM(\widetilde{c},\widetilde{c})$ [23] (where the subcase $\widetilde{c}=1$ is the exponential distribution $\zeta=EXP(1)$ with mean 1). Then for all $A>0$ and all $\mathbf{\Omega}\subset\mathbb{S}_{>0}^{K}$ with (7) there holds the BS minimizabilites (cf. (2))

$$\inf_{\mathbb{Q}\in A\cdot\mathbf{\Omega}} D_{\widetilde{c}\cdot\varphi_0}(\mathbb{Q},\mathbb{P})=-\lim_{n\to\infty}\frac{1}{n}\log\Pi\Big[\boldsymbol{\xi}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]+\widetilde{c}\cdot(A-1-\log A), \tag{92}$$

$$\inf_{\mathbb{Q}\in A\cdot\mathbf{\Omega}}\widetilde{I}(\mathbb{Q},\mathbb{P})=\inf_{\mathbb{Q}\in A\cdot\mathbf{\Omega}}\sum_{k=1}^{K}p_k\cdot\log\Big(\frac{p_k}{q_k}\Big)=-\lim_{n\to\infty}\frac{1}{n}\log\Pi\Big[\boldsymbol{\xi}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]-\log A.$$

(b) The special case $\varphi:=\widetilde{c}\cdot\varphi_{\gamma}\quad(\gamma=0)$ of Theorem 14 works analogously to (a), with the differences that we employ (i) additionally a sequence $(X_i)_{i\in\mathbb{N}}$ of random variables being independent of $(W_i)_{i\in\mathbb{N}}$ and satisfying condition (26) (resp. (30)), (ii) $A=1$ (instead of arbitrary $A>0$), (iii) $\Pi_{X_1^n}[\cdot]$ (instead of $\Pi[\cdot]$), and (iv) $\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\boldsymbol{\xi}_n^{w\mathbf{W}}$).

Proposition 24: (a) Consider the context of Remark 15(vi) for $\varphi:=\widetilde{c}\cdot\varphi_{\gamma}$ with $\gamma\in]0,1[$, and let $\mathbb{P}\in\mathbb{S}_{>0}^{K}$ as well as $\widetilde{c}>0$ be arbitrary but fixed. Furthermore, let $W:=(W_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of non-negative real-valued random variables with Compound-Poisson-Gamma distribution $\zeta=C(POI(\theta),GAM(\alpha,\beta))$ having parameters $\theta=\frac{\widetilde{c}}{\gamma}>0$, $\alpha=\frac{\widetilde{c}}{1-\gamma}>0$, $\beta=\frac{\gamma}{1-\gamma}>0$; in other words, the W_i are independent copies of a random variable $W_1:=\sum_{j=1}^{N}\overline{W}_j$ [24] constituted of some i.i.d. sequence $(\overline{W}_j)_{j\in\mathbb{N}}$ of $Gamma(\alpha,\beta)-$distributed random variables and some independent $POI(\theta)-$distributed random variable N. Then for all $A>0$ and all $\mathbf{\Omega}\subset\mathbb{S}^{K}$ with (7) there hold (83), (84), (88), (91) as well as

$$\sup_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}H_{\gamma}(\mathbf{Q},\mathbb{P})=\lim_{n\to\infty}A^{\gamma}\cdot\Big(1+\gamma\cdot\frac{1}{n}\cdot\log\Pi\Big[\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]\Big)^{1-\gamma}, \tag{93}$$

$$\sup_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}c_1\cdot\Big(H_{\gamma}(\mathbf{Q},\mathbb{P})^{c_2}-c_3\Big)=\lim_{n\to\infty}c_1\cdot\Big\{A^{c_2\cdot\gamma}\cdot\Big(1+\frac{\gamma}{n}\cdot\log\Pi\Big[\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]\Big)^{c_2\cdot(1-\gamma)}-c_3\Big\},\quad\text{if } c_1\cdot c_2>0,\ c_3\in\mathbb{R}, \tag{94}$$

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}c_1\cdot\Big(H_{\gamma}(\mathbf{Q},\mathbb{P})^{c_2}-c_3\Big)=\lim_{n\to\infty}c_1\cdot\Big\{A^{c_2\cdot\gamma}\cdot\Big(1+\frac{\gamma}{n}\cdot\log\Pi\Big[\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]\Big)^{c_2\cdot(1-\gamma)}-c_3\Big\},\quad\text{if } c_1\cdot c_2<0,\ c_3\in\mathbb{R}, \tag{95}$$

$$\sup_{\mathbf{Q}\in A\cdot\mathbf{\Omega}} c_1\cdot\Big(\Big(\sum_{k=1}^{K} q_k^{\gamma}\Big)^{c_2}-c_3\Big)=\lim_{n\to\infty} c_1\cdot\Big\{K^{c_2\cdot(1-\gamma)}\cdot A^{c_2\cdot\gamma}\cdot\Big(1+\frac{\gamma}{n}\cdot\log \Pi\Big[\breve{\breve{\boldsymbol{\xi}}}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]\Big)^{c_2\cdot(1-\gamma)}-c_3\Big\}, \quad \text{if } c_1\cdot c_2>0,\ c_3\in\mathbb{R}, \tag{96}$$

[22] the latter two notations will be also used in the following Propositions 23 to 27

[23] here and henceforth, we use the notation that a Gamma distribution $GAM(\alpha,\beta)$ with rate parameter (inverse scale parameter) $\alpha>0$ and shape parameter $\beta>0$ has (Lebesgue-)density $f(y):=\frac{\alpha^{\beta}\cdot y^{\beta-1}\cdot e^{-\alpha\cdot y}}{\Gamma(\beta)}\cdot 1_{]0,\infty[}(y),\ \ y\in\mathbb{R};$ its cumulant generating function is $\Lambda(z)=\beta\cdot\log(\frac{\alpha}{\alpha-z})$ for $z\in]-\infty,\alpha[$ (and $\Lambda(z)=\infty$ for $z\geq\alpha$).

[24] with the usual convention $\sum_{i=1}^{0}\overline{W}_i:=0$

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}} c_1\cdot\Big(\Big(\sum_{k=1}^{K} q_k^{\gamma}\Big)^{c_2}-c_3\Big)=\lim_{n\to\infty} c_1\cdot\Big\{K^{c_2\cdot(1-\gamma)}\cdot A^{c_2\cdot\gamma}\cdot\Big(1+\frac{\gamma}{n}\cdot\log \Pi\Big[\breve{\breve{\boldsymbol{\xi}}}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]\Big)^{c_2\cdot(1-\gamma)}-c_3\Big\},$$

$$\text{if } c_1\cdot c_2<0,\ c_3\in\mathbb{R}. \quad (97)$$

From this, the BS-minimizability/maximizability of the important norms/entropies/diversity indices (E1) to (E6) follows immediately as special cases.
(b) The special case $\varphi := \widetilde{c}\cdot\varphi_\gamma \quad (\gamma\in]0,1[)$ of Theorem 14 works analogously to (a), with the differences that we employ (i) additionally a sequence $(X_i)_{i\in\mathbb{N}}$ of random variables being independent of $(W_i)_{i\in\mathbb{N}}$ and satisfying condition (26) (resp. (30)), (ii) $A=1$ (instead of arbitrary $A>0$), (iii) $\Pi_{X_1^n}[\cdot]$ (instead of $\Pi[\cdot]$), (iv) $\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\boldsymbol{\xi}_n^{w\mathbf{W}}$), (v) $\breve{\boldsymbol{\xi}}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}}$), and (vi) $\breve{\breve{\boldsymbol{\xi}}}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\breve{\breve{\boldsymbol{\xi}}}_n^{w\mathbf{W}}$).

This follows from Theorem 14, Remark 15(vi), Lemma 16(a), (65), (69), (73), (74) and the below-mentioned $\mathfrak{C}-$concerning Example 48(b). Employing Lemma 16(b) and Example 50(a) instead, one ends up with the following proposition on the Kullback-Leibler divergence:

Proposition 25: (a) Consider the context of Remark 15(vi) for $\varphi := \widetilde{c}\cdot\varphi_\gamma$ with $\gamma=1$, and let $\mathbb{P}\in\mathbb{S}_{>0}^K$ as well as $\widetilde{c}>0$ be arbitrary but fixed. Furthermore, let $W:=(W_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of non-negative real-valued random variables with distribution $\mathfrak{C}=\frac{1}{\widetilde{c}}\cdot POI(\widetilde{c})$ being the "$\frac{1}{\widetilde{c}}-$fold Poisson distribution with mean $\widetilde{c}$" , which means that $W_1=\frac{1}{\widetilde{c}}\cdot Z$ for a Poissonian $POI(\widetilde{c})-$distributed random variable Z with mean $\widetilde{c}$ (where the subcase $\widetilde{c}=1$ amounts to $\mathfrak{C}=POI(1)$). Then for all $A>0$ and all $\mathbf{\Omega}\subset\mathbb{S}^K$ with (7) there holds

$$-\lim_{n\to\infty}\frac{1}{n}\log \Pi\Big[\boldsymbol{\xi}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big] \quad = \quad \inf_{\mathbb{Q}\in A\cdot\mathbf{\Omega}}\ \widetilde{c}\cdot\Big[1-A\cdot\exp\Big(-\frac{1}{A\cdot\widetilde{c}}\cdot D_{\widetilde{c}\cdot\varphi_1}(\mathbf{Q},\mathbb{P})+\frac{1}{A}-1\Big)\Big] \quad (98)$$

and the BS minimizabilities/maximizabilites (cf. Definition 1)

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}} D_{\widetilde{c}\cdot\varphi_1}(\mathbf{Q},\mathbb{P})=\lim_{n\to\infty}\widetilde{c}\cdot\Big\{1-A\cdot\Big[1+\log\Big(\frac{1}{A}\cdot\Big(1+\frac{1}{\widetilde{c}}\cdot\frac{1}{n}\cdot\log \Pi\Big[\boldsymbol{\xi}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]\Big)\Big)\Big]\Big\}, \quad (99)$$

$$\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}} I(\mathbf{Q},\mathbb{P})=\inf_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\sum_{k=1}^{K} q_k\cdot\log\Big(\frac{q_k}{p_k}\Big)=-\lim_{n\to\infty} A\cdot\log\Big(\frac{1}{A}\cdot\Big(1+\frac{1}{n}\cdot\log \Pi\Big[\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]\Big)\Big),$$

$$\max_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\mathcal{E}^{Sh}(\mathbf{Q})=\max_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}(-1)\cdot\sum_{k=1}^{K} q_k\cdot\log(q_k)=\lim_{n\to\infty}\frac{\log K}{K}+A\cdot\log\Big(\frac{1}{A}\cdot\Big(1+\frac{1}{n}\cdot\log \Pi\Big[\breve{\breve{\boldsymbol{\xi}}}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]\Big)\Big), \quad (100)$$

$$\max_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\mathcal{E}^{gSM2}(\mathbf{Q})=\max_{\mathbf{Q}\in A\cdot\mathbf{\Omega}}\frac{1}{1-s}\cdot\exp\Big\{(s-1)\cdot\sum_{k=1}^{K} q_k\cdot\log(q_k)-1\Big\}$$

$$=\lim_{n\to\infty}\frac{1}{1-s}\cdot\exp\Big\{(1-s)\cdot\Big[\frac{\log K}{K}+A\cdot\log\Big(\frac{1}{A}\cdot\Big(1+\frac{1}{n}\cdot\log \Pi\Big[\breve{\breve{\boldsymbol{\xi}}}_n^{w\mathbf{W}}\in\mathbf{\Omega}\Big]\Big)\Big)\Big]-1\Big\},\quad s\in]0,1[\cup]1,\infty[. \quad (101)$$

The special subcase $A=1$ in (100) (and thus, $\mathbf{Q}$ is a probability vector) corresponds to the *maximum entropy problem* for the Shannon entropy $\mathcal{E}^{Sh}(\cdot)$. This can hence be tackled by our bare-simulation approach for almost arbitrary sets $\mathbf{\Omega}$ of probability vectors.
(b) The special case $\varphi := \widetilde{c}\cdot\varphi_\gamma \quad (\gamma=1)$ of Theorem 14 works analogously to (a), with the differences that we employ (i) additionally a sequence $(X_i)_{i\in\mathbb{N}}$ of random variables being independent of $(W_i)_{i\in\mathbb{N}}$ and satisfying condition (26) (resp. (30)), (ii) $A=1$ (instead of arbitrary $A>0$), (iii) $\Pi_{X_1^n}[\cdot]$ (instead of $\Pi[\cdot]$), (iv) $\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\boldsymbol{\xi}_n^{w\mathbf{W}}$), (v) $\breve{\boldsymbol{\xi}}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}}$), and (vi) $\breve{\breve{\boldsymbol{\xi}}}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\breve{\breve{\boldsymbol{\xi}}}_n^{w\mathbf{W}}$).

For the sake of completeness, let us mention here that we do not deal with the case $\gamma\in]1,2[$, for which we conjecture that $\mathfrak{C}$ becomes a signed finite measure with total mass 1, i.e. it has a density (with respect to some dominating measure) with positive *and negative* values which "integrates to 1" ; accordingly, our BS method can not be applied to this situation.

To proceed with further $\gamma-$cases, a combination of Theorem 14 respectively Remark 15(vi), Lemma 16(a), (65), (69), (73), (74) and the below-mentioned $\mathfrak{C}-$concerning Example 48(c) leads to the following

Proposition 26: (a) Consider the context of Remark 15(vi) for $\varphi := \widetilde{c}\cdot\varphi_\gamma$ with $\gamma=2$, and let $\mathbb{P}\in\mathbb{S}_{>0}^K$ as well as $\widetilde{c}>0$ be arbitrary but fixed. Furthermore, let $W:=(W_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of real-valued random variables with probability distribution $\mathfrak{C}=NOR(1,\frac{1}{\widetilde{c}})$ being the Normal (Gaussian) law with mean 1 and variance $\frac{1}{\widetilde{c}}$. Then for all $A>0$ and $\mathbf{\Omega}\subset\mathbb{S}^K$ with (7) there hold all the BS-extremizabilites (83) to (91) as well as (116) (below) with plugging-in $\gamma=2$. From this, the BS-minimizability/maximizability of the important norms/entropies/diversity indices (E1) to (E6) follow immediately as special

cases. By Remark 18(c), one can even take $A<0$ in (83) to (91) and (116) as well as in (E1), (E2), (E4) and (E6).
(b) The special case $\varphi := \widetilde{c}\cdot\varphi_\gamma \quad (\gamma=2)$ of Theorem 14 works analogously to (a), with the differences that we employ (i) additionally a sequence $(X_i)_{i\in\mathbb{N}}$ of random variables being independent of $(W_i)_{i\in\mathbb{N}}$ and satisfying condition (26) (resp. (30)), (ii) $A=1$ (instead of arbitrary $A>0$), (iii) $\Pi_{X_1^n}[\cdot]$ (instead of $\Pi[\cdot]$), (iv) $\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\boldsymbol{\xi}_{n}^{w\mathbf{W}}$), (v) $\breve{\boldsymbol{\xi}}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\breve{\boldsymbol{\xi}}_{n}^{w\mathbf{W}}$), and (vi) $\breve{\breve{\boldsymbol{\xi}}}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\breve{\breve{\boldsymbol{\xi}}}_{n}^{w\mathbf{W}}$).

For instance, the BS minimizability (89) of Proposition 26(a) can be employed to solve the following discrete Monge-Kantorovich-type optimal mass transportation problem (optimal coupling problem) with *side (i.e. additional) constraints*: given two nonnegative-entries vectors $\boldsymbol{\mu} := (\mu_1,\ldots\mu_{K_1}) \in [0,\infty[^{K_1}$ and $\boldsymbol{\nu} := (\nu_1,\ldots\nu_{K_2}) \in [0,\infty[^{K_2}$ with equal total "mass" $\sum_{k=1}^{K_1}\mu_k = \sum_{k=1}^{K_2}\nu_k = A>0$, compute

$$\inf_{K_1\times K_2-\text{matrices }\pi} K_1\cdot K_2\cdot\sum_{u=1}^{K_1}\sum_{v=1}^{K_2}\left(\pi_{u,v}-\frac{1}{K_1\cdot K_2}\right)^2 \tag{102}$$

subject to

$$\sum_{v=1}^{K_2}\pi_{u,v}=\mu_u \quad \text{for all } u\in\{1,\ldots,K_1\}, \tag{103}$$

$$\sum_{u=1}^{K_1}\pi_{u,v}=\nu_v \quad \text{for all } v\in\{1,\ldots,K_2\}, \tag{104}$$

$$\pi_{u,v}\in[0,A] \quad \text{for all } u\in\{1,\ldots,K_1\},\, v\in\{1,\ldots,K_2\}, \tag{105}$$

$$\text{side constraints on } \pi,\ \boldsymbol{\mu},\ \boldsymbol{\nu}. \tag{106}$$

Indeed, this problem can be equivalently rewritten in terms $K_1\cdot K_2-$dimensional vectors as follows: given two nonnegative-entries vectors $\boldsymbol{\mu}$,$\boldsymbol{\nu}$ as above, compute

$$\inf_{\mathbf{Q}\in\boldsymbol{\Omega}} K_1\cdot K_2\cdot\sum_{k=1}^{K_1\cdot K_2}\left(q_k-\frac{1}{K_1\cdot K_2}\right)^2 \;=\; \inf_{\mathbf{Q}\in\boldsymbol{\Omega}} K_1\cdot K_2\cdot\sum_{k=1}^{K_1\cdot K_2} q_k^2+1-2A \tag{107}$$

where $\boldsymbol{\Omega}\subset\mathbb{R}^{K_1\cdot K_2}$ is the set of all vectors $\mathbf{Q}=(q_1,\ldots,q_{K_1\cdot K_2})$ which satisfy the constraints

$$\sum_{j=1}^{K_2} q_{(i-1)\cdot K_2+j} \;=\; \mu_i \quad \text{for all } i\in\{1,\ldots,K_1\}, \tag{108}$$

$$\sum_{i=1}^{K_1} q_{(i-1)\cdot K_2+j} \;=\; \nu_j \quad \text{for all } j\in\{1,\ldots,K_2\}, \tag{109}$$

$$q_k\in[0,A] \quad \text{for all } k\in\{1,\ldots,K_1\cdot K_2\}, \tag{110}$$

$$\text{side constraints on } \mathbf{Q},\ \boldsymbol{\mu},\ \boldsymbol{\nu}. \tag{111}$$

Clearly, via divisions by A, one can equivalently rewrite $\boldsymbol{\Omega}=A\cdot\boldsymbol{\mathfrak{Q}}$ for some $\boldsymbol{\mathfrak{Q}}\subset\mathbb{S}^{K_1\cdot K_2}$ in the $K_1\cdot K_2-$dimensional probability simplex. Hence, we can employ (89) with $c_1=K_1\cdot K_2$, $c_2=1$ and $c_3=1-2A$, provided that the side constraints (111) are such that $\boldsymbol{\Omega}$ satisfies the regularity property (7) and the finiteness property (9). Notice that (107) is equal to

$$\inf_{\mathbf{Q}\in\boldsymbol{\Omega}} D_{2\cdot\varphi_2}(\mathbf{Q},\mathbb{P}^{unif})$$

where $\mathbb{P}^{unif} := (\frac{1}{K_1\cdot K_2},\ldots,\frac{1}{K_1\cdot K_2})$ is the probability vector of frequencies of the uniform distribution on $\{1,\ldots,K_1\cdot K_2\}$, and $\widetilde{c}=2$. The special case $A=1$ with side constraint (111) of the form $K_1\cdot\min_{i\in\{1,\ldots,K_1\}}\mu_i+K_2\cdot\min_{j\in\{1,\ldots,K_2\}}\nu_j\geq 1$ was explicitly solved by e.g. Bertrand et al. [43], [44], who also give applications to cryptographic guessing problems (spy problems), task partitioning and graph clustering.

The importance of the case $\gamma=2$ stems also from the fact that one can equivalently rewrite *separable quadratic minimization problems* as minimization problems of Pearson chi-square divergences. Indeed, by straightforward calculations one can derive that

$$\inf_{\breve{\mathbf{Q}}\in\breve{\boldsymbol{\Omega}}}\sum_{k=1}^{K}(\,c_{1,k}+c_{2,k}\cdot\breve{q}_k+c_{3,k}\cdot\breve{q}_k^2\,)\,, \qquad c_{1,k}\in\mathbb{R},\ c_{2,k}\in\mathbb{R}\backslash\{0\},\ c_{3,k}\in]0,\infty[, \tag{112}$$

is equal to (recall that $\varphi_2(t) := \frac{(t-1)^2}{2}$, cf. (43))

$$c_4+\inf_{\mathbf{Q}\in\boldsymbol{\Omega}} D_{2\cdot\varphi_2}(\mathbf{Q},\mathbf{P})\,, \tag{113}$$

where $\mathbf{Q} := (q_1, \ldots, q_K)$ with $q_k := -c_{2,k} \cdot \breve{q}_k$, $\mathbf{P} := (p_1, \ldots, p_K)$ with $p_k := \frac{c_{2,k}^2}{2 \cdot c_{3,k}} > 0$, $c_4 := \sum_{k=1}^{K} \big(c_{1,k} - \frac{c_{2,k}^2}{4 \cdot c_{3,k}} \big)$, and $\mathbf{\Omega}$ is the corresponding reformulation of the constraint set $\breve{\mathbf{\Omega}}$. To achieve the applicability of our BS method, we further transform (113) into its equal form (cf. (12))

$$c_4 + \inf_{\widetilde{\mathbf{Q}} \in \mathbf{\Omega}/M_P} D_{2M_{\mathbf{P}} \cdot \varphi_2}(\mathbf{Q}, \widetilde{\mathbb{P}}) \tag{114}$$

with $M_{\mathbf{P}} := \sum_{k=1}^{K} p_k > 0$ and $\widetilde{\mathbb{P}} := \mathbf{P}/M_{\mathbf{P}}$. If $\mathbf{\Omega}/M_{\mathbf{P}}$ satisfies (7) and (9) (e.g. it may be highly disconnected), then we can apply Theorem 10. In contrast, if $\mathbf{\Omega}/M_{\mathbf{P}} = A \cdot \mathbf{\Omega\!\!\Omega}$ for some $A \in \mathbb{R}\backslash\{0\}$ and some $\mathbf{\Omega\!\!\Omega} \subset \mathbb{S}_{>0}^{K}$ satisfying (7), then we can apply Proposition 26(a) together with Remark 18(c); for instance, this may appear if $\breve{\mathbf{\Omega}}$ contains (amongst others) the original constraint $\sum_{k=1}^{K} \breve{q}_k = C$ for some constant $C > 0$, and $c_{2,k} = c_2$ does not depend on k, which leads to the choice $A = -\frac{c_2 \cdot C}{M_{\mathbf{P}}}$. Notice that $A < 0$ if $c_2 > 0$. For example, optimization problems (112) with $c_{1,k} > 0$, $c_{2,k} > 0$, $c_{3,k} > 0$ and constraints $\sum_{k=1}^{K} \breve{q}_k = C$, $\breve{q}_k \in [\underline{\breve{q}}_k, \overline{\breve{q}}_k]$ appear in distributed energy management as *economic dispatch problems* in smart grids of power generators, where $\breve{q}_k$ is the active power generation of the $k-$th generator, C is the total power demand, $\underline{\breve{q}}_k$ resp. $\overline{\breve{q}}_k$ represent the lower resp. upper bound of the $k-$th generator's output, and the cost of power generation is $c_{1,k} + c_{2,k} \cdot \breve{q}_k + c_{3,k} \cdot \breve{q}_k^2$ (cf. e.g. Yang et al. [412], Loia & Vaccaro [230], Wood et al. [394], Xu et al. [404]).
Another important special case of (112) to (114) is the omnipresent L_2-minimization; indeed, with the choices $c_{3,k} = 1$, $c_{2,k} = -2v_k$, and $c_{1,k} = v_k^2$ for some $\mathbf{V} = (v_1, \ldots, v_K)$, the minimization problem (112) is nothing but $\inf_{\breve{\mathbf{Q}} \in \breve{\mathbf{\Omega}}} ||\breve{\mathbf{Q}} - \mathbf{V}||_2^2$; if $\breve{\mathbf{\Omega}}$ depends on a pregiven $L-$dimensional vector $\mathbf{x}$ (with $L < K$), this can be regarded as a non-parametric regression problem in a wide sense.

To continue with our general investigations, by combining Theorem 14 respectively Remark 15(vi), Lemma 16(a), (65), (69), (73), (74) and the below-mentioned $\mathfrak{C}-$concerning Example 48(e), we arrive at the following

Proposition 27: (a) Consider the context of Remark 15(vi) for $\varphi := \widetilde{c} \cdot \varphi_\gamma$ with $\gamma > 2$, and let $\mathbb{P} \in \mathbb{S}_{>0}^{K}$ as well as $\widetilde{c} > 0$ be arbitrary but fixed. Furthermore, let $W := (W_i)_{i \in \mathbb{N}}$ be an i.i.d. sequence of real-valued random variables having density[25]

$$\frac{\exp\{\frac{y \cdot \widetilde{c}}{\gamma - 1}\}}{\exp\{\widetilde{c}/\gamma\}} \cdot f_Z(-y), \qquad y \in]-\infty, \infty[, \tag{115}$$

where f_Z is the density of a random variable Z which has stable law with parameter-quadruple $(\frac{\gamma}{\gamma-1}, 1, 0, \frac{\widetilde{c}^{1/(1-\gamma)} \cdot (\gamma-1)^{\gamma/(\gamma-1)}}{\gamma})$ in terms of the above-mentioned "form-B notation" in Zolotarev [428]. Then for all $A > 0$ and $\mathbf{\Omega\!\!\Omega} \subset \mathbb{S}^K$ with (7) there hold all the BS-extremizabilites (83) to (90) as well as

$$\sup_{\mathbf{Q} \in A \cdot \mathbf{\Omega\!\!\Omega}} \frac{1}{1-\gamma} \cdot \log\Big(\sum_{k=1}^{K} q_k^{\gamma} \Big) = \lim_{n \to \infty} \frac{1}{\gamma \cdot (\gamma - 1)} \cdot \Big[\log\Big(A^{\gamma} \cdot \Big(1 + \gamma \cdot \frac{1}{n} \cdot \log \mathbb{\Pi}\Big[\breve{\boldsymbol{\xi}}_n^{w\mathbf{W}} \in \mathbf{\Omega\!\!\Omega} \Big] \Big)^{1-\gamma} \Big) + (1-\gamma) \cdot \log(K) \Big]. \tag{116}$$

From this, the BS-minimizability/maximizability of the important norms/entropies/diversity indices (E1) to (E6) follow immediately as special cases.
(b) The special case $\varphi := \widetilde{c} \cdot \varphi_\gamma \quad (\gamma > 2)$ of Theorem 14 works analogously to (a), with the differences that we employ (i) additionally a sequence $(X_i)_{i \in \mathbb{N}}$ of random variables being independent of $(W_i)_{i \in \mathbb{N}}$ and satisfying condition (26) (resp. (30)), (ii) $A = 1$ (instead of arbitrary $A > 0$), (iii) $\mathbb{\Pi}_{X_1^n}[\cdot]$ (instead of $\mathbb{\Pi}[\cdot]$), (iv) $\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\boldsymbol{\xi}_{n}^{w\mathbf{W}}$), (v) $\breve{\boldsymbol{\xi}}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\breve{\boldsymbol{\xi}}_{n}^{w\mathbf{W}}$), and (vi) $\breve{\breve{\boldsymbol{\xi}}}_{n,\mathbf{X}}^{w\mathbf{W}}$ (instead of $\breve{\breve{\boldsymbol{\xi}}}_{n}^{w\mathbf{W}}$).

As mentioned above, in the Propositions 22 to 27 we have combined Theorem 14 respectively Remark 15(vi), Lemma 16 and explicitly solved representations (6). The latter, important step will be discussed in a structured, comprehensive manner in the Section V below.

By retransformation, we can even deal with optimizations of nonnegative *linear* objective functions with constraint sets on Euclidean γ-norm spheres. Indeed, for nonnegative $\breve{\mathbf{Q}} := (\breve{q}_1, \ldots, \breve{q}_K)$ and $\breve{\mathbf{P}} := (\breve{p}_1, \ldots, \breve{p}_K)$ one can rewrite their scalar product as $\gamma-$order Hellinger integrals

$$\sum_{k=1}^{K} \breve{q}_k \cdot \breve{p}_k \; = \; c_1 \cdot \sum_{k=1}^{K} q_k^{\gamma} \cdot p_k^{1-\gamma} \; = \; c_1 \cdot H_\gamma(\mathbf{Q}, \mathbb{P}) \qquad \text{where} \tag{117}$$

$$\gamma \in]0,1[\cup [2, \infty[\quad \text{if} \quad \breve{\mathbf{Q}} \in [0, \infty[^K, \; \breve{\mathbf{P}} \in]0, \infty[^K \qquad \text{respectively} \qquad \gamma \in]-\infty, 0[\quad \text{if} \quad \breve{\mathbf{Q}} \in]0, \infty[^K, \; \breve{\mathbf{P}} \in]0, \infty[^K, \tag{118}$$

$$q_k \; := \; \breve{q}_k^{1/\gamma}, \quad p_k \; := \; \frac{\breve{p}_k^{1/(1-\gamma)}}{\sum_{i=1}^{K} \breve{p}_i^{1/(1-\gamma)}}, \quad c_1 := \Big(\sum_{i=1}^{K} \breve{p}_i^{1/(1-\gamma)} \Big)^{1-\gamma} =: ||\breve{\mathbf{P}}||_{1-\gamma} \; . \tag{119}$$

[25] in the classical sense, with respect to Lebesgue measure

The required constraint $\sum_{k=1}^{K} q_k = A > 0$ retransforms to $||\breve{\mathbf{Q}}||_\gamma = A^{1/\gamma}$ and thus, $\breve{\mathbf{Q}}$ must lie on (the positive/nonnegative part of) the $\gamma-$norm-sphere $\partial B_\gamma(0, A^{1/\gamma})$ around the origin with radius $A^{1/\gamma}$. Accordingly, for $\gamma \in [2, \infty[$ we have

$$\inf_{\breve{\mathbf{Q}} \in \breve{\mathbf{\Omega}}} \sum_{k=1}^{K} \breve{q}_k \cdot \breve{p}_k \;=\; c_1 \cdot \inf_{\mathbf{Q} \in A \cdot \mathbf{\Omega}} H_\gamma(\mathbf{Q}, \mathbb{P}) \tag{120}$$

and we can apply (85) of Proposition 26(a) respectively Proposition 27(a) [26], as long as the original constraint set $\breve{\mathbf{\Omega}} \in \partial B_\gamma(0, A^{1/\gamma}) \cap [0, \infty[^K$ transforms (via $q_k = \breve{q}_k^{1/\gamma}$) into a constraint set $A \cdot \mathbf{\Omega}$ which satisfies the regularity assumption (7) in the relative topology (as a side remark, notice that $int(\partial B_\gamma(0, A^{1/\gamma})) = \emptyset$ in the full topology). For the case $\gamma \in]-\infty, 0[$ we also have (120) and apply (85) of Proposition 22(a) for any original constraint set $\breve{\mathbf{\Omega}} \in \partial B_\gamma(0, A^{1/\gamma}) \cap]0, \infty[^K$ which transforms into $A \cdot \mathbf{\Omega}$ satisfying (7) in the relative topology. In contrast, for the case $\gamma \in]0, 1[$ we get

$$\sup_{\breve{\mathbf{Q}} \in \breve{\mathbf{\Omega}}} \sum_{k=1}^{K} \breve{q}_k \cdot \breve{p}_k \;=\; c_1 \cdot \sup_{\mathbf{Q} \in A \cdot \mathbf{\Omega}} H_\gamma(\mathbf{Q}, \mathbb{P})$$

and apply (93) of Proposition 24(a) for any original constraint set $\breve{\mathbf{\Omega}} \in \partial B_\gamma(0, A^{1/\gamma}) \cap [0, \infty[^K$ which transforms into $A \cdot \mathbf{\Omega}$ satisfying (7) in the relative topology.

As a continuation of Remark 12, we can principally tackle all the optimization problems of this Subsection IV-C1 by basically *only employing a fast and accurate — pseudo, true, natural, quantum — random number generator*, provided that the constraint set $A \cdot \mathbf{\Omega}$ satisfies the mild assumptions (7) (in the relative topology) and (9). Recall that $A > 0$ (and for φ_2 even $A \in \mathbb{R}\backslash\{0\}$) and that $\mathbf{Q} \in A \cdot \mathbf{\Omega}$ implies in particular the constraint $\sum_{k=1}^{K} q_k = A$. The regularity assumption (7) allows for e.g. high-dimensional constraint sets $A \cdot \mathbf{\Omega}$ which are *non-convex* and even *highly disconnected*, and for which other minimization methods (e.g. pure enumeration, gradient or steepest descent methods, etc.) may be problematic or intractable. For example, (7) covers kind of "$K-$dimensional (not necessarily regular) polka dot pattern type" relaxations $A \cdot \mathbf{\Omega} := \dot{\bigcup}_{i=1}^{N} \mathcal{U}_i(Q_i^{dis})$ of finite discrete constraint sets $A \cdot \mathbf{\Omega}^{dis} := \{Q_1^{dis}, \ldots, Q_N^{dis}\}$ of high cardinality N (e.g. being exponential or factorial in a large K), where each $K-$dimensional vector Q_i^{dis} has total-sum-of-components equal to A and is surrounded by some small ("flat", i.e. in the relative topology) neighborhood $\mathcal{U}_i(Q_i^{dis})$. For the sake of brevity, in the following discussion we confine ourselves to the deterministic setup (e.g. Proposition 26(a) rather than (b)) which particularly involves $\Pi[\cdot]$ (rather than $\Pi_{X_1^n}[\cdot]$) and $\boldsymbol{\xi}_n^{w\mathbf{W}}$ (rather than $\xi_{n,\mathbf{X}}^{w\mathbf{W}}$). In such a context, all the optimization problems of this Subsection IV-C1, subsumed as (cf. (1) to (3))

$$\inf_{\mathbf{Q} \in A \cdot \mathbf{\Omega}} \Phi(\mathbf{Q}) \qquad \text{respectively} \qquad \sup_{\mathbf{Q} \in A \cdot \mathbf{\Omega}} \Phi(\mathbf{Q})$$

can be regarded as a "BS-tractable" relaxations of the corresponding nonlinear discrete (e.g. integer, combinatorial) programming problems

$$\inf_{\mathbf{Q} \in A \cdot \mathbf{\Omega}^{dis}} \Phi(\mathbf{Q}) \qquad \text{respectively} \qquad \sup_{\mathbf{Q} \in A \cdot \mathbf{\Omega}^{dis}} \Phi(\mathbf{Q})\,;$$

as examples take e.g. $\Phi(\mathbf{Q}) = c_1 \cdot \Big(\Big(\sum_{k=1}^{K} q_k^\gamma \Big)^{c_2} - c_3 \Big)$ (with $\gamma \neq 0, 1$) or $\Phi(\mathbf{Q}) = \Phi_{\mathbb{P}}(\mathbf{Q}) = D_{\widetilde{c} \cdot \varphi_\gamma}(\mathbf{Q}, \mathbb{P})$. For instance, $A \cdot \mathbf{\Omega}^{dis}$ may contain only $K-$dimensional vectors Q_i^{dis} $(i = 1, \ldots, N)$ whose components stem from a finite set $\mathcal{B}$ of nonnegative integers and add up to A. If $\mathcal{B} = \{0, 1\}$, then we can even deal with nonnegative *linear* objective functions $\Phi(\mathbf{Q}) = \sum_{k=1}^{K} \breve{p}_k \cdot q_k$ where $\mathbf{Q} := (q_1, \ldots, q_K)$ with $q_k \in \{0, 1\}$ and $\breve{\mathbf{P}} := (\breve{p}_1, \ldots, \breve{p}_K)$ has components $\breve{p}_k > 0$ which reflect e.g. the cost associated with the $k-$th state. Indeed, by noticing that $q_k^{1/\gamma} = q_k$ for $\gamma \in]0, 1[\cup [2, \infty[$, we can employ (117) and (119) to end up with

$$\inf_{\mathbf{Q} \in A \cdot \mathbf{\Omega}^{dis}} \sum_{k=1}^{K} q_k \cdot \breve{p}_k = ||\breve{\mathbf{P}}||_{1-\gamma} \cdot \inf_{\mathbf{Q} \in A \cdot \mathbf{\Omega}^{dis}} \sum_{k=1}^{K} q_k^\gamma \cdot \left(\frac{\breve{p}_k^{1/(1-\gamma)}}{\sum_{i=1}^{K} \breve{p}_i^{1/(1-\gamma)}} \right)^{1-\gamma} = ||\breve{\mathbf{P}}||_{1-\gamma} \cdot \inf_{\mathbf{Q} \in A \cdot \mathbf{\Omega}^{dis}} H_\gamma(\mathbf{Q}, \mathbb{P}), \quad \gamma \in [2, \infty[, \tag{121}$$

$$\sup_{\mathbf{Q} \in A \cdot \mathbf{\Omega}^{dis}} \sum_{k=1}^{K} q_k \cdot \breve{p}_k = ||\breve{\mathbf{P}}||_{1-\gamma} \cdot \sup_{\mathbf{Q} \in A \cdot \mathbf{\Omega}^{dis}} H_\gamma(\mathbf{Q}, \mathbb{P}), \quad \gamma \in]0, 1[\,. \tag{122}$$

The corresponding relaxations are

$$\inf_{\mathbf{Q} \in A \cdot \mathbf{\Omega}} \sum_{k=1}^{K} q_k \cdot \breve{p}_k \;=\; ||\breve{\mathbf{P}}||_{1-\gamma} \cdot \inf_{\mathbf{Q} \in A \cdot \mathbf{\Omega}} H_\gamma(\mathbf{Q}, \mathbb{P}), \quad \gamma \in [2, \infty[\,, \tag{123}$$

$$\sup_{\mathbf{Q} \in A \cdot \mathbf{\Omega}} \sum_{k=1}^{K} q_k \cdot \breve{p}_k \;=\; ||\breve{\mathbf{P}}||_{1-\gamma} \cdot \sup_{\mathbf{Q} \in A \cdot \mathbf{\Omega}} H_\gamma(\mathbf{Q}, \mathbb{P}), \quad \gamma \in]0, 1[\,; \tag{124}$$

[26] here and analogously henceforth, by this we mean the condition (85) as it appears in the Proposition 26(a) respectively Proposition 27(a)

for (123) we can apply (85) of Proposition 26(a) respectively Proposition 27(a), whereas for (124) we apply (93) of Proposition 24(a) — as long as the relaxation constraint set $A \cdot \mathbf{\Omega}$ satisfies (7) in the relative topology. For the sake of illustration, let us consider a sum-minimization-type *linear assignment problem* with side constraints (for a comprehensive book on assignment problems see e.g. Burkard et al. [69]). Suppose that there are K individuals (people, machines, etc.) to carry out K tasks (jobs, etc.). Each individual is assigned to carry out exactly one task. There is cost $c_{ij} > 0$ if individual i is assigned to (i.e., carries out) task j. We want to find the minimum total cost amongst all assignments. There may be side constraints, e.g. each assignment has a value $v_{ij} > 0$ and the total value of the assignment should be above a pregiven threshold. As usual, the problem can be formulated with the help of binary variables x_{ij} where $x_{ij} = 1$ if individual i is assigned to task j, and $x_{ij} = 0$ otherwise. Accordingly, we want to compute

$$\inf_{K\times K-\text{matrices } x=(x_{ij})} \sum_{i=1}^{K} \sum_{j=1}^{K} c_{ij} \cdot x_{ij} \tag{125}$$

subject to

$$\sum_{j=1}^{K} x_{ij} = 1 \quad \text{for all } i \in \{1,\ldots,K\}, \qquad \text{(i.e. each individual } i \text{ does one task)}, \tag{126}$$

$$\sum_{i=1}^{K} x_{ij} = 1 \quad \text{for all } j \in \{1,\ldots,K\}, \qquad \text{(i.e. each task } j \text{ is done by one individual)}, \tag{127}$$

$$x_{ij} \in \{0,1\} \quad \text{for all } i \in \{1,\ldots,K\},\, j \in \{1,\ldots,K\}, \tag{128}$$

$$\text{side (i.e. additional) constraints on } x = (x_{ij})_{i,j=1,\ldots,K}. \tag{129}$$

Analogously to (107), this problem can be equivalently rewritten in terms of K^2-dimensional vectors as follows: let $\mathbf{Q} := (q_1, \ldots, q_{K^2})$ and $\breve{\mathbf{P}} := (\breve{p}_1, \ldots, \breve{p}_{K^2})$ be such that $c_{ij} = \breve{p}_{(i-1)\cdot K+j}$ and $x_{ij} = q_{(i-1)\cdot K+j}$ for $i,j \in \{1,\ldots,K\}$ and compute

$$\inf_{\mathbf{Q}\in K\cdot\mathbf{\Omega}^{dis}} \sum_{k=1}^{K} q_k \cdot \breve{p}_k \tag{130}$$

where $K \cdot \mathbf{\Omega}^{dis} \subset \mathbb{R}^{K^2}$ is the set of all vectors $\mathbf{Q} = (q_1, \ldots, q_{K^2})$ which satisfy the constraints

$$\sum_{j=1}^{K} q_{(i-1)\cdot K+j} \;=\; 1 \quad \text{for all } i \in \{1,\ldots,K\}, \tag{131}$$

$$\sum_{i=1}^{K} q_{(i-1)\cdot K+j} \;=\; 1 \quad \text{for all } j \in \{1,\ldots,K\}, \tag{132}$$

$$q_k \in \{0,1\} \quad \text{for all } k \in \{1,\ldots,K^2\}, \tag{133}$$

$$\text{side constraints on } \mathbf{Q}. \tag{134}$$

As seen above, this can be rewritten as $\gamma-$order Hellinger-integral minimization problem (121), with $\gamma \geq 2$. We can obtain a *highly disconnected* "non-void-interior-type" relaxation of the binary integer programming problem (130) to (134) by replacing (133) with

$$q_k \in [0,\varepsilon_1] \cup [1-\varepsilon_2, 1] \quad \text{for all } k \in \{1,\ldots,K^2\}, \tag{135}$$

for some (possibly arbitrarily) small $\varepsilon_1, \varepsilon_2 > 0$ with $\varepsilon_1 + \varepsilon_2 < 1$. We denote by $K \cdot \mathbf{\Omega}$ the outcoming set manifested by the constraints (131), (132), (134) and (135), and accordingly we end up with a minimization problem of type (123), which we can tackle by (85) of Proposition 26(a) respectively Proposition 27(a), as long as (7) (in the relative topology) is satisfied. For instance, we can take $\gamma = 2$ and basically solve the corresponding optimization problem by basically simulating K^2-dimensional Gaussian random variables (even though the cardinality of $K \cdot \mathbf{\Omega}^{dis}$ may be high). As a side remark, let us mention that our relaxation (135) contrasts considerably to the frequently used continuous *linear programming (LP) relaxation*

$$q_k \in [0,1] \quad \text{for all } k \in \{1,\ldots,K^2\}.$$

Let us finally mention that an important special case of a minimization problem (125) to (129) is — the integer programming formulation of — the omnipresent (asymmetric) *traveling salesman problem (TSP)* with possible side constraints [27]. There, one has K cities and the cost of traveling from city i to city $j \neq i$ is given by $c_{ij} > 0$. Moreover, one sets $x_{ij} = 1$ if the traveler goes directly from city i to city j (in that order), and $x_{ij} = 0$ otherwise. For technical reasons, for $i = j$ we attribute a

[27] see e.g. Applegate et al. [14], Gutin & Punnen [147], Cook [92] for comprehensive books on TSP, its variations and its applications to logistics, machine scheduling, printed circuit board drilling, communication-network frequencing, genome sequencing, data clustering, and many others.

cost $c_{ii} > 0$ (e.g. hotel costs), but we require that always $x_{ii} = 0$ which we subsume as the first part of the constraints (129). Then, the constraint (126) means that the traveler leaves from city i exactly once, whereas (127) reflects that the traveler arrives at city j exactly once. The goal is to find a directed tour — i.e. a directed cycle/circuit that visits all K cities once — of minimum cost. Within this context, the second part of the constraints (129) should basically exclude solutions which consist of disconnected subtours (subtour elimination constraints (of e.g. the seminal Dantzig et al. [104]), connectivity constraints, cut-set constraints). Here, we also allow for additional/side constraints which we subsume as the third part (129) of the constraints. Hence, by the above-mentioned considerations we can principally tackle such kind of TSP problems with our BS method.

For sum-*maximization*-type linear assignment problems with side constraints, where e.g. c_{ij} is a profit (rather than a cost) and the ultimate goal is total profit maximization, we can proceed analogously, by employing (122) and (124) (instead of (121) and (123)).

Let us end this subsection with a comparison: suppose that we have a (sufficiently large) number n of *concrete* data observations $X_i = x_i$ $(i = 1, \ldots, n)$ from the unknown probability distribution $\mathbb{P}$ (in vector form), and from these we want to approximate/estimate the unknown distance $\inf_{\mathbb{Q}\in\boldsymbol{\Omega}} D_{\varphi_\gamma}(\mathbb{Q}, \mathbb{P})$ from a family of probability models (in vector form) $\boldsymbol{\Omega}$ (e.g. for model-adequacy evaluations, for goodness-of-fit testing purposes): by the above-mentioned Propositions 22 to 27 (and especially, by (84), (92), (99)) one can use

$$G\Big(-\frac{1}{n} \cdot \log \Pi_{x_1^n}\Big[\boldsymbol{\xi}_{n,\mathbf{x}}^{w\mathbf{W}} \in \boldsymbol{\Omega}\Big] \Big) \tag{136}$$

where $\Pi_{x_1^n}[\,\cdot\,] := \Pi[\,\cdot\,|\, X_1 = x_1, \ldots, X_n = x_n]$, $\mathbf{x} := (x_1, \ldots, x_n)$, and G (cf. (2)) is e.g. chosen as follows: $G(z) := -\frac{\widetilde{c}}{\gamma\cdot(\gamma-1)} \cdot \Big\{1 - \big(1 - \frac{\gamma}{\widetilde{c}} \cdot z\big)^{1-\gamma}\Big\}$ for the three cases $\gamma < 0$, $\gamma \in]0,1[$ and $\gamma \geq 2$, $G(z) := z$ for $\gamma = 0$ (reversed Kullback-Leibler divergence), and $G(z) := -\widetilde{c} \cdot \log(1 - \frac{1}{\widetilde{c}} \cdot z)$ for $\gamma = 1$ (Kullback-Leibler divergence). Notice that (136) contrasts to the alternative approximation (of $\inf_{\mathbb{Q}\in\boldsymbol{\Omega}} D_{\varphi_\gamma}(\mathbb{Q}, \mathbb{P})$) given by

$$\inf_{\mathbb{Q}\in\boldsymbol{\Omega}} D_{\varphi_\gamma}(\mathbb{Q}, \mathbb{P}_n^{emp,co}) \tag{137}$$

which is used in the context of "classical" statistical minimum distance estimation (MDE) with power divergences; in (137), we have employed $\mathbb{P}_n^{emp,co} = \frac{1}{n} \cdot \sum_{i=1}^n \delta_{x_i}$ to be the realization of the empirical distribution $\mathbb{P}_n^{emp} = \frac{1}{n} \cdot \sum_{i=1}^n \delta_{X_i}$.

Indeed, especially in complicated high-dimensional non-parametric or semi-parametric big-data contexts, we have substituted a quite difficult *optimization problem* (137) by a much easier solvable *counting problem* (136). The same holds analogously for Renyi distances/divergences, etc.

2) Construction principle for bounds of the minimum divergence in the general case :

Turning back to Theorem 14, we now consider the general case when the divergence $\varphi \in \Upsilon(]a, b[)$ is *not* of the power type (43). Recall from (34) the crucial terms (with $\mathbb{P} \in \mathbb{S}_{>0}$)

$$\inf_{m\neq 0} D_\varphi(m \cdot \boldsymbol{\Omega}, \mathbb{P}) := \inf_{m\neq 0} \inf_{\mathbb{Q}\in\boldsymbol{\Omega}} D_\varphi(m \cdot \mathbb{Q}, \mathbb{P}) = \inf_{\mathbb{Q}\in\boldsymbol{\Omega}} \inf_{m\neq 0} D_\varphi(m \cdot \mathbb{Q}, \mathbb{P}) \,<\, \infty \tag{138}$$

for all sets $\boldsymbol{\Omega}$ satisfying the regularity properties (7) and the convenient, more restrictive finiteness property

$$\inf_{\mathbb{Q}\in\boldsymbol{\Omega}} \inf_{k=1,\ldots,K} \frac{q_k}{p_k} \in dom(\varphi), \quad \sup_{\mathbb{Q}\in\boldsymbol{\Omega}} \sup_{k=1,\ldots,K} \frac{q_k}{p_k} \in dom(\varphi) \tag{139}$$

which implies (9); notice that $\inf_{k=1,\ldots,K} \frac{q_k}{p_k} \leq 1$, $\sup_{k=1,\ldots,K} \frac{q_k}{p_k} \geq 1$ with equalities if and only if $\mathbb{Q} = \mathbb{P}$. Since $\boldsymbol{\Omega} \neq \{\mathbb{P}\}$ (cf. the right-hand side of (7)), the double infimum (supremum) in (139) is strictly smaller (larger) than 1. In general, the inner minimization $\inf_{m\neq 0} D_\varphi(m \cdot \mathbb{Q}, \mathbb{P})$ in (138) can not be performed in explicit closed form, but e.g. in the specific case of power divergences (cf. (43), (44)) the optimization $\inf_{m\neq 0} D_{\widetilde{c}\cdot\varphi_\gamma}(m \cdot \mathbb{Q}, \mathbb{P})$ produces an explicit form, which in turn leads to a simple one-to-one correspondence between $D_{\widetilde{c}\cdot\varphi_\gamma}(\boldsymbol{\Omega}, \mathbb{P})$ and $\inf_{m\neq 0}\ D_{\widetilde{c}\cdot\varphi_\gamma}(m \cdot \boldsymbol{\Omega}, \mathbb{P})$ (cf. Lemma 16).

Under (7) and (139) it clearly holds

$$\inf_{m\neq 0} D_\varphi(m \cdot \boldsymbol{\Omega}, \mathbb{P}) \;\leq\; D_\varphi(\boldsymbol{\Omega}, \mathbb{P}) \;\leq\; D_\varphi(\mathbb{Q}, \mathbb{P}). \tag{140}$$

For transparency, we first investigate the (widely useable) subsetup where $dom(\varphi) =]0, \infty[$ (and thus, $int(dom(\varphi)) =]a, b[=]0, \infty[$) and $\boldsymbol{\Omega} \subset \mathbb{S}_{>0}^K$. Let us start with the lower bound $\inf_{m\neq 0} D_\varphi(m \cdot \boldsymbol{\Omega}, \mathbb{P})$. It can be proved that the minimizer in m is a well defined constant, which belongs to a compact set in $\mathbb{R}_{>0}$. To see this, let us first observe that, obviously, from (139) one can obtain

$$\left[\inf_{\mathbb{Q}\in\mathbf{\Omega}} \inf_{k=1,\ldots,K} \frac{m\cdot q_k}{p_k} \in dom(\varphi), \quad \sup_{\mathbb{Q}\in\mathbf{\Omega}} \sup_{k=1,\ldots,K} \frac{m\cdot q_k}{p_k} \in dom(\varphi) \right] \iff m \in]0,\infty[. \tag{141}$$

Moreover, for any fixed $\mathbb{Q}$ in $\mathbf{\Omega}$ there is a unique number $m = m(\mathbb{Q}) > 0$ which satisfies the first-order optimality condition for $m \in\,]0,\infty[$

$$\psi_{\mathbb{Q}}(m) := \frac{d}{dm} D_\varphi\left(m\cdot\mathbb{Q},\mathbb{P}\right) = \sum_{k=1}^{K} q_k \cdot \varphi'\left(\frac{m\cdot q_k}{p_k}\right) = 0 \tag{142}$$

and thus

$$D_\varphi\left(m(\mathbb{Q})\cdot\mathbb{Q},\mathbb{P}\right) = \inf_{m\neq 0} D_\varphi\left(m\cdot\mathbb{Q},\mathbb{P}\right); \tag{143}$$

indeed, the mapping $]0,\infty[\ \ni m \to D_\varphi\left(m\cdot\mathbb{Q},\mathbb{P}\right)$ is strictly convex and infinitely differentiable (which follows straightforwardly from (G5),(G6) in the below-mentioned Section V together with (C7ii), (C7iii) in Appendix D), and the strictly increasing function $\psi_{\mathbb{Q}}$ is such that $\psi_{\mathbb{Q}}(m)$ is strictly negative for all $m \in]0,1[$ for which $\sup_{k=1,\ldots,K} \frac{m\cdot q_k}{p_k} < 1$ whereas $\psi_{\mathbb{Q}}(m)$ is strictly positive for all $m > 1$ for which $\inf_{k=1,\ldots,K} \frac{m\cdot q_k}{p_k} > 1$ (recall the note right after (139) and $\varphi'(1) = 0$). Hence, for any $\mathbb{Q} \in \mathbf{\Omega}$ the unique zero $m(\mathbb{Q})$ of (142) (and hence, unique minimizer in (143)) is in the compact interval

$$\left[\frac{1}{\sup_{k=1,\ldots,K} \frac{q_k}{p_k}}, \frac{1}{\inf_{k=1,\ldots,K} \frac{q_k}{p_k}}\right] \subseteq \left[\frac{1}{\sup_{\mathbb{Q}\in\mathbf{\Omega}}\ \sup_{k=1,\ldots,K} \frac{q_k}{p_k}}, \frac{1}{\inf_{\mathbb{Q}\in\mathbf{\Omega}}\ \inf_{k=1,\ldots,K} \frac{q_k}{p_k}}\right] \subset \left]\frac{1}{b},\frac{1}{a}\right[=]0,\infty[.$$

When $\mathbf{\Omega}$ is closed in $\mathbb{S}^K$, then by continuity of the function $\mathbb{Q} \mapsto D_\varphi\left(m\left(\mathbb{Q}\right)\cdot\mathbb{Q},\mathbb{P}\right)$ there exists a $\mathbb{Q}^*$ in $\mathbf{\Omega}$ which achieves the infimum on $\mathbf{\Omega}$. When $\mathbf{\Omega}$ is not closed but satisfies (7), then the infimum exists anyway, possibly on the boundary $\partial\mathbf{\Omega}$. Anyhow, for such $\mathbb{Q}^*$ there holds

$$D_\varphi\left(m(\mathbb{Q}^*)\cdot\mathbb{Q}^*,\mathbb{P}\right) \leq D_\varphi\left(\mathbf{\Omega},\mathbb{P}\right) \leq D_\varphi\left(\mathbb{Q}^*,\mathbb{P}\right), \tag{144}$$

where we use the continuity of $\mathbb{Q} \mapsto D_\varphi\left(\mathbb{Q},\mathbb{P}\right)$ and (7) to obtain the last inequality above, even when $\mathbb{Q}^* \in \partial\mathbf{\Omega}$ and $\mathbb{Q}^* \notin \mathbf{\Omega}$.

That (144) provides sharp bounds can be seen through the case of power divergences. Indeed, for the latter one basically gets (cf. Appendix C) $m(\mathbb{Q}) = (1 + \frac{\gamma\cdot(\gamma-1)}{\widetilde{c}} \cdot D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbb{Q},\mathbb{P})\,)^{1/(1-\gamma)}$ and $D_{\varphi_\gamma}\left(m(\mathbb{Q})\cdot\mathbb{Q},\mathbb{P}\right) = \frac{\widetilde{c}}{\gamma}(1 - m(\mathbb{Q}))$ for the case $\gamma \in \mathbb{R}\backslash\{0,1\}$, respectively, $m(\mathbb{Q}) = \exp(-\frac{1}{\widetilde{c}} \cdot D_{\widetilde{c}\cdot\varphi_1}(\mathbb{Q},\mathbb{P}))$ and $D_{\widetilde{c}\cdot\varphi_1}\left(m(\mathbb{Q})\cdot\mathbb{Q},\mathbb{P}\right) = \frac{\widetilde{c}}{\gamma}(1 - m(\mathbb{Q}))$ for the case $\gamma = 1$, respectively, $m(\mathbb{Q}) = 1$ and $D_{\widetilde{c}\cdot\varphi_0}\left(m(\mathbb{Q})\cdot\mathbb{Q},\mathbb{P}\right) = D_{\widetilde{c}\cdot\varphi_0}\left(\mathbb{Q},\mathbb{P}\right)$ for the remaining case $\gamma = 0$. In all cases, $D_{\varphi_\gamma}\left(m(\mathbb{Q})\cdot\mathbb{Q},\mathbb{P}\right)$ is an increasing function of $D_{\varphi_\gamma}\left(\mathbb{Q},\mathbb{P}\right)$ and therefore, $\mathbb{Q}^* \in \arg\inf_{\mathbb{Q}\in\mathbf{\Omega}} D_\varphi\left(m(\mathbb{Q})\cdot\mathbb{Q},\mathbb{P}\right)$ also satisfies $\mathbb{Q}^* \in \arg\inf_{\mathbb{Q}\in\mathbf{\Omega}} D_\varphi\left(\mathbb{Q},\mathbb{P}\right)$. Hence, the right-hand side and the left-hand side of (144) coincide. Now due to (6), the LHS of (144) can be estimated since by Theorem 14 for each $\mathbb{P} \in \mathbb{S}^K_{>0}$ the divergence $\inf_{m\neq 0} D_\varphi\left(m\cdot\mathbb{Q},\mathbb{P}\right)$ is BS-minimizable on sets $\mathbf{\Omega} \subset \mathbb{S}^K$. We shall propose in Section V an algorithm to handle the estimation of the RHS of (144), whenever $\mathbb{P}$ is known (as in Remark 13(v)) or when $\mathbb{P}$ is approximated by the empirical distribution of the data set $(X_1,..,X_n)$. Also note that (144) holds also for $\mathbf{\Omega}$ substituted by $A\cdot\mathbf{\Omega}$ for any $A \neq 0$.

Other cases of interest include when $dom\left(\varphi\right)$ is not $]0,\infty[$. We list two cases which extend the above discussion. Firstly, consider φ with $dom\left(\varphi\right) = [0,\infty[$. Then — since $\varphi'(0) = -\infty$ in order that (6) should hold (see (G10ii) in Section V below) — we may extend (144) to cases when $\mathbf{\Omega} \subset \mathbb{S}^K$ instead of $\mathbf{\Omega} \subset \mathbb{S}^K_{>0}$, hence allowing for possible null entries in $\mathbf{\Omega}$. When $dom\left(\varphi\right) =]a,b[$ for some $a < 0$, then clearly the same argument leading to (144) holds; this case is of interest, for instance, when extending a statistical model to signed measures (see e.g. Broniatowski et al. [63] for the important task of testing the number of components in a parametric probability mixture model).

Example 28: Consider the (non-probability version of the) Jensen-Shannon divergence defined by

$$J(\mathbf{Q},\mathbf{P}) := I(\mathbf{Q},(\mathbf{Q}+\mathbf{P})/2) + I(\mathbf{P},(\mathbf{Q}+\mathbf{P})/2), \qquad \mathbf{P},\mathbf{Q} \in \mathbb{R}^K_{\geq 0},$$

where $I(\mathbf{Q},\mathbf{P})$ denotes the modified Kullback-Leibler information between $\mathbf{Q}$ and $\mathbf{P}$ (cf. (46) with $\mathbf{P}$ instead of $\mathbb{P}$). In (188) and (189) of Example 43 below, we shall show that $J(\mathbf{Q},\mathbf{P}) = D_{\varphi_{snKL}}(\mathbf{Q},\mathbf{P})$ with (basically) divergence generator $\varphi_{snKL}(t) :=\ t\cdot\log t + (t+1)\cdot\log\left(\frac{2}{t+1}\right)$ for $t > 0$. It is known that J^2 is a metric. We explore the sharpness of the bounds for $J(\mathbf{\Omega},\mathbb{P})$ as defined in (144). For this, we consider a given probability distribution $\mathbb{P}$ on $\mathcal{Y}$ with strictly positive entries; the set $\mathbf{\Omega}$ consists of all probability distributions $\mathbb{Q}$ on $\mathcal{Y}$ whose total variation distance $V(\mathbb{Q},\mathbb{P}) := \sum_{k=1}^{K} |q_k - p_k|$ [28] to $\mathbb{P}$ lies between v and $v+h$ for $v > 0$ and small h and which also satisfies

$$\sup\left(\sup_{k=1,\ldots,K} \frac{p_k}{q_k}, \sup_{k=1,\ldots,K} \frac{q_k}{p_k}\right) \leq L$$

[28] notice that $V(\mathbb{Q},\mathbb{P})$ always takes values in the interval $[0,2[$

for some strictly positive finite L. This set $\boldsymbol{\Omega}$ defines a class of distributions $\mathbb{Q}$ away from $\mathbb{P}$ still keeping some regularity w.r.t. $\mathbb{P}$. Also, $\boldsymbol{\Omega}$ satisfies (7). We will prove that the bounds in (120) are sharp in this case. Notice that $J(m\cdot\mathbb{Q},\mathbb{P})=\infty$ for $m<0$ and hence $\inf_{\mathbb{Q}\in\boldsymbol{\Omega}}\inf_{m\neq 0}J(m\cdot\mathbb{Q},\mathbb{P})=\inf_{\mathbb{Q}\in\boldsymbol{\Omega}}\inf_{m>0}J(m\cdot\mathbb{Q},\mathbb{P})$. We first provide a lower bound for the latter. It holds for all $m>0$ and $\mathbb{Q}$ in $\boldsymbol{\Omega}$

$$J(m\mathbb{Q},\mathbb{P})=(m+1)\cdot\log(2)-(m+1)\cdot\log(m+1)+m\log m+I^{\alpha}(\mathbb{P},\mathbb{Q})+m\cdot I^{1-\alpha}(\mathbb{Q},\mathbb{P})$$

where $\alpha:=1/(m+1)$ and $I^{\alpha}(\mathbb{P},\mathbb{Q})$ is the $\alpha-$skewed Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{Q}$ defined through

$$I^{\alpha}(\mathbb{P},\mathbb{Q}):=I(\mathbb{P},\alpha\mathbb{P}+(1-\alpha)\mathbb{Q}).$$

By Inequality (27) in Yamano [410]

$$I^{\alpha}(\mathbb{P},\mathbb{Q})\geq-\log\left(1-\frac{\alpha^{2}}{4}V(\mathbb{Q},\mathbb{P})^{2}\right).$$

Since $(m+1)\cdot\log(2)-(m+1)\cdot\log(m+1)+m\cdot\log m$ is non-negative for all $m>0$ and takes its minimal value 0 for $m=1$, we obtain

$$\inf_{m>0}J(m\mathbb{Q},\mathbb{P})\geq\inf_{m>0}K(m)$$

where

$$K(m):=-\log\left(1-\frac{1}{4\left(m+1\right)^{2}}\cdot V(\mathbb{Q},\mathbb{P})^{2}\right)-m\cdot\log\left(1-\frac{m^{2}}{4\left(m+1\right)^{2}}\cdot V(\mathbb{Q},\mathbb{P})^{2}\right).$$

Since $-\log(1-x)\geq x$ for all $x<1$ and both $\frac{1}{4(m+1)^{2}}\cdot V(\mathbb{Q},\mathbb{P})^{2}$ and $\frac{m^{2}}{4(m+1)^{2}}\cdot V(\mathbb{Q},\mathbb{P})^{2}$ are less than 1, it follows that

$$K(m)\geq\frac{V(\mathbb{Q},\mathbb{P})^{2}}{4}\cdot\frac{m^{3}+1}{\left(m+1\right)^{2}}$$

where the right-hand side attains its minimal value on $]0,\infty[$ at $m^{+}=\sqrt{3}-1\approx 0.73$. Hence, we obtain

$$\inf_{m>0}J(m\cdot Q,P)\geq\frac{V(\mathbb{Q},\mathbb{P})^{2}}{4}\cdot(2\sqrt{3}-3)>0.116\,v^{2}$$

Now by (19) in Yamano [410], for any $\mathbb{Q}$

$$J(\mathbb{Q},\mathbb{P})\leq\frac{1}{4}\underline{J}(\mathbb{Q},\mathbb{P})$$

where $\underline{J}(\mathbb{Q},\mathbb{P}):=I(\mathbb{Q},\mathbb{P})+I(\mathbb{P},\mathbb{Q})$ is the Jensen divergence (also called symmetrized Kullback-Leibler divergence) between $\mathbb{Q}$ and $\mathbb{P}$. Since (see Dragomir [116])

$$I(\mathbb{P},\mathbb{Q})\leq\sum_{k=1}^{K}\sqrt{\frac{p_{k}}{q_{k}}}\cdot\left|q_{k}-q_{k}\right|,$$

it follows that

$$J(\mathbb{Q}^{*},\mathbb{P})\leq\frac{1}{2}\sqrt{L}\cdot V(\mathbb{Q}^{*},\mathbb{P})$$

which provides

$$0.116\,v^{2}\;\leq\;\inf_{m>0}J(m\cdot\mathbb{Q}^{*},\mathbb{P})=J(m(\mathbb{Q}^{*})\cdot\mathbb{Q}^{*},\mathbb{P})\;\leq\;J(\boldsymbol{\Omega},\mathbb{P})\;\leq\;J(\mathbb{Q}^{*},\mathbb{P})\;\leq\;\frac{1}{2}\sqrt{L}\cdot(v+h).$$

For small v the difference between the RHS and the LHS in the above display is $cst\cdot v+o(v)+\frac{1}{2}\sqrt{L}\cdot h$ which proves that the bounds are sharp locally, with non-trivial lower bound. Other upper bounds can be adapted to sets $\boldsymbol{\Omega}$ defined through tighter conditions on $\sup_{\mathbb{Q}\in\boldsymbol{\Omega}}\sup_{k=1,\ldots,K}\frac{p_{k}}{q_{k}}$ and $\sup_{\mathbb{Q}\in\boldsymbol{\Omega}}\sup_{k=1,\ldots,K}\frac{q_{k}}{p_{k}}$ (of e.g. Dragomir [116]).

3) On the difference between minimization problems of deterministic nature and risk minimization:

In the context of minimization of the functional $\Phi_{\mathbf{P}}(\mathbf{Q})$ over $\mathbf{\Omega}\subset\mathbb{R}^{K}$ for *known* vector $\mathbf{P}$, due to Theorem 10 our bare simulation approach allows for the *approximate solution* for any divergence D_{φ} satisfying the basic representation (6). Indeed, any proxy of $\Pi\left[\boldsymbol{\xi}_{n}^{\widetilde{\mathbf{W}}}\in\Omega/M_{\mathbf{P}}\right]$ yields a corresponding proxy for $\Phi_{\mathbf{P}}[\mathbf{\Omega}]$. This paves the way to the solution of numerous optimization problems, where the divergence D_{φ} is specifically suited to the problem at hand.

In the statistical context, when the *probability distribution* (in its vector-form) $\mathbb{P}$ is *unknown* up to some indirect information provided by sampling or by any mean providing a sequence $(X_{i})_{i\in\mathbb{N}}$ satisfying condition (26) (resp. (30)), Theorem 14 adds a complementary step of complexity; indeed, the *estimation* of $\Phi_{\mathbf{P}}(\boldsymbol{\Omega})$ over $\boldsymbol{\Omega}\subset\mathbb{S}^{K}$ results as its subproduct through the optimization upon m which can be performed explicitly only in a number of specific divergences D_{φ}, e.g. the power divergences $D_{\varphi_{\gamma}}$, and which carries over also to their monotone transformations such as e.g. the Renyi divergences. It is of relevance to mention that — as already indicated above — these divergences cover a *very broad range* of statistical criteria, indeed most of them, from the (maximum-likelihood estimation connected) likelihood divergence ($\gamma = 0$) to the Kullback-Leibler one ($\gamma = 1$), the two standard Chi-square distances ($\gamma = 2, \ \ \gamma = -1$), the Hellinger distance ($\gamma = 1/2$), etc.; in contrast with deterministic minimization problems, the choice of a statistical criterion (or risk function) is not imposed by the modelling of the problem at hand, but is dictated by the need for sharp measures of fit. Other divergences are more difficult to handle and our general results in Section IV-C2 should still prove some usefulness, since estimation of upper and lower bounds for risk is of common use.

As a "preparatory" remark, recall first that each probability distribution (probability measure) $\mathbb{P}$ on $\mathcal{Y} = \{d_1, \ldots, d_K\}$ has been uniquely identified with the vector $\mathbb{IP} := (p_1, \ldots, p_K) \in \mathbb{S}^K$ of the corresponding probability masses (frequencies) $p_k = \mathbb{P}[\{d_k\}]$ via $\mathbb{P}[A] = \sum_{k=1}^K p_k \cdot 1_A(d_k)$ for each $A \subset \mathcal{Y}$; from this, we have measured the distance/divergence between two probability distributions $\mathbb{P}, \mathbb{Q}$ through the distance/divergence between their frequency vectors $\mathbb{IP}, \mathbb{Q}$:

$$D_\varphi(\mathbb{Q}, \mathbb{P}) := D_\varphi(\mathbb{Q}, \mathbb{IP}) \qquad \text{(cf. (28))}.$$

However, it has been noted in Kißlinger & Stummer [190] in a context of even more general divergences $D(\mathbf{Q}, \mathbf{P})$ between vectors $\mathbf{P}, \mathbf{Q}$ that — alternatively — the latter two may consist of components $p_k = \mathbb{P}[\{E_k\}]$, $q_k = \mathbb{Q}[\{E_k\}]$ which are probabilities of only some *selected* (e.g. increasing) events $(E_k)_{k \in \{1,\ldots,M\}}$ of *application-based concrete* interest (within *not necessarily discrete* probability models). Of course, we can apply our BS method to such a vector context.

As other alternatives, in the following we deal with divergences between *non-probabilistic* uncertainty quantifications.

D. Minimization problems with fuzzy sets

Our BS framework also covers the — imprecise/inexact/vague information describing — *fuzzy sets* (cf. Zadeh [417]) and optimization problems on divergences between those. Indeed, let $\mathcal{Y} = \{d_1, \ldots, d_K\}$ be a finite set (called the *universe (of discourse)*), $A \subset \mathcal{Y}$ and $M^A : \mathcal{Y} \mapsto [0, 1]$ be a corresponding *membership function*, where $M^A(d_k)$ represents the degree/grade of membership of the element d_k to the set A; accordingly, the object $A^* := \{\langle x, M^A(x)\rangle | x \in \mathcal{Y}\}$ is called a *fuzzy set in* $\mathcal{Y}$ (or fuzzy subset of $\mathcal{Y}$). Moreover, if $A \subset \mathcal{Y}$ and $B \subset \mathcal{Y}$ are unequal, then the corresponding membership functions M^A and M^B should be unequal. Furthermore, we model the *vector of membership degrees to* A by $\mathbf{P}^A := \big(p_k^A\big)_{k=1,\ldots,K} := \big(M^A(d_k)\big)_{k=1,\ldots,K}$ which satisfies the *key constraint* $0 \leq p_k^A \leq 1$ for all $k \in \{1, \ldots, K\}$ and, consequently, the *aggregated key constraint* $0 \leq \sum_{k=1}^K p_k^A \leq K$ (as a side remark, $\sum_{k=1}^K M^A(d_k)$ is called *power of the fuzzy set* A^*). For divergence generators φ in $\Upsilon(]a, b[)$ (resp. $\widetilde{\Upsilon}(]a, b[)$) with — say — $0 \leq a < 1 < b$ and for two sets $A, B \subset \mathcal{Y}$ we can apply (4) to the corresponding membership functions and define the φ*−divergence* $D_\varphi(B^*, A^*)$ *between the fuzzy sets* B^* *and* A^* (on the same universe $\mathcal{Y}$) as

$$D_\varphi(B^*, A^*) := D_\varphi(\mathbf{P}^B, \mathbf{P}^A) = \sum_{k=1}^K p_k^A \cdot \varphi\left(\frac{p_k^B}{p_k^A}\right) = \sum_{k=1}^K M^A(d_k) \cdot \varphi\left(\frac{M^B(d_k)}{M^A(d_k)}\right) \geq 0 \tag{145}$$

(depending on φ, zero degree values may have to be excluded for finiteness). For instance, we can take $\varphi(t) := \varphi_1(t) := t \cdot \log t + 1 - t \in [0, \infty[$ for $t \in [0, \infty[$ (cf. (43)) to end up with a generalized Kullback-Leibler divergence (generalized relative entropy) between B^* and A^*; this contrasts the choice $\varphi(t) := \breve{\varphi}(t) := t \cdot \log t \in [-\frac{1}{e}, \infty[$ of Bhandari & Pal [47] for which $D_{\breve{\varphi}}(B^*, A^*)$ (which they call *fuzzy expected information for discrimination in favor of* B *against* A) may become negative (cf. Stummer & Vajda [349] in a more general context). In terms of (145), as a special case of the above-mentioned BS concepts, we can tackle optimization problems of the type

$$\inf_{B^* \in \Omega^*} D_\varphi(B^*, A^*) := \inf_{\mathbf{P}^B \in \mathbf{\Omega}} D_\varphi(\mathbf{P}^B, \mathbf{P}^A)$$

where Ω^* is a collection of fuzzy sets (on the same universe $\mathcal{Y}$) whose membership-degree vectors form the set $\mathbf{\Omega}$ satisfying (7) and (9). Because of the inequality-type key constraint

$$0 \leq M^B(d_k) \leq 1 \quad \text{for all } k \in \{1, \ldots, K\}$$

which is contained in $\mathbf{\Omega}$ and which implies $0 \leq \sum_{k=1}^K p_k^B \leq K$, Theorem 10 and its consequences and derived examples will apply correspondingly — unless there is a more restrictive constraint which violates (7) such as e.g. $\sum_{k=1}^K p_k^B = C$ with $C \leq K$ for which Theorem 14 (and its consequences and derived examples) can be employed.

The above-mentioned considerations can be extended to the recent concept of ν*−rung orthopair fuzzy sets* (cf. Yager [406]) and divergences between those. Indeed, for $A \subset \mathcal{Y}$, besides a membership function $M^A : \mathcal{Y} \mapsto [0, 1]$ one additionally models a *non-membership function* $N^A : \mathcal{Y} \mapsto [0, 1]$, where $N^A(d_k)$ represents the degree/grade of non-membership of the element

d_k to the set A. Moreover, if $A \subset \mathcal{Y}$ and $B \subset \mathcal{Y}$ are unequal, then the corresponding non-membership functions N^A and N^B should be unequal. For fixed $\nu \in [1,\infty[$, the *key constraint*

$$0 \leq \big(M^A(d_k)\big)^\nu + \big(N^A(d_k)\big)^\nu \leq 1 \quad \text{for all } k \in \{1,\ldots,K\} \tag{146}$$

is required to be satisfied, too. Accordingly, the object $A^{**} := \{\langle x, M^A(x), N^A(x)\rangle | x \in \mathcal{Y}\}$ is called a *$\nu-$rung orthopair fuzzy set in $\mathcal{Y}$ (or ... subset of $\mathcal{Y}$)*. The object A^{**} is called *intuitionistic fuzzy set in $\mathcal{Y}$* (cf. Atanassov [20]) in case of $\nu = 1$, and *Pythagorean fuzzy set in $\mathcal{Y}$* (cf. Yager [405], [407]) in case of $\nu = 2$. For the choice $\nu = 1$ together with $N^A(x) := 1 - M^A(x)$, the object A^{**} can be regarded as an extended representation of the fuzzy set A^* in $\mathcal{Y}$. As is well known, there is a vast amount of recent literature on applications of fuzzy sets; for the sake of brevity, let us exemplarily mention the survey of Yanase & Triantaphyllou [411] on some recent uses in computer-aided medical diagnosis. For any $\nu-$rung orthopair fuzzy set A^{**} in $\mathcal{Y}$, we model the corresponding *vector of concatenated membership and non-membership degrees to A* by $\mathbf{P}^A := \big(p_k^A\big)_{k=1,\ldots,2K} := \big(M^A(d_1),\ldots M^A(d_K), N^A(d_1),\ldots N^A(d_K)\big)$ which due to (146) satisfies the *aggregated key constraint*

$$0 \leq \sum_{k=1}^{2K} \big(p_k^A\big)^\nu \leq K;$$

in other words, $\mathbf{P}^A$ lies (within the $2K-$dimensional Euclidean space) in the intersection of the first/positive orthant with the $\nu-$norm ball centered at the origin and with radius $K^{1/\nu}$. Analogously to (145), we can define the *$\varphi-$divergence $D_\varphi(B^{**}, A^{**})$ between the $\nu-$rung orthopair fuzzy sets B^{**} and A^{**}* (on the same universe $\mathcal{Y}$) as

$$D_\varphi(B^{**}, A^{**}) := D_\varphi(\mathbf{P}^B, \mathbf{P}^A) = \sum_{k=1}^{2K} p_k^A \cdot \varphi\left(\frac{p_k^B}{p_k^A}\right) = \sum_{k=1}^{K} \left\{ M^A(d_k)\cdot\varphi\left(\frac{M^B(d_k)}{M^A(d_k)}\right) + N^A(d_k)\cdot\varphi\left(\frac{N^B(d_k)}{N^A(d_k)}\right)\right\} \geq 0 \tag{147}$$

respectively as its variant

$$\begin{aligned} &D_\varphi^{var}(B^{**}, A^{**}) := D_\varphi\big(\big(\mathbf{P}^B\big)^\nu, \big(\mathbf{P}^A\big)^\nu\big) \\ &= \sum_{k=1}^{2K} \big(p_k^A\big)^\nu \cdot \varphi\left(\frac{\big(p_k^B\big)^\nu}{\big(p_k^A\big)^\nu}\right) = \sum_{k=1}^{K} \left\{ \big(M^A(d_k)\big)^\nu \cdot \varphi\left(\frac{\big(M^B(d_k)\big)^\nu}{\big(M^A(d_k)\big)^\nu}\right) + \big(N^A(d_k)\big)^\nu \cdot \varphi\left(\frac{\big(N^B(d_k)\big)^\nu}{\big(N^A(d_k)\big)^\nu}\right)\right\} \geq 0. \end{aligned} \tag{148}$$

For the special choice $\nu = 1$, $N^A(x) := 1 - M^A(x)$ and $\varphi(t) := \varphi_1(t) := t\cdot\log t + 1 - t \in [0,\infty[$ for $t \in [0,\infty[$ (cf. (43)), one can straightforwardly show that the outcoming divergence $D_{\varphi_1}(B^{**}, A^{**})$ coincides with $D_{\breve{\varphi}}(B^{**}, A^{**})$ where $\breve{\varphi}(t) := t \cdot \log t$; the latter divergence was used e.g. in Bhandari & Pal [47] under the name *average fuzzy information for discrimination in favor of B against A*). Moreover, the special choice $\nu = 1$ and $\varphi(t) := \varphi_{snKL,1}(t)$ (cf. (188)) leads to the *Jensen-Shannon divergence between B^{**} and A^{**}* given by $D_{\varphi_{snKL,1}}(B^{**}, A^{**}) := D_{\varphi_{snKL,1}}(\mathbf{P}^B, \mathbf{P}^A)$; from (147) and (189) one can see that this coincides with the *symmetric information measure between B^{**} and A^{**}* of Vlachos & Sergiadis [385].

In terms of the divergences (147) and (148), we can tackle — as a special case of the above-mentioned BS concepts — optimization problems of the type

$$\begin{aligned} &\inf_{B^{**}\in\Omega^{**}} D_\varphi(B^{**}, A^{**}) := \inf_{\mathbf{P}^B\in\mathbf{\Omega}} D_\varphi(\mathbf{P}^B, \mathbf{P}^A) \qquad \text{respectively} \\ &\inf_{B^{**}\in\Omega^{**}} D_\varphi^{var}(B^{**}, A^{**}) := \inf_{\mathbf{P}^B\in\mathbf{\Omega}} D_\varphi\big(\big(\mathbf{P}^B\big)^\nu, \big(\mathbf{P}^A\big)^\nu\big), \end{aligned}$$

where Ω^{**} is a collection of $\nu-$rung orthopair fuzzy sets whose concatenated-membership-nonmembership-degree vectors form the set $\mathbf{\Omega}$ satisfying (7) and (9) as well as (146) for B in place of A. Because of the latter, Theorem 10 and its consequences and derived examples will apply correspondingly — unless there is a more restrictive constraint which violates (7) such as e.g. $\sum_{k=1}^{2K} p_k^B = C$ with $C \leq K$ for which Theorem 14 (and its consequences and derived examples) can be employed; such a situation appears e.g. in the above-mentioned case $\nu = 1$ together with $N^A(x) := 1 - M^A(x)$ which leads to $C = K$.

For $\nu-$rung orthopair fuzzy sets A^{**} in $\mathcal{Y}$, we can also further "flexibilize" our divergences by additionally incorporating the *hesitancy degree* of the element d_k to A which is defined as

$$H^A(d_k) := \Big(1 - \big(M^A(d_k)\big)^\nu - \big(N^A(d_k)\big)^\nu\Big)^{1/\nu} \in [0,1]$$

(cf. Yager [406]) and which implies the *key constraint*

$$\big(H^A(d_k)\big)^\nu + \big(M^A(d_k)\big)^\nu + \big(N^A(d_k)\big)^\nu = 1 \quad \text{for all } k \in \{1,\ldots,K\}. \tag{149}$$

Accordingly, the object $A^{***} := \{\langle x, M^A(x), N^A(x), H^A(x)\rangle | x \in \mathcal{Y}\}$ can be regarded as an extended representation of the $\nu-$rung orthopair fuzzy set A^{**} in $\mathcal{Y}$. For A^{***}, we model the corresponding *vector of concatenated membership, non-membership and hesitancy degrees to* A by

$$\mathbf{P}^A := \big(p_k^A\big)_{k=1,\ldots,3K} := \big(M^A(d_1), \ldots M^A(d_K), N^A(d_1), \ldots N^A(d_K), H^A(d_1), \ldots H^A(d_K)\big)$$

which due to (149) satisfies the *aggregated key constraint*

$$\sum_{k=1}^{3K} \big(p_k^A\big)^{\nu} = K; \tag{150}$$

in other words, $\mathbf{P}^A$ lies (within the $3K-$dimensional Euclidean space) in the intersection of the first/positive orthant with the $\nu-$norm sphere centered at the origin and with radius $K^{1/\nu}$. Analogously to (147) and (148), we can define the *$\varphi-$divergence $D_{\varphi}(B^{***}, A^{***})$ between the extended-representation-type $\nu-$rung orthopair fuzzy sets B^{***} and A^{***}* (on the same universe $\mathcal{Y}$) as

$$\begin{aligned}
&D_{\varphi}(B^{***}, A^{***}) := D_{\varphi}(\mathbf{P}^B, \mathbf{P}^A) \\
&= \sum_{k=1}^{3K} p_k^A \cdot \varphi\left(\frac{p_k^B}{p_k^A}\right) = \sum_{k=1}^{K} \left\{ M^A(d_k) \cdot \varphi\left(\frac{M^B(d_k)}{M^A(d_k)}\right) + N^A(d_k) \cdot \varphi\left(\frac{N^B(d_k)}{N^A(d_k)}\right) + H^A(d_k) \cdot \varphi\left(\frac{H^B(d_k)}{H^A(d_k)}\right) \right\} \geq 0
\end{aligned}$$

respectively as its variant

$$\begin{aligned}
&D_{\varphi}^{var}(B^{***}, A^{***}) := D_{\varphi}(\big(\mathbf{P}^B\big)^{\nu}, \big(\mathbf{P}^A\big)^{\nu}) = \sum_{k=1}^{3K} \big(p_k^A\big)^{\nu} \cdot \varphi\left(\frac{\big(p_k^B\big)^{\nu}}{\big(p_k^A\big)^{\nu}}\right) \\
&= \sum_{k=1}^{K} \left\{ \big(M^A(d_k)\big)^{\nu} \cdot \varphi\left(\frac{\big(M^B(d_k)\big)^{\nu}}{\big(M^A(d_k)\big)^{\nu}}\right) + \big(N^A(d_k)\big)^{\nu} \cdot \varphi\left(\frac{\big(N^B(d_k)\big)^{\nu}}{\big(N^A(d_k)\big)^{\nu}}\right) + \big(H^A(d_k)\big)^{\nu} \cdot \varphi\left(\frac{\big(H^B(d_k)\big)^{\nu}}{\big(H^A(d_k)\big)^{\nu}}\right) \right\} \geq 0.
\end{aligned} \tag{151}$$

For instance, by taking the special choice $\nu = 2$ and $\varphi(t) := \varphi_{snKL,1}(t)$ (cf. (188)) in (151), we arrive at the *Jensen-Shannon divergence between B^{***} and A^{***}* of the form $D_{\varphi_{snKL,1}}^{var}(B^{***}, A^{***}) := D_{\varphi_{snKL,1}}^{var}(\mathbf{P}^B, \mathbf{P}^A)$ which — by the virtue of (151) and (189) — coincides with the *(squared) Pythagorean-fuzzy-set Jensen-Shannon divergence measure between B^{***} and A^{***}* of Xiao & Ding [399]. To continue with the general context, as a particular application of the above-mentioned BS concepts, we can tackle general optimization problems of the type

$$\begin{aligned}
&\inf_{B^{***} \in \Omega^{***}} D_{\varphi}(B^{***}, A^{***}) := \inf_{\mathbf{P}^B \in \mathbf{\Omega}} D_{\varphi}(\mathbf{P}^B, \mathbf{P}^A) \qquad \text{respectively} \\
&\inf_{B^{***} \in \Omega^{***}} D_{\varphi}^{var}(B^{***}, A^{***}) := \inf_{\mathbf{P}^B \in \mathbf{\Omega}} D_{\varphi}(\big(\mathbf{P}^B\big)^{\nu}, \big(\mathbf{P}^A\big)^{\nu})
\end{aligned}$$

where Ω^{***} is a collection of extended-representation-type $\nu-$rung orthopair fuzzy sets whose concatenated-membership-nonmembership-hesitancy-degree vectors form the set $\mathbf{\Omega}$ satisfying (7) and (9) as well as (149) for B in place of A. Because of the latter and the implied aggregated key constraint (150) for B in place of A, Theorem 14 (and its consequences and derived examples) can be employed.

Of course, we can also correspondingly adapt the *transformations* of $\varphi-$divergences and entropy-type special cases given in the sections above and below, to (classical respectively $\nu-$rung othopair) fuzzy sets, and apply our BS method for this. For the sake of brevity, we only give a short example, namely the *$\gamma-$order Renyi divergence between $\nu-$rung othopair fuzzy sets* which we define by (cf. (69))

$$\begin{aligned}
&R_{\gamma}^{var}(B^{***}, A^{***}) := \frac{1}{\gamma \cdot (\gamma - 1)} \cdot \left[\log\left(\sum_{k=1}^{3K} \Big(\big(p_k^B\big)^{\nu}\Big)^{\gamma} \cdot \Big(\big(p_k^A\big)^{\nu}\Big)^{1-\gamma}\right) - \log(K)\right] \\
&= \frac{1}{\gamma \cdot (\gamma - 1)} \cdot \Bigg[\log\Bigg(\sum_{k=1}^{K} \Bigg\{ \Big(\big(M^B(d_k)\big)^{\nu}\Big)^{\gamma} \cdot \Big(\big(M^A(d_k)\big)^{\nu}\Big)^{1-\gamma} + \Big(\big(N^B(d_k)\big)^{\nu}\Big)^{\gamma} \cdot \Big(\big(N^A(d_k)\big)^{\nu}\Big)^{1-\gamma} \\
&+ \Big(\big(H^B(d_k)\big)^{\nu}\Big)^{\gamma} \cdot \Big(\big(H^A(d_k)\big)^{\nu}\Big)^{1-\gamma} \Bigg\}\Bigg) - \log(K)\Bigg] \geq 0; \qquad \gamma \in]-\infty, 0[\cup]0, 1[\cup [1, \infty[,
\end{aligned} \tag{152}$$

depending on γ, zero degree values may have to be excluded for finiteness. As a side remark, let us mention that our divergence (152) contrasts to the recent (first) divergence of Verma [382] who basically uses a different scaling, the product $\prod_{k=1}^{K}$ instead of the sum $\sum_{k=1}^{K}$, as well as $\frac{(p_k^A)^{\nu} + (p_k^B)^{\nu}}{2}$ instead of $\big(p_k^A\big)^{\nu}$. By equivalently rewriting (152), we can use (88) with $A = 1$ together with the Propositions 22, 24, 26 and 27 to tackle for $\gamma \in]-\infty, 0[\cup]0, 1[\cup [2, \infty[$ the minimization problem

$$\inf_{B^{***} \in \Omega^{***}} R_{\gamma}^{var,nor}(B^{***}, A^{***}) = \inf_{\mathbb{P}^B \in \mathbf{\Omega}} R_{\gamma}(\mathbb{P}^B, \mathbb{P}^A)$$

with artificially generated probability vectors (cf. (150))

$$\mathbb{P}^C := \left(\frac{\big(M^C(d_1)\big)^\nu}{K}, \ldots, \frac{\big(M^C(d_K)\big)^\nu}{K}, \frac{\big(N^C(d_1)\big)^\nu}{K}, \ldots, \frac{\big(N^C(d_K)\big)^\nu}{K}, \frac{\big(H^C(d_1)\big)^\nu}{K}, \ldots, \frac{\big(H^C(d_K)\big)^\nu}{K} \right), \quad C = A, B;$$

here, Ω^{***} is a collection of extended-representation-type $\nu-$rung orthopair fuzzy sets B^{***} whose corresponding normalized concatenated-membership-nonmembership-hesitancy-degree vectors $\mathbb{P}^B$ form the set $\boldsymbol{\Omega}$ satisfying (7) (in the relative topology).

The above-mentioned considerations can be carried over to (classical, intuitionistic, Pythagorean, $\nu-$rung orthopair) $L-$fuzzy sets, where the range of the membership functions, non-membership functions and hesitancy functions is an appropriately chosen lattice L (rather than $L = [0, 1]$); for the sake of brevity, this is omitted here.

E. Minimization problems with basic belief assignments

Our BS framework also covers — imprecise/inexact/vague information describing — *basic belief assignments* from Dempster-Shafer evidence theory (cf. [109], [327]) and optimization problems on divergences between those. Indeed, let $\mathcal{Y} = \{d_1, \ldots, d_K\}$ be a finite set (called the *frame of discernment*) of mutally exclusive and collectively exhaustive events d_k. The corresponding power set of $\mathcal{Y}$ is denoted by $2^{\mathcal{Y}}$ and has 2^K elements; we enumerate this by $2^{\mathcal{Y}} := \{A_1, \ldots, A_{2^K}\}$ where for convenience we set $A_1 := \emptyset$. A mapping $M : 2^{\mathcal{Y}} \mapsto [0, 1]$ is called a *basic belief assignment (BBA)* [29] if it satisfies the two conditions

$$M(\emptyset) = 0 \qquad \text{and} \qquad \sum_{A \in 2^{\mathcal{Y}}} M(A) \, = \, 1. \tag{153}$$

Here, the belief mass $M(A)$ reflects e.g. the trust degree of evidence to proposition $A \in 2^{\mathcal{Y}}$. From this, one can build the belief function $Bel : 2^{\mathcal{Y}} \mapsto [0, 1]$ by $Bel(A) := \sum_{B: B \subseteq A} M(B)$ and the plausability function $Pl : 2^{\mathcal{Y}} \mapsto [0, 1]$ by $Pl(A) := \sum_{B: B \cap A \neq \emptyset} M(B)$. Moreover, we model the 2^K-dimensional *vector of* ($M-$*based) BBA values* (vector of ($M-$based) belief masses) by $\mathbf{P}^M := \big(p_k^M\big)_{k=1,\ldots,2^K} := (M(A_k))_{k=1,\ldots,2^K}$ which satisfies the *key constraint* $0 \leq p_k^M \leq 1$ for all $k \in \{1, \ldots, 2^K\}$ and, by virtue of (153), the *aggregated key constraint* $\sum_{k=1}^{2^K} p_k^M = 1$. Hence, $\mathbf{P}^M$ lies formally in the 2^K-dimensional simplex $\mathbb{S}^{2^K}$ (but generally not in the corresponding probability-vector-describing $\mathbb{S}^K$).

For divergence generators φ in $\Upsilon(]a, b[)$ with — say — $0 \leq a < 1 < b$ and for two BBAs M_1, M_2 on the same frame of discernment $\mathcal{Y}$, we can apply (4) to the corresponding vectors of BBA-values and define the *$\varphi-$divergence $D_\varphi(M_2, M_1)$ between the BBAs M_2 and M_1* (in short, *Belief$-\varphi-$divergence*) as

$$D_\varphi(M_2, M_1) := D_\varphi(\mathbf{P}^{M_2}, \mathbf{P}^{M_1}) = \sum_{k=1}^{2^K} p_k^{M_1} \cdot \varphi\left(\frac{p_k^{M_2}}{p_k^{M_1}}\right) = \sum_{k=1}^{2^K} M_1(A_k) \cdot \varphi\left(\frac{M_2(A_k)}{M_1(A_k)}\right) \, \geq 0 \tag{154}$$

(depending on φ, zero belief masses may have to be excluded for finiteness). For instance, we can take in (154) the special case $\varphi(t) := \varphi_{snKL,1}(t)$ (cf. (188)) to end up with the recent *Belief-Jensen-Shannon divergence* of Xiao [397], [398] who applies this to multi-sensor data fusion. As another special case we can take $\varphi(t) := \varphi_{1/2}(t)$ (cf. (43)) to end up with the $4-$times square of the recent *Hellinger distance of BBAs* of Li et al. [224], who use this for characterizing the degree of conflict between BBAs. To continue with the general context, as a particular application of the above-mentioned BS concepts, we can tackle general optimization problems of the type

$$\inf_{M_2 \in \Omega^{BBA}} D_\varphi(M_2, M_1) := \inf_{\mathbf{P}^{M_2} \in \boldsymbol{\Omega}} D_\varphi(\mathbf{P}^{M_2}, \mathbf{P}^{M_2}) \qquad \text{respectively}$$

where Ω^{BBA} is a collection of BBAs whose vectors of BBA-values form the set $\boldsymbol{\Omega} \in \mathbb{S}^{2^K}$ satisfying (7) and (9) as well as (149). for B in place of A. Hence, Theorem 14 (and its consequences and derived examples) can be employed.

We can also apply our BS method to "crossover cases " $D_\varphi(\mathbf{P}^M, \mathbf{P})$ (respectively with interchanged components) where $\mathbf{P}^M$ is a vector of $M-$based BBA values and $\mathbf{P}$ is a vector whose sum of components may not necessarily be 1. For instance, for the special choice $\varphi(t) := \varphi_1(t) := t \cdot \log t + 1 - t \in [0, \infty[$ (cf. 43), $\mathbf{P}^M := \big(p_k^M\big)_{k=1,\ldots,2^K} := (M(A_k))_{k=1,\ldots,2^K}$, $\mathbf{P} := (p_k)_{k=1,\ldots,2^K}$ with $p_k := 2^{|A_k|} - 1$ employing the cardinality $|A_k|$ of A_k, and the usual convention $0 \cdot \log(\frac{0}{0}) := 0$, we end up with (cf. (44))

$$D_{\varphi_1}(\mathbf{P}^M, \mathbf{P}) = \sum_{k=2}^{2^K} M(A_k) \cdot \log\left(\frac{M(A_k)}{2^{|A_k|} - 1}\right) - 1 + \sum_{k=2}^{2^K} (2^{|A_k|} - 1) =: -\mathcal{E}^{DE}(M) - 1 + \sum_{k=2}^{2^K} (2^{|A_k|} - 1)$$

[29] sometimes alternatively called basic probability assignment (BPA)

where $\mathcal{E}^{DE}(M) := -\sum_{k=2}^{2^K} M(A_k)\cdot\log\left(\frac{M(A_k)}{2^{|A_k|}-1}\right) \geq 0$ is nothing but (a multiple of) Deng's entropy of the BBA (BPA) M (cf. [110], see also e.g. Kang & Deng [181]).

Our BS method can also be applied to divergences between rescalings of BBAs. For instance, let $\breve{M}(A) := \frac{M(A)}{2^{|A|}-1}\quad (A \in 2^{\mathcal{Y}})$ with the convention that $\frac{0}{0} := 0$, and denote the corresponding vector $\mathbf{P}^{\breve{M}} := \left(p_k^{\breve{M}}\right)_{k=1,\ldots,2^K} := \left(\breve{M}(A_k)\right)_{k=1,\ldots,2^K}$. Accordingly, we define the $\varphi-$*divergence* $D_{\varphi}(\breve{M}_2,\breve{M}_1)$ *between the rescaled BBAs* $\breve{M}_2$ *and* $\breve{M}_1$ (in short, *rescaled Belief*$-\varphi-$*divergence*) as

$$D_{\varphi}(\breve{M}_2,\breve{M}_1) := D_{\varphi}(\mathbf{P}^{\breve{M}_2},\mathbf{P}^{\breve{M}_1}) = \sum_{k=1}^{2^K} p_k^{\breve{M}_1}\cdot\varphi\left(\frac{p_k^{\breve{M}_2}}{p_k^{\breve{M}_1}}\right) = \sum_{k=2}^{2^K}\breve{M}_1(A_k)\cdot\varphi\left(\frac{\breve{M}_2(A_k)}{\breve{M}_1(A_k)}\right) = \sum_{k=2}^{2^K}\frac{M_1(A_k)}{2^{|A_k|}-1}\cdot\varphi\left(\frac{M_2(A_k)}{M_1(A_k)}\right) \geq 0 \tag{155}$$

where for $A_1 := \emptyset$ we have used the convention that $0\cdot\varphi(\frac{0}{0}) := 0$ (depending on φ, other zero rescaled belief masses may have to be excluded for finiteness); notice that Remark 4 applies with $c_k := \frac{1}{2^{|A_k|}-1} > 0$. As an example, for the special choice $\varphi(t) := \varphi_1(t) := t\cdot\log t + 1 - t \in [0,\infty[$ (cf. 43), we derive from (44) and (155) the divergence

$$\begin{aligned}0 \leq D_{\varphi_1}(\breve{M}_2,\breve{M}_1) &= \sum_{k=2}^{2^K}\frac{M_2(A_k)}{2^{|A_k|}-1}\cdot\log\left(\frac{M_2(A_k)}{M_1(A_k)}\right) - \sum_{k=2}^{2^K}\frac{M_2(A_k)}{2^{|A_k|}-1} + \sum_{k=2}^{2^K}\frac{M_1(A_k)}{2^{|A_k|}-1}\\ &=: D^{SD}(M_2,M_1) - \sum_{k=2}^{2^K}\frac{M_2(A_k)}{2^{|A_k|}-1} + \sum_{k=2}^{2^K}\frac{M_1(A_k)}{2^{|A_k|}-1}\end{aligned}$$

where $D^{SD}(M_2,M_1)$ has been recently developed by Song & Deng [339]; notice that $D^{SD}(M_2,M_1)$ may be negative (cf. Stummer & Vajda [349]) and then it is not a divergence anymore. However, for applications to data fusion Song & Deng apply the symmetrization $\frac{1}{2}\cdot\left(D^{SD}(M_2,M_1) + D^{SD}(M_1,M_2)\right)$ which is equal to $\frac{1}{2}\cdot\left(D_{\varphi_1}(\breve{M}_2,\breve{M}_1) + D_{\varphi_1}(\breve{M}_1,\breve{M}_2)\right)$ and thus nonnegative.

Of course, we can also correspondingly adapt the *transformations* of $\varphi-$divergences (e.g. Renyi divergences) and entropy-type special cases given in the sections above and below, to BBAs as well as crossover cases and rescalings, and apply our BS method for this.

V. Finding/Constructing/On the distribution of the weights

Recall first that in Theorem 14, one crucial component is the sequence $(W_n)_{n\in\mathbb{N}}$ of weights being i.i.d. copies of a random variable W whose probability distribution is ζ (i.e. $\mathbb{\Pi}[W \in \cdot\,] = \zeta[\,\cdot\,]$), where the latter has to be connected with the divergence generator $\varphi \in \Upsilon(]a,b[)$ through the representation

$$\varphi(t) = \sup_{z\in\mathbb{R}}\left(z\cdot t - \log\int_{\mathbb{R}} e^{zy}\,d\zeta(y)\right), \qquad t\in\mathbb{R}, \qquad\qquad (cf.\ (6))$$

under the additional requirement that the function $z \mapsto MGF_{\zeta}(z) := \int_{\mathbb{R}} e^{zy}\,d\zeta(y)$ is finite on some open interval containing zero ("light-tailedness"); for Theorem 10, we need the corresponding variant (21) for $M_{\mathbf{P}}\cdot\varphi \in \Upsilon(]a,b[)$ (rather than φ).

Hence, finding such "BS-associated pairs (φ,ζ)" is an important issue. Subsequently, let us discuss the following direction: starting from a concrete optimization problem (27) — respectively (8) — with pregiven $\varphi \in \widetilde{\Upsilon}(]a,b[)$ (cf. Definition 3), as a first step one would like to verify whether indeed $\varphi \in \Upsilon(]a,b[)$ (i.e. it additionally satisfies (6)) — respectively $M_{\mathbf{P}}\cdot\varphi \in \Upsilon(]a,b[)$; as a second step, one would like to find the corresponding ζ explicitly.

As far as the above-mentioned first step is concerned, let us first present some fundamental properties of all $\varphi \in \Upsilon(]a,b[)$:

Proposition 29: Let $\varphi \in \Upsilon(]a,b[)$. Then the following assertions hold:

(G1) $\varphi :]-\infty,\infty[\to [0,\infty]$ is lower semicontinuous and convex;
(G2) $\varphi(1) = 0$;
(G3) $int(dom(\varphi)) =]a,b[$ for some $-\infty \leq a < 1 < b \leq \infty$;
(G4) φ is continuously differentiable on $]a,b[$ (i.e. $\varphi \in C^1(]a,b[)$;
(G5) φ is strictly convex *only* in a non-empty neighborhood $]t_-^{sc},t_+^{sc}[\subseteq]a,b[$ of one ($t_-^{sc} < 1 < t_+^{sc}$);
(G6) φ is infinitely differentiable on $]t_-^{sc},t_+^{sc}[$ (i.e. $\varphi \in C^{\infty}(]t_-^{sc},t_+^{sc}[)$, and hence, $\varphi'(1) = 0$, $\varphi''(t) > 0$ for all $t \in]t_-^{sc},t_+^{sc}[$; notice that the left-hand second derivative and the right-hand second derivative of φ may not coincide at t_-^{sc} respectively at t_+^{sc} (i.e. possible non-second-differentiability at these two points);

(G7) if $a > -\infty$, then $a = t_{-}^{sc}$;
if $a = -\infty$, then either $t_{-}^{sc} = -\infty$ or $\varphi(t) = \varphi(t_{-}^{sc}) + \varphi'(t_{-}^{sc}) \cdot (t - t_{-}^{sc})$ for all $t \in]-\infty, t_{-}^{sc}[$ (affine-linearity); notice that $\varphi'(t_{-}^{sc}) < 0$;

(G8) if $b < \infty$, then $b = t_{+}^{sc}$;
if $b = \infty$, then either $t_{+}^{sc} = \infty$ or $\varphi(t) = \varphi(t_{+}^{sc}) + \varphi'(t_{+}^{sc}) \cdot (t - t_{+}^{sc})$ for all $t \in]t_{+}^{sc}, \infty[$ (affine-linearity); notice that $\varphi'(t_{+}^{sc}) > 0$;

(G9) the *Fenchel-Legendre transform* (also called *convex conjugate*) of φ

$$\varphi^{*}(z) = \sup_{t \in \mathbb{R}} \left(z \cdot t - \varphi(t)\right) = \sup_{t \in]a,b[} \left(z \cdot t - \varphi(t)\right), \qquad z \in \mathbb{R}, \tag{156}$$

has the following properties:

(G9i) $int(dom(\varphi^{*})) =]\lambda_{-}, \lambda_{+}[$, where $dom(\varphi^{*}) := \{z \in \mathbb{R} : -\infty < \varphi^{*}(z) < \infty\}$,
$\lambda_{-} := \inf_{t \in]a,b[} \varphi'(t) = \lim_{t \downarrow a} \varphi'(t) =: \varphi'(a) < 0$ and
$\lambda_{+} := \sup_{t \in]a,b[} \varphi'(t) = \lim_{t \uparrow b} \varphi'(t) =: \varphi'(b) > 0$;

(G9ii) if $a > -\infty$, then
- $\lambda_{-} = -\infty$;
- the function $z \mapsto e^{-a \cdot z + \varphi^{*}(z)} =: M(z)$ is *absolutely monotone* on $]-\infty, 0[$, i.e. all derivatives exist and satisfy $\frac{\partial^k}{\partial z^k} M(z) \geq 0$ ($k \in \mathbb{N}_0$, $z \in]-\infty, 0[$);
- $\lim_{z \to 0-} M(z) = 1$;

(G9iii) if $b < \infty$, then
- $\lambda_{+} = \infty$;
- the function $z \mapsto e^{b \cdot z + \varphi^{*}(-z)} =: M(z)$ is absolutely monotone on $]-\infty, 0[$;
- $\lim_{z \to 0-} M(z) = 1$;

(G9iv) if $a = -\infty$ and $b = -\infty$, then
- the function $z \mapsto e^{\varphi^{*}(z)} =: M(z)$ is exponentially convex on $]\lambda_{-}, \lambda_{+}[$, i.e. $M(\cdot)$ is continuous and satisfies

$$\sum_{i,j=1}^{n} c_i \cdot c_j \cdot M\Big(\frac{z_i + z_j}{2}\Big) \geq 0 \qquad \text{for all } n \in \mathbb{N},\ c_i, c_j \in \mathbb{R} \text{ and } z_i, z_j \in]\lambda_{-}, \lambda_{+}[;$$

- $\lim_{z \to 0-} M(z) = 1$;
as a side remark, notice the well-known fact that exponential-convexity is stronger than the usual log-convexity.

(G10) the endpoints of $int(dom(\varphi)) =]a, b[$ have the following important "functioning" for the underlying probability distribution $\mathbb{\zeta}$ (cf. (6)) respectively of an associated random variable W with $\mathbb{\zeta}[\cdot\,] := \mathbb{\Pi}[W \in \cdot\,]$:

(G10i) $a = \inf supp(\mathbb{\zeta}) = \inf supp(W)$, $b = \sup supp(\mathbb{\zeta}) = \sup supp(W)$, where $supp(\mathbb{\zeta})$ respectively $supp(W)$ denotes the support of $\mathbb{\zeta}$ respectively W; consequently, $]a, b[= int(conv(supp(\mathbb{\zeta}))) = int(conv(supp(W)))$ where $conv(A)$ denotes the convex hull of a set A;

(G10ii) if $a > -\infty$, then $\varphi(a) = -\log \mathbb{\zeta}[\{a\}] = -\log \mathbb{\Pi}[W = a\,]$; consequently, there holds:
$a = \min supp(\mathbb{\zeta}) = \min supp(W)$ if and only if $\mathbb{\zeta}[\{a\}] = \mathbb{\Pi}[W = a\,] > 0$ if and only if $\varphi(a) < \infty$ if and only if $a \in dom(\varphi)$;

(G10iii) if $b < \infty$, then $\varphi(b) = -\log \mathbb{\zeta}[\{b\}] = -\log \mathbb{\Pi}[W = b\,]$; consequently, there holds:
$b = \max supp(\mathbb{\zeta}) = \max supp(W)$ if and only if $\mathbb{\zeta}[\{b\}] = \mathbb{\Pi}[W = b\,] > 0$ if and only if $\varphi(b) < \infty$ if and only if $b \in dom(\varphi)$.

(G11) the first two derivatives of φ at the point 1 have the following important "functioning" for the underlying probability distribution $\mathbb{\zeta}$ (cf. (6)) respectively of an associated random variable W:

(G11i) $1 = \varphi'^{-1}(0) = \int_{\mathbb{R}} y \, d\mathbb{\zeta}(y) = E_{\mathbb{\Pi}}[W]$ where $\varphi'^{-1}(\cdot)$ denotes the inverse of the first derivative $\varphi'(\cdot)$ of $\varphi(\cdot)$,

(G11ii) $\frac{1}{\varphi''(1)} = \int_{\mathbb{R}} \left(y - \int_{\mathbb{R}} \widetilde{y} \, d\mathbb{\zeta}(\widetilde{y})\right)^2 d\mathbb{\zeta}(y) = E_{\mathbb{\Pi}}[W^2] - (E_{\mathbb{\Pi}}[W])^2 = Var_{\mathbb{\Pi}}[W]$;
in particular, scaling $\widetilde{c} \cdot \varphi$ ($\widetilde{c} > 0$) does not change the mean 1 but the variance of W.

The corresponding proof of Proposition 29 will be given in Appendix D, except for the second items of (G9ii) and (G9iii) as well as the first item of (G9iv). Those will be treated in the second next paragraph below, because the corresponding line of argumentation builds an insightful start for subsequently performed procedures to further track down the weight distribution $\mathbb{\zeta}$.

The properties (G1) to (G9iv) constitute necessary conditions for a pregiven function φ to belong to $\Upsilon(]a, b[)$; accordingly, these should be verified first, in concrete situations where one aims to apply the BS approach. An important role is played by the boundary points a and b of $int(dom(\varphi))$ through (G10i) to (G10iii), because their finiteness opens the gate to apply — via some straightforward transformations — a rich class of real-valued characterization theorems for probability distributions whose support lies in the positive real line $[0, \infty[$. In contrast, there exist much less real-valued characterization theorems for probability distributions whose support is the whole real line $]-\infty, \infty[$; typically, the involved conditions are also more difficult to verify.

Indeed, if $\varphi \in \Upsilon(]a, b[)$ then one can deduce straightforwardly from the representation (6) that

$$e^{\varphi^*(z)} = \int_{-\infty}^{\infty} e^{z \cdot y} \, \mathrm{d}\mathfrak{C}(y) = E_{\mathbb{\Pi}}[e^{z \cdot W}], \quad z \in]\lambda_-, \lambda_+[, \tag{157}$$

where W is a random variable whose distribution is $\mathbb{\Pi}[W \in \cdot\,] = \mathfrak{C}[\,\cdot\,]$; under the additional knowledge $a > -\infty$ (and consequently $\lambda_- = -\infty$) employed together with (G10i) and thus $\mathbb{\Pi}[W \geq a] = \mathfrak{C}[\,[a, \infty[\,] = 1$, one arrives at

$$e^{\varphi^*(z) - a \cdot z} = \int_{a}^{\infty} e^{z \cdot (y-a)} \, \mathrm{d}\mathfrak{C}(y) = \int_{0}^{\infty} e^{z \cdot \widetilde{y}} \, \mathrm{d}\widetilde{\widetilde{\mathfrak{C}}}(\widetilde{y}) = E_{\mathbb{\Pi}}[e^{z \cdot (W-a)}], \quad z \in]-\infty, \lambda_+[, \tag{158}$$

where the probability distribution $\widetilde{\widetilde{\mathfrak{C}}}[\,\cdot\,] := \mathfrak{C}[\,\cdot + a\,]$ is the $a-$shifted companion of $\mathfrak{C}$; recall that $\lambda_+ > 0$. Put in other words, $\mathbb{\Pi}[\widetilde{W} \in \cdot\,] = \widetilde{\widetilde{\mathfrak{C}}}[\,\cdot\,]$ is the probability distribution of the (a.s.) nonnegative random variable $\widetilde{W} := W - a$. Naturally, in this context, the interesting case is $-\infty < a \leq 0$. Similarly, if $\varphi \in \Upsilon(]a, b[)$ and $b < \infty$ (and hence $\lambda_+ = \infty$), one can derive from (G10i) and its consequence $\mathbb{\Pi}[W \leq b] = \mathfrak{C}[\,]-\infty, b]\,] = 1$ that

$$e^{\varphi^*(-z) + b \cdot z} = \int_{-\infty}^{b} e^{z \cdot (b-y)} \, \mathrm{d}\mathfrak{C}(y) = \int_{0}^{\infty} e^{z \cdot \widetilde{y}} \, \mathrm{d}\widetilde{\widetilde{\mathfrak{C}}}(\widetilde{y}) = E_{\mathbb{\Pi}}[e^{z \cdot (b-W)}], \quad z \in]-\infty, -\lambda_-[, \tag{159}$$

where $-\lambda_- > 0$ and the probability distribution $\widetilde{\widetilde{\mathfrak{C}}}[\,\cdot\,] := \mathfrak{C}[\,b - \cdot\,]$ is the mirrored$-b-$shifted companion of $\mathfrak{C}$. This means that $\mathbb{\Pi}[\widetilde{W} \in \cdot\,] = \widetilde{\widetilde{\mathfrak{C}}}[\,\cdot\,]$ is the probability distribution of the (a.s.) nonnegative random variable $\widetilde{W} := b - W$. Naturally, the interesting case is $0 < b \leq \infty$.

As already indicated above, the considerations (157) to (159) open the gate to the adaption of well-known real-valued (rather than complex-valued) characterizations from probability theory. To begin with, the following assertions are very prominent:

Theorem 30: (a) Let $M :]-\infty, 0] \mapsto]0, \infty[$ be continuous on $]-\infty, 0]$ with $M(0) = 1$. Then one has

$$\text{M is absolutely monotone on }]-\infty, 0[\iff \exists \text{ unique prob. distr. } \widetilde{\widetilde{\mathfrak{C}}} \text{ on } [0, \infty[\text{ s.t. } M(z) = \int_{0}^{\infty} e^{z \cdot y} \mathrm{d}\widetilde{\widetilde{\mathfrak{C}}}(y) \text{ for all } z \in]-\infty, 0[.$$

(b) Let I be an open interval which contains 0, and $M : I \mapsto [0, \infty[$ be continuous with $M(0) = 1$. Then one gets

$$\text{M is exponentially convex } \iff \exists \text{ unique prob. distr. } \widetilde{\widetilde{\mathfrak{C}}} \text{ on }]-\infty, \infty[\text{ such that } M(z) = \int_{-\infty}^{\infty} e^{z \cdot y} \mathrm{d}\widetilde{\widetilde{\mathfrak{C}}}(y) \text{ for all } z \in I.$$

Assertion (a) of Theorem 30 is known as (probability-version of) *Bernstein's theorem* [42] (see e.g. also Schilling et al. [322]), whereas assertion (b) is known as (probability-version of) *Widder's theorem* [391] [30] (see e.g. also Widder [392], Akhiezer [9], Shucker [333], Jaksetic & Pecaric [163], Kotelina & Pevny [196]).

From Theorem 30(b) and (157), the first item in (G9iv) follows immediately by using the choice $I =]\lambda_-, \lambda_+[$. Moreover, Theorem 30(a) together with (158) (respectively (159)) implies the second item of (G9ii) (respectively of (G9iii)). In fact, with the help of Theorem 30 and some further considerations e.g. on light-tailedness, one even gets assertions on the sufficiency of (G9ii), (G9iii) and (G9iv) for a "candidate generator" φ to belong to the BS-suitable class $\Upsilon(]a, b[)$. More precisely, we obtain

Proposition 31: Suppose that $\varphi :]-\infty, \infty[\mapsto [0, \infty]$ satisfies (G1) to (G8), and recall the notations in (G9i). Then, $\varphi \in \Upsilon(]a, b[)$ if one of the following three conditions holds:
(a) $a > -\infty$, $\lambda_- = -\infty$, and the function $z \mapsto e^{-a \cdot z + \varphi^*(z)}$ is absolutely monotone on $]-\infty, 0[$,
(b) $b < \infty$, $\lambda_+ = \infty$, and the function $z \mapsto e^{b \cdot z + \varphi^*(-z)}$ is absolutely monotone on $]-\infty, 0[$,

[30] for the relevant conversion between the involved Riemann-Stieltjes integral with nondecreasing (but not necessarily right-continuous) integrator into a measure integral, one can apply the general theory in e.g. Chapter 6 of Chow & Teicher [88].

(c) $a = -\infty$, $b = -\infty$, and the function $z \mapsto e^{\varphi^*(z)}$ is exponentially convex on $]\lambda_-, \lambda_+[$.
If one of the three conditions (a) to (c) holds, then[31] the associated probability distribution ζ (cf. (6)) has expectation $\int_{\mathbb{R}} y d\zeta(y) = 1$ and finite moments of all orders, i.e. $\int_{\mathbb{R}} y^j d\zeta(y) < \infty$ for all $j \in \mathbb{N}_0$; in terms of $\zeta[\cdot] := \Pi[W \in \cdot]$ this means that $E_{\Pi}[W] = 1$ and $E_{\Pi}[W^j] < \infty$.

The proof of Proposition 31 will be given in Appendix D. As far as applicability is concerned, it is well known that, in general, verifying absolute monotonicity is typically more comfortable than verifying exponential convexity. Fortunately, one can often use the former, since for many known divergence generators there holds $a > -\infty$ (often $a = 0$) or $b < \infty$ or both, which by virtue of (G10i) is directly linked with the (endpoints of the) support of the potentially existing probability distribution ζ.

For a pregiven divergence generator φ, once its membership in $\Upsilon(]a, b[)$ (and in particular, the representability (6)) is verified, one would like to concretely find the underlying probability distribution ζ. This may be quite challenging, but can be made more comfortable by systematically narrowing down the family of distributions where ζ belongs to. In fact, we have already performed a first down-narrowing, in terms of identifying the endpoints of the support of ζ to be the endpoints of the effective domain of φ (cf. (G10i)). A further down-narrowing can be achieved from (157) to (159) in combination with real-valued characterization theorems which are more specific than Theorem 30. This will be shown exemplarily for a few important sub-setups, in the following.

For the identification of light-tailed semi/half-lattice distributions, we obtain the following two sets of sufficient conditions, which even allow for the desired explicit determination of ζ:

Proposition 32: Suppose that $\varphi :]-\infty, \infty[\mapsto [0, \infty]$ satisfies (G1) to (G8), with some $a > -\infty$. Furthermore, assume that there exists some constant $\breve{c} > 0$ as well as some function $H : [0, \infty[\mapsto [0, \infty[$ which is continuous on $[0, 1]$ with $H(1) = 1$ and absolutely monotone on $]0, 1[$, such that

$$e^{\varphi^*(\frac{z}{\breve{c}}) - a \cdot \frac{z}{\breve{c}}} = H(e^z), \qquad z \in]-\infty, \breve{c} \cdot \lambda_+[.$$

Then one has $\varphi \in \Upsilon(]a, b[)$ and

$$\zeta = \sum_{n=0}^{\infty} p_n \cdot \delta_{a + \breve{c} \cdot n} \qquad \text{with} \;\; p_n := \frac{1}{n!} \cdot \frac{\mathrm{d}^n H}{\mathrm{d}t}(0),$$

i.e. $\Pi[W = a + \breve{c} \cdot n] = p_n \quad (n \in \mathbb{N}_0)$.

Proposition 33: Suppose that $\varphi :]-\infty, \infty[\mapsto [0, \infty]$ satisfies (G1) to (G8), with some $b < \infty$. Furthermore, assume that there exists some constant $\breve{c} > 0$ as well as some function $H : [0, \infty[\mapsto [0, \infty[$ which is continuous on $[0, 1]$ with $H(1) = 1$ and absolutely monotone on $]0, 1[$, such that

$$e^{\varphi^*(-\frac{z}{\breve{c}}) + b \cdot \frac{z}{\breve{c}}} = H(e^z), \qquad z \in]-\infty, -\breve{c} \cdot \lambda_-[.$$

Then one has $\varphi \in \Upsilon(]a, b[)$ and

$$\zeta = \sum_{n=0}^{\infty} p_n \cdot \delta_{b - \breve{c} \cdot n} \qquad \text{with} \;\; p_n := \frac{1}{n!} \cdot \frac{\mathrm{d}^n H}{\mathrm{d}t}(0),$$

i.e. $\Pi[W = b - \breve{c} \cdot n] = p_n \quad (n \in \mathbb{N}_0)$.

The Propositions 32 respectively 33 follow from (158) respectively (159), some straightforward transformations, and a well-known characterization of probability generating functions H (see e.g. in Theorem 1.2.10 of Stroock [343]).

As an incentive for the following investigations, let us recall the discussion in the surroundings of Condition 9 pertaining to the minimization problem (13), where we have addressed possible connections between the two representabilities (6) (needed e.g. for Theorem 14) and (21) (needed e.g. for Theorem 10); this strongly relates to the question, for which constants $\widetilde{c} > 0$ the validity $\varphi \in \Upsilon(]a, b[)$ triggers the validity of $\widetilde{c} \cdot \varphi \in \Upsilon(]a, b[)$. To begin with, it is straightforward to see that $\varphi \in \Upsilon(]a, b[)$ *always implies* $\widetilde{c} \cdot \varphi \in \Upsilon(]a, b[)$ for all integers $\widetilde{c} \in \mathbb{N}$; indeed, if φ satisfies (6) for some $\zeta = \Pi[W \in \cdot]$, then for each integer $\widetilde{c} \in \mathbb{N}$ one gets that $\widetilde{c} \cdot \varphi$ satisfies (6) for $\widetilde{\zeta} = \Pi[\sum_{j=1}^{\widetilde{c}} \frac{W_j}{\widetilde{c}} \in \cdot]$; in the latter, the W_j's are i.i.d. copies from W. Clearly, $MGF_{\widetilde{\zeta}}$ is then finite on some open interval containing zero (differing from the one for MGF_{ζ} only by a scaling with $1/\widetilde{c}$).

For the following family of distributions, one can even trigger $\widetilde{c} \cdot \varphi \in \Upsilon(]a, b[)$ for *all* $\widetilde{c} > 0$: for the sake of a corresponding precise formulation, recall first the common knowledge that, generally speaking, a probability distribution ζ on $\mathbb{R}$ with light

[31] basically by Theorem 30 with $M(\cdot)$ defined in G9(ii),(iii) or (iv); see Appendix D.

tails — in the sense that its moment generating function $z \mapsto MGF_{\zeta}(z) := \int_{\mathbb{R}} e^{z\cdot y} d\zeta(y)$ is finite on some open interval $]\lambda_-, \lambda_+[$ containing zero — is (said to be) *infinitely divisible* if there holds

$$\text{for each } n \in \mathbb{N} \text{ there exists a probability distribution } \zeta_n \text{ on } \mathbb{R} \text{ such that } \int_{-\infty}^{\infty} e^{z\cdot y} d\zeta(y) = \Big(\int_{-\infty}^{\infty} e^{z\cdot y} d\zeta_n(y) \Big)^n, \quad z \in]\lambda_-, \lambda_+[; \tag{160}$$

in fact, (160) means that the (light-tailed) moment generating function MGF_{ζ} is *infinitely divisible* in the sense that each $n-$th root $(MGF_{\zeta})^{1/n}$ must be the moment generating function of some (light-tailed) probability distribution (denoted here by ζ_n). In particular, (160) implies that ζ_n is unique, and that ζ must necessarily have (one-sided or two-sided) unbounded support $supp(\zeta)$. The latter may differ from $supp(\zeta_n)$. In our BS context (6), (160) equivalently means that the associated random variable W is *infinitely divisible* (with light-tailed distribution), in the sense that

$$\text{for each } n \in \mathbb{N} \text{ there exists a sequence of i.i.d. random variables } Y_{n,1}, \cdots, Y_{n,n} \text{ such that } W \stackrel{\mathrm{d}}{=} Y_{n,1} + \cdots + Y_{n,n}, \tag{161}$$

where $\stackrel{\mathrm{d}}{=}$ means "have equal probability distributions" and $\Pi[W \in \cdot\,] = \zeta[\,\cdot\,]$, $\Pi[Y_{n,1} \in \cdot\,] = \zeta_n[\,\cdot\,]$.

For the above-mentioned context, we obtain the useful

Proposition 34: Suppose that $\varphi \in \Upsilon(]a,b[)$, with connected probability distribution ζ from (6) (recall that this implies that ζ is not a one-point distribution, cf. Remark 5). Then there holds:

$$\widetilde{c} \cdot \varphi \in \Upsilon(]a,b[) \text{ for all } \widetilde{c} > 0 \quad \Longleftrightarrow \quad \zeta \text{ is infinitely divisible.}$$

The proof of Proposition 34 is given in Appendix E.

Notice that Proposition 34 covers especially the important prominent *power divergences* (cf. Examples 39 and 40 below) for which we provide the corresponding infinitely divisible distributions explicitly in the Examples 48 and 50 below, and for which the subsequent form of estimators (cf. Chapter VI below) can be simplified.

For the identification of light-tailed infinitely divisible distributions, we obtain the following three sets of sufficient conditions:

Proposition 35: Suppose that $\varphi :]-\infty, \infty[\mapsto [0,\infty]$ satisfies (G1) to (G8), and recall the notations in (G9i) as well as $a = \inf supp(\zeta)$, $b = \sup supp(\zeta)$ (cf. (G10i)). Then, $\varphi \in \Upsilon(]a,b[)$ and the associated probability distribution ζ is infinitely divisible, if one of the following three conditions holds:
(a) $a > -\infty$, $\lambda_- = -\infty$, and the function $z \mapsto \varphi^{*\prime}(z) - a = (\varphi')^{-1}(z) - a$ is absolutely monotone on $]-\infty, 0[$,
(b) $b < \infty$, $\lambda_+ = \infty$, and the function $z \mapsto -\varphi^{*\prime}(-z) + b = -(\varphi')^{-1}(-z) + b$ is absolutely monotone on $]-\infty, 0[$,
(c) $a = -\infty$, $b = -\infty$, and the function $z \mapsto \frac{\varphi^{*\prime\prime}(z)}{\varphi^{*\prime\prime}(0)} = \frac{\varphi''(1)}{\varphi''((\varphi')^{-1}(z))}$ is exponentially convex on $]\lambda_-, \lambda_+[$.
In the first case (a) there automatically follows $b = \infty$, whereas in the second case (b) one automatically gets $a = -\infty$.

The proof of Proposition 35 is given in Appendix E.

So far, in the current section we have started from a given divergence generator $\varphi \in \widetilde{\Upsilon}(]a,b[)$ having some additional properties, switched to its Fenchel-Legendre transform φ^* and some exponentially-linear transforms thereof, and presented some sufficient conditions for verifying that the outcome is a moment-generating function MGF_{ζ} of a unique probability distribution ζ which has light tails. For finding the concrete ζ, one typically should know the explicit form of φ^*. However, it is well known that it can sometimes be hard to determine the explicit form of the Fenchel-Legendre transform of a convex function. This issue also applies for the reverse direction of starting from a concrete probability distribution ζ with light tails, computing its log-moment-generating function (called cumulant-generating function) $z \mapsto \Lambda_{\zeta}(z) := \log MGF_{\zeta}(z)$ and the corresponding Fenchel-Legendre transform Λ_{ζ}^* which is nothing but the associated divergence generator φ (cf. (6)). As will be illuminated in several examples below, the — "kind of intermediate" — construction method given in the below-mentioned Theorem 36 can help to ease these two tasks. To formulate this, we employ the class $\mathfrak{F}$ of functions $F :]-\infty, \infty[\mapsto [-\infty, \infty]$ with the following properties:
(F1) $int(dom(F)) =]a_F, b_F[$ for some $-\infty \leq a_F < 1 < b_F \leq \infty$;
(F2) F is smooth (infinitely continuously differentiable) on $]a_F, b_F[$;
(F3) F is strictly increasing on $]a_F, b_F[$.

Clearly, for any $F \in \mathfrak{F}$ one gets the existence of $F(a_F) := \lim_{t\downarrow a_F} F(t) \in [-\infty, \infty[$ and $F(b_F) := \lim_{t\uparrow b_F} F(t) \in]-\infty, \infty]$; moreover, its inverse $F^{-1} : \mathcal{R}(F) \mapsto [a_F, b_F]$ exists, where $\mathcal{R}(F) := \{F(t) : t \in dom(F)\}$. Furthermore, F^{-1} is strictly increasing and smooth (infinitely continuously differentiable) on the open interval $int(\mathcal{R}(F)) = \{F(t) : t \in]a_F, b_F[\} =]F(a_F), F(b_F)[$, and $F^{-1}(int(\mathcal{R}(F))) =]a_F, b_F[$. Within such a context, we obtain

Theorem 36: Let $F \in \mathfrak{F}$ and fix an arbitrary point $c \in int(\mathcal{R}(F))$. Moreover, introduce the notations[32] $]\lambda_-, \lambda_+[:= int(\mathcal{R}(F)) - c$ and $]t_-^{sc}, t_+^{sc}[:=]1 + a_F - F^{-1}(c), 1 + b_F - F^{-1}(c)[$ (which implies $\lambda_- < 0 < \lambda_+$ and $t_-^{sc} < 1 < t_+^{sc}$). Furthermore, define the functions $\Lambda :]-\infty, \infty[\mapsto [-\infty, \infty]$ and $\varphi :]-\infty, \infty[\mapsto [0, \infty]$ by

$$\Lambda(z) := \Lambda^{(c)}(z) := \begin{cases} \int_0^z F^{-1}(u+c)\,du + z \cdot (1 - F^{-1}(c)) \ \in]-\infty, \infty[, & \text{if } z \in]\lambda_-, \lambda_+[, \\ \int_0^{\lambda_-} F^{-1}(u+c)\,du + \lambda_- \cdot (1 - F^{-1}(c)) \ \in [-\infty, \infty], & \text{if } z = \lambda_- > -\infty, \\ \int_0^{\lambda_+} F^{-1}(u+c)\,du + \lambda_+ \cdot (1 - F^{-1}(c)) \ \in \ [-\infty, \infty], & \text{if } z = \lambda_+ < \infty, \\ \infty, & \text{else,} \end{cases} \tag{162}$$

where the second respectively third line are meant as $\lim_{z \downarrow \lambda_-} \big(\int_0^z F^{-1}(u+c)\,du + z \cdot (1 - F^{-1}(c)) \big)$ respectively $\lim_{z \uparrow \lambda_+} \big(\int_0^z F^{-1}(u+c)\,du + z \cdot (1 - F^{-1}(c)) \big)$, and

$$\varphi(t) := \varphi^{(c)}(t) := \begin{cases} (t + F^{-1}(c) - 1) \cdot [F\big(t + F^{-1}(c) - 1\big) - c] - \int_0^{F(t+F^{-1}(c)-1)-c} F^{-1}(u+c)du \ \in \ [0, \infty[, \\ \qquad \text{if } t \in]t_-^{sc}, t_+^{sc}[, \\ (t_-^{sc} + F^{-1}(c) - 1) \cdot [F\big(t_-^{sc} + F^{-1}(c) - 1\big) - c] - \int_0^{F(t_-^{sc}+F^{-1}(c)-1)-c} F^{-1}(u+c)du \ \in]0, \infty], \\ \qquad \text{if } t = t_-^{sc} > -\infty, \\ (t_+^{sc} + F^{-1}(c) - 1) \cdot [F\big(t_+^{sc} + F^{-1}(c) - 1\big) - c] - \int_0^{F(t_+^{sc}+F^{-1}(c)-1)-c} F^{-1}(u+c)du \ \in]0, \infty], \\ \qquad \text{if } t = t_+^{sc} < \infty, \\ \varphi(t_-^{sc}) + \lambda_- \cdot (t - t_-^{sc}) \ \in]0, \infty], \qquad \text{if } t_-^{sc} > -\infty, \text{ and } t \in]-\infty, t_-^{sc}[, \\ \varphi(t_+^{sc}) + \lambda_+ \cdot (t - t_+^{sc}) \ \in]0, \infty], \qquad \text{if } t_+^{sc} < \infty, \text{ and } t \in]t_+^{sc}, \infty[, \\ \infty, \qquad \text{else,} \end{cases} \tag{163}$$

where the second respectively third line are again meant as lower respectively upper limit.
Then, Λ and φ have the following properties:
(i) On $]\lambda_-, \lambda_+[$, the function Λ is smooth and strictly convex and consequently, $\exp(\Lambda))$ is smooth and strictly log-convex; moreover, there holds $\Lambda(0) = 0$, $\Lambda'(0) = 1$;
(ii) $\varphi \in \widetilde{\Upsilon}(]a, b[)$, where $a := t_-^{sc} \cdot 1_{\{-\infty\}}(\lambda_-) - \infty \cdot 1_{]-\infty,0[}(\lambda_-)$, $b := t_+^{sc} \cdot 1_{\{\infty\}}(\lambda_+) + \infty \cdot 1_{]0,\infty[}(\lambda_+)$, and φ has the properties (G1) to (G8).
(iii) $\varphi(t) = \sup_{z \in]-\infty,\infty[} (z \cdot t - \Lambda(z)) = \sup_{z \in]\lambda_-,\lambda_+[} (z \cdot t - \Lambda(z))$ for all $t \in \mathbb{R}$.
(iv) $\Lambda(z) = \varphi^*(z) = \sup_{t \in]-\infty,\infty[} (t \cdot z - \varphi(t)) = \sup_{t \in]a,b[} (t \cdot z - \varphi(t))$ for all $z \in \mathbb{R}$.

The proof of Theorem 36 will be given Appendix F.

Remark 37: Theorem 36 indicates that the $F-$constructed function $z \mapsto \exp(\Lambda(z)) = \exp(\varphi^*(z))$ is a good candidate for a moment generating function of a probability distribution $\mathfrak{C}$, and hence for the representability (6). However, one still needs to verify one of the conditions (a) to (c) of Proposition 31. This may go wrong, as the case of power divergences φ_γ with $\gamma \in]1, 2[$ indicates (cf. the conjecture of Example 48(f) below).

Notice that the newly constructed Λ and φ (cf. (162), (163)) depend on the choice of the anchor point c; this is e.g. illustrated in Example 40(b) below. Hence, as a side effect, by using whole families $(F_\vartheta)_\vartheta$ together with different anchor points c, via Theorem 36 one can generate new classes (and new classifications) of $\varphi-$divergence generators — and thus of corresponding $\varphi-$divergences — which can be of great use, even in other contexts beyond our BS optimization framework.

If F satisfies $F(1) = 0$ and thus $F^{-1}(0) = 1$, then the natural choice $c := 0$ induces $]\lambda_-, \lambda_+[= int(\mathcal{R}(F))$ and $]t_-^{sc}, t_+^{sc}[=]a_F, b_F[$, and consequently (due to $F^{-1}(c) - 1 = 0$) leads to the simplification of "the first lines of" (162) and (163) to

$$\Lambda(z) := \Lambda^{(0)}(z) := \int_0^z F^{-1}(u)du, \qquad z \in int(\mathcal{R}(F)), \tag{164}$$

$$\varphi(t) := \varphi^{(0)}(t) := t \cdot F(t) - \int_0^{F(t)} F^{-1}(u)du, \qquad t \in]a_F, b_F[; \tag{165}$$

the simplifications of the respective other lines of (162) and (163) are straightforward.

[32] for the sake of brevity, we avoid here the more complete notation $\lambda_-^{F,c}$, $\lambda_+^{F,c}$, $t_-^{sc,F,c}$, $t_+^{sc,F,c}$ indicating the dependence on F and c.

Remark 38: Let $F \in \mathfrak{F}$ with $a_F = 0$, $b_F = \infty$, $F(1) = 0$ and hence, $int(\mathcal{R}(F)) =]F(0), F(\infty)[$. Then the transformation

$$\widetilde{F}(t) \quad := \quad \begin{cases} -\int_0^{F(\frac{1}{t})} F^{-1}(u)\, du, & \text{if } \ t \in]0,\infty[, \\ -\int_0^{F(\infty)} F^{-1}(u)\, du, & \text{if } \ t = 0, \\ -\infty, & \text{if } \ t \in]-\infty, 0[, \end{cases} \tag{166}$$

satisfies $\widetilde{F} \in \mathfrak{F}$ with $a_{\widetilde{F}} = 0$, $b_{\widetilde{F}} = \infty$, $\widetilde{F}(1) = 0$ and $int(\mathcal{R}(\widetilde{F})) = \big] -\int_0^{F(\infty)} F^{-1}(u)\, du, -\int_0^{F(0)} F^{-1}(u)\, du \,\big[$. By choosing the natural anchor point $c = 0$ (for both F and $\widetilde{F}$) and by using the relations $\widetilde{F}(t) = -\Lambda(F(\frac{1}{t}))$, $\widetilde{F}^{-1}(z) = \frac{1}{F^{-1}(\Lambda^{-1}(-z))}$, as well as (264) in combination with (266) (for both contexts), it is straightforward to see that the corresponding quantities $\widetilde{\Lambda}$ and $\widetilde{\varphi}$ satisfy $\widetilde{\Lambda}(z) = -(-\Lambda)^{-1}(z) \quad (z \in int(\mathcal{R}(\widetilde{F})))$ and $\widetilde{\varphi}(t) = t \cdot \varphi(\frac{1}{t}) \quad (t \in]0,\infty[)$. Hence, the corresponding divergences (cf. (4)) are "reciprocal to each other" in the sense that $D_{\widetilde{\varphi}}(\mathbf{Q},\mathbf{P}) = D_{\varphi}(\mathbf{P},\mathbf{Q})$ for all $\mathbf{P},\mathbf{Q} \in \mathbb{S}^K_{>0}$, and in case that Λ and $\widetilde{\Lambda}$ are indeed cumulant generating functions of some light-tailed distributions $\mathfrak{C}$ and $\widetilde{\mathfrak{C}}$ (cf. Remark 37), then the latter two are said to be *inverse to each other* in the sense of Tweedie [368] (see also e.g. Folks [127]).

As already indicated above, from Theorem 36 one can comfortably generate various interesting examples, which we demonstrate in the following.

Example 39: (a) For $\gamma \in \mathbb{R}\backslash\{1,2\}$, $\widetilde{c} \in]0,\infty[$ and $]a_{F_{\gamma,\widetilde{c}}}, b_{F_{\gamma,\widetilde{c}}}[=]0,\infty[$ we define

$$F_{\gamma,\widetilde{c}}(t) \quad := \quad \begin{cases} \frac{\widetilde{c}}{\gamma-1} \cdot (t^{\gamma-1} - 1), & \text{if } \ t \in]0,\infty[, \\ -\frac{\widetilde{c}}{\gamma-1}, & \text{if } \ t = 0 \text{ and } \ \gamma \in]1,2[\cup]2,\infty[, \\ -\infty, & \text{if } \ t = 0 \text{ and } \ \gamma < 1, \\ -\infty, & \text{if } \ t \in]-\infty,0[, \end{cases}$$

Clearly, $\mathcal{R}(F_{\gamma,\widetilde{c}}) = \big[-\frac{\widetilde{c}}{\gamma-1}, \infty \big[$ for $\gamma \in]1,2[\cup]2,\infty[$, respectively $\mathcal{R}(F_{\gamma,\widetilde{c}}) = \big] -\infty, \frac{\widetilde{c}}{1-\gamma} \big[$ for $\gamma < 1$; notice that $0 \in int(\mathcal{R}(F_{\gamma,\widetilde{c}}))$ for all $\gamma \in \mathbb{R}\backslash\{1,2\}$. Furthermore, $F_{\gamma,\widetilde{c}}(\cdot)$ is strictly increasing and smooth on $]0,\infty[$, and thus, $F_{\gamma,\widetilde{c}} \in \mathfrak{F}$. Since $F_{\gamma,\widetilde{c}}(1) = 0$, let us choose the natural anchor point $c := 0$, which leads to $]\lambda_-, \lambda_+[= int(\mathcal{R}(F_{\gamma,\widetilde{c}}))$ and $]t^{sc}_-, t^{sc}_+[=]0,\infty[$. By using $F^{-1}_{\gamma,\widetilde{c}}(x) = (1 + \frac{(\gamma-1)\cdot x}{\widetilde{c}})^{\frac{1}{\gamma-1}}$ for $x \in int(\mathcal{R}(F_{\gamma,\widetilde{c}}))$, we can derive from formula (162) (see also (164)) for all $\gamma \in \mathbb{R}\backslash\{0,1,2\}$ and $z \in \mathbb{R}$

$$\Lambda_{\gamma,\widetilde{c}}(z) := \Lambda^{(0)}_{\gamma,\widetilde{c}}(z) \quad = \quad \begin{cases} \frac{\widetilde{c}}{\gamma} \cdot \Big\{ \big(\frac{\gamma-1}{\widetilde{c}} \cdot z + 1 \big)^{\frac{\gamma}{\gamma-1}} - 1 \Big\}, & \text{if } \ \gamma \in]1,2[\cup]2,\infty[\text{ and } \ z \in \big] -\frac{\widetilde{c}}{\gamma-1}, \infty \big[\\ & \text{or if } \ \gamma \in]-\infty,0[\cup]0,1[\text{ and } \ z \in \big] -\infty, \frac{\widetilde{c}}{1-\gamma} \big[, \\ -\frac{\widetilde{c}}{\gamma} < 0, & \text{if } \ \gamma \in]1,2[\cup]2,\infty[\text{ and } \ z = -\frac{\widetilde{c}}{\gamma-1}, \\ -\frac{\widetilde{c}}{\gamma} > 0, & \text{if } \ \gamma < 0 \text{ and } \ z = \frac{\widetilde{c}}{1-\gamma}, \\ \infty, & \text{if } \ \gamma \in]0,1[\text{ and } \ z = \frac{\widetilde{c}}{1-\gamma}, \\ \infty, & \text{else.} \end{cases} \tag{167}$$

Notice that $\Lambda_{\gamma,\widetilde{c}}(0) = 0$ for all $\gamma \in \mathbb{R}\backslash\{0,1,2\}$. Moreover, for $\gamma \in]1,2[\cup]2,\infty[$ one has $\Lambda_{\gamma,\widetilde{c}}(\infty) = \infty$, $\Lambda'_{\gamma,\widetilde{c}}(-\frac{\widetilde{c}}{\gamma-1}) = 0$ and $\Lambda'_{\gamma,\widetilde{c}}(\infty) = \infty$. For $\gamma < 0$ one gets $\Lambda_{\gamma,\widetilde{c}}(-\infty) = -\infty$, $\Lambda'_{\gamma,\widetilde{c}}(\frac{\widetilde{c}}{1-\gamma}) = \infty$ and $\Lambda'_{\gamma,\widetilde{c}}(-\infty) = 0$. In contrast, if $\gamma \in]0,1[$ then $\Lambda_{\gamma,\widetilde{c}}(-\infty) = -\frac{\widetilde{c}}{\gamma} < 0$, $\Lambda'_{\gamma,\widetilde{c}}(\frac{\widetilde{c}}{1-\gamma}) = \infty$ and $\Lambda'_{\gamma,\widetilde{c}}(-\infty) = 0$. To proceed, from formula (163) (see also (165)) we can deduce for all $\gamma \in \mathbb{R}\backslash\{0,1,2\}$ and $t \in \mathbb{R}$

$$\varphi_{\gamma,\widetilde{c}}(t) := \varphi^{(0)}_{\gamma,\widetilde{c}}(t) \ = \ \begin{cases} \widetilde{c} \cdot \frac{t^\gamma - \gamma\cdot t + \gamma - 1}{\gamma\cdot(\gamma-1)} \ \in \ [0,\infty[, & \text{if } t \in]0,\infty[, \\ \frac{\widetilde{c}}{\gamma} > 0, & \text{if } \ \gamma \in]1,2[\cup]2,\infty[\text{ and } \ t = 0, \\ \infty, & \text{if } \ \gamma < 0 \text{ and } \ t = 0, \\ \frac{\widetilde{c}}{\gamma} > 0, & \text{if } \ \gamma \in]0,1[\text{ and } \ t = 0, \\ \frac{\widetilde{c}}{\gamma} - \frac{\widetilde{c}}{\gamma-1} \cdot t \ \in \ ]0,\infty[, & \text{if } \ \gamma \in]1,2[\cup]2,\infty[\text{ and } \ t < 0, \\ \infty, & \text{else,} \end{cases} \tag{168}$$

which coincides with $\widetilde{c} \cdot \varphi_\gamma(t)$ for $\varphi_\gamma(t)$ from (43) and which generates the $\gamma-$corresponding power divergences given in (44); the first line in (168) can be proved by

$$\varphi_{\gamma,\widetilde{c}}(t) := \varphi^{(0)}_{\gamma,\widetilde{c}}(t) := t \cdot F_{\gamma,\widetilde{c}}(t) - \int_0^{F_{\gamma,\widetilde{c}}(t)} F^{-1}_{\gamma,\widetilde{c}}(u)du$$
$$= \frac{t \cdot \widetilde{c}}{\gamma - 1} \cdot (t^{\gamma-1} - 1) - \frac{\widetilde{c}}{\gamma} \cdot \left\{ \left(\frac{\gamma - 1}{\widetilde{c}} \cdot \left[\frac{\widetilde{c}}{\gamma - 1} \cdot (t^{\gamma-1} - 1) \right] + 1 \right)^{\frac{\gamma}{\gamma-1}} - 1 \right\}$$
$$= \widetilde{c} \cdot \frac{t^\gamma - \gamma \cdot t + \gamma - 1}{\gamma \cdot (\gamma - 1)}, \qquad t \in]0, \infty[\,. \tag{169}$$

Notice that for all $\gamma \in \mathbb{R}\backslash\{0, 1, 2\}$ one has $\varphi_{\gamma,\widetilde{c}}(1) = 0$, $\varphi'_{\gamma,\widetilde{c}}(1) = 0$ and $\varphi_{\gamma,\widetilde{c}}(\infty) = \infty$. Moreover, for $\gamma \in]1, 2[\cup]2, \infty[$ one has $\varphi'_{\gamma,\widetilde{c}}(0) = -\frac{\widetilde{c}}{\gamma-1} < 0$ and $\varphi'_{\gamma,\widetilde{c}}(\infty) = \infty$. In contrast, for $\gamma < 0$ and $\gamma \in]0, 1[$ one gets $\varphi'_{\gamma,\widetilde{c}}(0) = -\infty$ and $\varphi'_{\gamma,\widetilde{c}}(\infty) = \frac{\widetilde{c}}{1-\gamma} > 0$.

(b) For $\gamma = 2$, $\widetilde{c} \in]0, \infty[$ and $]a_{F_{\gamma,\widetilde{c}}}, b_{F_{\gamma,\widetilde{c}}}[=] - \infty, \infty[$ we define

$$F_{2,\widetilde{c}}(t) \;\; := \;\; \widetilde{c} \cdot (t - 1), \qquad t \in] - \infty, \infty[,$$

Clearly, $\mathcal{R}(F_{2,\widetilde{c}}) =] - \infty, \infty[$, $0 \in int(\mathcal{R}(F_{2,\widetilde{c}}))$, and $F_{2,\widetilde{c}}(\cdot)$ is strictly increasing as well as smooth on $] - \infty, \infty[$. Hence, $F_{2,\widetilde{c}} \in \mathfrak{F}$. Since $F_{2,\widetilde{c}}(1) = 0$, let us choose the natural anchor point $c := 0$, which leads to $]\lambda_-, \lambda_+[= int(\mathcal{R}(F_{2,\widetilde{c}})) =]-\infty, \infty[$ and $]t^{sc}_-, t^{sc}_+[=] - \infty, \infty[$. By using $F^{-1}_{2,\widetilde{c}}(x) = 1 + \frac{x}{\widetilde{c}}$ for $x \in int(\mathcal{R}(F_{2,\widetilde{c}}))$, we can derive from formula (162) (see also (164))

$$\Lambda_{2,\widetilde{c}}(z) := \Lambda^{(0)}_{2,\widetilde{c}}(z) \;\; = \;\; \frac{\widetilde{c}}{2} \cdot \left\{ \left(\frac{1}{\widetilde{c}} \cdot z + 1 \right)^2 - 1 \right\} = \frac{z^2}{2\widetilde{c}} + z, \qquad z \in] - \infty, \infty[. \tag{170}$$

Notice that $\Lambda_{2,\widetilde{c}}(0) = 0$, $\Lambda_{2,\widetilde{c}}(-\infty) = \Lambda_{2,\widetilde{c}}(\infty) = \infty$, $\Lambda'_{2,\widetilde{c}}(-\infty) = -\infty$ and $\Lambda'_{2,\widetilde{c}}(\infty) = \infty$. From formula (163) (see also (165)) we can deduce analogously to (169)

$$\varphi_{2,\widetilde{c}}(t) := \varphi^{(0)}_{2,\widetilde{c}}(t) \;\; = \;\; \widetilde{c} \cdot \frac{(t - 1)^2}{2} \;\; \in \;\; [0, \infty[, \qquad t \in] - \infty, \infty[, \tag{171}$$

which coincides with $\widetilde{c} \cdot \varphi_2(t)$ for $\varphi_2(t)$ from (43) which generates the corresponding power divergence given in the sixth line of (44). Notice that $\varphi_{2,\widetilde{c}}(1) = 0$, $\varphi'_{2,\widetilde{c}}(1) = 0$ and $\varphi_{2,\widetilde{c}}(-\infty) = \varphi_{2,\widetilde{c}}(\infty) = \infty$.

As an application of the reciprocity considerations of Remark 38, it is straightforward to see from the above-mentioned considerations (a) and (b) that for all $\gamma \in \mathbb{R}\backslash\{0, 1\}$ one has $\widetilde{F}_{\gamma,\widetilde{c}}(t) = -\Lambda_{\gamma,\widetilde{c}}(F_{\gamma,\widetilde{c}}(\frac{1}{t})) = F_{1-\gamma,\widetilde{c}}(t)$ for all $t \in]0, \infty[$.

(c) Let us now continue with the remaining case $\gamma = 0$ (recall the natural anchor point $c := 0$). By using $F^{-1}_{0,\widetilde{c}}(x) = \frac{1}{1-\frac{x}{\widetilde{c}}}$ for $x \in int(\mathcal{R}(F_{0,\widetilde{c}})) =] - \infty, \widetilde{c}[$, we can derive from formula (162) (see also (164))

$$\Lambda_{0,\widetilde{c}}(z) := \Lambda^{(0)}_{0,\widetilde{c}}(z) \;\; = \;\; \begin{cases} -\widetilde{c} \cdot \log\left(1 - \frac{z}{\widetilde{c}}\right), & \text{if } \; z \in \left] - \infty, \widetilde{c}\right[, \\ \infty, & \text{if } \; z \in \left[\widetilde{c}, \infty\right[. \end{cases} \tag{172}$$

Notice that $\Lambda_{0,\widetilde{c}}(0) = 0$, $\Lambda_{0,\widetilde{c}}(-\infty) = -\infty$, $\Lambda'_{0,\widetilde{c}}(\widetilde{c}) = \infty$ and $\Lambda'_{0,\widetilde{c}}(-\infty) = 0$. Moreover, from formula (163) (see also (165)) we can deduce

$$\varphi_{0,\widetilde{c}}(t) := \varphi^{(0)}_{0,\widetilde{c}}(t) \; = \; \begin{cases} \widetilde{c} \cdot (-\log t + t - 1) \;\; \in \;\; [0, \infty[, & \text{if } t \in]0, \infty[, \\ \infty, & \text{if } \; t \in] - \infty, 0], \end{cases} \tag{173}$$

which coincides with $\widetilde{c} \cdot \varphi_0(t)$ for the generator $\varphi_0(t)$ from (43) which generates the reverse Kullback-Leibler divergence (reverse relative entropy) given in (44) with $\widetilde{c} = 1$; the first line in (173) can be proved by

$$\varphi_{0,\widetilde{c}}(t) := \varphi^{(0)}_{0,\widetilde{c}}(t) := t \cdot F_{0,\widetilde{c}}(t) - \int_0^{F_{0,\widetilde{c}}(t)} F^{-1}_{0,\widetilde{c}}(u)du$$
$$= t \cdot \widetilde{c} \cdot \left(1 - \frac{1}{t}\right) - (-\widetilde{c}) \cdot \log\left(1 - \frac{1}{\widetilde{c}} \cdot \left[\widetilde{c} \cdot \left(1 - \frac{1}{t}\right)\right]\right) = \widetilde{c} \cdot (-\log t + t - 1)\,, \qquad t \in]0, \infty[\,. \tag{174}$$

Notice that one has $\varphi_{0,\widetilde{c}}(1) = 0$, $\varphi_{0,\widetilde{c}}(\infty) = \infty$, $\varphi'_{0,\widetilde{c}}(1) = 0$, $\varphi'_{0,\widetilde{c}}(0) = -\infty$ and $\varphi'_{0,\widetilde{c}}(\infty) = \widetilde{c}$.

Example 40: (a) For the remaining case $\gamma = 1$, $\widetilde{c} \in]0,\infty[$ and $]a_{F_{1,\widetilde{c}}}, b_{F_{1,\widetilde{c}}}[=]0,\infty[$ we define

$$F_{1,\widetilde{c}}(t) \quad := \quad \begin{cases} \widetilde{c} \cdot \log t = \lim_{\gamma\to 1} F_{\gamma,\widetilde{c}}(t), & \text{if} \quad t \in]0,\infty[, \\ -\infty, & \text{if} \quad t \in]-\infty, 0]. \end{cases}$$

Clearly, $\mathcal{R}(F_{1,\widetilde{c}}) =]-\infty,\infty[$. Moreover, $F_{1,\widetilde{c}}(\cdot)$ is strictly increasing and smooth on $]0,\infty[$, and hence, $F_{\gamma,\widetilde{c}} \in \boldsymbol{\mathfrak{F}}$. Since $F_{1,\widetilde{c}}(1) = 0$, let us *first* choose the natural anchor point $c := 0$, which leads to $]\lambda_-, \lambda_+[= int(\mathcal{R}(F_{1,\widetilde{c}})) =]-\infty,\infty[$ and $]t^{sc}_-, t^{sc}_+[=]0,\infty[$. By using $F^{-1}_{1,\widetilde{c}}(x) = \exp(\frac{x}{\widetilde{c}})$ for $x \in \mathcal{R}(F_{1,\widetilde{c}})$, we can derive from formula (162) (see also (164))

$$\Lambda_{1,\widetilde{c}}(z) := \Lambda^{(0)}_{1,\widetilde{c}}(z) := \int_0^z F^{-1}_{1,\widetilde{c}}(u)\, du = \widetilde{c} \cdot \left(\exp\left(\frac{z}{\widetilde{c}}\right) - 1\right), \quad z \in]-\infty,\infty[. \tag{175}$$

Notice that $\Lambda_{1,\widetilde{c}}(0) = 0$, $\Lambda_{1,\widetilde{c}}(-\infty) = -\widetilde{c}$, $\Lambda_{1,\widetilde{c}}(\infty) = \infty$, $\Lambda'_{1,\widetilde{c}}(-\infty) = 0$ and $\Lambda'_{0,\widetilde{c}}(\infty) = \infty$. Moreover, from formula (163) (see also (165)) we can deduce

$$\varphi_{1,\widetilde{c}}(t) := \varphi^{(0)}_{1,\widetilde{c}}(t) \; := \; \begin{cases} \widetilde{c} \cdot (t \cdot \log t + 1 - t) \; \in \; [0,\infty[, & \text{if } t \in]0,\infty[, \\ 1, & \text{if} \quad t = 0, \\ \infty, & \text{if} \quad t \in]-\infty, 0[, \end{cases} \tag{176}$$

which coincides with $\widetilde{c} \cdot \varphi_1(t)$ for the generator $\varphi_1(t)$ from (43) which generates the Kullback-Leibler divergence (relative entropy) given in (44) with $\widetilde{c} = 1$; the first line in (176) can be proved by

$$\begin{aligned} \varphi_{1,\widetilde{c}}(t) &:= \varphi^{(0)}_{1,\widetilde{c}}(t) := t \cdot F_{1,\widetilde{c}}(t) - \int_0^{F_{1,\widetilde{c}}(t)} F^{-1}_{1,\widetilde{c}}(u) du \\ &= t \cdot \widetilde{c} \cdot \log t - \widetilde{c} \cdot \left(\exp\left(\frac{1}{\widetilde{c}} \cdot [\widetilde{c} \cdot \log t]\right) - 1\right) = \widetilde{c} \cdot (t \cdot \log t + 1 - t), \qquad t \in]0,\infty[\,, \end{aligned} \tag{177}$$

Notice that one has $\varphi_{1,\widetilde{c}}(1) = 0$, $\varphi_{1,\widetilde{c}}(\infty) = \infty$, $\varphi'_{1,\widetilde{c}}(1) = 0$, $\varphi'_{1,\widetilde{c}}(0) = -\infty$ and $\varphi'_{1,\widetilde{c}}(\infty) = \infty$.

As an application of the reciprocity considerations of Remark 38, it is straightforward to see that $\widetilde{F}_{1,\widetilde{c}}(t) = -\Lambda_{1,\widetilde{c}}(F_{1,\widetilde{c}}(\frac{1}{t})) = F_{0,\widetilde{c}}(t)$ for all $t \in]0,\infty[$.

(b) For the choice $\widetilde{c} = 1$, let us now fix a general anchor point $c \in \mathcal{R}(F_{1,\widetilde{c}}) =]-\infty,\infty[$ (rather than $c = 0$), which leads to $]\lambda_-, \lambda_+[= int(\mathcal{R}(F_{1,1})) - c =]-\infty,\infty[$ and $]t^{sc}_-, t^{sc}_+[=]1 + a_{F_{1,1}} - F^{-1}_{1,1}(c), 1 + b_{F_{1,1}} - F^{-1}_{1,1}(c)[=]1 - e^c,\infty[$. Accordingly, the formula (162) (see also (164)) leads to

$$\begin{aligned} \Lambda_{1,1}(z) := \Lambda^{(c)}_{1,1}(z) &:= \int_0^z F^{-1}_{1,1}(u + c) du + z \cdot (1 - F^{-1}_{1,1}(c)) \\ &= e^c \cdot (e^z - 1) + z \cdot (1 - e^c), \qquad z \in]-\infty,\infty[, \end{aligned} \tag{178}$$

for which there holds $\Lambda^{(c)}_{1,1}(0) = 0$, $\Lambda^{(c)}_{1,1}(-\infty) = \infty \cdot 1_{]0,\infty[}(c) - \infty \cdot 1_{]-\infty,0[}(c) - 1 \cdot 1_{\{0\}}(c)$, $\Lambda^{(c)}_{1,1}(\infty) = \infty$, $\Lambda^{(c)\prime}_{1,1}(-\infty) = 1 - e^c$ and $\Lambda^{(c)\prime}_{1,1}(\infty) = \infty$. Moreover, from formula (163) (see also (165)) we can deduce

$$\varphi_{1,1}(t) := \varphi^{(c)}_{1,1}(t) \; := \; \begin{cases} (t + e^c - 1) \cdot [\log(t + e^c - 1) - c] + 1 - t \; \in \; [0,\infty[, & \text{if } t \in]1 - e^c,\infty[, \\ e^c, & \text{if} \quad t = 1 - e^c, \\ \infty, & \text{if} \quad t \in]-\infty, 1 - e^c[; \end{cases} \tag{179}$$

the first line in (179) can be proved by

$$\begin{aligned} \varphi_{1,1}(t) &:= \varphi^{(c)}_{1,1}(t) := (t + F^{-1}_{1,1}(c) - 1) \cdot [F_{1,1}\left(t + F^{-1}_{1,1}(c) - 1\right) - c] - \int_0^{F_{1,1}\left(t + F^{-1}_{1,1}(c) - 1\right) - c} F^{-1}_{1,1}(u + c) du \\ &= (t + e^c - 1) \cdot [\log(t + e^c - 1) - c] - e^c \cdot \Big\{ \exp[\log(t + e^c - 1) - c] - 1 \Big\} \\ &= (t + e^c - 1) \cdot [\log(t + e^c - 1) - c] + 1 - t, \qquad t \in]1 - e^c,\infty[\,. \end{aligned} \tag{180}$$

Clearly, one has $\varphi^{(c)}_{1,1}(1)=0$, $\varphi^{(c)}_{1,1}(\infty)=\infty$, $\varphi^{(c)\prime}_{1,1}(1)=0$, $\varphi^{(c)\prime}_{1,1}(1-e^{c})=-\infty$ and $\varphi^{(c)\prime}_{1,1}(\infty)=\infty$.
The corresponding divergence $D_{\varphi^{(c)}_{1,1}}(\mathbb{Q},\mathbb{P})$ has been recently used in Broniatowski et al. [63] for the important task of testing mixtures of probability distributions; in fact, in order to get considerable comfort in testing mixture-type hypotheses against corresponding marginal-type alternatives, they employ choices $c>0$ since then $\varphi^{(c)}_{1,1}(t)$ is finite especially for some range of negative values $t<0$. The latter feature is also valid for the divergence generator $\varphi_{bw,\beta,\widetilde{c}}$ in the next example (cf. (182) below).

Example 41: For $\beta\in]0,1]$, $\widetilde{c}\in]0,\infty[$ and $]a_{F_{bw,\beta,\widetilde{c}}},b_{F_{bw,\beta,\widetilde{c}}}[=]1-\frac{1}{\beta},\infty[$ we define

$$F_{bw,\beta,\widetilde{c}}(t) \;:=\; \begin{cases} \frac{\widetilde{c}}{2\beta}\cdot\Big(1-\frac{1}{(\beta\cdot t+1-\beta)^2}\Big), & \text{if } \; t\in]1-\frac{1}{\beta},\infty[,\\ -\infty, & \text{if } \; t\in]-\infty,1-\frac{1}{\beta}]. \end{cases}$$

Clearly, $\mathcal{R}(F_{bw,\beta,\widetilde{c}})=\,]-\infty,\frac{\widetilde{c}}{2\beta}[$ and $0\in int(\mathcal{R}(F_{bw,\beta,\widetilde{c}}))$. Moreover, $F_{bw,\beta,\widetilde{c}}(\cdot)$ is strictly increasing and smooth on $]1-\frac{1}{\beta},\infty[$, and thus, $F_{bw,\beta,\widetilde{c}}\in\mathfrak{F}$. Since $F_{bw,\beta,\widetilde{c}}(1)=0$, let us choose the natural anchor point $c:=0$, which leads to $]\lambda_-,\lambda_+[=int(\mathcal{R}(F_{bw,\beta,\widetilde{c}}))=\,]-\infty,\frac{\widetilde{c}}{2\beta}[$ and $]t^{sc}_-,t^{sc}_+[=]a_{F_{bw,\beta,\widetilde{c}}},b_{F_{bw,\beta,\widetilde{c}}}[=]1-\frac{1}{\beta},\infty[$. By using $F^{-1}_{bw,\beta,\widetilde{c}}(x)=\frac{1}{\beta}\cdot\Big\{\frac{1}{\sqrt{1-2\beta\cdot x/\widetilde{c}}}+\beta-1\Big\}$ for $x\in int(\mathcal{R}(F_{bw,\beta,\widetilde{c}}))$, we can derive from formula (162) (see also (164)) for all $\beta\in]0,1]$ and $z\in\mathbb{R}$

$$\Lambda_{bw,\beta,\widetilde{c}}(z):=\Lambda^{(0)}_{bw,\beta,\widetilde{c}}(z) \;=\; \begin{cases} -(\frac{1}{\beta}-1)\cdot z+\frac{\widetilde{c}}{\beta^2}\cdot\Big\{1-\sqrt{1-\frac{2\beta}{\widetilde{c}}\cdot z}\;\Big\}, & \text{if } z\in\,]-\infty,\frac{\widetilde{c}}{2\beta}],\\ \infty, & \text{else}. \end{cases} \tag{181}$$

Notice that $\Lambda_{bw,\beta,\widetilde{c}}(0)=0$. Moreover, $\Lambda_{bw,\beta,\widetilde{c}}(-\infty)=\infty$, $\Lambda_{bw,\beta,\widetilde{c}}(\frac{\widetilde{c}}{2\beta})=\frac{\widetilde{c}\cdot(\beta+1)}{2\beta^2}$, $\Lambda'_{bw,\beta,\widetilde{c}}(-\infty)=-\frac{1-\beta}{\beta}<0$ and $\Lambda'_{bw,\beta,\widetilde{c}}(\frac{\widetilde{c}}{2\beta})=\infty$. Furthermore, from formula (163) (see also (165)) we can straightforwardly deduce for all $t\in\mathbb{R}$

$$\varphi_{bw,\beta,\widetilde{c}}(t):=\varphi^{(0)}_{bw,\beta,\widetilde{c}}(t)\;:=\;\begin{cases}\widetilde{c}\cdot\frac{(t-1)^2}{2(\beta\cdot t+1-\beta)}\;\in\;[0,\infty[, & \text{if } t\in]1-\frac{1}{\beta},\infty[,\\ \infty, & \text{if } \; t\in]-\infty,1-\frac{1}{\beta}].\end{cases} \tag{182}$$

Note that $1-\frac{1}{\beta}<0$ so that negative t are allowed here. For $t\geq 0$, $\varphi_{bw,\beta,\widetilde{c}}(t)$ is known as Rukhin's generator (cf. [312], see e.g. also Marhuenda et al. [247], Pardo [282]). Obviously, one has $\varphi_{bw,\beta,\widetilde{c}}(1)=0$, $\varphi'_{bw,\beta,\widetilde{c}}(1)=0$, $\varphi'_{bw,\beta,\widetilde{c}}(1-\frac{1}{\beta})=-\infty$ and $\varphi'_{bw,\beta,\widetilde{c}}(\infty)=\frac{\widetilde{c}}{2\beta}$. From the generator $\varphi_{bw,\beta,\widetilde{c}}$ given in (182), we build the corresponding divergence (cf. (4))

$$\begin{aligned} D_{\varphi_{bw,\beta,\widetilde{c}}}(\mathbf{Q},\mathbf{P}) &= \widetilde{c}\cdot\sum_{k=1}^{K}p_k\cdot\frac{(\frac{q_k}{p_k}-1)^2}{2(\beta\cdot\frac{q_k}{p_k}+1-\beta)}\\ &=\frac{\widetilde{c}}{2}\cdot\sum_{k=1}^{K}\frac{(q_k-p_k)^2}{\beta\cdot q_k+(1-\beta)\cdot p_k}, \qquad \text{if } \mathbf{P}\in\mathbb{R}^K \text{ and } \mathbf{Q}\in\mathbb{R}^K \text{ with } \mathbf{Q}\in]\,\mathbf{P}\cdot(1-\frac{1}{\beta}),\infty[\text{ component-wise;} \end{aligned} \tag{183}$$

for the special subcase $\widetilde{c}=1$ and $\mathbf{Q}\in\mathbb{R}^K_{>0}$, $D_{\varphi_{bw,\beta,1}}(\mathbf{Q},\mathbf{P})$ can be interpreted as — "non-probability version" of — the well-known blended weight chi-square divergence of Lindsay [221] (see e.g. also Basu & Lindsay [35], Györfy & Vajda [148], Basu et al. [36]). The special case $\widetilde{c}=1$ and $\beta=\frac{1}{2}$ for probability vectors, i.e. $D_{\varphi_{bw,1/2,1}}(\mathbb{Q},\mathbb{P})$, is equal to (a multiple of the matrix-vector-converted (cf. Remark 19)) *Sanghvi's genetic difference measure* [316] and equal to the double of the so-called *(squared) Vincze-Le Cam distance* (cf. Vincze [384], Le Cam [212], see also e.g. Topsoe [360] — who used the alternative naming *triangular discrimination* — and Vajda [373]); this divergence $D_{\varphi_{bw,1/2,1}}(\mathbb{Q},\mathbb{P})$ has been used e.g. in Liu et al. [227] for a machine learning context of detecting salient objects, where $\mathbb{Q}$ and $\mathbb{P}$ are appropriate histograms of RGB color.

Remark 42: (a) By straightforward calculations, one can show that $\varphi_{bw,1,\widetilde{c}}$ (i.e. with the choice $\beta=1$) is equal to the $\widetilde{c}-fold$ power-divergence generator $\varphi_{\gamma,\widetilde{c}}=\widetilde{c}\cdot\varphi_{\gamma}$ (cf. (43)) with $\gamma=-1$; the corresponding divergence $D_{\varphi_{bw,1,\widetilde{c}}}(\mathbf{Q},\mathbf{P})$ is thus equal to the power divergence $D_{\widetilde{c}\cdot\varphi_{-1}}(\mathbf{Q},\mathbf{P})$ (cf. (44)) which is nothing but the — "non-probability version" — of Neyman's chi-square divergence.
(b) For the case $\beta=0$ — which has been excluded in Example 41 for technical brevity — the divergence generator $\varphi_{bw,0,\widetilde{c}}$ corresponds to $\widetilde{c}-fold$ power-divergence generator $\varphi_{\gamma,\widetilde{c}}$ with $\gamma=2$; the corresponding divergence $D_{\varphi_{bw,0,\widetilde{c}}}(\mathbf{Q},\mathbf{P})$ is thus equal to the power divergence $D_{\widetilde{c}\cdot\varphi_2}(\mathbf{Q},\mathbf{P})$ (cf. (44)) which is nothing but the — "non-probability version" — of Pearson's (i.e. the *classical*) chi-square divergence.

Example 43: Let us give an interesting *generalization* of the Kullback-Leibler case of Example 40(a). For $\widetilde{c} > 0$ and $\alpha \in]-1,0[\cup]0,\infty[$ let us define

$$F_{gKL,\alpha,\widetilde{c}}(t) \quad := \quad \begin{cases} \widetilde{c} \cdot \log\left(\frac{(1+\alpha)\cdot t}{1+\alpha\cdot t}\right), & \text{if } \{ \ \alpha \in]0,\infty[\text{ and } t \in]0,\infty[\ \} \text{ or } \{ \ \alpha \in]-1,0[\text{ and } t \in]0,-\frac{1}{\alpha}[\ \}, \\ -\infty, & \text{if } \alpha \in]-1,0[\cup]0,\infty[\text{ and } t \in]-\infty,0], \\ \infty, & \text{if } \alpha \in]-1,0[\text{ and } t \in \ [-\frac{1}{\alpha},\infty[, \end{cases}$$

(notice that $\lim_{\alpha\to 0_+} F_{gKL,\alpha,\widetilde{c}}(t) = F_{1,\widetilde{c}}(t)$, cf. Example 40(a)). Clearly, $]a_{F_{gKL,\alpha,\widetilde{c}}}, b_{F_{gKL,\alpha,\widetilde{c}}}[:=]0,\infty[$ for $\alpha \in]0,\infty[$ and $]a_{F_{gKL,\alpha,\widetilde{c}}}, b_{F_{gKL,\alpha,\widetilde{c}}}[:=]0,-\frac{1}{\alpha}[$ for $\alpha \in]-1,0[$. Moreover, $\mathcal{R}(F_{gKL,\alpha,\widetilde{c}}) =]-\infty, \widetilde{c}\cdot\log(1+\frac{1}{\alpha})[$ for $\alpha \in]0,\infty[$ and $\mathcal{R}(F_{gKL,\alpha,\widetilde{c}}) =]-\infty,\infty[$ for $\alpha \in]-1,0[$. Furthermore, $F_{gKL,\alpha,\widetilde{c}}(\cdot)$ is strictly increasing and smooth on the respective $]a_{F_{gKL,\alpha,\widetilde{c}}}, b_{F_{gKL,\alpha,\widetilde{c}}}[$, and thus, $F_{gKL,\alpha,\widetilde{c}} \in \mathfrak{F}$. Since $F_{gKL,\alpha,\widetilde{c}}(1) = 0$, let us choose the natural anchor point $c := 0$, which leads to $]\lambda_-,\lambda_+[= int(\mathcal{R}(F_{gKL,\alpha,\widetilde{c}})) =]-\infty, \widetilde{c}\cdot\log(1+\frac{1}{\alpha})[$ and $]t_-^{sc}, t_+^{sc}[=]0,\infty[$ for the case $\alpha \in]0,\infty[$, respectively, to $]\lambda_-,\lambda_+[= int(\mathcal{R}(F_{gKL,\alpha,\widetilde{c}})) =]-\infty,\infty[$ and $]t_-^{sc}, t_+^{sc}[=]0,-\frac{1}{\alpha}[$ for the case $\alpha \in]-1,0[$. By employing $F^{-1}_{gKL,\alpha,\widetilde{c}}(x) = \frac{1}{(1+\alpha)\cdot e^{-x/\widetilde{c}}-\alpha}$ for $x \in]\lambda_-,\lambda_+[$, one can deduce from formula (162) (see also (164))

$$\begin{aligned} &\Lambda_{gKL,\alpha,\widetilde{c}}(z) := \Lambda^{(0)}_{gKL,\alpha,\widetilde{c}}(z) \\ &:= \begin{cases} \int_0^z F^{-1}_{gKL,\alpha,\widetilde{c}}(u)\,du = -\frac{\widetilde{c}}{\alpha}\cdot\log((1+\alpha)-\alpha\cdot e^{z/\widetilde{c}}), & \text{if } \alpha \in]0,\infty[\text{ and } z \in]-\infty, \widetilde{c}\cdot\log(1+\frac{1}{\alpha})[, \\ \int_0^z F^{-1}_{gKL,\alpha,\widetilde{c}}(u)\,du = -\frac{\widetilde{c}}{\alpha}\cdot\log((1+\alpha)-\alpha\cdot e^{z/\widetilde{c}}), & \text{if } \alpha \in]-1,0[\text{ and } z \in]-\infty,\infty[, \\ \infty, & \text{if } \alpha \in]0,\infty[\text{ and } z \in \ [\widetilde{c}\cdot\log(1+\frac{1}{\alpha}),\infty[, \end{cases} \end{aligned} \tag{184}$$

for which there holds $\Lambda_{gKL,\alpha,\widetilde{c}}(0) = 0$ and $\Lambda_{gKL,\alpha,\widetilde{c}}(-\infty) = -\frac{\widetilde{c}}{\alpha}\cdot\log(1+\alpha)$ for $\alpha \in]-1,0[\cup]0,\infty[$, as well as $\Lambda_{gKL,\alpha,\widetilde{c}}(\widetilde{c}\cdot\log(1+\frac{1}{\alpha})) = \infty$ for $\alpha \in]0,\infty[$ and $\Lambda_{gKL,\alpha,\widetilde{c}}(\infty) = \infty$ for $\alpha \in]-1,0[$. The corresponding derivative satisfies $\Lambda'_{gKL,\alpha,\widetilde{c}}(-\infty) = 0$ for $\alpha \in]-1,0[\cup]0,\infty[$, as well as $\Lambda'_{gKL,\alpha,\widetilde{c}}(\widetilde{c}\cdot\log(1+\frac{1}{\alpha})) = \infty$ for $\alpha \in]0,\infty[$ and $\Lambda'_{gKL,\alpha,\widetilde{c}}(\infty) = -\frac{1}{\alpha}$ for $\alpha \in]-1,0[$. Furthermore, from formula (163) (see also (165)) one can derive

$$\begin{aligned} &\varphi_{gKL,\alpha,\widetilde{c}}(t) := \varphi^{(0)}_{gKL,\alpha,\widetilde{c}}(t) \\ &:= \begin{cases} \widetilde{c}\cdot\left[t\cdot\log t + (t+\frac{1}{\alpha})\cdot\log\left(\frac{1+\alpha}{1+\alpha\cdot t}\right)\right] \ \in \ [0,\infty[, & \text{if } \{ \ \alpha \in]0,\infty[\text{ and } t \in]0,\infty[\ \} \text{ or } \{ \ \alpha \in]-1,0[\text{ and } t \in]0,-\frac{1}{\alpha}[\ \}, \\ \frac{\widetilde{c}}{\alpha}\cdot\log(1+\alpha) \ \in]0,\infty[, & \text{if } \alpha \in]-1,0[\cup]0,\infty[\text{ and } t = 0, \\ \infty, & \text{if } \alpha \in]-1,0[\cup]0,\infty[\text{ and } t \in]-\infty,0[, \\ \infty, & \text{if } \alpha \in]-1,0[\text{ and } t \in \ [-\frac{1}{\alpha},\infty[; \end{cases} \end{aligned} \tag{185}$$

the first line in (185) can be proved by

$$\begin{aligned} \varphi_{gKL,\alpha,\widetilde{c}}(t) &:= \varphi^{(0)}_{gKL,\alpha,\widetilde{c}}(t) := t\cdot F_{gKL,\alpha,\widetilde{c}}(t) - \int_0^{F_{gKL,\alpha,\widetilde{c}}(t)} F^{-1}_{gKL,\alpha,\widetilde{c}}(u)\,du \\ &= \widetilde{c}\cdot t\cdot\log\left(\frac{(1+\alpha)\cdot t}{1+\alpha\cdot t}\right) + \frac{\widetilde{c}}{\alpha}\cdot\log\left((1+\alpha)-\alpha\cdot\exp\left[\log\left(\frac{(1+\alpha)\cdot t}{1+\alpha\cdot t}\right)\right]\right) \\ &= \widetilde{c}\cdot\left[t\cdot\log t + t\cdot\log\left(\frac{1+\alpha}{1+\alpha\cdot t}\right) + \frac{1}{\alpha}\cdot\log\left(\frac{1+\alpha}{1+\alpha\cdot t}\right)\right]. \end{aligned} \tag{186}$$

Obviously, one has $\varphi_{gKL,\alpha,\widetilde{c}}(1) = 0$, $\varphi'_{gKL,\alpha,\widetilde{c}}(1) = 0$, $\varphi'_{gKL,\alpha,\widetilde{c}}(0) = -\infty$ for $\alpha \in]-1,0[\cup]0,\infty[$. Moreover, for $\alpha \in]0,\infty[$ there holds $\varphi_{gKL,\alpha,\widetilde{c}}(\infty) = \infty$, and $\varphi'_{gKL,\alpha,\widetilde{c}}(\infty) = \widetilde{c}\cdot\log(1+\frac{1}{\alpha})$, whereas for $\alpha \in]-1,0[$ we obtain $\varphi_{gKL,\alpha,\widetilde{c}}(-\frac{1}{\alpha}) = \infty$, and $\varphi'_{gKL,\alpha,\widetilde{c}}(-\frac{1}{\alpha}) = \infty$.

From the generator $\varphi_{gKL,\alpha,\widetilde{c}}$ given in (185), we build the corresponding divergence (cf. (4))

$$D_{\varphi_{gKL,\alpha,\widetilde{c}}}(\mathbf{Q},\mathbf{P}) = \widetilde{c}\cdot\Big\{\sum_{k=1}^K q_k\cdot\log\Big(\frac{q_k}{(1-\frac{1}{1+\alpha})\cdot q_k + \frac{1}{1+\alpha}\cdot p_k}\Big) + \frac{1}{\alpha}\cdot\sum_{k=1}^K p_k\cdot\log\Big(\frac{p_k}{(1-\frac{1}{1+\alpha})\cdot q_k + \frac{1}{1+\alpha}\cdot p_k}\Big)\Big\}, \tag{187}$$

$$\text{if } \{ \ \alpha \in]0,\infty[, \ \mathbf{P} \in \mathbb{R}^K_{>0} \text{ and } \mathbf{Q} \in \mathbb{R}^K_{\geq 0} \ \} \text{ or } \{ \ \alpha \in]-1,0[, \ \mathbf{P} \in \mathbb{R}^K_{>0} \text{ and } \mathbf{Q} \in \mathbb{R}^K_{\geq 0} \text{ with } \mathbf{Q} \leq -\frac{1}{\alpha}\cdot\mathbf{P} \ \}.$$

Notice that the symmetry $D_{\varphi_{gKL,\alpha,\widetilde{c}}}(\mathbf{Q},\mathbf{P}) = D_{\varphi_{gKL,\alpha,\widetilde{c}}}(\mathbf{P},\mathbf{Q})$ generally holds only if $\mathbf{P},\mathbf{Q} \in \mathbb{R}^K_{>0}$ and $\alpha = 1$; indeed, this special case leads to

$$\varphi_{snKL,\widetilde{c}}(t) := \varphi_{gKL,1,\widetilde{c}}(t) := \begin{cases} \widetilde{c} \cdot \left[t \cdot \log t + (t+1) \cdot \log \left(\frac{2}{t+1} \right) \right] \in [0,\infty[, & \text{if } t \in]0,\infty[, \\ \widetilde{c} \cdot \log 2, & \text{if } t = 0, \\ \infty, & \text{if } t \in]-\infty, 0[, \end{cases} \tag{188}$$

and

$$D_{\varphi_{snKL,\widetilde{c}}}(\mathbf{Q},\mathbf{P}) := D_{\varphi_{gKL,1,\widetilde{c}}}(\mathbf{Q},\mathbf{P}) = \widetilde{c} \cdot \Big\{ \sum_{k=1}^{K} q_k \cdot \log \Big(\frac{2q_k}{q_k + p_k} \Big) + \sum_{k=1}^{K} p_k \cdot \log \Big(\frac{2p_k}{q_k + p_k} \Big) \Big\}, \qquad \mathbf{P} \in \mathbb{R}^K_{>0}, \mathbf{Q} \in \mathbb{R}^K_{\geq 0}. \tag{189}$$

For the special subcase that $\widetilde{c} = 1$ and that $\mathbf{P} = \mathbb{P}$, $\mathbf{Q} = \mathbb{Q}$ are probability vectors, the divergence (189) can be rewritten as sum of two Kullback-Leibler divergences (cf. (44))

$$D_{\varphi_{snKL,1}}(\mathbb{Q},\mathbb{P}) = D_{\varphi_1}(\mathbb{Q},(\mathbb{Q}+\mathbb{P})/2) + D_{\varphi_1}(\mathbb{P},(\mathbb{Q}+\mathbb{P})/2), \qquad \mathbb{P} \in \mathbb{S}^K_{>0}, \mathbb{Q} \in \mathbb{S}^K_{\geq 0}, \tag{190}$$

which is the well-known (cf. Burbea & Rao [68], Lin [219], Pardo & Vajda [284], Topsoe [360], Endres & Schindelin [120], Vajda [373], Sason [317]) *Jensen-Shannon divergence* (being also called symmetrized and normalized Kullback-Leibler divergence, symmetrized and normalized relative entropy, capacitory discrimination); this is equal to the $(2\log 2)-$fold of a special (namely, equally-weighted two-population) case of the Sibson information radius of order 1 (cf. [334]) which has also been addressed e.g. by Rao [301] for genetic cluster analysis. By the way, for $\alpha > 0$ the divergence $D_{\varphi_{gKL,\alpha,\widetilde{c}}}(\mathbb{Q},\mathbb{P})$ can also be interpreted as a multiple of a special non-equally-weighted Sibson information radius of order 1. In a context of comparison of — not necessarily connected — networks where $\mathbb{Q}$, $\mathbb{P}$ are probability vectors derived from matrices (cf. Remark 19) which are transforms of corresponding graph invariants (e.g. network portraits), the (matrix-equivalent of the) Jensen-Shannon divergence $D_{\varphi_{snKL,1}}(\mathbb{Q},\mathbb{P})$ is also called the *network portrait divergence*, cf. Bagrow and Bollt [28].

There is a vast literature on recent applications of the Jensen-Shannon divergence, for instance it appears exemplarily in Kvitsiani et al. [208] for finding connections between the circuit-level function of different interneuron types in regulating the flow of information and the behavioural functions served by the cortical circuits, in Xu et al. (2014) for browsing and exploration of video sequences, in Jenkinson et al. [168] for the fundamental understanding of the epigenome that leads to a powerful approach for studying its role in disease and aging, in Martin et al. [250] for the implementation of an evolutionary-based global localization filter for mobile robots, in Suo et al. [354] for the revelation of critical regulators of cell identity in mice, in Abante et al. [2] for the detection of biologically significant differences in DNA methylation between alleles associated with local changes in genetic sequences — for a better understanding of the mechanism of complex human diseases, in Afek et al. [5] for revealing mechanisms by which mismatches can recruit transcription factors for modulating replication and repair activities in cells, in Alaiz-Rodriguez & Parnell [10] for the quantification of stability in feature selection and ranking algorithms, in Biau et al. [53] for generative adversarial networks (GANs) in artificial intelligence and machine learning, in Carre et al. [74] for the standardization of brain magnetic resonance (MR) images, in Chakraborty et al. [75] for hierarchical clustering in foreign exchange FOREX markets (e.g. in periods of major international crises), in Chong et al. [87] as part of a web-based platform for comprehensive analysis of microbiome data outputs, in Cui et al. [101] for modelling latent friend recommendation in online social media, in Gholami & Hodtani [134] for refinements of safety-and-security-targeted location verification systems in wireless communication networks (e.g in Intelligent Transportation Systems (ITSs) and vehicular technology), in Guo & Yuan [146] for accurate abnormality classification in semi-supervised Wireless Capsule Endoscopy (WCE) for digestive system cancer diagnosis, in Jiang et al. [169] for the training of deep neural discriminative and generative networks used for designing and evaluating photonic devices, in Kartal et al. [186] for uncovering the relationship between some genomic features and cell type-specific methylome diversity, in Laszlovszky et al. [210] for investigating mechanisms of basal forebrain neurons which modulate synaptic plasticity,cortical processing, brain states and oscillations, in Lawson et al. [211] for the improved understanding of some genetic circuits that allow cancer cells to evade destruction by the host immune system, in Li et al. [215] for the search of causes of the progressive neurodevelopmental disorder Rett syndrome, in Machado et al. [239] for discovering relations between distinct RNA viruses (including SARS-CoV-2), in Mohammadi et al. [261] for the identification of cell states and their underlying topology, in Mohanty et al. [262] for the design of implantable nanophotonic (i.e. chip-scale optical circuit type) silicon probes for sub-millisecond deep-brain optical stimulation — e.g. for the purpose of gaining a deeper understanding of the neural code, in Perera et al. [290] for the quantification of the level of rationality in supply chain networks, in Pierri et al. [294] for the study of growth of malicious/misleading information in some social media diffusion networks, in Rabadan et al. [299] for the identification of gene mutations that lead to the genesis and progression of tumors, in Reiter et al. [306] for quantifying metastatic phylogenetic diversity, in Van de Sande et al. [378] as part of a computational toolbox for single-cell gene regulatory network analysis, in Skinnider et al. [337] for the prediction of the chemical structures of genomically encoded antibiotics — in order to find means against the looming global crisis of antibiotic resistance, in Tuo et al. [367] for the detection of high-order single nucleotide polymorphism (SNP) interactions, in Uttam et al. [370] for predicting the risk of colorectal cancer recurrence and inferring associated tumor microenvironment networks, in Zhang et al. [421] for incipient fault (namely, crack) detection, in Zhi et al. [425] for the strengthening of information-centric networks against interest flooding attack (IFAs), in Acera Mateos et al. [3] for deep-learning classification of SARS-CoV-2 and co-infecting RNA viruses, in Avsec et al. [24] for uncovering the motifs and syntax of cis-regulatory sequences in genomics data, in Barennes et al. [32] for comparing the accuracy of current T cell receptor sequencing methods employed for the understanding of adaptive immune responses, in Chen et al. [79] for clustering high-dimensional microbial data from RNA sequencing, in Chen et al. [84] for investigating key aspects of effective vocal social communication, in Koldobskiy et al. [193] for investigations of genetic and epigenetic drivers of paediatric acute lymphoblastic leukaemia, in McGinnis et al. [258] for evaluating RNA sequencing of pooled blood cell samples, in Mühlroth & Grottke [268] for the detection of emerging trends and technologies through artificial intelligence techniques, in Necci et al. [272] for the assessment of protein intrinsic disorder predictions, in Okada et al. [277] for the identification of genetic factors that cause individual differences in whole lymphocyte profiles and their changes after vaccination, and in Zhang et al. [422] for the learning of functional magnetic resonance imaging (fMRI) time-series in a brain disease diagnosis context.

Remark 44: Let us transform $\varphi_{gSH,\alpha}(t) := \frac{1-t}{\alpha} \cdot \log(1+\alpha) - \varphi_{gKL,\alpha,1}(t) = -t \cdot \log t + \frac{1}{\alpha} \cdot (1+\alpha \cdot t) \cdot \log(1+\alpha \cdot t) - \frac{1}{\alpha} \cdot (1+\alpha) \cdot t \cdot \log(1+\alpha)$ (for $t \in [0,1]$). The function $\varphi_{gSH,\alpha}(\cdot)$ is strictly concave on $[0,1]$ with $\varphi_{gSH,\alpha}(0) = \varphi_{gSH,\alpha}(1) = 0$. Hence, for probability vectors $\mathbb{Q} = (q_k)_{k=1,\ldots,K}$, the $\varphi-$entropy $\sum_{k=1}^{K} \varphi_{gSH,\alpha}(q_k)$ is Kapur's [183] generalization of the Shannon entropy (which corresponds to $\alpha = 0$ in the limit) whose maximization has been connected with generalizations of the Bose-Einstein statistics and the Fermi-Dirac statistics e.g. in Kapur & Kesavan [185].

Example 45: Let us fix any $z_1, z_2 \in \mathbb{R}$, $p \in]0,1[$ which satisfy $z_1 < 1 < z_2$ and $z_1 \cdot p + z_2 \cdot (1-p) = 1$ (and thus $p = \frac{z_2-1}{z_2-z_1}$). On $]a_{F_{twop}}, b_{F_{twop}}[:=]z_1, z_2[$ we define

$$\begin{aligned} F_{twop}(t) \quad &:= \quad \frac{1}{z_2 - z_1} \cdot \log\left(\frac{(t-z_1)\cdot p}{(z_2-t)\cdot(1-p)}\right) \\ &= \quad \frac{1}{z_2 - z_1} \cdot \log\left(\frac{(t-z_1)\cdot(z_2-1)}{(z_2-t)\cdot(1-z_1)}\right), \qquad t \in]z_1, z_2[, \end{aligned}$$

where for the last equality we have used the above constraint (in order to obtain a two-parameter representation). Straightforwardly, we have $\mathcal{R}(F_{twop}) =]-\infty, \infty[$. Moreover, $F_{twop}(\cdot)$ is strictly increasing and smooth on $]0,\infty[$, and thus, $F_{twop} \in \mathfrak{F}$. Since $F_{twop}(1) = 0$, let us choose the natural anchor point $c := 0$, which leads to $]\lambda_-, \lambda_+[= int(\mathcal{R}(F_{snKL,\widetilde{c}})) =]-\infty, \infty[$ and $]t_-^{sc}, t_+^{sc}[=]z_1, z_2[$. By using

$$F_{twop}^{-1}(x) = \frac{p \cdot z_1 + (1-p)\cdot z_2 \cdot e^{(z_2-z_1)\cdot x}}{p + (1-p)\cdot e^{(z_2-z_1)\cdot x}}, \qquad x \in]-\infty, \infty[,$$

we derive from formula (162) (see also (164))

$$\Lambda_{twop}(z) := \Lambda_{twop}^{(0)}(z) := \int_0^z F_{twop}^{-1}(u)du = \log\left(p \cdot e^{z_1 \cdot z} + (1-p)\cdot e^{z_2 \cdot z}\right), \qquad z \in]-\infty, \infty[, \tag{191}$$

which has the properties $\Lambda_{twop}(0) = 0$, $\Lambda_{twop}(-\infty) = \infty \cdot 1_{]-\infty,0[}(z_1) - \infty \cdot 1_{]0,\infty[}(z_1) + \log p \cdot 1_{\{0\}}(z_1)$, $\Lambda_{twop}(\infty) = \infty$, $\Lambda'_{twop}(-\infty) = z_1$ and $\Lambda'_{twop}(\infty) = z_2$. Furthermore, from formula (163) (see also (165)) we deduce

$$\varphi_{twop}(t) := \varphi_{twop}^{(0)}(t) := \begin{cases} \frac{t-z_1}{z_2-z_1} \cdot \log\left(\frac{(t-z_1)\cdot(z_2-1)}{(z_2-t)\cdot(1-z_1)}\right) - \log\left(\frac{z_2-1}{z_2-t}\right) \in [0,\infty[, & \text{if } t \in]0,\infty[, \\ \log\left(\frac{z_2-z_1}{z_2-1}\right), & \text{if } t = z_1, \\ \log\left(\frac{z_2-z_1}{1-z_1}\right), & \text{if } t = z_2, \\ \infty, & \text{if } t \in]-\infty, z_1[\cup]z_2,\infty[; \end{cases} \tag{192}$$

the first line in (192) can be proved by

$$\varphi_{twop}(t) := \varphi_{twop}^{(0)}(t) := t \cdot F_{twop}(t) - \int_0^{F_{twop}(t)} F_{twop}^{-1}(u) du$$
$$= \frac{t}{z_2 - z_1} \cdot \log\left(\frac{(t-z_1)\cdot p}{(z_2-t)\cdot(1-p)}\right)$$
$$- \log\left(p \cdot \left(\frac{(t-z_1)\cdot p}{(z_2-t)\cdot(1-p)}\right)^{\frac{z_1}{z_2-z_1}} + (1-p)\cdot\left(\frac{(t-z_1)\cdot p}{(z_2-t)\cdot(1-p)}\right)^{\frac{z_2}{z_2-z_1}}\right)$$
$$= \frac{t-z_1}{z_2-z_1} \cdot \log\left(\frac{(t-z_1)\cdot p}{(z_2-t)\cdot(1-p)}\right) - \log\left(\frac{(z_2-z_1)\cdot p}{z_2-t}\right) \tag{193}$$
$$= \frac{t-z_1}{z_2-z_1} \cdot \log\left(\frac{(t-z_1)\cdot(z_2-1)}{(z_2-t)\cdot(1-z_1)}\right) - \log\left(\frac{z_2-1}{z_2-t}\right), \qquad t \in]z_1, z_2[\,,$$

where for the last equality we have used the above constraint (to obtain a two-parameter representation). Straightforwardly, one has $\varphi_{twop}(1) = 0$, $\varphi'_{twop}(1) = 0$, $\varphi'_{twop}(z_1) = -\infty$ and $\varphi'_{twop}(z_2) = \infty$.

From the generator φ_{twop} given in (192), we build the corresponding divergence (cf. (4))

$$D_{\varphi_{twop}}(\mathbf{Q},\mathbf{P}) = \sum_{k=1}^{K} \frac{q_k - z_1\cdot p_k}{z_2 - z_1} \cdot \log\left(\frac{(z_2-1)\cdot(q_k - z_1\cdot p_k)}{(1-z_1)\cdot(z_2\cdot p_k - q_k)}\right) - \sum_{k=1}^{K} p_k \cdot \log\left(\frac{(z_2-1)\cdot p_k}{z_2\cdot p_k - q_k}\right). \tag{194}$$

It is known that some types of robustness properties of minimum-divergence estimators are connected with the *boundedness* of the derivative φ' of the divergence generator φ; this property is satisfied for the next Example 46 (and its $W-$concerning continuation in Example 55), which leads to the new classes of divergences (199), (203) and (208):

Example 46: (a) For any parameter-quadrupel $\alpha, \beta_1, \beta_2, \widetilde{c} \in]0,\infty[$ with $\beta_1 < \beta_2$, we choose

$$]a_F, b_F[\ :=]a_{F_{\alpha,\beta_1,\beta_2,\widetilde{c}}}, b_{F_{\alpha,\beta_1,\beta_2,\widetilde{c}}}[\ := \ \left]1 - \alpha \cdot \frac{(\beta_1-\beta_2)^2 + \beta_1^2 + \beta_1\cdot\beta_2}{2\beta_1\cdot\beta_2\cdot(\beta_2-\beta_1)}, \infty\right[\ \ni 1$$

and define with $\breve{\theta} := 1 + \alpha \cdot \left(\frac{1}{\beta_2} - \frac{1}{\beta_1}\right) < 1$

$$F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t) \ := \ \begin{cases} \widetilde{c}\cdot\frac{\beta_1-\beta_2}{2} + \frac{\widetilde{c}}{\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1}} \cdot \left(1 - \frac{1}{2}\cdot\sqrt{4 + \left(\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1}\right)^2\cdot(\beta_1+\beta_2)^2}\right), & \text{if } \ t \in]a_F, b_F[\backslash\{\breve{\theta}\}, \\ \widetilde{c}\cdot\frac{\beta_1-\beta_2}{2}, & \text{if } \ t = \breve{\theta} \in]a_F, b_F[, \\ -\widetilde{c}\cdot\beta_1, & \text{if } \ t = a_F, \\ -\infty, & \text{if } \ t \in]-\infty, a_F[. \end{cases} \tag{195}$$

Notice that $\breve{\theta} \in]a_F, b_F[$ if and only if $\beta_1 \in]\frac{\beta_2}{3}, \beta_2[$; if (say) the latter holds, then one has the continuity $\lim_{t\to\breve{\theta}} F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t) = \widetilde{c}\cdot\frac{\beta_1-\beta_2}{2}$. For $\beta_1 \leq \frac{\beta_2}{3}$ one gets $]a_F, b_F[\backslash\{\breve{\theta}\} =]a_F, b_F[$. Returning to the general case, one can see in a straightforward way that $F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\cdot)$ is strictly increasing and that $\mathcal{R}(F_{\alpha,\beta_1,\beta_2,\widetilde{c}}) = [-\widetilde{c}\cdot\beta_1, \widetilde{c}\cdot\beta_1[$. Furthermore, $F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\cdot)$ is smooth on $]a_F, b_F[$, and thus $F_{\alpha,\beta_1,\beta_2,\widetilde{c}} \in \mathfrak{F}$. Since $F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(1) = 0$, let us choose the natural anchor point $c := 0$, which leads to $]\lambda_-, \lambda_+[= int(\mathcal{R}(F_{\alpha,\beta_1,\beta_2,\widetilde{c}}) =]-\widetilde{c}\cdot\beta_1, \widetilde{c}\cdot\beta_1[$ and $]t_-^{sc}, t_+^{sc}[=]a_F, b_F[$. Moreover, it is straightforward to see that the corresponding inverse is

$$F^{-1}_{\alpha,\beta_1,\beta_2,\widetilde{c}}(x) = 1 + \alpha\cdot\left(\frac{1}{\beta_2} - \frac{1}{\beta_1}\right) - \alpha \cdot \frac{\frac{1}{\beta_2} - \frac{1}{\beta_1} - \frac{2x}{\widetilde{c}\cdot\beta_1\cdot\beta_2}}{1 + \frac{x}{\widetilde{c}}\cdot\left(\frac{1}{\beta_2}-\frac{1}{\beta_1}\right) - \frac{x^2}{\widetilde{c}^2\cdot\beta_1\cdot\beta_2}}, \qquad x \in int(\mathcal{R}(F_{\alpha,\beta_1,\beta_2,\widetilde{c}})); \tag{196}$$

from this, we can derive from formula (162) (see also (164)) for all $z \in \mathbb{R}$

$$\Lambda_{\alpha,\beta_1,\beta_2,\widetilde{c}}(z) := \Lambda^{(0)}_{\alpha,\beta_1,\beta_2,\widetilde{c}}(z) \quad = \quad \begin{cases} \breve{\theta}\cdot z - \widetilde{c}\cdot\alpha\cdot\log\left(1 + \frac{z}{\widetilde{c}}\cdot\left(\frac{1}{\beta_2}-\frac{1}{\beta_1}\right) - \frac{z^2}{\widetilde{c}^2\cdot\beta_1\cdot\beta_2}\right), & \text{if } \ z \in]-\widetilde{c}\cdot\beta_1, \widetilde{c}\cdot\beta_1[, \\ -\widetilde{c}\cdot\breve{\theta}\cdot\beta_1 - \widetilde{c}\cdot\alpha\cdot\log\left(2 - 2\frac{\beta_1}{\beta_2}\right), & \text{if } \ z = -\widetilde{c}\cdot\beta_1, \\ \infty, & \text{else.} \end{cases} \tag{197}$$

Notice that $\Lambda_{\alpha,\beta_1,\beta_2,\widetilde{c}}(0)=0$ and $\lim_{z\to\widetilde{c}\cdot\beta_1}\Lambda_{\alpha,\beta_1,\beta_2,\widetilde{c}}(z)=\infty$. Moreover, $\Lambda'_{\alpha,\beta_1,\beta_2,\widetilde{c}}(-\widetilde{c}\cdot\beta_1)=a_F$ and $\Lambda'_{-\widetilde{c}\cdot\beta_1}(\widetilde{c}\cdot\beta_1)=\infty=b_F$ (which have to be interpreted as limits, as usual). To proceed, from formula (163) (see also (165)) we can deduce for all $t\in\mathbb{R}$

$$\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t):=\varphi^{(0)}_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t)=\begin{cases}\widetilde{c}\cdot\alpha\cdot\Big\{\frac{\sqrt{4+(\beta_1+\beta_2)^2\cdot(\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2}-(\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})\cdot(\beta_1-\beta_2)-2}{2}\\ +\log\frac{\sqrt{4+(\beta_1+\beta_2)^2\cdot(\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2}-2}{\beta_1\beta_2\cdot(\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2}\Big\}\ \in[0,\infty[, & \text{if } t\in]a_F,\infty[,\\ \widetilde{c}\cdot\alpha\cdot\Big\{\frac{3\beta_1-\beta_2}{2(\beta_2-\beta_1)}+\log\frac{2(\beta_2-\beta_1)}{\beta_2}\Big\}-\widetilde{c}\cdot\beta_1\cdot(t-a_F)\ \in]0,\infty[, & \text{if } \ t\in]-\infty,a_F].\end{cases}\tag{198}$$

The first subcase in (198) can be proved by

$$\begin{aligned}
&\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t):=\varphi^{(0)}_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t):=t\cdot F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t)\ -\int_0^{F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t)}F^{-1}_{\alpha,\beta_1,\beta_2,\widetilde{c}}(u)\,du\\
&=(t-\breve{\theta})\cdot F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t)+\widetilde{c}\cdot\alpha\cdot\log\Big(1+\frac{F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t)}{\widetilde{c}}\cdot\Big(\frac{1}{\beta_2}-\frac{1}{\beta_1}\Big)-\frac{(F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t))^2}{\widetilde{c}^2\cdot\beta_1\cdot\beta_2}\Big)\\
&=\widetilde{c}\cdot\alpha\cdot\frac{\sqrt{4+(\beta_1+\beta_2)^2\cdot(\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2}\ -\ (\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})\cdot(\beta_1-\beta_2)\ -\ 2}{2}\\
&\quad+\widetilde{c}\cdot\alpha\cdot\log\Bigg(1+\Big[\frac{\beta_1-\beta_2}{2}+\frac{1}{\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1}}\cdot\Big(1-\frac{1}{2}\cdot\sqrt{4+(\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2\cdot(\beta_1+\beta_2)^2}\Big)\Big]\cdot\frac{\beta_1-\beta_2}{\beta_1\cdot\beta_2}\\
&\quad-\Big[\frac{\beta_1-\beta_2}{2}+\frac{1}{\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1}}\cdot\Big(1-\frac{1}{2}\cdot\sqrt{4+(\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2\cdot(\beta_1+\beta_2)^2}\Big)\Big]^2\cdot\frac{1}{\beta_1\cdot\beta_2}\Bigg)
\end{aligned}$$

and some straightforward calculations. The second line in (198) follows by computing $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(a_F)=\widetilde{c}\cdot\alpha\cdot\Big\{\frac{3\beta_1-\beta_2}{2(\beta_2-\beta_1)}+\log\frac{2(\beta_2-\beta_1)}{\beta_2}\Big\}$. Notice that $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(1)=0$, $\varphi'_{\alpha,\beta_1,\beta_2,\widetilde{c}}(1)=0$, $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(-\infty)=\infty$ and $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\infty)=\infty$. Moreover, $\varphi'_{\alpha,\beta_1,\beta_2,\widetilde{c}}(-\infty)=\varphi'_{\alpha,\beta_1,\beta_2,\widetilde{c}}(a_F)=-\widetilde{c}\cdot\beta_1$ and $\varphi'_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\infty)=\widetilde{c}\cdot\beta_1$.
From the generator $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}$ given in (198), we construct the corresponding divergence (cf. (4))

$$\begin{aligned}
&D_{\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}}(\mathbf{Q},\mathbf{P})=\sum_{k=1}^{K}p_k\cdot\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}\Big(\frac{q_k}{p_k}\Big)\\
&=\sum_{k=1}^{K}p_k\cdot\Bigg[\mathbb{1}_{]a_F,\infty[}(\frac{q_k}{p_k})\cdot\widetilde{c}\cdot\alpha\cdot\Big\{\frac{\sqrt{4+(\beta_1+\beta_2)^2\cdot(\frac{1-\frac{q_k}{p_k}}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2}\ -\ (\frac{1-\frac{q_k}{p_k}}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})\cdot(\beta_1-\beta_2)\ -\ 2}{2}\\
&\quad+\log\frac{\sqrt{4+(\beta_1+\beta_2)^2\cdot(\frac{1-\frac{q_k}{p_k}}{\alpha}+\frac{1}{\beta_2}\ -\ \frac{1}{\beta_1})^2}\ -\ 2}{\beta_1\beta_2\cdot(\frac{1-\frac{q_k}{p_k}}{\alpha}+\frac{1}{\beta_2}\ -\ \frac{1}{\beta_1})^2}\Big\}\\
&\quad+\mathbb{1}_{]-\infty,a_F]}(\frac{q_k}{p_k})\cdot\widetilde{c}\cdot\Big\{\alpha\cdot\Big\{\frac{3\beta_1-\beta_2}{2(\beta_2-\beta_1)}+\log\frac{2(\beta_2-\beta_1)}{\beta_2}\Big\}-\beta_1\cdot(\frac{q_k}{p_k}-a_F)\Big\}\Bigg],\qquad \mathbf{P}\in\mathbb{R}^K_{\geq 0},\mathbf{Q}\in\mathbb{R}^K.
\end{aligned}\tag{199}$$

Notice that we can particularly include the case where $p_k=0$ in combination with $q_k\neq 0$, since $\lim_{t\to 0_+}t\cdot\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\frac{1}{t})=\widetilde{c}\cdot\beta_1$ and $\lim_{t\to 0_-}t\cdot\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\frac{1}{t})=-\widetilde{c}\cdot\beta_1$ are both finite, and hence $p_k\cdot\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\frac{q_k}{p_k})=q_k\cdot\frac{p_k}{q_k}\cdot\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\frac{q_k}{p_k})$ stays finite as p_k tends to zero.

(b) For any parameter-quadrupel $\alpha,\beta_1,\beta_2,\widetilde{c}\in]0,\infty[$ with $\beta_1>\beta_2$, one can proceed analogously to (a). Let us start by choosing

$$]a_F,b_F[\ :=]a_{F_{\alpha,\beta_1,\beta_2,\widetilde{c}}},b_{F_{\alpha,\beta_1,\beta_2,\widetilde{c}}}[\ :=\ \Big]-\infty,\,1+\alpha\cdot\frac{(\beta_1-\beta_2)^2+\beta_1\cdot\beta_2+\beta_2^2}{2\beta_1\cdot\beta_2\cdot(\beta_1-\beta_2)},\ \Big[\ \ni 1$$

and defining with the same $\breve{\theta} := 1 + \alpha \cdot \left(\frac{1}{\beta_2} - \frac{1}{\beta_1}\right) > 1$

$$F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t) \ := \ \begin{cases} \widetilde{c} \cdot \frac{\beta_1-\beta_2}{2} + \frac{\widetilde{c}}{\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1}} \cdot \left(1 - \frac{1}{2} \cdot \sqrt{4 + \left(\frac{1-t}{\alpha} + \frac{1}{\beta_2} - \frac{1}{\beta_1}\right)^2 \cdot (\beta_1+\beta_2)^2}\right), & \text{if } \ t \in]a_F, b_F[\backslash\{\breve{\theta}\}, \\ \widetilde{c} \cdot \frac{\beta_1-\beta_2}{2}, & \text{if } \ t = \breve{\theta} \in]a_F, b_F[, \\ \widetilde{c} \cdot \beta_2, & \text{if } \ t = b_F, \\ \infty, & \text{if } \ t \in]b_F, \infty[. \end{cases} \tag{200}$$

Clearly, $\breve{\theta} \in]a_F, b_F[$ if and only if $\beta_1 \in]\beta_2, 3\beta_2[$; if (say) the latter holds, then one gets the continuity $\lim_{t\to\breve{\theta}} F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t) = \widetilde{c} \cdot \frac{\beta_1-\beta_2}{2}$. For $\beta_1 \leq 3\beta_2$ there holds $]a_F, b_F[\backslash\{\breve{\theta}\} =]a_F, b_F[$. Returning to the general case, one can show comfortably that $F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\cdot)$ is strictly increasing and that $\mathcal{R}(F_{\alpha,\beta_1,\beta_2,\widetilde{c}}) =]-\widetilde{c}\cdot\beta_2, \widetilde{c}\cdot\beta_2]$. Moreover, $F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\cdot)$ is smooth on $]a_F, b_F[$, and hence $F_{\alpha,\beta_1,\beta_2,\widetilde{c}} \in \mathfrak{F}$. In face of the validity of $F_{\alpha,\beta_1,\beta_2,\widetilde{c}}(1) = 0$, let us choose the natural anchor point $c := 0$, which amounts to $]\lambda_-, \lambda_+[= int(\mathcal{R}(F_{\alpha,\beta_1,\beta_2,\widetilde{c}}) =]-\widetilde{c}\cdot\beta_2, \widetilde{c}\cdot\beta_2[$ and $]t_-^{sc}, t_+^{sc}[=]a_F, b_F[$. Since the first line in (200) coincides formally with that of (195) (with different $]a_F, b_F[$), the corresponding inverse is formally the same as (196) (with different $]a_F, b_F[$), and hence

$$\Lambda_{\alpha,\beta_1,\beta_2,\widetilde{c}}(z) := \Lambda^{(0)}_{\alpha,\beta_1,\beta_2,\widetilde{c}}(z) \ = \ \begin{cases} \breve{\theta}\cdot z - \widetilde{c}\cdot\alpha\cdot\log\left(1 + \frac{z}{\widetilde{c}}\cdot\left(\frac{1}{\beta_2} - \frac{1}{\beta_1}\right) - \frac{z^2}{\widetilde{c}^2\cdot\beta_1\cdot\beta_2}\right), & \text{if } \ z \in]-\widetilde{c}\cdot\beta_2, \widetilde{c}\cdot\beta_2[, \\ \widetilde{c}\cdot\breve{\theta}\cdot\beta_2 - \widetilde{c}\cdot\alpha\cdot\log\left(2 - 2\frac{\beta_2}{\beta_1}\right), & \text{if } \ z = \widetilde{c}\cdot\beta_2, \\ \infty, & \text{else}. \end{cases} \tag{201}$$

Notice that $\Lambda_{\alpha,\beta_1,\beta_2,\widetilde{c}}(0) = 0$ and $\lim_{z\to-\widetilde{c}\cdot\beta_2} \Lambda_{\alpha,\beta_1,\beta_2,\widetilde{c}}(z) = -\infty$. Furthermore, $\Lambda'_{\alpha,\beta_1,\beta_2,\widetilde{c}}(-\widetilde{c}\cdot\beta_2) = -\infty = a_F$ and $\Lambda'_{-\widetilde{c}\cdot\beta_1}(\widetilde{c}\cdot\beta_2) = b_F$. To proceed, from formula (163) (see also (165)) we can derive — analogously to (198) — for all $t \in \mathbb{R}$

$$\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t) := \varphi^{(0)}_{\alpha,\beta_1,\beta_2,\widetilde{c}}(t) = \begin{cases} \widetilde{c}\cdot\alpha\cdot\Big\{\frac{\sqrt{4+(\beta_1+\beta_2)^2\cdot(\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2} - (\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})\cdot(\beta_1-\beta_2) - 2}{2} & \\ + \log\frac{\sqrt{4+(\beta_1+\beta_2)^2\cdot(\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2} - 2}{\beta_1\beta_2\cdot(\frac{1-t}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2}\Big\} \ \in [0,\infty[, & \text{if } t \in]-\infty, b_F[, \\ \widetilde{c}\cdot\alpha\cdot\Big\{\frac{3\beta_2-\beta_1}{2(\beta_1-\beta_2)} + \log\frac{2(\beta_1-\beta_2)}{\beta_1}\Big\} + \widetilde{c}\cdot\beta_2\cdot(t-b_F) \ \in]0,\infty[, & \text{if } \ t \in [b_F, \infty[, \end{cases} \tag{202}$$

where the last line in (202) follows by calculating $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(b_F) = \widetilde{c}\cdot\alpha\cdot\Big\{\frac{3\beta_2-\beta_1}{2(\beta_1-\beta_2)} + \log\frac{2(\beta_1-\beta_2)}{\beta_1}\Big\}$. Notice that $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(1) = 0$, $\varphi'_{\alpha,\beta_1,\beta_2,\widetilde{c}}(1) = 0$, $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(-\infty) = \infty$ and $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\infty) = \infty$. Furthermore, $\varphi'_{\alpha,\beta_1,\beta_2,\widetilde{c}}(-\infty) = -\widetilde{c}\cdot\beta_2$ and $\varphi'_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\infty) = \varphi'_{\alpha,\beta_1,\beta_2,\widetilde{c}}(b_F) = \widetilde{c}\cdot\beta_2$.
From the generator $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}$ given in (202), we construct the corresponding divergence (cf. (4))

$$\begin{aligned} & D_{\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}}(\mathbf{Q},\mathbf{P}) = \sum_{k=1}^{K} p_k \cdot \varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}\Big(\frac{q_k}{p_k}\Big) \\ & = \sum_{k=1}^{K} p_k \cdot \Bigg[\mathbb{1}_{]-\infty,b_F[}\Big(\frac{q_k}{p_k}\Big)\cdot\widetilde{c}\cdot\alpha\cdot\Bigg\{\frac{\sqrt{4+(\beta_1+\beta_2)^2\cdot(\frac{1-\frac{q_k}{p_k}}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2} - (\frac{1-\frac{q_k}{p_k}}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})\cdot(\beta_1-\beta_2) - 2}{2} \\ & + \log\frac{\sqrt{4+(\beta_1+\beta_2)^2\cdot(\frac{1-\frac{q_k}{p_k}}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2} - 2}{\beta_1\beta_2\cdot(\frac{1-\frac{q_k}{p_k}}{\alpha}+\frac{1}{\beta_2}-\frac{1}{\beta_1})^2}\Bigg\} \\ & + \mathbb{1}_{[b_F,\infty[}\Big(\frac{q_k}{p_k}\Big)\cdot\widetilde{c}\cdot\Big\{\alpha\cdot\Big\{\frac{3\beta_2-\beta_1}{2(\beta_1-\beta_2)} + \log\frac{2(\beta_1-\beta_2)}{\beta_1}\Big\} + \beta_2\cdot\Big(\frac{q_k}{p_k} - b_F\Big)\Big\}\Bigg], \qquad \mathbf{P} \in \mathbb{R}^K_{\geq 0}, \mathbf{Q} \in \mathbb{R}^K. \end{aligned} \tag{203}$$

As above, we can particularly include the case where $p_k = 0$ in combination with $q_k \neq 0$, since $\lim_{t\to 0_+} t\cdot\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\frac{1}{t}) = \widetilde{c}\cdot\beta_2$ and $\lim_{t\to 0_-} t\cdot\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}(\frac{1}{t}) = -\widetilde{c}\cdot\beta_2$ are both finite.

(c) The analysis for the case $\beta_1 = \beta_2 =: \beta$ can be obtained by taking $\lim_{\beta_1\to\beta_2}$ in (a) respectively (b). Alternatively, one can start afresh. Due to its importance and its particularities, we nevertheless state the corresponding results explicitly. To begin with, for any parameter-triple $\alpha, \beta, \widetilde{c} \in]0,\infty[$ we choose

$$]a_F, b_F[\ :=]a_{F_{\alpha,\beta,\widetilde{c}}}, b_{F_{\alpha,\beta,\widetilde{c}}}[\ :=]-\infty, \infty[$$

and define with $\breve{\theta} := 1$

$$F_{\alpha,\beta,\widetilde{c}}(t) \ := \ \begin{cases} \frac{\widetilde{c}\cdot\alpha}{1-t} \cdot \Big(1 - \sqrt{1 + \big(\frac{1-t}{\alpha}\big)^2 \cdot \beta^2}\Big), & \text{if} \ \ t \in]a_F, b_F[\backslash\{\breve{\theta}\}, \\ 0, & \text{if} \ \ t = \breve{\theta}. \end{cases} \tag{204}$$

Clearly, one has the continuity $\lim_{t\to\breve{\theta}} F_{\alpha,\beta,\widetilde{c}}(t) = 0$. Moreover, one can see in a straightforward way that $F_{\alpha,\beta,\widetilde{c}}(\cdot)$ is strictly increasing and that $\mathcal{R}(F_{\alpha,\beta,\widetilde{c}}) =]-\widetilde{c}\cdot\beta, \widetilde{c}\cdot\beta[$. Furthermore, $F_{\alpha,\beta,\widetilde{c}}(\cdot)$ is smooth on $]a_F, b_F[$, and thus $F_{\alpha,\beta,\widetilde{c}} \in \mathfrak{F}$. Since $F_{\alpha,\beta,\widetilde{c}}(1) = 0$, let us choose the natural anchor point $c := 0$, which leads to the choice $]\lambda_-, \lambda_+[= int(\mathcal{R}(F_{\alpha,\beta,\widetilde{c}}) =]-\widetilde{c}\cdot\beta, \widetilde{c}\cdot\beta[$ and $]t_-^{sc}, t_+^{sc}[=]a_F, b_F[=]-\infty, \infty[$. The inverse in (196) collapses to

$$F^{-1}_{\alpha,\beta,\widetilde{c}}(x) = 1 + \alpha \cdot \frac{\frac{2x}{\widetilde{c}\cdot\beta^2}}{1 - \frac{x^2}{\widetilde{c}^2\cdot\beta^2}}, \qquad x \in int(\mathcal{R}(F_{\alpha,\beta,\widetilde{c}})); \tag{205}$$

from this, we can derive from formula (162) (see also (164)) for all $z \in \mathbb{R}$

$$\Lambda_{\alpha,\beta,\widetilde{c}}(z) := \Lambda^{(0)}_{\alpha,\beta,\widetilde{c}}(z) \ = \ \begin{cases} \breve{\theta}\cdot z - \widetilde{c}\cdot\alpha\cdot\log\Big(1 - \frac{z^2}{\widetilde{c}^2\cdot\beta^2}\Big), & \text{if} \ \ z \in]-\widetilde{c}\cdot\beta, \widetilde{c}\cdot\beta[, \\ \infty, & \text{else.} \end{cases} \tag{206}$$

Notice that $\Lambda_{\alpha,\beta,\widetilde{c}}(0) = 0$, $\lim_{z\to-\widetilde{c}\cdot\beta} \Lambda_{\alpha,\beta,\widetilde{c}}(z) = -\infty =$, and $\lim_{z\to\widetilde{c}\cdot\beta} \Lambda_{\alpha,\beta,\widetilde{c}}(z) = \infty$. Furthermore, $\lim_{z\to-\widetilde{c}\cdot\beta} \Lambda'_{\alpha,\beta,\widetilde{c}}(z) = -\infty = a_F$, and $\lim_{z\to\widetilde{c}\cdot\beta} \Lambda'_{\alpha,\beta,\widetilde{c}}(z) = \infty = b_F$. To proceed, the formula (198) (respectively, (202)) collapses to

$$\varphi_{\alpha,\beta,\widetilde{c}}(t) := \varphi^{(0)}_{\alpha,\beta,\widetilde{c}}(t) = \widetilde{c}\cdot\alpha\cdot\Big\{\sqrt{1 + \beta^2\cdot\Big(\frac{1-t}{\alpha}\Big)^2} - 1 + \log\frac{2\cdot\Big(\sqrt{1+\beta^2\cdot\Big(\frac{1-t}{\alpha}\Big)^2} - 1\Big)}{\beta^2\cdot\Big(\frac{1-t}{\alpha}\Big)^2}\Big\} \in [0,\infty[, \ \ t \in]-\infty,\infty[=]a_F, b_F[. \tag{207}$$

Notice that $\varphi_{\alpha,\beta,\widetilde{c}}(1) = 0$, $\varphi'_{\alpha,\beta}(1) = 0$, $\varphi_{\alpha,\beta,\widetilde{c}}(-\infty) = \infty$ and $\varphi_{\alpha,\beta,\widetilde{c}}(\infty) = \infty$. Moreover, $\varphi'_{\alpha,\beta,\widetilde{c}}(-\infty) = \varphi'_{\alpha,\beta,\widetilde{c}}(a_F) = -\widetilde{c}\cdot\beta$ and $\varphi'_{\alpha,\beta,\widetilde{c}}(\infty) = \varphi'_{\alpha,\beta,\widetilde{c}}(b_F) = \widetilde{c}\cdot\beta$.
From the generator $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}$ given in (207), we construct the corresponding divergence (cf. (4))

$$\begin{aligned} D_{\varphi_{\alpha,\beta,\widetilde{c}}}(\mathbf{Q},\mathbf{P}) &= \sum_{k=1}^{K} p_k \cdot \varphi_{\alpha,\beta,\widetilde{c}}\Big(\frac{q_k}{p_k}\Big) \\ &= \widetilde{c}\cdot\alpha\cdot\sum_{k=1}^{K} p_k\cdot\Big\{\sqrt{1+\beta^2\cdot\Big(\frac{1-\frac{q_k}{p_k}}{\alpha}\Big)^2} - 1 + \log\frac{2\cdot\Big(\sqrt{1+\beta^2\cdot\Big(\frac{1-\frac{q_k}{p_k}}{\alpha}\Big)^2} - 1\Big)}{\beta^2\cdot\Big(\frac{1-\frac{q_k}{p_k}}{\alpha}\Big)^2}\Big\}, \qquad \mathbf{P} \in \mathbb{R}^K_{\geq 0}, \mathbf{Q} \in \mathbb{R}^K. \end{aligned} \tag{208}$$

As above, we can particularly include the case where $p_k = 0$ in combination with $q_k \neq 0$, since $\lim_{t\to 0_+} t\cdot\varphi_{\alpha,\beta,\widetilde{c}}(\frac{1}{t}) = \widetilde{c}\cdot\beta$ and $\lim_{t\to 0_-} t\cdot\varphi_{\alpha,\beta,\widetilde{c}}(\frac{1}{t}) = -\widetilde{c}\cdot\beta$ are both finite. This ends the current Example 46.

As a side effect in the above-mentioned Example 46, for fixed $\beta_2, \alpha, \widetilde{c}$ notice the interesting behaviour (e.g. with respect to $int(dom(F)) =]a_F, b_F[$ and the range of φ') as β_1 moves from $]0, \beta_2[$ to β_2 and further to $]\beta_2, \infty[$.

Remark 47: The characterization of the probability distribution ζ in (6) which may result from Theorem 36 — as seen through the above examples — considerably improves other approaches which make use of their identification through the concept of power variance functions of Natural Exponential Families, as developed by Tweedie [369], Morris [267], Letac & Mora [214], and others. This approach has been used in Broniatowski [58] in a similar perspective as developed here, but can not be extended outside the range of power divergences, in contrast with the Examples 41, 43, 45 and 46 which can only be handled as a consequence of Theorem 36.

To continue with our general procedure, suppose now that for a divergence generator φ of interest we have concretely/explicitly found (e.g. by direct calculations or via our $F-$connection in Theorem 36, see also Remark 37) its Fenchel-Legendre transform $\Lambda = \varphi^*$; for this "candidate", in order to achieve the desired representability (6) it remains to verify that

$$\exp(\Lambda(z)) = \int_{\mathbb{R}} e^{z\cdot y}\, d\zeta(y), \qquad z \in \mathbb{R}, \tag{209}$$

for some probability distribution/measure ζ on the real line (the light-tailedness in the sense of finiteness on some open interval containing zero, will be typically guaranteed automatically by the assumptions on φ); of course, this is equivalent to "the existence " of a random variable W whose moment generating function is equal to $\exp(\Lambda)$ (and thus, its cumulant generating function (log moment generating function) is Λ), i.e.

$$\exp(\Lambda(z)) = E_{\mathbb{\Pi}}[\exp(z \cdot W)] \qquad z \in \mathbb{R}, \tag{210}$$

with $\mathbb{\Pi}[W \in \cdot\,] = \mathfrak{C}[\,\cdot\,]$); recall that from this, we need to simulate a sequence $(W_i)_{i\in\mathbb{N}}$ of i.i.d. copies of W which are the crucial building ingredients of $\boldsymbol{\xi}_n^{\mathbf{W}}$ in Theorem 10, respectively, of $\xi_{n,\mathbf{X}}^{w\mathbf{W}}$ in Theorem 14.

For the above-mentioned Examples 39 to 46, we can give explicit solutions to the representabilities (209) respectively (210); this is achieved in the following Examples 48 to 55 (notice that the corresponding supports of $\mathfrak{C}$ are explicitly mentioned in the summarizing Table 1 above):

Example 48: for the power-divergence context of Example 39 we obtain:

(a) Case $\gamma = 0$, $\widetilde{c} > 0$: $\Lambda_{0,\widetilde{c}}(z) = -\widetilde{c} \cdot \log\left(1 - \frac{z}{\widetilde{c}}\right)$ (cf. (172)) is the cumulant generating function of the Gamma distribution $\mathfrak{C} = GAM(\widetilde{c}, \widetilde{c})$ with rate parameter (inverse scale parameter) $\widetilde{c}$ and shape parameter $\widetilde{c}$; hence, $\varphi_{0,\widetilde{c}} \in \Upsilon(]0,\infty[)$.

- Prominent special case $\widetilde{c} = 1$: $\mathfrak{C} = GAM(1,1) = EXP(1)$ is the exponential distribution with mean 1.
- Type: $\mathfrak{C}$ is an infinitely divisible (cf. Proposition 34) continuous distribution with density $f(y) := \frac{\widetilde{c}^{\widetilde{c}} \cdot y^{\widetilde{c}-1} \cdot e^{-\widetilde{c}\cdot y}}{\Gamma(\widetilde{c})} \cdot \mathbb{1}_{]0,\infty[}(y)$ $(y \in \mathbb{R})$.
- Behaviour at zero: $\mathfrak{C}[\,]0,\infty[\,] = \mathbb{\Pi}[W > 0] = 1$.
- Corresponding generator: $\varphi_{0,\widetilde{c}} = \widetilde{c} \cdot \varphi_0$ (cf. (173), (43)) of the $\widetilde{c}-$fold of the reversed Kullback-Leibler divergence (reversed relative entropy) given in the second line of (44).
- Sums: for i.i.d. copies $(W_i)_{i\in\mathbb{N}}$ of W, the probability distribution of $\breve{W} := \sum_{i\in I_k^{(n)}} W_i$ (cf. Remark 11(ii)) is $GAM(\widetilde{c}, \widetilde{c} \cdot card(I_k^{(n)}))$.

(b) Case $\gamma \in]0,1[$, $\widetilde{c} > 0$: $\Lambda_{\gamma,\widetilde{c}}^{(0)}(z) = \frac{\widetilde{c}}{\gamma} \cdot \left\{ \left(\frac{\gamma-1}{\widetilde{c}} \cdot z + 1\right)^{\frac{\gamma}{\gamma-1}} - 1 \right\}$ (cf. (167)) is the cumulant generating function of the Compound-Poisson-Gamma distribution $\mathfrak{C} = C(POI(\theta), GAM(\alpha,\beta))$ with $\theta = \frac{\widetilde{c}}{\gamma} > 0$, rate parameter (inverse scale parameter) $\alpha = \frac{\widetilde{c}}{1-\gamma} > 0$, and shape parameter $\beta = \frac{\gamma}{1-\gamma} > 0$. In other words, W has the comfortably simulable form $W = \sum_{i=1}^{N} \widetilde{W}_i$ [33] for some i.i.d. sequence $(\widetilde{W}_i)_{i\in\mathbb{N}}$ of Gamma $GAM(\alpha,\beta)$ distributed random variables (with parameter-pair (α,β)) and some independent $POI(\theta)-$distributed random variable N. Hence, $\varphi_{\gamma,\widetilde{c}} \in \Upsilon(]0,\infty[)$.

- Type: $\mathfrak{C}$ is an infinitely divisible distribution (cf. Proposition 34), mixture of a one-point distribution at zero and a continuous distribution on $[0,\infty[$, with $\mathfrak{C}[\{0\}] = \mathbb{\Pi}[W = 0] = e^{-\theta}$ and $\mathfrak{C}[B] = \mathbb{\Pi}[W \in B] = \int_B f_{\widetilde{c},\gamma}(u)\, du$ for every (measurable) subset of $]0,\infty[$ having density

$$f_{C(POI(\theta),GAM(\alpha,\beta))}(y) := \frac{\exp\left(-\alpha \cdot y - \theta\right)}{y} \cdot \sum_{k=1}^{\infty} \frac{\theta^k \cdot (\alpha y)^{k\beta}}{k! \cdot \Gamma(k\beta)} \cdot \mathbb{1}_{]0,\infty[}(y) \tag{211}$$

$$= \frac{1}{y} \cdot \exp\left(-\widetilde{c} \cdot \left(\frac{y}{1-\gamma} + \frac{1}{\gamma}\right)\right) \cdot \sum_{k=1}^{\infty} \frac{a_k}{k!} \cdot \widetilde{c}^{k/(1-\gamma)} \cdot \gamma^{-k} \cdot (1-\gamma)^{-k\gamma/(1-\gamma)} \cdot y^{k\gamma/(1-\gamma)} \cdot \mathbb{1}_{[0,\infty[}(y) =: f_{\widetilde{c},\gamma}(y), \quad y \in \mathbb{R},$$

  where $a_k := 1/\Gamma(\frac{k\cdot\gamma}{1-\gamma})$ (see e.g. Aalen [1] with a different parametrization).
- Behaviour at zero: $\mathfrak{C}[\,[0,\infty[\,] = \mathbb{\Pi}[W \geq 0] = 1$, $\mathfrak{C}[\,\{0\}\,] = \mathbb{\Pi}[W = 0] = e^{-\theta}$.
- Corresponding generator: $\varphi_{\gamma,\widetilde{c}}^{(0)} = \widetilde{c} \cdot \varphi_\gamma$ (cf. (168), (43)) of the power divergence given in the third line of (44); recall that the special case $\gamma = \frac{1}{2}$ corresponds to the prominent (multiple of the squared) Hellinger distance.
- Sums: for i.i.d. copies $(W_i)_{i\in\mathbb{N}}$ of W, the probability distribution of $\breve{W} := \sum_{i\in I_k^{(n)}} W_i$ (cf. Remark 11(ii)) is $C(POI(\breve{\theta}), GAM(\alpha,\beta))$ with $\breve{\theta} = \frac{\widetilde{c}\cdot card(I_k^{(n)})}{\gamma} > 0$, $\alpha = \frac{\widetilde{c}}{1-\gamma} > 0$, $\beta = \frac{\gamma}{1-\gamma} > 0$.

(c) Case $\gamma = 2$, $\widetilde{c} > 0$: $\Lambda_{2,\widetilde{c}}^{(0)}(z) = \frac{z^2}{2\widetilde{c}} + z$ (cf. (170)) is the well-known cumulant generating function of the Normal distribution (Gaussian distribution) $\mathfrak{C} = N(1, \frac{1}{\widetilde{c}})$ with mean 1 and variance $\frac{1}{\widetilde{c}}$. Thus, $\varphi_{2,\widetilde{c}} \in \Upsilon(]-\infty,\infty[)$.

- Type: $\mathfrak{C}$ is an infinitely divisible (cf. Proposition 34) continuous distribution with density $f_{N(1,\frac{1}{\widetilde{c}})}(y) := \sqrt{\frac{\widetilde{c}}{2\pi}} \cdot \exp(-\frac{\widetilde{c}\cdot(y-1)^2}{2})$, $(y \in \mathbb{R})$.
- Behaviour at zero: $\mathfrak{C}[\,]0,\infty[\,] = \mathbb{\Pi}[W > 0] = \int_0^\infty f_{N(1,\frac{1}{\widetilde{c}})}(u)\, du \in]0,1[$, $\quad \mathfrak{C}[\,\{0\}\,] = \mathbb{\Pi}[W = 0] = 0$.
- Corresponding generator: $\varphi_{2,\widetilde{c}}^{(0)} = \widetilde{c} \cdot \varphi_2$ (cf. (168), (43)) is the generator of the $\widetilde{c}-$fold of the half Pearson-chisquare divergence given in the sixth line of (44).
- Sums: for i.i.d. copies $(W_i)_{i\in\mathbb{N}}$ of W, the probability distribution of $\breve{W} := \sum_{i\in I_k^{(n)}} W_i$ (cf. Remark 11(ii)) is $N(card(I_k^{(n)}), \frac{card(I_k^{(n)})}{\widetilde{c}})$.

[33] with the usual convention $\sum_{i=1}^{0} \widetilde{W}_i := 0$

(d) Case $\gamma < 0$, $\widetilde{c} > 0$: $\Lambda^{(0)}_{\gamma,\widetilde{c}}(z) = \frac{\widetilde{c}}{\gamma} \cdot \Big\{ \big(\frac{\gamma-1}{\widetilde{c}} \cdot z + 1 \big)^{\frac{\gamma}{\gamma-1}} - 1 \Big\}$ (cf. (167)) is the cumulant generating function of a "tilted (i.e. negatively distorted) stable distribution" $\mathfrak{Q}[\,\cdot\,] = \Pi[W \in \cdot\,]$ of a random variable W, which can be constructed as follows: let Z be an auxiliary random variable (having density f_Z and support $supp(Z) = [0,\infty[$) of a stable law with parameter-quadruple $(\frac{-\gamma}{1-\gamma}, 1, 0, -\frac{\widetilde{c}^{1/(1-\gamma)} \cdot (1-\gamma)^{-\gamma/(1-\gamma)}}{\gamma})$ in terms of the "form-B notation" on p.12 in Zolotarev [428]; by applying a general Laplace-transform result on p.112 of the same text we can deduce

$$M_Z(z) := E_{\Pi}[\exp(z \cdot Z)] = \int_0^{\infty} \exp(z \cdot y) \cdot f_Z(y)\, dy \;=\; \begin{cases} \exp\Big(\frac{\widetilde{c}^{1/(1-\gamma)} \cdot (1-\gamma)^{-\gamma/(1-\gamma)}}{\gamma} \cdot (-z)^{\alpha} \Big), & \text{if } \; z \in]-\infty, 0], \\ \infty, & \text{if } \; z \in]0, \infty[, \end{cases} \tag{212}$$

where $\alpha := -\frac{\gamma}{1-\gamma} \in]0,1[$. Since $0 \notin int(dom(M_Z))$ (and thus, Z does not have light-tails) we have to tilt (dampen) the density in order to extend the effective domain. Accordingly, let W be a random variable having density

$$f_W(y) \;:=\; \frac{\exp\{-\frac{y \cdot \widetilde{c}}{1-\gamma}\}}{\exp\{\widetilde{c}/\gamma\}} \cdot f_Z(y) \cdot \mathbb{1}_{]0,\infty[}(y), \qquad y \in \mathbb{R}, \qquad \text{(cf. (82))}.$$

Then one can straightforwardly deduce from (212) that $\int_0^{\infty} f_W(y)\, dy = 1$ and that

$$M_W(z) := E_{\Pi}[\exp(z \cdot W)] = \int_0^{\infty} \exp(z \cdot y) \cdot f_W(y)\, dy \;=\; \begin{cases} \exp\Big(\frac{\widetilde{c}}{\gamma} \cdot \Big\{ \big(\frac{\gamma-1}{\widetilde{c}} \cdot z + 1 \big)^{\frac{\gamma}{\gamma-1}} - 1 \Big\} \Big), & \text{if } \; z \in]-\infty, \frac{\widetilde{c}}{1-\gamma}], \\ \infty, & \text{if } \; z \in]\frac{\widetilde{c}}{1-\gamma}, \infty[. \end{cases}$$

Hence, $\varphi_{\gamma,\widetilde{c}} \in \Upsilon(]0,\infty[)$.

- Type: $\mathfrak{Q}$ is an infinitely divisible (cf. Proposition 34) continuous distribution with density f_W.
- Behaviour at zero: $\mathfrak{Q}[\,]0,\infty[\,] = \Pi[W > 0] = 1$.
- Corresponding generator: $\varphi^{(0)}_{\gamma,\widetilde{c}} = \widetilde{c} \cdot \varphi_{\gamma}$ (cf. (168), (43)) of the power divergence given in the first line of (44).
- Sums: for i.i.d. copies $(W_i)_{i \in \mathbb{N}}$ of W, the probability distribution of $\breve{W} := \sum_{i \in I_k^{(n)}} W_i$ (cf. Remark 11(ii)) has density

$$f_{\breve{W}}(y) \;:=\; \frac{\exp\{-\frac{y \cdot \widetilde{c}}{1-\gamma}\}}{\exp\{\widetilde{c} \cdot card(I_k^{(n)})/\gamma\}} \cdot f_{\breve{Z}}(y) \cdot \mathbb{1}_{]0,\infty[}(y), \qquad y \in \mathbb{R}, \tag{213}$$

where $\breve{Z}$ is a random variable with density $f_{\breve{Z}}$ of a stable law with parameter-quadruple $(\frac{-\gamma}{1-\gamma}, 1, 0, -card(I_k^{(n)}) \cdot \frac{\widetilde{c}^{1/(1-\gamma)} \cdot (1-\gamma)^{-\gamma/(1-\gamma)}}{\gamma})$.

(e) Case $\gamma > 2$, $\widetilde{c} > 0$: $\Lambda^{(0)}_{\gamma,\widetilde{c}}(z) = \frac{\widetilde{c}}{\gamma} \cdot \Big\{ \big(\frac{\gamma-1}{\widetilde{c}} \cdot z + 1 \big)^{\frac{\gamma}{\gamma-1}} - 1 \Big\}$ (cf. (167)) is the cumulant generating function of a "distorted stable distribution" $\mathfrak{Q}[\,\cdot\,] = \Pi[W \in \cdot\,]$ of a random variable W, which can be constructed as follows: let Z be an auxiliary random variable (having density f_Z and support $supp(Z) =]-\infty,\infty]$) of a stable law with parameter-quadruple $(\frac{\gamma}{\gamma-1}, 1, 0, \frac{\widetilde{c}^{1/(1-\gamma)} \cdot (\gamma-1)^{\gamma/(\gamma-1)}}{\gamma})$ in terms of the above-mentioned "form-B notation" ; by applying a general Laplace-transform result on p. 112 of Zolotarev [428], we can derive

$$M_Z(z) := E_{\Pi}[\exp(z \cdot Z)] = \int_0^{\infty} \exp(z \cdot y) \cdot f_Z(y)\, dy \;=\; \begin{cases} \exp\Big(\frac{\widetilde{c}^{1/(1-\gamma)} \cdot (\gamma-1)^{\gamma/(\gamma-1)}}{\gamma} \cdot (-z)^{\alpha} \Big), & \text{if } \; z \in]-\infty, 0], \\ \infty, & \text{if } \; z \in]0, \infty[, \end{cases} \tag{214}$$

where $\alpha := \frac{\gamma}{\gamma-1} \in]1,2[$. Since $0 \notin int(dom(M_Z))$ (and thus, Z does not have light-tails) we have to distort the density in order to extend the effective domain. Accordingly, let W be a random variable having density

$$f_W(y) \;:=\; \frac{\exp\{\frac{y \cdot \widetilde{c}}{\gamma-1}\}}{\exp\{\widetilde{c}/\gamma\}} \cdot f_Z(-y), \qquad y \in \mathbb{R}, \qquad \text{(cf. (115))}.$$

Then one can straightforwardly deduce from (214) that $\int_{-\infty}^{\infty} f_W(y)\, dy = 1$ and that

$$M_W(z) := E_{\Pi}[\exp(z \cdot W)] = \int_{-\infty}^{\infty} \exp(z \cdot y) \cdot f_W(y)\, dy \;=\; \begin{cases} \exp\Big(\frac{\widetilde{c}}{\gamma} \cdot \Big\{ \big(\frac{\gamma-1}{\widetilde{c}} \cdot z + 1 \big)^{\frac{\gamma}{\gamma-1}} - 1 \Big\} \Big), & \text{if } \; z \in [-\frac{\widetilde{c}}{\gamma-1}, \infty[, \\ \infty, & \text{if } \; z \in]-\infty, -\frac{\widetilde{c}}{\gamma-1}[. \end{cases}$$

Thus, $\varphi_{\gamma,\widetilde{c}} \in \Upsilon(]-\infty,\infty[)$.

- Type: $\mathfrak{Q}$ is an infinitely divisible (cf. Proposition 34) continuous distribution with density f_W.
- Behaviour at zero: $\mathfrak{Q}[\,]0,\infty[\,] = \Pi[W > 0] = \int_0^{\infty} f_W(u)\, du \in]0,1[$, $\quad \mathfrak{Q}[\,\{0\}\,] = \Pi[W = 0] = 0$.
- Corresponding generator: $\varphi^{(0)}_{\gamma,\widetilde{c}} = \widetilde{c} \cdot \varphi_{\gamma}$ (cf. (168), (43)) of the power divergence given in the seventh line of (44).

- Sums: for i.i.d. copies $(W_i)_{i\in\mathbb{N}}$ of W, the probability distribution of $\breve{W} := \sum_{i\in I_k^{(n)}} W_i$ (cf. Remark 11(ii)) has density

$$f_{\breve{W}}(y) \;:=\; \frac{\exp\{\frac{y\cdot\widetilde{c}}{\gamma-1}\}}{\exp\{\widetilde{c}\cdot card(I_k^{(n)})/\gamma\}}\cdot f_{\breve{Z}}(-y), \qquad y\in\mathbb{R},$$

where $\breve{Z}$ is a random variable with density $f_{\breve{Z}}$ of a stable law with parameter-quadruple $(\frac{\gamma}{\gamma-1},1,0,card(I_k^{(n)})\cdot\frac{\widetilde{c}^{1/(1-\gamma)}\cdot(\gamma-1)^{\gamma/(\gamma-1)}}{\gamma})$.

(f) Case $\gamma\in]1,2[$, $\widetilde{c}>0$: one still has the (cumulant-generating-function) candidate $\Lambda_{\gamma,\widetilde{c}}^{(0)}(z)=\frac{\widetilde{c}}{\gamma}\cdot\Big\{\big(\frac{\gamma-1}{\widetilde{c}}\cdot z+1\big)^{\frac{\gamma}{\gamma-1}}-1\Big\}$ (cf. (167)), but for the crucial exponent there holds $\frac{\gamma}{\gamma-1}>2$. From this, we conjecture that $\mathfrak{C}$ becomes a *signed* finite measure with total mass 1, i.e. it has a density (with respect to some dominating measure) with positive *and negative* values which "integrates to 1" ; accordingly, our BS method can not be applied to this situation.

Remark 49: As a continuation of Remark 38 and the note in the third line after (169), we have shown as a side effect that for $\gamma\in]-\infty,-1]\cup]0,1[\cup[2,\infty[$ the distributions $\mathfrak{C}_\gamma$ and $\mathfrak{C}_{1-\gamma}$ of Example 48(b)-(e) are inverse to each other.

Example 50: for the power-divergence context of Example 40 we obtain:
(a) Case $\gamma=1$, $\widetilde{c}>0$, anchor point $c=0$: $\Lambda_{1,\widetilde{c}}(z)=\widetilde{c}\cdot\big(\exp(\frac{z}{\widetilde{c}})-1\big)$ (cf. (175)) is the cumulant generating function of $\mathfrak{C}=\frac{1}{\widetilde{c}}\cdot POI(\widetilde{c})$ being the "$\frac{1}{\widetilde{c}}-$fold Poisson distribution with mean $\widetilde{c}$" which means that $W=\frac{1}{\widetilde{c}}\cdot Z$ for a $POI(\widetilde{c})-$distributed random variable Z. Thus, $\varphi_{1,\widetilde{c}}\in\Upsilon(]0,\infty[)$.

- Prominent special case $\widetilde{c}=1$: $\mathfrak{C}=POI(1)$ is the Poisson distribution with mean 1.
- Type: $\mathfrak{C}$ is an infinitely divisible (cf. Proposition 34) discrete distribution with frequencies: $\Pi[W=\ell\cdot\frac{1}{\widetilde{c}}]=\exp(-\widetilde{c})\cdot\frac{\widetilde{c}^\ell}{\ell!}$ for all nonnegative integers $\ell\in\mathbb{N}_0$ (and zero elsewhere).
- Behaviour at zero: $\Pi[W\geq 0]=1$, $\Pi[W=0]=\exp(-\widetilde{c})$.
- Corresponding generator: $\varphi_{1,\widetilde{c}}=\widetilde{c}\cdot\varphi_1$ (cf. (176), (43)) of the $\widetilde{c}-$fold of the Kullback-Leibler divergence (relative entropy) given in the fourth line of (44).
- Sums: for i.i.d. copies $(W_i)_{i\in\mathbb{N}}$ of W, the probability distribution of $\breve{W} := \sum_{i\in I_k^{(n)}} W_i$ (cf. Remark 11(ii)) is $\frac{1}{\widetilde{c}}\cdot POI(\widetilde{c}\cdot card(I_k^{(n)}))$.

(b) Case $\gamma=1$, $\widetilde{c}=1$, anchor point $c\in\mathbb{R}$: $\Lambda_{1,1}^{(c)}(z)=e^c\cdot(e^z-1)+z\cdot(1-e^c)$ (cf. (178)) is the well-known cumulant generating function of the "shifted Poisson distribution" $\mathfrak{C}=POI(e^c)+1-e^c$, i.e. $W:=Z+1-e^c$ with a $POI(e^c)-$distributed random variable Z. Hence, $\varphi_{1,1}^{(c)}\in\Upsilon(]1-e^c,\infty[)$.

- Type: $\mathfrak{C}$ is a discrete distribution with frequencies: $\Pi[W=\ell+1-e^c]=\exp(-e^c)\cdot\frac{e^{c\cdot\ell}}{\ell!}$ for all $\ell\in\mathbb{N}_0$ (and zero elsewhere).
- Behaviour at zero: $\Pi[W>0]=1$ iff $c<0$, $\Pi[W<0]>0$ iff $c>0$, $\Pi[W=0]\neq 0$ iff "$c=\log(1+k)$ for some $k\in\mathbb{N}_0$" .
- Corresponding generator: $\varphi_{1,1}^{(c)}$ (cf. (179)) of the divergence

$$D_{\varphi_{1,1}^{(c)}}(\mathbf{Q},\mathbf{P}):=\sum_{k=1}^K\Big(q_k+p_k\cdot(e^c-1)\Big)\cdot\Big\{\log\Big(\frac{q_k}{p_k}+e^c-1\Big)-c\Big\}-\sum_{k=1}^K q_k+\sum_{k=1}^K p_k,$$

$$\text{if }\;\mathbf{P}\in\mathbb{R}_{\neq 0}^K\text{ and }\mathbf{Q}\in\mathbb{R}^K\text{with }\mathbf{Q}\in\;[(1-e^c)\cdot\mathbf{P},\infty[\text{ component-wise,}\qquad(215)$$

which for $c=0$ coincides with the Kullback-Leibler divergence (relative entropy) given in the fourth line of (44).
- Sums: for i.i.d. copies $(W_i)_{i\in\mathbb{N}}$ of W, the probability distribution of $\breve{W} := \sum_{i\in I_k^{(n)}} W_i$ (cf. Remark 11(ii)) is $POI(card(I_k^{(n)})\cdot e^c)+(1-e^c)\cdot card(I_k^{(n)})$.

Remark 51: (a) One can see from the Examples 48 and 50 the interesting effect that the "homogeneous" class of power-divergence generators $(\varphi_\gamma)_{\gamma\in\mathbb{R}}$ are connected to a "very inhomogeneous" family $(\mathfrak{C}_\gamma)_{\gamma\in\mathbb{R}}$ of $W-$distributions: discrete, continuous, mixture of discrete and continuous, as the parameter γ varies.
Moreover, some cases satisfy $\Pi[W=0]=0$ and some $\Pi[W=0]>0$, some $\Pi[W>0]=1$ and some $\Pi[W>0]\in]0,1[$.
(b) As a continuation of Remark 38 and the note in the last line of Example 40(a), we have shown as a side effect that for the the natural-anchor-point choice $c=0$, the distributions $\mathfrak{C}_1$ of of Example 50(a) and $\mathfrak{C}_0$ of Example 48(a) are inverse to each other.

Example 52: for the context of Example 41 we obtain:
Case $\widetilde{c}>0$, anchor point $c=0$: $\Lambda_{bw,\beta,\widetilde{c}}(z)=-(\frac{1}{\beta}-1)\cdot z+\frac{\widetilde{c}}{\beta^2}\cdot\Big\{1-\sqrt{1-\frac{2\beta}{\widetilde{c}}\cdot z}\Big\}$ (cf. (181)) is the cumulant generating function of a probability distribution $\mathfrak{C}[\cdot]=\Pi[\breve{W}\in\cdot]$ of a random variable $\breve{W}$, which can be constructed as

follows: $\breve{W} := \frac{W}{\beta} - (\frac{1}{\beta} - 1)$, where W is the random variable constructed in Example 48(d) with $\gamma = -1$ and with $\widetilde{c}$ replaced by $\frac{\widetilde{c}}{\beta^2}$ (recall that W has a tilted stable distribution). In other words, $\mathfrak{C}$ is a special kind of *modified tilted stable distribution*.

- Type: $\mathfrak{C}$ is an infinitely divisible (cf. Proposition 34) continuous distribution with density $f_{\breve{W}}(u) := \beta \cdot f_W(\beta \cdot u + 1 - \beta) \cdot \mathbf{1}_{]-(\frac{1}{\beta}-1),\infty[}(u)$ $(u \in \mathbb{R})$, where $f_W(\cdot)$ is given in (82) with $\gamma = -1$ and with $\widetilde{c}$ replaced by $\frac{\widetilde{c}}{\beta^2}$.
- Behaviour at zero: $\mathfrak{C}[]0,\infty[] = \Pi[\breve{W} > 0] > 0$.
- Corresponding generator: $\varphi^{(0)}_{bw,\beta,\widetilde{c}}$ (cf. (182)) of the — "non-probability version" of — the well-known *blended weight chi-square divergence* given in (183).
- Sums: for i.i.d. copies $(\breve{W}_i)_{i\in\mathbb{N}}$ of $\breve{W}$, the probability distribution of $\breve{\breve{W}} := \sum_{i\in I_k^{(n)}} \breve{W}_i = \frac{1}{\beta} \cdot \sum_{i\in I_k^{(n)}} W_i - n_k \cdot (\frac{1}{\beta} - 1)$ (cf. Remark 11(ii)) has density $f_{\breve{\breve{W}}}(u) := \beta \cdot f_{\breve{W}}(\beta \cdot u + (1-\beta)\cdot n_k) \cdot \mathbf{1}_{]-n_k\cdot(\frac{1}{\beta}-1),\infty[}(u)$ $(u \in \mathbb{R})$, where $f_{\breve{W}}(\cdot)$ is given in (213) (cf. Example 48(d)) with $\gamma = -1$ and with $\widetilde{c}$ replaced by $\frac{\widetilde{c}}{\beta^2}$.

Example 53: for the context of Example 43 we obtain:
(a) Case $\alpha \in]0,\infty[$, $\widetilde{c} > 0$, anchor point $c = 0$: $\Lambda_{gKL,\alpha,\widetilde{c}}(z) = -\frac{\widetilde{c}}{\alpha} \cdot \log((1+\alpha) - \alpha \cdot e^{z/\widetilde{c}})$ (cf. (184)) is the cumulant generating function of $\mathfrak{C} = \frac{1}{\widetilde{c}} \cdot NB(\frac{\widetilde{c}}{\alpha}, \frac{1}{1+\alpha})$ being the "$\frac{1}{\widetilde{c}}-$fold Negative-Binomial distribution with parameters $\frac{\widetilde{c}}{\alpha}$ and $\frac{1}{1+\alpha}$" which means that $W = \frac{1}{\widetilde{c}} \cdot Z$ for a $NB(\frac{\widetilde{c}}{\alpha}, \frac{1}{1+\alpha})-$distributed random variable Z. Thus, $\varphi_{gKL,\alpha,\widetilde{c}} \in \Upsilon(]0,\infty[)$.

- Prominent special case $\widetilde{c} = 1$, $\alpha = 1$ (see below): $\mathfrak{C} = NB(1, \frac{1}{2})$ is the Negative-Binomial distribution with parameters 1 and $\frac{1}{2}$.
- Type: $\mathfrak{C}$ is an infinitely divisible (cf. Proposition 34) discrete distribution with frequencies: $\Pi[W = \ell \cdot \frac{1}{\widetilde{c}}] = (-1)^\ell \cdot \binom{-\frac{\widetilde{c}}{\alpha}}{\ell} \cdot \alpha^\ell \cdot (1+\alpha)^{-\ell-\widetilde{c}/\alpha}$ for all nonnegative integers $\ell \in \mathbb{N}_0$ (and zero elsewhere).
- Behaviour at zero: $\Pi[W \geq 0] = 1$, $\Pi[W = 0] = \frac{1}{(1+a)^{\widetilde{c}/\alpha}}$.
- Corresponding generator: $\varphi_{gKL,\alpha,\widetilde{c}}$ (cf. (185)) of the divergence (187); the special case $\widetilde{c} = 1$, $\alpha = 1$ — i.e. $\varphi_{gKL,1,1} =: \varphi_{snKL,1}$ (cf. (188)) — corresponds to the generator of the — "non-probability version" of the — Jensen-Shannon divergence (symmetrized and normalized Kullback-Leibler divergence, symmetrized and normalized relative entropy) given in (189).
- Sums: for i.i.d. copies $(W_i)_{i\in\mathbb{N}}$ of W, the probability distribution of $\breve{W} := \sum_{i\in I_k^{(n)}} W_i$ (cf. Remark 11(ii)) is $\frac{1}{\widetilde{c}} \cdot NB(\frac{\widetilde{c}}{\alpha} \cdot card(I_k^{(n)}), \frac{1}{1+\alpha})$.

(b) Case $\alpha \in]-1,0[$, $\widetilde{c} > 0$, anchor point $c = 0$: for any integer $m \in \mathbb{N}$ being strictly larger than $\widetilde{c}$ and the choice $\alpha = -\frac{\widetilde{c}}{m}$, we obtain $\Lambda_{gKL,-\widetilde{c}/m,\widetilde{c}}(z) = m \cdot \log((1 - \frac{\widetilde{c}}{m}) + \frac{\widetilde{c}}{m} \cdot e^{z/\widetilde{c}})$ (cf. (184)) which is the cumulant generating function of $\mathfrak{C} = \frac{1}{\widetilde{c}} \cdot BIN(m, \frac{\widetilde{c}}{m})$ being the "$\frac{1}{\widetilde{c}}-$fold Binomial distribution with parameters m and $\frac{\widetilde{c}}{m}$" which means that $W = \frac{1}{\widetilde{c}} \cdot Z$ for a $BIN(m, \frac{\widetilde{c}}{m})-$distributed random variable Z. Thus, $\varphi_{gKL,-\widetilde{c}/m,\widetilde{c}} \in \Upsilon(]0,\infty[)$.

- Type: $\mathfrak{C}$ is a non-infinitely divisible discrete distribution with frequencies: $\Pi[W = \ell \cdot \frac{1}{\widetilde{c}}] = \binom{m}{\ell} \cdot (\frac{\widetilde{c}}{m})^\ell \cdot (1 - \frac{\widetilde{c}}{m})^{m-\ell}$ for $\ell \in \{0, 1, \ldots, m\}$ (and zero elsewhere).
- Behaviour at zero: $\Pi[W \geq 0] = 1$, $\Pi[W = 0] = (1 - \frac{\widetilde{c}}{m})^m$.
- Corresponding generator: $\varphi_{gKL,\alpha,\widetilde{c}}$ (cf. (185)) of the divergence (187).
- Sums: for i.i.d. copies $(W_i)_{i\in\mathbb{N}}$ of W, the probability distribution of $\breve{W} := \sum_{i\in I_k^{(n)}} W_i$ (cf. Remark 11(ii)) is $\frac{1}{\widetilde{c}} \cdot BIN(m \cdot card(I_k^{(n)}), \frac{\widetilde{c}}{m})$.

Example 54: for the context of Example 45 we obtain:
Case of anchor point $c = 0$: $\Lambda_{twop}(z) = \log\Big(p \cdot e^{z_1 \cdot z} + (1-p) \cdot e^{z_2 \cdot z}\Big)$ (cf. (191)) is the well-known cumulant generating function of the two-point probability distribution $\mathfrak{C} = p \cdot \delta_{z_1} + (1-p) \cdot \delta_{z_2}$, where $z_1 < 1 < z_2$ and $p = \frac{z_2 - 1}{z_2 - z_1}$. Hence, $\varphi_{twop} \in \Upsilon(]z_1, z_2[)$.

- Type: $\mathfrak{C}$ is a discrete distribution with frequencies: $\Pi[W = z_1] = p$, $\Pi[W = z_2] = 1 - p$ (and zero elsewhere).
- Behaviour at zero: $\Pi[W > 0] = 1$ iff $z_1 > 0$, $\Pi[W = 0] \neq 0$ iff $z_1 = 0$.
- Corresponding generator: φ_{twop} (cf. (192)) of the divergence given in (194).
- Sums: for i.i.d. copies $(W_i)_{i\in\mathbb{N}}$ of W, the probability distribution of $\breve{W} := \sum_{i\in I_k^{(n)}} W_i$ (cf. Remark 11(ii)) is the distribution of the $card(I_k^{(n)})-$th step of a generalized random walk starting at zero; this has a nice explicit ("binomial-type") expression in the special case $z_1 = -z_2$, namely $\sum_{\ell=0}^{card(I_k^{(n)})} \binom{card(I_k^{(n)})}{\ell} \cdot p^{card(I_k^{(n)})-\ell} \cdot (1-p)^\ell \cdot \delta_{z_2\cdot(2\ell - card(I_k^{(n)}))}$.

Example 55: for the context of Example 46 we obtain:
Case $\alpha, \beta_1, \beta_2, \widetilde{c} \in]0,\infty[$, anchor point $c = 0$: by using $\breve{\theta} := 1 + \alpha \cdot \Big(\frac{1}{\beta_2} - \frac{1}{\beta_1}\Big)$ one can see that
$\Lambda_{\alpha,\beta_1,\beta_2,\widetilde{c}}(z) = \breve{\theta} \cdot z - \widetilde{c} \cdot \alpha \cdot \log\Big(1 + \frac{z}{\widetilde{c}} \cdot \Big(\frac{1}{\beta_2} - \frac{1}{\beta_1}\Big) - \frac{z^2}{\widetilde{c}^2 \cdot \beta_1 \cdot \beta_2}\Big)$ for $z \in]-\widetilde{c} \cdot \min\{\beta_1, \beta_2\}, \widetilde{c} \cdot \min\{\beta_1, \beta_2\}[$
— with different boundary behaviour for the three subcases $\beta_1 < \beta_2$ resp. $\beta_1 > \beta_2$ resp. $\beta_1 = \beta_2$ (cf. (197),(201),(206)) —

is the cumulant generating function of a *generalized asymmetric Laplace distribution* $\mathfrak{C}[\,\cdot\,] = \Pi[W \in \cdot\,]$ of a random variable $W := \breve{\theta} + Z_1 - Z_2$, where Z_1 respectively Z_2 are auxiliary random variables which are independent and $GAM(\widetilde{c} \cdot \beta_1, \widetilde{c} \cdot \alpha)-$distributed respectively $GAM(\widetilde{c} \cdot \beta_2, \widetilde{c} \cdot \alpha)-$distributed. In particular, $E_{\Pi}[W] = \breve{\theta} + \frac{\widetilde{c}\cdot\alpha}{\widetilde{c}\cdot\beta_1} - \frac{\widetilde{c}\cdot\alpha}{\widetilde{c}\cdot\beta_2} = 1$ (as required). Thus, $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}} \in \Upsilon(]-\infty,\infty[)$.

- Prominent special case $\widetilde{c} = 1$, $\alpha = 1$, $\beta_1 = \beta_2 =: \beta$ (and hence, $\breve{\theta} = 1$): $\mathfrak{C}$ is a *classical Laplace distribution* (two-tailed exponential distribution, bilateral exponential law) with location parameter 1 and scale parameter $\frac{1}{\beta}$.
- Type: $\mathfrak{C}$ is an infinitely divisible (cf. Proposition 34) continuous distribution with density

$$f(u) := \frac{\sqrt{2}\cdot\exp\{\frac{1}{\sigma\cdot\sqrt{2}}\cdot(\frac{1}{\kappa}-\kappa)\cdot(u-\theta)\}}{\sqrt{\pi}\cdot\sigma^{\tau+1/2}\cdot\Gamma(\tau)} \cdot \left(\frac{\sqrt{2}\cdot|u-\theta|}{\kappa+\frac{1}{\kappa}}\right)^{\tau-1/2} \cdot K_{\tau-1/2}\left(\frac{1}{\sigma\cdot\sqrt{2}}\cdot\left(\kappa+\frac{1}{\kappa}\right)\cdot|u-\theta|\right), \quad u \neq \theta, \tag{216}$$

  where $(\theta,\kappa,\sigma,\tau)$ is given in Remark 56 below and K_λ is the modified Bessel function of the third kind with index λ. For the above-mentioned special case of the classical Laplace distribution, this considerably simplifies to $f(u) := \frac{\beta}{2}\exp\{-\beta\cdot|u-1|\}$.
- Behaviour at zero: $\mathfrak{C}[]0,\infty[] = \Pi[W > 0] = \int_0^\infty f(u)\,du \in]0,1[$, $\quad \mathfrak{C}[\{0\}] = \Pi[W = 0] = 0$.
- Corresponding generator: $\varphi_{\alpha,\beta_1,\beta_2,\widetilde{c}}$ (cf. (198) respectively (202) respectively (207)) of the divergence given in (199) respectively (203) respectively (208).
- Sums: for i.i.d. copies $(W_i)_{i\in\mathbb{N}}$ of W, the probability distribution of $\breve{W} := \sum_{i\in I_k^{(n)}} W_i$ (cf. Remark 11(ii)) is the same as that of a random variable $\breve{\breve{W}} := \breve{\theta}\cdot card(I_k^{(n)}) + \breve{Z}_1 - \breve{Z}_2$, where $\breve{Z}_1$ respectively $\breve{Z}_2$ are auxiliary random variables which are independent and $GAM(\widetilde{c}\cdot\beta_1, \widetilde{c}\cdot\alpha\cdot card(I_k^{(n)}))-$distributed respectively $GAM(\widetilde{c}\cdot\beta_2, \widetilde{c}\cdot\alpha\cdot card(I_k^{(n)}))-$distributed.

Remark 56: In the book of Kotz et al. [197] one can find a very comprehensive study on generalized asymmetric Laplace distributions (also known as Bessel function distributions, McKay distributions), their close relatives (such as e.g. the financial-econometric variance gamma model of Madan & Seneta [240]) as well as their applications; see also e.g. Klar [192] for connections with some other Gamma difference distributions. [197] use a different parametrization $(\theta,\kappa,\sigma,\tau)$ which is one-to-one with our parametrization $(\breve{\theta},\alpha,\beta_1,\beta_2,\widetilde{c}=1)$, as follows: $\theta = \breve{\theta}$, $\tau = \widetilde{c}\cdot\alpha$, $\sigma = \frac{1}{\widetilde{c}}\cdot\sqrt{\frac{2}{\beta_1\cdot\beta_2}}$, $\kappa = \frac{\sqrt{\frac{4}{\widetilde{c}^2}+(\beta_1-\beta_2)^2}+\beta_2-\beta_1}{2\cdot\sqrt{\beta_1\cdot\beta_2}} > 0$. In particular, this implies that we cover *all* generalized asymmetric Laplace distributions with mean 1. For better comparability, we have used the parametrization $(\theta,\kappa,\sigma,\tau)$ in the above-mentioned representation (216) of the density (due to [197]).

Let us end this section by giving some further comments on the task of finding concretely the probability distribution (if existent) $\mathfrak{C}[\,\cdot\,] = \Pi[W \in \cdot\,]$ from the Fenchel-Legendre transform $\Lambda = \varphi^*$ of a pregiven divergence generator φ, which should satisfy

$$\exp(\Lambda(z)) = \int_{\mathbb{R}} e^{z\cdot y}\,d\mathfrak{C}(y) = E_{\Pi}[\exp(z\cdot W)], \qquad z \in \mathbb{R}, \qquad \text{(cf. (209), (210))}.$$

Recall that this is used for the simulation of the weights $(W_i)_{i\in\mathbb{N}}$ which are i.i.d. copies of W and which are the crucial building ingredients of $\boldsymbol{\xi}_n^{\mathbf{W}}$ in Theorem 10, respectively, of $\xi_{n,\mathbf{X}}^{w\mathbf{W}}$ in Theorem 14. The search for $\mathfrak{C}$ can be done e.g. by inversion of the moment generating function MGF, or by search in tables or computer software which list distributions and their MGF. As already indicated above, we have eased/narrowed down this task by giving (additional) sufficient conditions for some deriving *principal* properties of $\mathfrak{C}$. Also notice that $\mathfrak{C}$ needs not necessarily to be explicitly known in full detail (e.g. in terms of a computationally tractable density or frequency); for instance, as well known from insurance applications, for — comfortably straightforwardly simulable — doubly-random sums $W := \sum_{i=1}^N A_i$ of nonnegative i.i.d. random variables $(A_i)_{i\in\mathbb{N}}$ with known law $\Pi_A[\,\cdot\,] := \Pi[A \in \cdot\,]$ being independent of a counting-type random variable N with known law Π_N, one can mostly compute explicitly $MGF_{\mathfrak{C}}(z) = PGF_{\Pi_N}(MGF_{\Pi_A}(z))$ in terms of $\mathfrak{C} := \Pi_W$ and the probability generating function PGF_{Π_N} of Π_N, but the corresponding density/frequency of $\mathfrak{C}$ may not be known explicitly in a tractable form. The above-mentioned Example 48(b) of power divergences with generator φ_γ ($\gamma \in]0,1[$) manifests such a situation.

In the end, if no explicit distribution $\mathfrak{C}$ and no comfortably simulable $W-$construction are available, one can still try to simulate an i.i.d. sequence $(W_i)_{i\in\mathbb{N}}$ from the pregiven moment generating function (which is $\exp(\Lambda(z))$ here); see e.g. McLeish [259] and references therein which also contains saddle point methods approximation techniques.

VI. ESTIMATORS

In the following, we demonstrate how one can *principally* implement our BS approach; a further, deeper analysis will be given in a follow-up paper.

A. Estimators for the deterministic minimization problem

We address the minimization problem

$$D_{\varphi}(\mathbf{\Omega}, \mathbf{P}) := \inf_{\mathbf{Q}\in\mathbf{\Omega}} D_{\varphi}(\mathbf{Q}, \mathbf{P}) = \inf_{\widetilde{\mathbf{Q}}\in\widetilde{\mathbf{\Omega}}} D_{\widetilde{\varphi}}(\widetilde{\mathbf{Q}}, \widetilde{\mathbb{P}}) =: D_{\widetilde{\varphi}}(\widetilde{\mathbf{\Omega}}, \widetilde{\mathbb{P}}) \qquad \text{with } \widetilde{\mathbf{\Omega}} := \mathbf{\Omega}/M_{\mathbf{P}} \qquad \text{(cf. (8) and (13))}, \tag{217}$$

whose numerical solution is based on Theorem 10 which basically states that for large integer $n \in \mathbb{N}$ one has

$$\inf_{\mathbf{Q}\in\mathbf{\Omega}} D_{\varphi}(\mathbf{Q}, \mathbf{P}) \approx -\frac{1}{n} \log \Pi\Big[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\mathbf{\Omega}}\Big] \tag{218}$$

in terms of $\widetilde{\varphi} := M_{\mathbf{P}} \cdot \varphi$ and the random vectors

$$\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} = \Big(\frac{1}{n} \sum_{i\in I_1^{(n)}} \widetilde{W}_i, \ldots, \frac{1}{n} \sum_{i\in I_K^{(n)}} \widetilde{W}_i\Big) \qquad \text{(cf. (23))}$$

with $n_k := \lfloor n \cdot \widetilde{p}_k \rfloor$ leading to the disjoint index blocks $I_1^{(n)} := \{1, \ldots, n_1\}$, $I_2^{(n)} := \{n_1 + 1, \ldots, n_1 + n_2\}$, …, $I_K^{(n)} := \{\sum_{k=1}^{K-1} n_k + 1, \ldots, n\}$. Recall that $\widetilde{\mathbf{W}} := (\widetilde{W}_1, \ldots, \widetilde{W}_n)$ is a random vector consisting of components $\widetilde{W}_i$ which are i.i.d. copies of the random variable $\widetilde{W}$ whose distribution is $\Pi[\widetilde{W} \in \cdot\,] = \widetilde{\mathfrak{l}}[\,\cdot\,]$ obeying the representation

$$\widetilde{\varphi}(t) = \sup_{z\in\mathbb{R}} \left(z \cdot t - \log \int_{\mathbb{R}} e^{zy} d\widetilde{\mathfrak{l}}(y) \right), \qquad t \in \mathbb{R}, \qquad \text{(cf. (21))}.$$

Hence, the estimation of $D_{\varphi}(\mathbf{\Omega}, \mathbf{P})$ amounts to the estimation of $\Pi\Big[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\mathbf{\Omega}}\Big]$. For the rest of this subsection, we assume that $\widetilde{\mathbb{P}} \in \mathbb{S}_{>0}^K$, that n is chosen such that all $n \cdot \widetilde{p}_k$ are integers (and hence, $n = \sum_{k=1}^K n_k$), and that $\widetilde{\mathbf{\Omega}} \subset \mathbb{R}^K$ satisfies the regularity property

$$cl(\widetilde{\mathbf{\Omega}}) = cl\left(int\left(\widetilde{\mathbf{\Omega}}\right)\right), \qquad int\left(\widetilde{\mathbf{\Omega}}\right) \neq \emptyset$$

which implies that the same condition holds for $\mathbf{\Omega}$; moreover, we suppose that $D_{\widetilde{\varphi}}(\widetilde{\mathbf{\Omega}}, \widetilde{\mathbb{P}})$ is finite. For the ease of the following discussions, we introduce the notations

$$T\left(\mathbf{x}\right) := \left(\frac{1}{n_1} \sum_{i\in I_1^{(n)}} x_i, \ldots, \frac{1}{n_K} \sum_{i\in I_K^{(n)}} x_i\right) \qquad \text{for any } \mathbf{x} := (x_1, .., x_n) \in \mathbb{R}^n,$$

as well as $\mathfrak{D}$ for the diagonal matrix with diagonal entries $1/\widetilde{p}_1, \ldots, 1/\widetilde{p}_K$ and null entries off the diagonal. Accordingly, the probability in (218) becomes

$$\Pi\Big[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\mathbf{\Omega}}\Big] \;=\; \Pi\Big[T(\widetilde{\mathbf{W}}) \in \mathbf{\Lambda}\Big]$$

where

$$\mathbf{\Lambda} := \mathfrak{D} \cdot \widetilde{\mathbf{\Omega}}$$

is a set of vectors in $\mathbb{R}^K$ which is known/derived from the concrete context. The *naive estimator* $\widehat{\widehat{\Pi}}_L^{naive}$ of $\Pi\Big[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\mathbf{\Omega}}\Big]$ is constructed through the following procedure: simulate independently L copies $\widetilde{\mathbf{W}}^1, \ldots, \widetilde{\mathbf{W}}^L$ of the vector $\widetilde{\mathbf{W}} := \Big(\widetilde{W}_1, \ldots, \widetilde{W}_n\Big)$, with independent entries under $\widetilde{\mathfrak{l}}$, and define (with a slight abuse of notation)

$$\widehat{\widehat{\Pi}}_L^{naive} := \frac{1}{L} \sum_{\ell=1}^{L} \mathbf{1}_{\mathbf{\Lambda}} \left(T\left(\widetilde{\mathbf{W}}^{\ell}\right)\right);$$

however this procedure is time costly, since this estimate has a very bad hit rate. Thus, in the following, a so-called "efficient Importance Sampling (IS)" scheme — in the sense of Sadowsky & Bucklew [313] (denoted [SB] hereunder) — is adapted for the sophisticated (i.e. non-naive) estimation of $\Pi[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\mathbf{\Omega}}]$. The main property of IS schemes lays in the fact that the runtime for an estimate with a controlled relative error does *not* increase at exponential rate as n increases, in contrast to $\widehat{\widehat{\Pi}}_L^{naive}$ which has exponential increase. In detail, let $\delta > 0$ be a given relative precision for an estimator $P_{\widetilde{S}}^{L_n}$ of $\Pi[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\boldsymbol{\Omega}}]$, based on a number L_n of simulated samples generated under some distribution $\widetilde{S}$, so that

$$\delta := \frac{var_{\widetilde{S}} P_{\widetilde{S}}^{L_n}}{\left(\Pi\left[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\boldsymbol{\Omega}}\right]\right)^2} .$$

Then L_n will grow exponentially as n tends to infinity if and only if $\widetilde{S}$ is not "asymptotically optimal" , the derivation of which is the scope of the current section.

To start with the details, for the sake of brevity (to avoid certain substantial discussions on potential technical relaxations) we shall employ the following additional Assumption (OM) on the set $\widetilde{\boldsymbol{\Omega}}$:

(OM) For any $\widetilde{\boldsymbol{\omega}} \in cl(\widetilde{\boldsymbol{\Omega}})$ there exists a vector $\mathbf{x} = (x_1, \ldots, x_n) \in \left]t_-^{sc}, t_+^{sc}\right[^n$ such that $\widetilde{\boldsymbol{\omega}} = \left(\frac{1}{n}\sum_{i \in I_1^{(n)}} x_i, \ldots, \frac{1}{n}\sum_{i \in I_K^{(n)}} x_i\right)$, or equivalently, for any $\boldsymbol{\lambda} \in cl(\boldsymbol{\Lambda})$ there exists a vector $\mathbf{x} = (x_1, \ldots, x_n) \in \left]t_-^{sc}, t_+^{sc}\right[^n$ such that $\boldsymbol{\lambda} = T(\mathbf{x})$.

For instance, in the common case $dom(\widetilde{\varphi}) = dom(\varphi) =]a,b[= \left]t_-^{sc}, t_+^{sc}\right[=]0,\infty[$ (e.g. for the power-divergence generators $\widetilde{\varphi} = \widetilde{c} \cdot \varphi_\gamma$, $\gamma \leq 0$, cf. Example 39) the Assumption (OM) is always feasible.

To proceed, for any distribution $\widetilde{S}$ on $\mathbb{R}^n$ with support included in the support of the product measure $\widetilde{\mathfrak{C}}^{\otimes n}$ it holds

$$\Pi\left[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\boldsymbol{\Omega}}\right] = E_{\widetilde{\mathfrak{C}}^{\otimes n}}\left[\mathbf{1}_{\boldsymbol{\Lambda}}(T(\widetilde{\mathbf{W}}))\right] = E_{\widetilde{S}}\left[\mathbf{1}_{\boldsymbol{\Lambda}}\left(T\left(\widetilde{\mathbf{V}}\right)\right) \cdot \frac{d\widetilde{\mathfrak{C}}^{\otimes n}}{d\widetilde{S}}\left(\widetilde{\mathbf{V}}\right)\right]$$

from where the *improved IS estimator* of $\Pi\left[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\boldsymbol{\Omega}}\right]$ is obtained by sampling L i.i.d. replications $\widetilde{\mathbf{V}}^1, \ldots, \widetilde{\mathbf{V}}^L$ of the random vector $\widetilde{\mathbf{V}}$ with distribution $\widetilde{S}$ and by defining

$$\widehat{\widetilde{\Pi}}_L^{improved} := \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}_{\boldsymbol{\Lambda}}(T(\widetilde{\mathbf{V}}^{(\ell)})) \cdot \frac{d\widetilde{\mathfrak{C}}^{\otimes n}}{d\widetilde{S}}\left(\widetilde{\mathbf{V}}^{(\ell)}\right) . \tag{219}$$

The precise form of the efficient IS distribution $\widetilde{S}^{opt}$ relies on the definition of a "dominating point" of $\boldsymbol{\Lambda}$, which we recall now. For $\mathbf{x} := (x_1, .., x_n)$ in $\mathbb{R}^n$ we define

$$I_{\widetilde{\mathbf{W}}}(\mathbf{x}) := \sup_{\mathbf{z} \in \mathbb{R}^n}\left(\langle \mathbf{z}, \mathbf{x}\rangle - \log E_{\widetilde{\mathfrak{C}}}[\exp(\langle \mathbf{z}, \widetilde{\mathbf{W}}\rangle)]\right) ,$$

and for $\boldsymbol{\lambda}$ in $\boldsymbol{\Lambda}$ we let

$$I(\boldsymbol{\lambda}) := \inf\left\{I_{\widetilde{\mathbf{W}}}(\mathbf{x}) : T(\mathbf{x}) = \boldsymbol{\lambda}\right\} .$$

Let $\underline{\boldsymbol{\lambda}} := (\underline{\lambda}_1, \ldots, \underline{\lambda}_K) \in \partial\boldsymbol{\Lambda}$. We call $\underline{\boldsymbol{\lambda}}$ a *minimal rate point (mrp) of* $\boldsymbol{\Lambda}$ if

$$I(\underline{\boldsymbol{\lambda}}) \leq I(\boldsymbol{\lambda}) \quad \text{ for all } \boldsymbol{\lambda} \in \boldsymbol{\Lambda}.$$

A minimal rate point $\underline{\boldsymbol{\lambda}}$ is called a *dominating point of* $\boldsymbol{\Lambda}$ if a) $\underline{\boldsymbol{\lambda}} \in \partial\boldsymbol{\Lambda}$, and b) $I\left(\underline{\boldsymbol{\lambda}}\right) \leq I(\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ with attainment, namely there exists a vector $\underline{\mathbf{x}} \in \left]t_-^{sc}, t_+^{sc}\right[^n$ such that $I_{\widetilde{W}}\left(\underline{\mathbf{x}}\right) = I\left(\underline{\boldsymbol{\lambda}}\right)$ with $\underline{\boldsymbol{\lambda}} = T\left(\underline{\mathbf{x}}\right)$. The characterization of the dominating point $\underline{\lambda}$ is settled in the following

Lemma 57: Let $\underline{\boldsymbol{\lambda}}$ be a mrp of $\boldsymbol{\Lambda}$. Then, under Assumption (OM), $\underline{\boldsymbol{\lambda}}$ is a dominating point, and $\inf\left\{I_{\widetilde{\mathbf{W}}}\left(\mathbf{x}\right), T(\mathbf{x}) = \underline{\boldsymbol{\lambda}}\right\}$ is reached at some vector $\underline{\mathbf{x}}$ in $\left]t_-^{sc}, t_+^{sc}\right[^n$ such that for all $k \in \{1, \ldots, K\}$ and all $i \in I_k^{(n)}$ there holds $\underline{x}_i = \underline{\lambda}_k$ and $I_{\widetilde{\mathbf{W}}}(\underline{\mathbf{x}}) = n \cdot \sum_{k=1}^{K} \widetilde{p}_k \cdot \widetilde{\varphi}\left(\underline{\lambda}_k\right)$.

The proof Lemma 57 is given in Appendix G. Notice that (OM) implies the *existence* of a dominating point $\underline{\boldsymbol{\lambda}}$, but *uniqueness* may not hold. In the latter case, one can try to proceed as in Theorem 2 of [SB] and the discussion thereafter.

However, we assume now uniqueness of $\underline{\boldsymbol{\lambda}}$; this allows for the identification of $\widetilde{S}^{opt}$. By Theorem 1 of [SB] and Theorem 3.1 of Csiszar [96], the asymptotically optimal IS distribution $\widetilde{S}^{opt}$ is obtained as the Kullback-Leibler projection of $\widetilde{\mathfrak{C}}^{n\otimes}$ on the set of all probability distributions on $\mathbb{R}^n$ centered at point $\underline{\mathbf{x}}$, whose coordinates are — according to Lemma 57 — functions of the coordinates of $\underline{\widetilde{\mathbf{Q}}} := \mathfrak{D}^{-1}\underline{\boldsymbol{\lambda}}$ such that $T\left(\underline{\mathbf{x}}\right) = \mathfrak{D}\underline{\widetilde{\mathbf{Q}}}$.

The above definition of $\widetilde{S}^{opt}$ presumes the knowledge of $\underline{\boldsymbol{\lambda}}$, which cannot be assumed (otherwise the minimization problem is solved in advance). The aim of the following construction is to provide a proxy $\widetilde{S}$ to $\widetilde{S}^{opt}$, where $\widetilde{S}$ is the Kullback-Leibler projection of $\widetilde{\mathfrak{C}}^{\otimes n}$ on the set of all probability distributions on $\mathbb{R}^n$ centered at some point $\mathbf{x}^*$ which is close to $\underline{\mathbf{x}}$. For this sake, we need to have at hand a *proxy* of $\underline{\boldsymbol{\lambda}}$ or, equivalently, a *preliminary guess* $\widetilde{\mathbf{Q}}^*$ of $\underline{\widetilde{\mathbf{Q}}} := \arg\inf_{\widetilde{\mathbf{Q}} \in \widetilde{\boldsymbol{\Omega}}} \sum_{k=1}^{K} \widetilde{p}_k \cdot \widetilde{\varphi}(\widetilde{q}_k / \widetilde{p}_k)$. This guess is by no means produced in order to provide a direct estimate of $D_{\widetilde{\varphi}}(\widetilde{\mathbf{\Omega}}, \widetilde{\mathbb{P}})$ but merely to provide the IS distribution $\widetilde{S}$ which in turn leads to a sharp estimate of $D_{\widetilde{\varphi}}(\widetilde{\mathbf{\Omega}}, \widetilde{\mathbb{P}})$.

Proxy method 1: in some cases we might have at hand some particular point $\widetilde{\mathbf{Q}}^* := (\widetilde{q}_1^*, .., \widetilde{q}_K^*)$ in $\widetilde{\mathbf{\Omega}}$; the resulting IS distribution $\widetilde{S}$ with $\underline{\widetilde{\mathbf{Q}}}$ substituted by $\widetilde{\mathbf{Q}}^*$ is not optimal in the sense of [SB], but anyhow produces an estimator with good hitting rate, possibly with a loss in the variance. A simple way to obtain such a point $\widetilde{\mathbf{Q}}^*$ in $\widetilde{\mathbf{\Omega}}$ is to simulate runs of (say) $M-$variate i.i.d. vectors $\widetilde{\mathbf{W}}$ under $\widetilde{\mathfrak{C}}^{\otimes M}$ until the first time where $\boldsymbol{\xi}_M^{\widetilde{\mathbf{W}}}$ belongs to $\widetilde{\mathbf{\Omega}}$; then we set $\widetilde{\mathbf{Q}}^* := \boldsymbol{\xi}_M^{\widetilde{\mathbf{W}}}$ for the succeeding realization $\widetilde{\mathbf{W}}$. Before we proceed, it is useful to mention that the need for a drastic fall in the number of simulation runs pertains for cases when $D_{\widetilde{\varphi}}(\widetilde{\mathbf{\Omega}}, \widetilde{\mathbb{P}})$ is large. The following construction is suited to this case, which is of relevance in applications both in optimization and in statistics when choosing between competing models none of which is assumed to represent the true one, but merely less inadequate ones.

Proxy method 2: when $D_{\widetilde{\varphi}}(\widetilde{\mathbf{\Omega}}, \widetilde{\mathbb{P}})$ is presumably large, we make use of asymptotic approximation to get a proxy of $\underline{\widetilde{\mathbf{Q}}}$. For this, we define a sampling distribution on $\mathbb{R}^K$ fitted to the divergence through

$$f(\widetilde{\mathbf{Q}}) := C \cdot \exp\left(-\sum_{k=1}^{K} \widetilde{p}_k \cdot \widetilde{\varphi}(\widetilde{q}_k/\widetilde{p}_k)\right) \ = \ C \cdot \exp\left(-D_{\widetilde{\varphi}}\left(\widetilde{\mathbf{Q}}, \widetilde{\mathbb{P}}\right)\right) \tag{220}$$

where C is a normalizing constant. Let $\mathbf{T}$ be a $K-$variate random variable with density f. The distribution of $\mathbf{T}$ given $\left(\mathbf{T} \in \widetilde{\mathbf{\Omega}}\right)$ concentrates around $\arg\inf_{\widetilde{\mathbf{Q}} \in \widetilde{\mathbf{\Omega}}} D_{\widetilde{\varphi}}(\widetilde{\mathbf{Q}}, \widetilde{\mathbb{P}})$ when $D_{\widetilde{\varphi}}(\widetilde{\mathbf{\Omega}}, \widetilde{\mathbb{P}})$ is large. Indeed, for any $\widetilde{\mathbf{Q}} \in \widetilde{\mathbf{\Omega}}$ denote by $\mathbf{V}_\varepsilon(\widetilde{\mathbf{Q}})$ a small neighborhood of $\widetilde{\mathbf{Q}}$ in $\mathbb{R}^K$ with radius ε; clearly, the probability of the event $\left(\mathbf{T} \in \mathbf{V}_\varepsilon(\widetilde{\mathbf{Q}})\right)$ when restricted to $\widetilde{\mathbf{Q}} \in \widetilde{\mathbf{\Omega}}$ is maximum when $\widetilde{\mathbf{Q}} = \underline{\widetilde{\mathbf{Q}}}$, where $\underline{\widetilde{\mathbf{Q}}}$ is the "dominating point of $\widetilde{\mathbf{\Omega}}$ " *in the sense that* $\underline{\widetilde{\mathbf{Q}}} := \mathfrak{D}^{-1}\underline{\boldsymbol{\lambda}}$ is the above-defined transform of the dominating point $\underline{\boldsymbol{\lambda}}$ (assuming uniqueness); a precise argumentation under adequate conditions is postponed to Appendix G. Accordingly, we obtain a proxy $\widetilde{\mathbf{Q}}^*$ of $\underline{\widetilde{\mathbf{Q}}}$ by simulating a sequence of independent $K-$variate random variables $\mathbf{T}_1, \ldots$ with distribution (220) until (say) $\mathbf{T}_m$ belongs to $\widetilde{\mathbf{\Omega}}$ and set $\widetilde{\mathbf{Q}}^* := \mathbf{T}_m$.

To proceed with the derivation of the IS sampling distribution $\widetilde{S}$ on $\mathbb{R}^n$, we fix $\widetilde{\mathbf{Q}}^* := (\widetilde{q}_1^*, .., \widetilde{q}_K^*)$ to be a proxy of $\underline{\widetilde{\mathbf{Q}}}$ or an initial guess in $\widetilde{\mathbf{\Omega}}$. As an intermediate step, we construct the probability distribution $\widetilde{U}_k$ on $\mathbb{R}$ given by

$$d\widetilde{U}_k(v) \ := \ \exp\left(\tau_k \cdot v - \Lambda_{\widetilde{\mathfrak{C}}}(\tau_k)\right) d\widetilde{\mathfrak{C}}(v) = \frac{\exp\left(\tau_k \cdot v\right)}{MGF_{\widetilde{\mathfrak{C}}}(\tau_k)}\, d\widetilde{\mathfrak{C}}(v) \tag{221}$$

where $\tau_k \in int(dom(MGF_{\widetilde{\mathfrak{C}}}))$ is the unique solution of the equation $\Lambda'_{\widetilde{\mathfrak{C}}}(\tau_k) = \frac{\widetilde{q}_k^*}{\widetilde{p}_k}$ and thus — by relation (281) of Appendix F — we can compute explicitly

$$\tau_k = \ \widetilde{\varphi}'\left(\frac{\widetilde{q}_k^*}{\widetilde{p}_k}\right) \ .$$

Therefore, $\widetilde{U}_k$ is the Kullback-Leibler projection of $\widetilde{\mathfrak{C}}$ on the class of all probability distributions on $\mathbb{R}$ whose expectation is $\widetilde{q}_k^*$. As a side remark, notice that one possible way of obtaining an explicit form of the probability distribution $\widetilde{U}_k$ may be by identification through its moment generating function

$$dom(MGF_{\widetilde{\mathfrak{C}}}) - \tau_k \ \ni \ z \ \mapsto MGF_{\widetilde{U}_k}(z) = \frac{MGF_{\widetilde{\mathfrak{C}}}(z + \tau_k)}{MGF_{\widetilde{\mathfrak{C}}}(\tau_k)} \tag{222}$$

of which all ingredients are principally available. For instance, this will be used in Example 58 below. From (221), we define

$$\widetilde{S}_k := \underbrace{\widetilde{U}_k \otimes \cdots \otimes \widetilde{U}_k}_{n_k \text{times}} \qquad \text{for all } k \in \{1, \ldots, K\},$$

whence

$$d\widetilde{S}_k\left(v_{k,1}, \ldots, v_{k,n_k}\right) = \exp\Big(\sum_{i \in I_k^{(n)}} \tau_k \cdot v_{k,i} - n_k \cdot \Lambda_{\widetilde{\mathfrak{C}}}(\tau_k)\Big)\ d\widetilde{\mathfrak{C}}\left(v_{k,1}\right) \cdots d\widetilde{\mathfrak{C}}\left(v_{k,n_k}\right), \tag{223}$$

(224)

which manifests $\widetilde{S}_k$ as the Kullback-Leibler projection of $\underbrace{\widetilde{\mathfrak{C}} \otimes \cdots \otimes \widetilde{\mathfrak{C}}}_{n_k \text{times}}$ on the class of all probability distributions on $\mathbb{R}^k$ whose expectation vector is $\widetilde{\mathbf{Q}}^* = (\widetilde{q}_1^*, \ldots, \widetilde{q}_K^*) \in \mathbb{R}^k$. Let now

$$\widetilde{S} := \widetilde{S}_1 \otimes \cdots \otimes \widetilde{S}_K \ , \tag{225}$$

which therefore satisfies (recall that $\sum_{k=1}^{K} n_k = n$)

$$d\widetilde{S}\,(v_{1,1},\ldots v_{1,n_1},\ldots,v_{K,1},\ldots v_{K,n_K}) = \exp\Big(\sum_{k=1}^{K}\sum_{i\in I_k^{(n)}}\big(\tau_k\cdot v_{k,i} - n_k\cdot\Lambda_{\widetilde{\mathbb{C}}}(\tau_k)\big)\Big)\, d\widetilde{\mathbb{C}}^{\otimes n}\,(v_{1,1},\ldots v_{1,n_1},\ldots,v_{K,1},\ldots v_{K,n_K})\,. \tag{226}$$

The same procedure with all $\widetilde{q}_k^*$ substituted by the coordinates $\underline{\widetilde{q}}_k$ of $\underline{\widetilde{\mathbf{Q}}}$ produces S^{opt}. Therefore, $\widetilde{S}$ is a substitute for S^{opt} with the change in the centering from the unknown vector $\underline{\widetilde{\mathbf{Q}}}$ to its proxy $\widetilde{\mathbf{Q}}^*$.

As a straightforward consequence of (219) and (226), we obtain the improved IS estimator of $\Pi\Big[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\boldsymbol{\Omega}}\Big]$ as

$$\widehat{\widetilde{\Pi}}_L^{improved} \;=\; \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}_{\boldsymbol{\Lambda}}(T(\widetilde{\mathbf{V}}^{(\ell)}))\cdot\prod_{k=1}^{K} IS_k(\widetilde{\mathbf{V}}_k^{(\ell)}) \tag{227}$$

where $\widetilde{\mathbf{V}}_k^{(\ell)} := \Big(\widetilde{V}_i^{(\ell)}\Big)_{i\in I_k^{(n)}}$ is the $k-$th block of the $\ell-$th replication $\widetilde{\mathbf{V}}^{(\ell)}$ of $\widetilde{\mathbf{V}}$ under S, and the $k-$th importance-sampling factor is

$$\widetilde{IS}_k(v_{k,1},\ldots,v_{k,n_k}) \;:=\; \frac{d\widetilde{\mathbb{C}}^{\otimes n_k}}{d\widetilde{S}_k}\,(v_{k,1},\ldots,v_{k,n_k}) = \exp\Big(n_k\cdot\Lambda_{\widetilde{\mathbb{C}}}(\tau_k) \;-\; \tau_k\cdot\sum_{i=1}^{n_k} v_{k,i}\Big)$$

with $n_k = card(I_k^{(n)})$.

Summing up things, we arrive at the following algorithm in case that $\widetilde{\boldsymbol{\Omega}}$ has a unique dominating point (in the above-defined sense):

Step D1
Exemplarily, we start with proxy method 2 (the other proxy method 1 works analogously): get a proxy $\widetilde{\mathbf{Q}}^*$ of $\underline{\widetilde{\mathbf{Q}}}$ by simulating a sequence of independent $K-$variate random variables $\mathbf{T}_1,\ldots$ with distribution (220) until (say) $\mathbf{T}_m$ belongs to $\widetilde{\boldsymbol{\Omega}}$ and set $\widetilde{\mathbf{Q}}^* := \mathbf{T}_m$.
Step D2
For all k in $\{1,\ldots,K\}$ compute $\tau_k = \;\widetilde{\varphi}'\left(\frac{\widetilde{q}_k^*}{\widetilde{p}_k}\right)$.
Step D3
For all ℓ in $\{1,\ldots,L\}$ perform a run of $\widetilde{\mathbf{V}}^{(\ell)}$ under $\widetilde{S}$ as follows:
For all k in $\{1,\ldots,K\}$ simulate n_k i.i.d. random variables $\widetilde{V}_{k_1}^{(\ell)},\ldots,\widetilde{V}_{k_{n_k}}^{(\ell)}$ with common distribution $\widetilde{U}_k$ defined in (221). Set $\widetilde{\mathbf{V}}_k^{(\ell)} := (\widetilde{V}_{k_1}^{(\ell)},\ldots,\widetilde{V}_{k_{n_k}}^{(\ell)})$ to be the corresponding row vector.
Construct $\widetilde{\mathbf{V}}^{(\ell)}$ as the row vector obtained by concatenating the $\widetilde{\mathbf{V}}_k^{(\ell)}$, i.e.

$$\widetilde{\mathbf{V}}^{(\ell)} := \Big(\widetilde{\mathbf{V}}_1^{(\ell)},\ldots,\widetilde{\mathbf{V}}_K^{(\ell)}\Big)\;,$$

and make use of $\widehat{\widetilde{\Pi}}_L^{improved}$ given in (227) with the τ_k's obtained in **Step D2** above to define (in the light of (217), (218)) the *BS minimum-distance estimator*

$$\widehat{D_{\varphi}}(\boldsymbol{\Omega},\mathbf{P}) \;:=\; \widehat{D_{\widetilde{\varphi}}}(\widetilde{\boldsymbol{\Omega}},\widetilde{\mathbb{P}}) \;:=\; -\frac{1}{n}\log\widehat{\widetilde{\Pi}}_L^{improved}\;. \tag{228}$$

For many cases, the simulation burden needed for the computation of $\widehat{\widetilde{\Pi}}_L^{improved}$ — and thus of $\widehat{D_{\varphi}}(\boldsymbol{\Omega},\mathbf{P})$ — can be drastically reduced, especially for high dimensions K and large sample size $n\cdot L$. In fact, in terms of the notations $n_k := card(I_k^{(n)})$, $\widehat{W}_k^{(\ell)} := \sum_{i\in I_k^{(n)}}\widetilde{V}_i^{(\ell)}$ and

$$\widetilde{ISF}_k(x) \;:=\; \frac{d\widetilde{\mathbb{C}}^{*n_k}}{d\widetilde{U}_k^{*n_k}}(x) \;=\; \exp(n_k\cdot\Lambda_{\widetilde{\mathbb{C}}}(\tau_k) \;-\; x\cdot\tau_k) \tag{229}$$

(where $\widetilde{\mathbb{C}}^{*n_k}$ is the n_k-convolution of the measure $\widetilde{\mathbb{C}}$), one can rewrite (227) as

$$\widehat{\widetilde{\Pi}}_L^{improved} = \frac{1}{L}\sum_{\ell=1}^{L}\mathbf{1}_{\boldsymbol{\Lambda}}\big(\,(\frac{1}{n_1}\,\widehat{W}_1^{(\ell)},\ldots,\frac{1}{n_K}\,\widehat{W}_K^{(\ell)})\,\big)\cdot\prod_{k=1}^{K}\widetilde{ISF}_k(\widehat{W}_k^{(\ell)}). \tag{230}$$

with $K-$vector $\big(\frac{1}{n_1}\,\widehat{W}_1^{(\ell)},\ldots,\frac{1}{n_K}\,\widehat{W}_K^{(\ell)}\big)$. Clearly, the random variable $\widehat{W}_k^{(\ell)}$ $(k=1,\ldots,K)$ has distribution $\widetilde{U}_k^{*n_k}$. Hence, if $\widetilde{U}_k^{*n_k}$ can be *explicitly* constructed, then for the computation of $\widehat{\widetilde{\Pi}}_L^{improved}$ it suffices to simulate the $K\cdot L$ random variables

$\widehat{W}_k^{(\ell)}$ rather than the $n \cdot L$ random variables $\widetilde{V}_i^{(\ell)}$; notice that according to the right-hand side of (229), one can explicitly compute $ISF_k(\cdot)$ which can be interpreted as *Importance Sampling Factor pertaining to the block k*. In the case that $\widetilde{\zeta}$ is infinitely divisible, simulation issues may become especially comfortable. In the following, we exemplarily demonstrate the tractability of this reduction effect, for the BS minimization of the important power divergences (for which the infinite divisibility holds):

Example 58: Let φ_γ ($\gamma \in \mathbb{R}\backslash]1,2[$) be the power divergence generator from the Examples 39 and 40, $\mathbf{P} \in \mathbb{R}_{>0}^K$, $M_{\mathbf{P}} := \sum_{i=1}^K p_i > 0$ and $n_k = n \cdot p_k \in \mathbb{N}$ where we have employed our notation $n_k = n \cdot p_k \in \mathbb{N}$ for all $k \in \{1, \ldots, K\}$. Moreover, let $\widetilde{\mathbf{Q}}^* := (\widetilde{q}_1^*, \ldots, \widetilde{q}_K^*)$ be a proxy obtained by either proxy method 1 or 2.

Case 1: Example 48(a): $\gamma = 0, \widetilde{c} > 0$. There holds $\widetilde{U}_k^{*n_k} = GAM(\widetilde{c} \cdot M_{\mathbf{P}} - \tau_k, n_k \cdot \widetilde{c} \cdot M_{\mathbf{P}})$, with $\tau_k = \widetilde{c} \cdot M_{\mathbf{P}} \cdot (1 - \frac{p_k}{M_{\mathbf{P}} \cdot \widetilde{q}_k^*})$ for $\widetilde{q}_k^* > 0$ (the latter is equivalent to $\tau_k < \widetilde{c} \cdot M_{\mathbf{P}}$). Moreover, for all $x > 0$ one gets $\widetilde{ISF}_k(x) = \left(\frac{\widetilde{c} \cdot M_{\mathbf{P}}}{\widetilde{c} \cdot M_{\mathbf{P}} - \tau_k}\right)^{n_k \cdot \widetilde{c} \cdot M_{\mathbf{P}}} \cdot e^{-\tau_k \cdot x}$.

Case 2: 48(b): $\gamma \in (0,1), \widetilde{c} > 0$. We derive $\widetilde{U}_k^{*n_k} = C\big(POI(n_k \cdot \breve{\theta}), GAM\big(\frac{\widetilde{c} \cdot M_{\mathbf{P}}}{1-\gamma} - \tau_k, \frac{\gamma}{1-\gamma}\big)\big)$ with $\breve{\theta} := \frac{\widetilde{c} \cdot M_{\mathbf{P}}}{\gamma} \cdot \big(\frac{(\gamma-1) \cdot \tau_k}{\widetilde{c} \cdot M_{\mathbf{P}}} + 1\big)^{\gamma/(\gamma-1)}$ and $\tau_k = \widetilde{c} \cdot M_{\mathbf{P}} \cdot \frac{1 - \big(\frac{\widetilde{q}_k^* \cdot M_{\mathbf{P}}}{p_k}\big)^{\gamma-1}}{1-\gamma}$ for $\widetilde{q}_k^* > 0$. Furthermore,

$$\widetilde{ISF}_k(x) = e^{-\tau_k x} \cdot \exp\left(\frac{n_k \cdot \widetilde{c} \cdot M_{\mathbf{P}}}{\gamma} \cdot \left(\left(1 + \frac{\gamma - 1}{\widetilde{c} \cdot M_{\mathbf{P}}} \cdot \tau_k\right)^{\frac{\gamma}{\gamma-1}} - 1\right)\right), \qquad x \geq 0,$$

(where $x = 0$ covers the atom at zero).

Case 3: Example 48(c): $\gamma = 2, \widetilde{c} > 0$. One gets $\widetilde{U}_k^{*n_k} = N(n_k \cdot (1 + \frac{\tau_k}{\widetilde{c} \cdot M_{\mathbf{P}}}), \frac{n_k}{\widetilde{c} \cdot M_{\mathbf{P}}})$ with $\tau_k = \widetilde{c} \cdot M_{\mathbf{P}} \cdot (\frac{\widetilde{q}_k^* \cdot M_{\mathbf{P}}}{p_k} - 1)$ for $\widetilde{q}_k^* \in \mathbb{R}$. Moreover, for all $x \in \mathbb{R}$ one obtains $\widetilde{ISF}_k(x) = \exp\big(\frac{n_k \cdot \tau_k^2}{2\widetilde{c} \cdot M_{\mathbf{P}}} - (x - n_k) \cdot \tau_k\big)$.

Case 4: Example 48(d): $\gamma < 0, \widetilde{c} > 0$. It holds that $\widetilde{U}_k^{*n_k}$ has the (Lebesgue-)density

$$f_{\widetilde{U}_k^{*n_k}}(x) \; := \; \frac{\exp((\tau_k - \frac{\widetilde{c} \cdot M_{\mathbf{P}}}{1-\gamma}) \cdot x)}{\exp\Big(n_k \cdot \frac{\widetilde{c} \cdot M_{\mathbf{P}}}{\gamma} \cdot (1 + \frac{\gamma-1}{\widetilde{c} \cdot M_{\mathbf{P}}} \cdot \tau_k)^{\gamma/(\gamma-1)}\Big)} \cdot f_{\breve{\breve{Z}}}(x) \cdot 1_{]0,\infty[}(x), \qquad x \in \mathbb{R},$$

where $\tau_k = \widetilde{c} \cdot M_{\mathbf{P}} \cdot \frac{1 - \big(\frac{\widetilde{q}_k^* \cdot M_{\mathbf{P}}}{p_k}\big)^{\gamma-1}}{1-\gamma}$ for $\widetilde{q}_k^* > 0$, and $\breve{\breve{Z}}$ is a random variable with density $f_{\breve{\breve{Z}}}$ of a stable law with parameter-quadruple $(\frac{-\gamma}{1-\gamma}, 1, 0, -n_k \cdot \frac{(\widetilde{c} \cdot M_{\mathbf{P}})^{1/(1-\gamma)} \cdot (1-\gamma)^{-\gamma/(1-\gamma)}}{\gamma})$ (analogously to $\breve{Z}$ of Example 40 (d) but with $\widetilde{c}$ replaced by $\widetilde{c} \cdot M_{\mathbf{P}}$). Also,

$$\widetilde{ISF}_k(x) = e^{-\tau_k x} \cdot \exp\left(\frac{n_k \cdot \widetilde{c} \cdot M_{\mathbf{P}}}{\gamma} \cdot \left(\left(1 + \frac{\gamma - 1}{\widetilde{c} \cdot M_{\mathbf{P}}} \cdot \tau_k\right)^{\frac{\gamma}{\gamma-1}} - 1\right)\right), \qquad x > 0.$$

Case 5 : Example 48(e): $\gamma > 2, \widetilde{c} > 0$. We derive that $\widetilde{U}_k^{*n_k}$ has the (Lebesgue-)density

$$f_{\widetilde{U}_k^{*n_k}}(x) \; := \; \frac{\exp((\tau_k + \frac{\widetilde{c} \cdot M_{\mathbf{P}}}{\gamma-1}) \cdot x)}{\exp\Big(n_k \cdot \frac{\widetilde{c} \cdot M_{\mathbf{P}}}{\gamma} \cdot (1 + \frac{\gamma-1}{\widetilde{c} \cdot M_{\mathbf{P}}} \cdot \tau_k)^{\gamma/(\gamma-1)}\Big)} \cdot f_{\breve{\breve{Z}}}(-x), \qquad x \in \mathbb{R},$$

where $\tau_k = -\frac{\widetilde{c} \cdot M_{\mathbf{P}}}{\gamma-1} \cdot \big(1 - \big(\frac{\widetilde{q}_k^* \cdot M_{\mathbf{P}}}{p_k}\big)^{\gamma-1} \cdot 1_{]0,\infty[}(\widetilde{q}_k^*)\big)$ for $\widetilde{q}_k^* \in \mathbb{R}$, and $\breve{\breve{Z}}$ is a random variable with density $f_{\breve{\breve{Z}}}$ of a stable law with parameter-quadruple $(\frac{\gamma}{\gamma-1}, 1, 0, n_k \cdot \frac{(\widetilde{c} \cdot M_{\mathbf{P}})^{1/(1-\gamma)} \cdot (\gamma-1)^{\gamma/(\gamma-1)}}{\gamma})$ (analogously to $\breve{Z}$ of Example 40 (e) but with $\widetilde{c}$ replaced by $\widetilde{c} \cdot M_{\mathbf{P}}$). Furthermore,

$$\widetilde{ISF}_k(x) = e^{-\tau_k x} \cdot \exp\left(\frac{n_k \cdot \widetilde{c} \cdot M_{\mathbf{P}}}{\gamma} \cdot \left(\left(1 + \frac{\gamma - 1}{\widetilde{c} \cdot M_{\mathbf{P}}} \cdot \tau_k\right)^{\frac{\gamma}{\gamma-1}} - 1\right)\right), \qquad x \in \mathbb{R}.$$

Case 6: Example 50(a): $\gamma = 1, \widetilde{c} > 0$, anchor point $c = 0$. It holds that $\widetilde{U}_k^{*n_k}$ is the probability distribution $\frac{1}{\widetilde{c} \cdot M_{\mathbf{P}}} \cdot POI\Big(n_k \cdot \widetilde{c} \cdot M_{\mathbf{P}} \cdot \exp(\frac{\tau_k}{\widetilde{c} \cdot M_{\mathbf{P}}})\Big)$ with support on the lattice $\Big\{\frac{j}{\widetilde{c} \cdot M_{\mathbf{P}}}, \, j \in \mathbb{N}_0\Big\}$, where $\tau_k = \widetilde{c} \cdot \log\Big(\frac{M_{\mathbf{P}} \cdot \widetilde{q}_k^*}{p_k}\Big)$ for $\widetilde{\omega}_k > 0$. Moreover, for all $j \in \mathbb{N}_0$ we obtain (by setting $x := \frac{j}{\widetilde{c} \cdot M_{\mathbf{P}}}$)

$$\widetilde{ISF}_k\left(\frac{j}{\widetilde{c} \cdot M_{\mathbf{P}}}\right) = \exp\left(n_k \cdot \widetilde{c} \cdot M_{\mathbf{P}} \cdot \left(\exp\left(\frac{\tau_k}{\widetilde{c} \cdot M_{\mathbf{P}}}\right) - 1\right) - m \cdot \frac{\tau_k}{\widetilde{c} \cdot M_{\mathbf{P}}}\right).$$

Case 7: Example 50(b): $\gamma = 1, \widetilde{c} = 1$, anchor point $c \in \mathbb{R}$. For $M_P = 1$, $\widetilde{U}_k^{*n_k}$ is the shifted Poisson distribution $POI\left(n_k \cdot e^{c+\tau_k}\right) + n_k \cdot (1-e^c)$ with support on the lattice $\{j + n_k \cdot (1-e^c),\, j \in \mathbb{N}_0\}$, where $\tau_k = \log\left(\frac{\widetilde{q}_k^*}{p_k} + e^c - 1\right) - c$ for $\widetilde{q}_k^* > p_k \cdot (1-e^c)$. Furthermore, for all $j \in \mathbb{N}_0$ we obtain (by setting $x := j + n_k \cdot (1-e^c)$)

$$\widetilde{ISF}_k\left(j + n_k \cdot (1-e^c)\right) = \exp\left(n_k \cdot e^c \cdot (e^{\tau_k} - 1) - j \cdot \tau_k\right).$$

Notice that the mass of $\widetilde{U}_k^{*n_k}$ at zero depends on the value of the anchor point c, since $\widetilde{U}_k^{*n_k}[\{0\}] > 0$ if and only if $c = \log(1 + \frac{\ell}{n_k})$ for some $\ell \in \mathbb{N}_0$; moreover, $\widetilde{U}_k^{*n_k}\left[\,]0,\infty[\,\right] = 1$ if $c < 0$ and $\widetilde{U}_k^{*n_k}\left[\,]-\infty,0[\,\right] > 0$ if $c > 0$.

Remark 59: (a) One can explicitly see in all cases of the above Example 58 that all ingredients for computation are at hand. (b) For both Cases 4 and 5 in the above Example 58, algorithms for simulation can be obtained by adapting e.g. the works of Devroye [111] and Devroye & James [112].

B. Estimators for the statistical minimization problem

1) General case, part 1: In the previous Subsection VI-A, as a first step we have estimated

$$\Pi\left[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\boldsymbol{\Omega}}\right]$$

in terms of the improved IS estimator $\widehat{\widehat{\Pi}}_L^{improved}$. From this, as a second step, we have derived — on the basis of Theorem 10 — the estimator

$$\widehat{D_\varphi}(\boldsymbol{\Omega}, \mathbf{P}) \;:=\; -\frac{1}{n} \log \widehat{\widehat{\Pi}}_L^{improved} \qquad \text{(cf. (228))}$$

of the minimum distance $D_\varphi(\boldsymbol{\Omega}, \mathbf{P}) := \inf_{\mathbf{Q} \in \boldsymbol{\Omega}} D_\varphi(\mathbf{Q}, \mathbf{P})$, where $\mathbf{P} \in \mathbb{R}_{>0}^K$ and $\boldsymbol{\Omega} \subset \mathbb{R}^K$. Recall that $\widetilde{\boldsymbol{\Omega}} := \boldsymbol{\Omega}/M_{\mathbf{P}}$ with $M_{\mathbf{P}} := \sum_{i=1}^K p_i > 0$, and that

$$\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} = \Big(\frac{1}{n} \sum_{i \in I_1^{(n)}} \widetilde{W}_i, \ldots, \frac{1}{n} \sum_{i \in I_K^{(n)}} \widetilde{W}_i\Big) \qquad \text{(cf. (23))}$$

where $\widetilde{\mathbf{W}} := (\widetilde{W}_1, \ldots, \widetilde{W}_n)$ is a random vector consisting of components $\widetilde{W}_i$ which are i.i.d. copies of the random variable $\widetilde{W}$ whose distribution is $\Pi[\widetilde{W} \in \cdot\,] = \widetilde{\zeta}[\,\cdot\,]$ obeying the representation (21).

In contrast, we now proceed as follows: as a first step, we derive an improved estimator $\widehat{\Pi}_L^{improved}$ of

$$\Pi_{X_1^n}\left[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \boldsymbol{\Omega}\right]$$

where $\boldsymbol{\Omega} \in \mathbb{S}^K$ is a set of probability vectors which satisfies the regularity properties (7) and the finiteness property (9). Recall that

$$\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \;:=\; \begin{cases} \left(\frac{\sum_{i \in I_1^{(n)}} W_i}{\sum_{k=1}^K \sum_{i \in I_k^{(n)}} W_i}, \ldots, \frac{\sum_{i \in I_K^{(n)}} W_i}{\sum_{k=1}^K \sum_{i \in I_k^{(n)}} W_i}\right), & \text{if } \sum_{j=1}^n W_j \neq 0, \\ (\infty, \ldots, \infty) =: \boldsymbol{\infty}, & \text{if } \sum_{j=1}^n W_j = 0, \end{cases} \qquad \text{(cf. (33))}$$

where $(X_i)_{i \in \mathbb{N}}$ is a sequence of random variables with values in $\mathcal{Y} := \{d_1, \cdots, d_K\}$ such that

$$\lim_{n \to \infty} \left(\frac{n_1}{n}, \ldots, \frac{n_K}{n}\right) = (p_1, \ldots, p_K) \qquad \text{a.s.} \qquad \text{cf. ((30))}$$

holds for some probability vector $\mathbb{P} := (p_1, \ldots, p_K) \in \mathbb{S}_{>0}^K$, by employing the notation

$$n_k \;:=\; card\big(\{i \in \{1, \ldots, n\} : \; X_i = d_k\}\big) \;=:\; card(I_k^{(n)}) \qquad \text{(cf. (29))};$$

hence, on the k-th block of indexes $I_k^{(n)}$ all the X_i's share the same value d_k. Moreover, recall that $(W_i)_{i \in \mathbb{N}}$ is a family of independent and identically distributed $\mathbb{R}-$valued random variables with probability distribution $\zeta[\cdot] := \Pi[W_1 \in \cdot\,]$ being connected with the divergence generator $\varphi \in \Upsilon(]a,b[)$ via the representability (6), such that $(W_i)_{i \in \mathbb{N}}$ is independent of $(X_i)_{i \in \mathbb{N}}$.

As a second step (see Subsubsection VI-B2 below), for the important special case of the power-divergence generators φ_γ (cf. (43)) we employ the Propositions 22 to 27 in order to deduce via the corresponding $\widehat{\Pi}_L^{improved}$ the estimators (e.g. for $\gamma < 0$)

$$\widehat{D_{\widetilde{c} \cdot \varphi_\gamma}(\boldsymbol{\Omega}, \mathbb{P})} \;:=\; \frac{\widetilde{c}}{\gamma \cdot (\gamma - 1)} \cdot \left\{ \left(1 + \frac{\gamma}{\widetilde{c}} \cdot \frac{1}{n} \cdot \log \widehat{\Pi}_L^{improved}\right)^{1-\gamma} - 1 \right\},$$

of the minimum power divergences

$$D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbf{\Omega}, \mathbb{P}) := \inf_{\mathbb{Q}\in\mathbf{\Omega}} D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbb{Q}, \mathbb{P})$$

as well as connected estimators of important deterministic transformations thereof.

As a third step (see Subsubsection VI-B3 below), on the basis of Subsubsection IV-C2 we derive estimators of bounds of $D_\varphi(\mathbf{\Omega}, \mathbb{P})$ for more general divergence generators φ.

Let us start with the above-mentioned first step, by remarking that the development of the estimator $\widehat{\Pi}_L^{improved}$ works quite analogously to that of $\widehat{\widehat{\Pi}}_L^{improved}$ in the previous Subsection VI-A. To make this even more transparent, we employ the notation $p_{n,k}^{emp} := n_k/n$ (cf. (29)) and label all random vectors of length n in the same way as above: we sort the already given and thus fixed data X_i's in such a way that the first n_1 of them share the same value d_1, and so on, until the last block with length n_K in which the data have common value d_K.

In the light of the above considerations, we could achieve a naive estimate $\widehat{\Pi}_L^{naive}$ of $\Pi_{X_1^n}[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \mathbf{\Omega}]$ through the following procedure. We simulate independently L replicates $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}$ of the vector $\mathbf{W} := (W_1, \ldots, W_n)$, with independent entries under $\mathbb{Z}$ (cf. (6)); those realizations do not depend on the X_i's. Then we construct

$$\widehat{\Pi}_L^{naive} := \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}_{\mathbf{\Omega}}\left(\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}^{(\ell)}}\right). \tag{231}$$

However, this procedure is time costly, since the estimate given in (231) has a very bad hit rate. Hence, analogously to Subsection VI-A we apply again an "efficient Importance Sampling (IS)" scheme in the sense of Sadowsky & Bucklew [313]. This will involve the simulation of L independent $n-$tuples $\mathbf{V}^{(\ell)} := \left(V_n^{(\ell)}, \ldots, V_n^{(\ell)}\right)$ with common distribution S on $\mathbb{R}^n$, such that $\mathbb{Z}^{\otimes n}$ is (measure-)equivalent with respect to S. In fact, we rewrite $\Pi_{X_1^n}[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \mathbf{\Omega}]$ as

$$\Pi_{X_1^n}[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \mathbf{\Omega}] = E_S\Big[\frac{\mathrm{d}\mathbb{Z}^{\otimes n}}{\mathrm{d}S}(V_1, \ldots, V_n) \cdot \mathbf{1}_{\mathbf{\Omega}}(\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}})\Big] \tag{232}$$

where S designates any IS distribution of the vector $\mathbf{V} := (V_1, \ldots, V_n)$, and $E_S[\,\cdot\,]$ denotes the corresponding expectation operation. Notice that S is a *random* probability distribution on $\mathbb{R}^n$; in fact, S is a conditional probability distribution given X_1^n, and thus it would be more precise to write $S|X_1^n$ instead of S; for the sake of brevity, we omit $|X_1^n$.

As a consequence of (232), for adequately chosen S, an improved estimator of $\Pi_{X_1^n}[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \mathbf{\Omega}]$ is given by

$$\widehat{\Pi}_L^{improved} := \frac{1}{L}\sum_{\ell=1}^{L} \frac{\mathrm{d}\mathbb{Z}^{\otimes n}}{\mathrm{d}S}(V_1^{(\ell)}, \ldots, V_n^{(\ell)}) \cdot \mathbf{1}_{\mathbf{\Omega}}(\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}^{(\ell)}}), \tag{233}$$

which also estimates $\inf_{\mathbb{Q}\in\mathbf{\Omega}}\ \inf_{m\neq 0} D_\varphi(m\cdot\mathbb{Q}, \mathbb{P})$ by the virtue of (36).

Let us now deal with the concrete construction of a reasonable S. Given some (typically) large integer M, we start with the realization $\mathbf{W}^* := (W_1^*, \ldots, W_M^*)$ such that $\mathbf{Q}^* := \boldsymbol{\xi}_{M,\mathbf{X}}^{w\mathbf{W}^*} \in int(\mathbf{\Omega})$. This may be given in advance or it may be achieved by drawing replicates $\mathbf{W} = (W_1, \ldots, W_M)$ under $\mathbb{Z}^{\otimes M}$ until the first time where $\boldsymbol{\xi}_{M,\mathbf{X}}^{w\mathbf{W}}$ belongs to $int(\mathbf{\Omega})$. Notice that by the nature of $\mathbf{\Omega}$, $\mathbf{Q}^*$ is a probability vector which has the K components

$$q_k^* := \sum_{i=1}^{M} \frac{W_i^*}{\sum_{j=1}^M W_j^*}\mathbf{1}_{\{d_k\}}(X_i), \qquad\qquad k = 1, \ldots, K. \tag{234}$$

Before we proceed, let us give the substantial remark that changing $(V_1, \ldots, V_n)$ drawn under S to $(c\cdot V_1, \ldots, c\cdot V_n)$ for any $c \neq 0$ yields $\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}} = \boldsymbol{\xi}_{n,\mathbf{X}}^{w\,c\cdot\mathbf{V}}$ so that the distribution S is not uniquely determined. Amongst all candidates, we choose the — uniquely determined — S which is the Kullback-Leibler projection of $\mathbb{Z}^{\otimes n}$ on the set of all probability distributions on $\mathbb{R}^n$ such that the K "non-normalized" moment constraints

$$E_S[\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{V}}] = \boldsymbol{\xi}_{M,\mathbf{X}}^{\mathbf{W}^*} \tag{235}$$

(rather than the normalized $E_S[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}}] = \boldsymbol{\xi}_{M,\mathbf{X}}^{w\mathbf{W}^*}$) are satisfied, with the non-normalized vectors

$$\boldsymbol{\xi}_{M,\mathbf{X}}^{\mathbf{W}^*} := \left(\frac{1}{M}\sum_{j=1}^{M} W_j^*\right)\cdot Q^* =: \overline{W^*}\cdot Q^*, \qquad \boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{V}} := \left(\frac{1}{n}\sum_{j=1}^{n} V_j\right)\cdot \boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}}\ .$$

As already indicated above, this projection S is a well-determined unique distribution on $\mathbb{R}^n$ and — as we shall see in Proposition 60 below — it is such that $\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}}$ belongs to $\boldsymbol{\Omega}$ with probability bounded away from 0 as n increases, when $(V_1, \ldots, V_n)$ are drawn under S. Therefore, this IS distribution produces an estimate of $\Pi_{X_1^n}[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \boldsymbol{\Omega}]$.

In order to justify the above construction of S, we give the following result, which states that the IS sampling distribution S yields a good hitting rate. Its proof will be given in Appendix H.

Proposition 60: With the above definition of S, $\liminf_{n\to\infty} S\left[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}} \in \boldsymbol{\Omega}\right]$ is bounded away from 0.

We now come to the detailed construction of S. The constraints (235) can be written in explicit form as

$$E_S\Big[\frac{1}{n_k}\sum_{i\in I_k^{(n)}} V_i\Big] \;=\; \overline{W^*}\cdot\frac{q_k^*}{p_{n,k}^{emp}}\,, \qquad k=1,\ldots,K. \tag{236}$$

The distribution S can be obtained by blocks. Indeed, let us define S^k as the Kullback-Leibler (KL) projection of $\zeta^{\otimes n_k}$ on the set of all distributions on $\mathbb{R}^{n_k}$ such that (236) holds. We define the resulting S as the product distribution of those S^k 's. To obtain the latter, we start by defining U_k as the KL projection of ζ on the set of all measures Q on $\mathbb{R}$ under (236). Then,

$$dU_k(v) = \exp(\tau_k v - \Lambda_\zeta(\tau_k))\, d\zeta(v)\,, \tag{237}$$

where $\tau_k \in int(dom(MGF_\zeta))$ is the unique solution of the equation

$$\Lambda'_\zeta(\tau_k) = \overline{W^*}\cdot\frac{q_k^*}{p_{n,k}^{emp}}$$

and thus — by relation (281) of Appendix F — we can compute explicitly

$$\tau_k = \varphi'\left(\frac{\overline{W^*}\cdot q_k^*}{p_{n,k}^{emp}}\right).$$

The distribution S^k is then defined by

$$S^k := \underbrace{U_k\otimes\cdots\otimes U_k}_{n_k\text{times}}$$

from which we obtain

$$S := S^1\otimes\cdots\otimes S^K.$$

With this construction, it holds

$$\frac{dS}{d\zeta^{\otimes n}}(v_1,\ldots,v_n) = \exp\left(\sum_{k=1}^K\left(\sum_{i\in I_k^{(n)}}\tau_k\cdot v_i - \Lambda_\zeta(\tau_k)\right)\right)$$

which proves that S is indeed the KL projection of $\zeta^{\otimes n}$ we aimed at.

Therefore, $\mathbf{V}$ is composed of K independent blocks of length n_k each, and the $k-$th subvector $\mathbf{V}_k$ consists of all the random variables V_i whose index i satisfies $X_i = d_k$. Within $\mathbf{V}_k$, all components are i.i.d. with same distribution U_k on $\mathbb{R}$ defined through

$$\frac{dU_k}{d\zeta}(u) = \exp\{\tau_k\cdot u - \Lambda_\zeta(\tau_k)\} = \frac{\exp\{\tau_k\cdot u\}}{MGF_\zeta(\tau_k)},$$

which leads to the moment generating function

$$dom(MGF_\zeta) - \tau_k \;\ni\; z \;\mapsto MGF_{U_k}(z) := \int_{\mathbb{R}} e^{zy}dU_k(y) = \frac{MGF_\zeta(z+\tau_k)}{MGF_\zeta(\tau_k)}.$$

Let us remark that U_k can be interpreted as the distorted distribution of ζ with the distortion parameter τ_k (in some cases, this distortion even becomes a tilting/dampening).

The estimator $\widehat{\Pi}_L^{improved}$ defined in (233) can be implemented through the following algorithm:

Step S1
Choose some (typically large) M and simulate repeatedly i.i.d. vectors $(W_1, .., W_M)$ — whose independent components have common distribution ζ — until $\boldsymbol{\xi}_{M,\mathbf{X}}^{w\mathbf{W}}$ belongs to $\boldsymbol{\Omega}$. Call $(W_1^*, .., W_M^*)$ the corresponding vector and $\overline{W^*}$ the arithmetic mean

of its components. Moreover, denote by $\boldsymbol{\xi}_{M,\mathbf{X}}^{w\mathbf{W}^*}$ the corresponding normalized weighted empirical measure, identified with the $K-$component vector $Q^* := (q_1^*, .., q_K^*)$ with q_k^* defined in (234).

Step S2

For all $k \in \{1, \ldots, K\}$ compute $\tau_k = \varphi'\left(\frac{\overline{W^*}\cdot q_k^*}{p_{n,k}^{emp}}\right)$.

Step S3

For all $\ell \in \{1, \ldots, L\}$ simulate independently for all $k \in \{1, \ldots, K\}$ a row vector $\mathbf{V}_k^{(\ell)} := \left(V_{k_1}^{(\ell)}, ..., V_{k_{n_k}}^{(\ell)}\right)$ with independent components with common distribution U_k defined in (221). Concatenate these vectors to define the row vector $\mathbf{V}^{(\ell)}$.

Step S4

Compute the estimator $\widehat{\Pi}_L^{improved}$ by making use of the formula (233) which turns into the explicit form

$$\widehat{\Pi}_L^{improved} = \frac{1}{L}\sum_{\ell=1}^{L} \exp\left(\sum_{k=1}^{K}\left(n_k \cdot \Lambda_{\mathfrak{C}}(\tau_k) - \tau_k \cdot \sum_{i \in I_k^{(n)}} V_i^{(\ell)}\right)\right) \cdot \mathbf{1}_{\mathbf{\Omega}}\left(\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}^{(\ell)}}\right) \tag{238}$$

Analogously to the paragraph right after (228) of the previous Subsection VI-A, in many cases we may improve the simulation burden needed for the computation of the estimator $\widehat{\Pi}_L^{improved}$. In fact, in terms of the notations $\widehat{W}_k^{(\ell)} := \sum_{i \in I_k^{(n)}} V_i^{(\ell)}$ we can rewrite (238) as

$$\widehat{\Pi}_L^{improved} = \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}_{\mathbf{\Omega}}\left(\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}^{(\ell)}}\right) \cdot \prod_{k=1}^{K} ISF_k\left(\widehat{W}_k^{(\ell)}\right) \tag{239}$$

with

$$ISF_k(x) \;:=\; \exp(n_k \cdot \Lambda_{\mathfrak{C}}(\tau_k) \,-\, x \cdot \tau_k) \tag{240}$$

and

$$\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}^{(\ell)}} \quad = \quad \begin{cases} \left(\frac{\widehat{W}_1^{(\ell)}}{\sum_{k=1}^K \widehat{W}_k^{(\ell)}}, \ldots, \frac{\widehat{W}_K^{(\ell)}}{\sum_{k=1}^K \widehat{W}_k^{(\ell)}}\right), & \text{if } \sum_{k=1}^K \widehat{W}_k^{(\ell)} \neq 0, \\ (\infty, \ldots, \infty) =: \boldsymbol{\infty}, & \text{if } \sum_{k=1}^K \widehat{W}_k^{(\ell)} = 0\,. \end{cases} \tag{241}$$

Clearly, the random variable $\widehat{W}_k^{(\ell)}$ $(k = 1, \ldots, K)$ has distribution $U_k^{*n_k}$. Hence, if $U_k^{*n_k}$ can be *explicitly* constructed, then for the computation of $\widehat{\Pi}_L^{improved}$ it suffices to independently simulate the $K \cdot L$ random variables $\widehat{W}_k^{(\ell)}$ (rather than the $n \cdot L$ random variables $V_i^{(\ell)}$). In the following subsubsection, we exemplarily demonstrate the tractability of this reduction effect.

2) BS minimization of power divergences and related quantities: ´

Consider the special case of power divergence generators $\varphi := \widetilde{c} \cdot \varphi_\gamma$ $(\gamma \in \mathbb{R}\backslash]1, 2[)$ of the Examples 39 and 40. The corresponding estimators $\widehat{\Pi}_L^{improved}$ can be obtained as follows:

(i) within the results of Example 58, set $M_{\mathbf{P}} = 1$, and replace $\widetilde{q}_k^*$ by $\overline{W^*} \cdot q_k^*$ as well as p_k by $p_{n,k}^{emp}$; accordingly, $\widetilde{U}_k^{*n_k}$ turns into $U_k^{*n_k}$ and $\widetilde{ISF}_k$ into ISF_k;

(ii) simulate independently the random variables $\widehat{W}_k^{(\ell)}$ from $U_k^{*n_k}$ $\quad(k \in \{1, \ldots, K\}, \ell \in \{1, \ldots, L\})$;

(iii) plug in the results of (i),(ii) into (239), (240), and (241) in order to concretely compute $\widehat{\Pi}_L^{improved}$.

From this, we can easily generate improved estimators of the power divergences $\inf_{\mathbf{Q}\in\mathbf{\Omega}} D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbf{Q}, \mathbb{P})$ — and more generally, improved estimators of all the infimum-quantities (e.g. Renyi divergences) respectively supremum-quantities in the parts (b) of the Propositions 22, 23, 24, 25, 26 and 27 with $A = 1$ — by simply replacing $\Pi_{X_1^n}[\boldsymbol{\xi}_n^{w\mathbf{W}} \in \mathbf{\Omega}]$ (respectively, its variants) by the corresponding estimator $\widehat{\Pi}_L^{improved}$. If — in the light of Remark 15(vi) — the $\mathbb{P} = (p_1, \ldots, p_K)$ is a pregiven known probability vector [34] (rather than the limit of the vector of empirical frequencies/masses of a sequence of random variables X_i, cf. (30)), then we proceed analogously as above by replacing $p_{n,k}^{emp}$ with p_k; correspondingly, we obtain improved estimators of all the infimum-quantities respectively supremum-quantities (e.g. Renyi entropies, diversity indices) in the parts (a) of the Propositions 22, 23, 24, 25, 26 and 27 with $A = 1$.

For the sake of brevity, in the following we only present explicitly the outcoming improved estimators for the power divergences (in the "X_i-context"). Indeed, we simply replace the $\Pi_{X_1^n}[\boldsymbol{\xi}_n^{w\mathbf{W}} \in \mathbf{\Omega}]$ in the formulas (84), (92), (99) (with $A = 1$) by

[34] e.g. the uniform distribution $\mathbb{P}^{unif}$ on $\{1, \ldots, K\}$

the improved estimator $\widehat{\Pi}_L^{improved}$ obtained through (i) to (iii); for arbitrarily fixed $\widetilde{c} > 0$, this leads to the *improved power-divergence estimators (BS estimators of power divergences)*

$$D_{\widehat{\widetilde{c}\cdot\varphi_\gamma}(\boldsymbol{\Omega}}, \mathbb{P}) := -\frac{\widetilde{c}}{\gamma(\gamma-1)} \left\{ 1 - \left(1 + \frac{\gamma}{\widetilde{c}} \cdot \frac{1}{n} \cdot \log \widehat{\Pi}_L^{improved} \right)^{1-\gamma} \right\}, \qquad \gamma \in]-\infty, 0[\cup]0,1[\cup[2,\infty[, \tag{242}$$

$$D_{\widehat{\widetilde{c}\cdot\varphi_0}(\boldsymbol{\Omega}}, \mathbb{P}) := -\frac{1}{n} \log \widehat{\Pi}_L^{improved}, \qquad \gamma = 0, \tag{243}$$

$$D_{\widehat{\widetilde{c}\cdot\varphi_1}(\boldsymbol{\Omega}}, \mathbb{P}) := -\widetilde{c} \cdot \log \left(1 + \frac{1}{\widetilde{c}} \cdot \frac{1}{n} \cdot \log \widehat{\Pi}_L^{improved} \right), \qquad \gamma = 1. \tag{244}$$

Let us finally remark that from the above-mentioned Steps S1 to S4 (and analogously D1 to D4) one can see that for our BS method we basically need only a fast and accurate — pseudo, true, natural, quantum — random number generator. The corresponding computations can be principally run in parallel, and require relatively moderate computer memory/storage; a detailed discussion is beyond the scope of this paper, given its current length.

3) General case, part 2: The algorithm which is presented in this section aims at the evaluation of the bounds

$$\inf_{m\neq 0} D_\varphi(m \cdot \boldsymbol{\Omega}, \mathbb{P}) = \inf_{\mathbb{Q}\in\boldsymbol{\Omega}} D_\varphi(m(\mathbb{Q}) \cdot \mathbb{Q}, \mathbb{P}) \overset{(1)}{=} D_\varphi(m(\mathbb{Q}^*) \cdot \mathbb{Q}^*, \mathbb{P}) \leq D_\varphi(\boldsymbol{\Omega}, \mathbb{P}) \leq D_\varphi(\mathbb{Q}^*, \mathbb{P}) \tag{245}$$

obtained in Section IV-C2, where $\mathbb{Q}^*$ satisfies the above equality (1). The estimator of the lower bound in (245) is $\widehat{D} := -\frac{1}{n} \log \widehat{\Pi_L^{improved}}$ defined in (238).

We now turn to an estimate of the upper bound. Consider for any fixed $\mathbb{Q} := (q_1, .., q_K)$ in $\mathbb{S}_{>0}^K$ the real number $m_n(\mathbb{Q})$ which satisfies

$$D_\varphi(m_n(\mathbb{Q}) \cdot \mathbb{Q}, \mathbb{P}_n^{emp}) = \inf_{m\neq 0} D_\varphi(m \cdot \boldsymbol{\Omega}, \mathbb{P}_n^{emp})$$

where $\mathbb{P}_n^{emp}$ was defined in the course of (29). Such $m_n(\mathbb{Q})$ is well defined for all $\mathbb{Q}$ since it satisfies the equation (in m)

$$\frac{d}{dm} D_\varphi(m \cdot \mathbb{Q}, \mathbb{P}_n^{emp}) = \sum_{k=1}^{K} q_k \cdot \varphi' \left(\frac{m \cdot q_k}{p_{n,k}^{emp}} \right) = 0. \tag{246}$$

Since the mapping $m \to D_\varphi(m \cdot \mathbb{Q}, \mathbb{P})$ is convex and differentiable, existence and uniqueness of $m_n(\mathbb{Q})$ hold; furthermore, $m_n(\mathbb{Q}) \in \left] \min_k p_{n,k}^{emp}/q_k, \max_k p_{n,k}^{emp}/q_k \right[$ since $\frac{d}{dm} D_\varphi(m \cdot \mathbb{Q}, \mathbb{P}_n^{emp})$ is negative when $m = \min_k p_{n,k}^{emp}/q_k$ and positive when $m = \max_k p_{n,k}^{emp}/q_k$.

An estimate of the distribution $\mathbb{Q}^*$ is required. This can be achieved as follows:

- Estimate $\inf_{m\neq 0} D_\varphi(m \cdot \boldsymbol{\Omega}, \mathbb{P})$ through $\widehat{D} := -\frac{1}{n} \log \widehat{\Pi}_L^{improved}$ defined in (238).
- Set $i = 0$.
- Get some $\mathbb{Q}_i := (q_{i,1}, \ldots, q_{i,K})$ in $\boldsymbol{\Omega}$; this can be obtained by simulating runs of vectors $(W_1, .., W_n)$ through i.i.d. sampling under $\mathbb{Ϛ}$. Evaluate $m_n(\mathbb{Q}_i)$ by solving (246) (with $q_{i,k}$ instead of q_k) for m, which is a fast calculation by the bisection method.
- If $D_\varphi(m_n(\mathbb{Q}_i) \cdot \mathbb{Q}_i, \mathbb{P}_n^{emp}) < \widehat{D} + \eta$ for some small $\eta > 0$, then the proxy of $\mathbb{Q}^*$ is $\mathbb{Q}_i$, denoted by $\widehat{\mathbb{Q}^*}$.
- Else set $i \leftarrow i + 1$ and get $\mathbb{Q}_i$ in $\boldsymbol{\Omega} \cap \{\mathbb{Q} : D_\varphi(\mathbb{Q}, \mathbb{P}_n^{emp}) < D_\varphi(\mathbb{Q}_{i-1}, \mathbb{P}_n^{emp})\}$ and iterate.

That this algorithm converges in the sense that it produces some $\widehat{\mathbb{Q}^*}$ is clear. Since by (245)

$$D_\varphi(m(\mathbb{Q}^*) \cdot \mathbb{Q}^*, \mathbb{P}) \leq D_\varphi(\boldsymbol{\Omega}, \mathbb{P}) \leq D_\varphi(\mathbb{Q}^*, \mathbb{P}),$$

we have obtained both estimated lower and upper bounds for $D_\varphi(\boldsymbol{\Omega}, \mathbb{P})$.

That the upper bound is somehow optimal can be seen from the power case developed in Section VI-B2. Indeed, in this case the solution of equation (246) is explicit and produces $m(\mathbb{Q})$ as a function of $D_\varphi(\mathbb{Q}, \mathbb{P})$ through a Hellinger integral, and the mapping $\mathbb{Q} \to D_\varphi(m(\mathbb{Q}) \cdot \mathbb{Q}, \mathbb{P})$ is increasing with respect to $D(\mathbb{Q}, \mathbb{P})$. Hence, $\mathbb{Q} \to \inf_{m\neq 0} D_\varphi(m \cdot \mathbb{Q}, \mathbb{P})$ is minimal when $D_\varphi(\mathbb{Q}, \mathbb{P})$ is minimal as $\mathbb{Q} \in \boldsymbol{\Omega}$. Therefore, $\mathbb{Q}^* \in \arg\inf_{\mathbb{Q}\in\boldsymbol{\Omega}} D_\varphi(m(\mathbb{Q}) \cdot \mathbb{Q}, \mathbb{P})$ also satisfies $\mathbb{Q}^* \in \arg\inf_{\mathbb{Q}\in\boldsymbol{\Omega}} D_\varphi(\mathbb{Q}, \mathbb{P})$.

Appendix A
Proofs — Part 1

Proof of Theorem 10.
This is a straightforward application of the classical Cramer-type Large Deviation Theorem in the vector case (see Theorem 2.2.30 and Corollary 6.1.6 in Dembo & Zeitouni [108]). Recall that above we have transformed the original problem into

a context where the second argument in $D_\varphi(\cdot,\cdot)$ is a probability vector, as follows: in terms of $M_{\mathbf{P}} := \sum_{i=1}^{K} p_i > 0$ we normalized $\widetilde{\mathbb{P}} := \mathbf{P}/M_{\mathbf{P}}$, and $\widetilde{\mathbf{Q}} := \mathbf{Q}/M_{\mathbf{P}}$ for $\mathbf{Q}$ in $\mathbf{\Omega}$. With $\widetilde{\varphi} \in \Upsilon(]a,b[)$ defined through $\widetilde{\varphi} := M_{\mathbf{P}} \cdot \varphi$, we have obtained

$$D_\varphi(\mathbf{Q},\mathbf{P}) = \sum_{k=1}^{K} p_k \cdot \varphi\left(\frac{q_k}{p_k}\right) = \sum_{k=1}^{K} M_{\mathbf{P}} \cdot \widetilde{p}_k \cdot \frac{\varphi\left(\frac{M_{\mathbf{P}}\cdot\widetilde{q}_k}{M_{\mathbf{P}}\cdot\widetilde{p}_k}\right)}{M_{\mathbf{P}}} = D_{\widetilde{\varphi}}(\widetilde{\mathbf{Q}},\widetilde{\mathbb{P}}) \qquad \text{(cf. (12))}.$$

It has followed that the solution of (8) coincides with the one of the problem of finding

$$\widetilde{\Phi}_{\widetilde{\mathbb{P}}}(\widetilde{\mathbf{\Omega}}) := \inf_{\widetilde{\mathbf{Q}}\in\widetilde{\mathbf{\Omega}}} D_{\widetilde{\varphi}}(\widetilde{\mathbf{Q}},\widetilde{\mathbb{P}}), \qquad \text{with } \widetilde{\mathbf{\Omega}} := \mathbf{\Omega}/M_{\mathbf{P}} \qquad \text{(cf. (13))}.$$

So let us continue by tackling (13). From the assumptions on $\widetilde{\varphi}$ and the requirement (21) one can see that

$$\widetilde{W}_1 \text{ has moment generating function } t \to E_{\Pi}[e^{z\cdot\widetilde{W}_1}] = MGF_{\widetilde{\zeta}}(z) \text{ which is finite on a non-void neighborhood of } 0, \tag{247}$$

$$E_{\Pi}[\widetilde{W}_1] = 1, \tag{248}$$

since $\widetilde{\varphi}(1) = 0 = \widetilde{\varphi}'(1)$. With the help of these, we obtain the following

Proposition 61: Under the assumptions of Theorem 10, for any set $\widetilde{\mathbf{\Omega}} \subset \mathcal{M} := \mathbb{R}^K$ with (7) one has

$$\begin{aligned} -\inf_{\widetilde{\mathbf{Q}}\in int(\widetilde{\mathbf{\Omega}})} D_\varphi\left(\widetilde{\mathbf{Q}},\widetilde{\mathbb{P}}\right) &\leq \liminf_{n\to\infty} \frac{1}{n}\log \Pi\left[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\mathbf{\Omega}}\right] \\ &\leq \limsup_{n\to\infty} \frac{1}{n}\log \Pi\left[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in \widetilde{\mathbf{\Omega}}\right] \leq -\inf_{\widetilde{\mathbf{Q}}\in cl(\widetilde{\mathbf{\Omega}})} D_\varphi\left(\widetilde{\mathbf{Q}},\widetilde{\mathbb{P}}\right). \end{aligned} \tag{249}$$

Proof of Proposition 61.
Recall from Remark 15(v) that $I_k^{(n)} := \{i \in \{1,\ldots,n\} : \widetilde{x}_i = d_k\}$ and $n_k := card(I_k^{(n)})$ denotes the number of elements therein ($k \in \{1,\ldots,K\}$), i.e. n_k is the number of the $\widetilde{x}_i$'s which equal d_k. We follow the line of proof of Theorem 2.2.30 in Dembo & Zeitouni [108], which states the large deviation principle (LDP) for the vector of partial sums of random vectors in $\mathbb{R}^K$, where we also use Corollary 6.1.6 in [108] in relation with condition (247). Indeed, since the $k-$th component of the vector $\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}}$ is the $1/n-$fold of the sum of the $\widetilde{W}_i$'s for which the corresponding $\widetilde{x}_i$'s equal d_k $\left(\text{i.e., } \frac{1}{n}\sum_{i\in I_k^{(n)}} \widetilde{W}_i\right)$ the proof will follow from a similar treatment as for the standard Cramer LDP in $\mathbb{R}^K$. The only difference lies in two facts: the number of the summands for the coordinate k is n_k , the number of $\widetilde{x}_i$'s which equal d_k, instead of n in the standard case. Furthermore we will need to substitute n_k by its equivalent $n \cdot \widetilde{p}_k$, which adds an approximation step. For the upper bound, the proof is based on the corresponding result for $B = B_1 \times \cdots \times B_K$ where the B_k's are open bounded intervals on $\mathbb{R}^+$. Since the sequence $(\widetilde{x}_1,\ldots)$ satisfies

$$\lim_{n\to\infty} \frac{n_k}{n} = \widetilde{p}_k, \qquad \text{(cf. (22))}$$

there holds

$$\begin{aligned} \frac{1}{n}\log \Pi\left[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in B\right] &= \frac{1}{n}\log \Pi\left[\bigcap_{k=1}^{K}\left(\frac{1}{n}\sum_{i\in I_k^{(n)}} \widetilde{W}_i \in B_k\right)\right] \\ &= \frac{1}{n}\sum_{k=1}^{K} \log \Pi\left[\frac{1+o(1)}{n_k}\sum_{i\in I_k^{(n)}} \widetilde{W}_i \in \frac{1}{\widetilde{p}_k} B_k\right], \end{aligned} \tag{250}$$

and hence

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n}\log \Pi\left[\boldsymbol{\xi}_n^{\widetilde{\mathbf{W}}} \in B\right] &\leq \sum_{k=1}^{K} \widetilde{p}_k \cdot \limsup_{n_k\to\infty} \frac{1}{n_k}\log \Pi\left[\frac{1}{n_k}\sum_{i\in I_k^{(n)}} \widetilde{W}_i \in \frac{1}{p_k} B_k\right] \\ &\leq -\sum_{k=1}^{K} \inf_{x_k\in cl(B_k)} \widetilde{p}_k \cdot \varphi\left(\frac{x_k}{\widetilde{p}_k}\right). \end{aligned} \tag{251}$$

To deduce (251) from (250), we have used (i) the fact that for all k the random variables $\frac{1}{n_k}\left(1+o(1)\right)) \cdot \sum_{i\in I_k^{(n)}} \widetilde{W}_i$ and $\frac{1}{n_k}\sum_{i\in I_k^{(n)}} \widetilde{W}_i$ are exponentially equivalent in the sense that their difference Δ_{n_k} satisfies

$$\limsup_{n_k\to\infty} \frac{1}{n_k}\log \Pi[\,|\Delta_{n_k}| > \eta\,] = -\infty,$$

making use of the Chernoff inequality for all positive η, as well as (ii) Theorem 4.2.13 in [108]. Now the summation and the inf-operations can be permuted in (251) which proves the claim for the rectangle B.

As in [108], for a compact set Ω we consider its finite covering by such open sets B and conclude; for Ω being a closed set, a tightness argument holds, following [108] Theorem 2.2.30 verbatim. For the lower bound consider the same rectangle B. The argument which locates the tilted distribution at the center of B, together with the use of the LLN for the corresponding r.v's as in [108], in combination with the same approximations as above to handle the approximation of n_k by $n \cdot \widetilde{p}_k$, complete the proof of Proposition 61. We omit the details. ■

Let us continue with the proof of Theorem 10, by giving the following two helpful lemmas for

$$\Phi_{\mathbb{P}}(\mathbf{A}) := \inf_{\mathbf{Q} \in \mathbf{A}} D_{\varphi}(\mathbf{Q}, \mathbb{P}), \qquad \mathbf{A} \subset \mathcal{M} := \mathbb{R}^K, \tag{252}$$

Lemma 62: For any open set $\mathbf{A} \subset \mathcal{M} := \mathbb{R}^K$ one has $\Phi_{\mathbb{P}}(\mathbf{A}) = \Phi_{\mathbb{P}}(cl(\mathbf{A}))$.

This is clear from the continuity of $\Phi_{\mathbb{P}}$.

Lemma 63: For any $\mathbf{A} \subset \mathcal{M} := \mathbb{R}^K$ satisfying (7) one has $\Phi_{\mathbb{P}}(cl(\mathbf{A})) = \Phi_{\mathbb{P}}(\mathbf{A}) = \Phi_{\mathbb{P}}(int(\mathbf{A}))$.

Proof of Lemma 63. Assume first that $\Phi_{\mathbb{P}}(\mathbf{A})$ is finite. Then suppose that $\mathbf{A}$ satisfies (7) and $\Phi_{\mathbb{P}}(cl(\mathbf{A})) < \Phi_{\mathbb{P}}(int(\mathbf{A}))$. The latter implies the existence of a point $a \in cl(\mathbf{A})$ such that $a \notin int(\mathbf{A})$ and $\Phi_{\mathbb{P}}(a) = \Phi_{\mathbb{P}}(cl(\mathbf{A}))$. But then, by Lemma 62 and (7) one gets $\Phi_{\mathbb{P}}(int(\mathbf{A})) = \Phi_{\mathbb{P}}(cl(int(\mathbf{A}))) = \Phi_{\mathbb{P}}(cl(\mathbf{A})) = \Phi_{\mathbb{P}}(a)$ which leads to a contradiction. When $\Phi_{\mathbb{P}}(\mathbf{A}) = \infty$ then $\Phi_{\mathbb{P}}(cl(\mathbf{A})) = \Phi_{\mathbb{P}}(int(\mathbf{A})) = \Phi_{\mathbb{P}}(\mathbf{A}) = \infty$. ■

Putting things together, the required asymptotic assertion (24) follows from (249), (7) and Lemma 63. This completes the proof of Theorem 10. ■

Appendix B
Proofs — Part 2

Before we tackle the proof of Theorem 14, let us introduce the following

Lemma 64: If $\mathbf{\Omega} \subset \mathbb{S}^K$ satisfies condition (7), then $\widetilde{\widetilde{\mathbf{\Omega}}} := \bigcup_{m \neq 0} cl(m \cdot \mathbf{\Omega})$ has the property (7).

This can be deduced in a straightforward way: the assumption implies that $cl(\mathbf{\Omega})$ satisfies (7), and thus also $m \cdot cl(\mathbf{\Omega})$ satisfies (7). But this implies the validity of (7) for the "cone" $\bigcup_{m \neq 0} m \cdot cl(\mathbf{\Omega})$ which is nothing but $\bigcup_{m \neq 0} cl(m \cdot \mathbf{\Omega})$.

Proof of Theorem 14.
Recall the interpretations of the two vectors $\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}}$ respectively $\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}}$ given in (31) respectively (33), and that the sum of their k components are $\sum_{k=1}^{K} \frac{1}{n} \sum_{i \in I_k^{(n)}} W_i = \frac{1}{n} \sum_{i=1}^{n} W_i$ respectively $\sum_{k=1}^{K} \frac{\sum_{i \in I_k^{(n)}} W_i}{\sum_{k=1}^{K} \sum_{i \in I_k^{(n)}} W_i} = 1$ (in case of $\sum_{i=1}^{n} W_i \neq 0$). In the light of these, for $\mathbf{\Omega} \subset \mathbb{S}^K$ one gets the set identification

$$\left\{\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \mathbf{\Omega}\right\} = \bigcup_{m \neq 0} \left\{\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}} \in m \cdot \mathbf{\Omega}, \frac{1}{n} \sum_{i=1}^{n} W_i = m\right\}$$

since $\{\sum_{i=1}^{n} W_i = 0\}$ amounts to $m = 0$, which cannot hold when $\left\{\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \mathbf{\Omega}\right\}$. Now

$$\begin{aligned} \Pi_{X_1^n}\left[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \mathbf{\Omega}\right] &= \Pi_{X_1^n}\left[\bigcup_{m \neq 0} \left\{\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}} \in m \cdot \mathbf{\Omega}, \frac{1}{n} \sum_{i=1}^{n} W_i = m\right\}\right] \\ &= \Pi_{X_1^n}\left[\bigcup_{m \neq 0} \left\{\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}} \in m \cdot \mathbf{\Omega}\right\}\right] = \Pi_{X_1^n}\left[\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}} \in \bigcup_{m \neq 0} m \cdot \mathbf{\Omega}\right] \end{aligned}$$

since $\left\{\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}} \in m \cdot \mathbf{\Omega}\right\} \subset \left\{\frac{1}{n} \sum_{i=1}^{n} W_i = m\right\}$. Therefore

$$\frac{1}{n} \log \Pi_{X_1^n}\left[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \mathbf{\Omega}\right] = \frac{1}{n} \log \Pi_{X_1^n}\left[\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}} \in \bigcup_{m \neq 0} m \cdot \mathbf{\Omega}\right]. \tag{253}$$

Because of Proposition 61 — applied to $\widetilde{\boldsymbol{\Omega}} := \bigcup_{m\neq 0} m \cdot \boldsymbol{\Omega}$ — one gets in terms of (252)

$$\begin{aligned} -\Phi_{\mathbb{P}}\Big(int\Big(\bigcup_{m\neq 0} m\cdot\boldsymbol{\Omega}\Big)\Big) &\leq \liminf_{n\to\infty}\frac{1}{n}\log \Pi_{X_1^n}\Big[\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}} \in \bigcup_{m\neq 0} m\cdot\boldsymbol{\Omega}\Big] \\ &\leq \limsup_{n\to\infty}\frac{1}{n}\log \Pi_{X_1^n}\Big[\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}} \in \bigcup_{m\neq 0} m\cdot\boldsymbol{\Omega}\Big] \leq -\Phi_{\mathbb{P}}\Big(cl\Big(\bigcup_{m\neq 0} m\cdot\boldsymbol{\Omega}\Big)\Big). \end{aligned} \tag{254}$$

But $$\Phi_{\mathbb{P}}\Big(int\Big(\bigcup_{m\neq 0} m\cdot\boldsymbol{\Omega}\Big)\Big) \leq \Phi_{\mathbb{P}}\Big(\bigcup_{m\neq 0} int\Big(m\cdot\boldsymbol{\Omega}\Big)\Big) = \inf_{m\neq 0}\Phi_{\mathbb{P}}(int(m\cdot\boldsymbol{\Omega})) \tag{255}$$

and $$\Phi_{\mathbb{P}}\Big(cl\Big(\bigcup_{m\neq 0} m\cdot\boldsymbol{\Omega}\Big)\Big) \geq \Phi_{\mathbb{P}}\Big(\bigcup_{m\neq 0} cl\Big(m\cdot\boldsymbol{\Omega}\Big)\Big) = \inf_{m\neq 0}\Phi_{\mathbb{P}}(cl(m\cdot\boldsymbol{\Omega})). \tag{256}$$

In fact, the inequality in (255) is straightforward because of $\bigcup_{m\neq 0} int(m\cdot\boldsymbol{\Omega}) \subset int(\bigcup_{m\neq 0} m\cdot\boldsymbol{\Omega})$ (since the latter is the largest open set contained in $\bigcup_{m\neq 0} m\cdot\boldsymbol{\Omega}$); the inequality in (256) follows from

$$\Phi_{\mathbb{P}}\Big(cl\Big(\bigcup_{m\neq 0} m\cdot\boldsymbol{\Omega}\Big)\Big) \geq \Phi_{\mathbb{P}}\Big(cl\Big(\bigcup_{m\neq 0} cl\Big(m\cdot\boldsymbol{\Omega}\Big)\Big)\Big) = \Phi_{\mathbb{P}}\Big(\bigcup_{m\neq 0} cl\Big(m\cdot\boldsymbol{\Omega}\Big)\Big)$$

An application of Lemma 63 yields $\Phi_{\mathbb{P}}(int(m\cdot\boldsymbol{\Omega})) = \Phi_{\mathbb{P}}(m\cdot\boldsymbol{\Omega}) = \Phi_{\mathbb{P}}(cl(m\cdot\boldsymbol{\Omega}))$ for all $m\neq 0$, and hence

$$\inf_{m\neq 0}\Phi_{\mathbb{P}}(int(m\cdot\boldsymbol{\Omega})) = \inf_{m\neq 0}\Phi_{\mathbb{P}}(m\cdot\boldsymbol{\Omega}) = \inf_{m\neq 0}\Phi_{\mathbb{P}}(cl(m\cdot\boldsymbol{\Omega})). \tag{257}$$

By combining (253), (254), (255), (256) and (257), one arrives at

$$\begin{aligned} &\lim_{n\to\infty}\frac{1}{n}\log \Pi_{X_1^n}\Big[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{W}} \in \boldsymbol{\Omega}\Big] = \lim_{n\to\infty}\frac{1}{n}\log \Pi_{X_1^n}\Big[\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{W}} \in \bigcup_{m\neq 0} m\cdot\boldsymbol{\Omega}\Big] \\ &= -\inf_{m\neq 0}\Phi_{\mathbb{P}}(m\cdot\boldsymbol{\Omega}) = -\inf_{m\neq 0}\inf_{\mathbf{Q}\in m\cdot\boldsymbol{\Omega}} D_{\varphi}(\mathbf{Q},\mathbb{P}) = -\inf_{m\neq 0}\inf_{\mathbb{Q}\in\boldsymbol{\Omega}} D_{\varphi}(m\cdot\mathbb{Q},\mathbb{P}), \end{aligned}$$

where in the second last equality we have "reverted" the notation (252). Note that we did not assume (7) for $\bigcup_{m\neq 0} m\cdot\boldsymbol{\Omega}$. ■

Appendix C
Proofs — Part 3

Proof of Lemma 16.
From (44) one gets straightforwardly for arbitrary $\widetilde{c} > 0$

$$D_{\widetilde{c}\cdot\varphi_\gamma}(m\cdot\mathbf{Q},\mathbb{P}) := \begin{cases} \frac{\widetilde{c}\cdot(m^\gamma\cdot H_\gamma - m\cdot A\cdot\gamma+\gamma-1)}{\gamma\cdot(\gamma-1)}, & \text{if } \gamma\in]-\infty,0[,\ \mathbb{P}\in\mathbb{S}_{\geq 0}^K,\ \mathbf{Q}\in A\cdot\mathbb{S}_{>0}^K \text{ and } m>0, \\ \widetilde{c}\cdot(-\log m + \widetilde{I} - 1 + m\cdot A), & \text{if } \gamma=0,\ \mathbb{P}\in\mathbb{S}_{\geq 0}^K,\ A\cdot\mathbf{Q}\in\mathbb{S}_{>0}^K \text{ and } m>0, \\ \frac{\widetilde{c}\cdot(m^\gamma\cdot H_\gamma - m\cdot A\cdot\gamma+\gamma-1)}{\gamma\cdot(\gamma-1)}, & \text{if } \gamma\in]0,1[,\ \mathbb{P}\in\mathbb{S}_{\geq 0}^K,\ \mathbf{Q}\in A\cdot\mathbb{S}_{\geq 0}^K \text{ and } m\geq 0, \\ \widetilde{c}\cdot(A\cdot m\cdot\log m + m\cdot(I-A)+1), & \text{if } \gamma=1,\ \mathbb{P}\in\mathbb{S}_{>0}^K,\ \mathbf{Q}\in A\cdot\mathbb{S}_{\geq 0}^K \text{ and } m\geq 0, \\ \frac{\widetilde{c}\cdot\left(m^\gamma\cdot H_\gamma\cdot 1_{[0,\infty[}(m) - m\cdot A\cdot\gamma+\gamma-1\right)}{\gamma\cdot(\gamma-1)}, & \text{if } \gamma\in]1,2[,\ \mathbb{P}\in\mathbb{S}_{>0}^K,\ \mathbf{Q}\in A\cdot\mathbb{S}^K \text{ and } m\in]-\infty,\infty[, \\ \frac{\widetilde{c}\cdot\left(m^2\cdot H_2 - m\cdot A\cdot 2+2-1\right)}{2\cdot(2-1)}, & \text{if } \gamma=2,\ \mathbb{P}\in\mathbb{S}_{>0}^K,\ \mathbf{Q}\in A\cdot\mathbb{S}^K \text{ and } m\in]-\infty,\infty[, \\ \frac{\widetilde{c}\cdot\left(m^\gamma\cdot H_\gamma\cdot 1_{[0,\infty[}(m) - m\cdot A\cdot\gamma+\gamma-1\right)}{\gamma\cdot(\gamma-1)}, & \text{if } \gamma\in]2,\infty[,\ \mathbb{P}\in\mathbb{S}_{>0}^K,\ \mathbf{Q}\in A\cdot\mathbb{S}^K \text{ and } m\in]-\infty,\infty[, \\ \infty, & \text{else}, \end{cases} \tag{258}$$

where we have used the three $m-$independent abbreviations

$$H_\gamma := \sum_{k=1}^K (q_k)^\gamma\cdot(p_k)^{1-\gamma} = 1+\gamma\cdot(A-1) + \frac{\gamma\cdot(\gamma-1)}{\widetilde{c}}\cdot D_{\widetilde{c}\cdot\varphi_\gamma}(\mathbf{Q},\mathbb{P}), \qquad \text{(cf. (45))}$$

$$I := \sum_{k=1}^K q_k\cdot\log\left(\frac{q_k}{p_k}\right) = \frac{1}{\widetilde{c}}\cdot D_{\widetilde{c}\cdot\varphi_1}(\mathbf{Q},\mathbb{P}) + A - 1, \qquad \text{(cf. (46))}$$

$$\widetilde{I} := \sum_{k=1}^{K} p_k \cdot \log\left(\frac{p_k}{q_k}\right) \;=\; \frac{1}{\widetilde{c}} \cdot D_{\widetilde{c}\cdot\varphi_0}(\mathbf{Q},\mathbb{P}) + 1 - A. \qquad \text{(cf. (47))}$$

To proceed, let us fix an arbitrary constant $\widetilde{c} > 0$.
(i) Case $\gamma \cdot (1-\gamma) \neq 0$.
(ia) Let us start with the subcase $\gamma \in]-\infty, 0[$. From the first and the last line of (258), it is clear that the corresponding $m-$infimum can not be achieved for $m \leq 0$; since $H_\gamma > 0$ one gets the unique minimizer $m_{min} = \left(\frac{H_\gamma}{A}\right)^{1/(1-\gamma)} > 0$ and the minimum $D_{\widetilde{c}\cdot\varphi_\gamma}(m_{min}\cdot\mathbf{Q},\mathbb{P}) = \frac{\widetilde{c}}{\gamma}\cdot(1 - \frac{H^{1/(1-\gamma)}}{A^{\gamma/(1-\gamma)}})$. Hence, (48) is established. The assertions (49) and (50) follow immediately by monotonicity inspection of $x \to \frac{\widetilde{c}}{\gamma}\cdot\left[1 - \frac{1}{A^{\gamma/(1-\gamma)}}\cdot\left[1+\gamma\cdot(A-1)+\frac{\gamma\cdot(\gamma-1)}{\widetilde{c}}\cdot x\right]^{-1/(\gamma-1)}\right]$ for $x \geq 0$ such that $1+\gamma\cdot(A-1)+\frac{\gamma\cdot(\gamma-1)}{\widetilde{c}}\cdot x \geq 0$.
(ib) The subcase $\gamma \in]0,1[$ (cf. the third line of (258)) works analogously if $H_\gamma > 0$; furthermore, if $H_\gamma = 0$ — which can only appear when $\mathbb{P}$, $\mathbf{Q}$ have disjoint supports (singularity)— then $\inf_{m>0} D_{\widetilde{c}\cdot\varphi_\gamma}(m\cdot\mathbf{Q},\mathbb{P}) = \frac{\widetilde{c}}{\gamma}$ which is (the corresponding special case of) (48).
(ic) In the subcase $\gamma \in]1,\infty[$ (cf. the fifth, sixth and seventh line of (258)) it is straightforward to see that the desired infimum can not be achieved for $m < 0$. Hence, one can proceed analogously to subcase (ia).
(id) The assertions (51) to (54) are straightforward.
(ii) Case $\gamma = 1$. From the fourth line of (258), one obtains the unique minimizer $m_{min} = \exp\{-I/A\}$ and the minimum $D_{\widetilde{c}\cdot\varphi_1}(m_{min}\cdot\mathbf{Q},\mathbb{P}) = \widetilde{c}\cdot(1 - A\cdot m_{min})$, which leads to (55). The monotonicity of $x \to \widetilde{c}\cdot(1-\exp\{-x/\widetilde{c}\})$ for $x \geq 0$ implies immediately (56) and (57); moreover, (58) and (59) are immediate.
(iii) Case $\gamma = 0$. The second line of (258) implies the unique minimizer $m_{min} = 1/A$, the minimum $D_{\widetilde{c}\cdot\varphi_0}(m_{min}\cdot\mathbf{Q},\mathbb{P}) = \widetilde{c}\cdot(\widetilde{I}+\log A)$, and hence (60). The assertions (61) to (64) are obvious. ■

Appendix D
Proofs — Part 4

Proof of Proposition 29. Clearly, (G1) and (G2) are part of the definition of $\widetilde{\Upsilon}(]a,b[)$. Recall our required representability (6). The therein involved Laplace-Stieltjes transform (Laplace-Lebesgue transform)

$$z \mapsto MGF_{\zeta}(z) := \int_{\mathbb{R}} e^{z\cdot y}\, d\zeta(y) = E_{\Pi}[e^{z\cdot W}] \tag{259}$$

of a probability measure ζ on the real line respectively of an associated random variable W (with $\zeta[\cdot\,] := \Pi[W \in \cdot\,]$) has the following fundamental properties, according to well-known general theory:

(M1) MGF_{ζ} takes values in $]0,\infty]$;
(M2) the effective domain $dom(MGF_{\zeta})$ is an interval which contains 0 and which may be degenerated or even the whole real line; correspondingly, we denote its interior by $]\lambda_-,\lambda_+[:= int(dom(MGF_{\zeta}))$ which may be the empty set (in case that $dom(MGF_{\zeta}) = \{0\}$, i.e. $\lambda_- = \lambda_+ = 0$); clearly, there holds $\lambda_- \in [-\infty,0]$ and $\lambda_+ \in [0,\infty]$;
(M3) MGF_{ζ} is continuous on $dom(MGF_{\zeta})$ and lower semicontinuous on $\mathbb{R}$;
(M4) if $\lambda_- \neq \lambda_+$, then MGF_{ζ} is real analytic and thus infinitely differentiable on $]\lambda_-,\lambda_+[$;
(M5) if MGF_{ζ} is finite in a neighborhood of zero, i.e. $0 \in]\lambda_-,\lambda_+[$, then for all $k \in \mathbb{N}_0$ the $k-$th moment of ζ respectively W exists and is finite and can be computed in terms of the $k-$th derivative $MGF_{\zeta}^{(k)}$ as

$$MGF_{\zeta}^{(k)}(0) = \int_{\mathbb{R}} y^k\, d\zeta(y) = E_{\Pi}[W^k], \tag{260}$$

which, by the way, then allows the interpretation of MGF_{ζ} as "moment generating function of ζ resp. W"[35];
(M6) if $\lambda_- \neq \lambda_+$, then MGF_{ζ} is strictly convex on $]\lambda_-,\lambda_+[$.

Hence, the logarithm of the Laplace-Stieltjes transform

$$z \mapsto \Lambda_{\zeta}(z) := \log MGF_{\zeta}(z) := \log \int_{\mathbb{R}} e^{z\cdot y}\, d\zeta(y) = \log E_{\Pi}[e^{z\cdot W}] \tag{261}$$

[35]since we assume $0 \in]\lambda_-,\lambda_+[$, we have already used the meaningful abbreviation MGF (rather than LST) in (259)

(which in case of $0 \in]\lambda_-, \lambda_+[$ can be interpreted as cumulant generating function) "carries over" (M1) to (M6), which partially can be even refined:

(C1) $\Lambda_{\mathbb{\zeta}}$ takes values in $]-\infty, \infty]$;
(C2) $dom(\Lambda_{\mathbb{\zeta}}) = dom(MGF_{\mathbb{\zeta}})$ and thus $int(dom(\Lambda_{\mathbb{\zeta}})) =]\lambda_-, \lambda_+[$;
(C3) $\Lambda_{\mathbb{\zeta}}$ is continuous on $dom(\Lambda_{\mathbb{\zeta}})$ and lower semicontinuous on $\mathbb{R}$;
(C4) if $\lambda_- \neq \lambda_+$, then $\lambda_{\mathbb{\zeta}}$ is infinitely differentiable on $]\lambda_-, \lambda_+[$;
(C5) if $0 \in]\lambda_-, \lambda_+[$, then

$$\Lambda_{\mathbb{\zeta}}(0) = 0, \quad \Lambda'_{\mathbb{\zeta}}(0) = \int_{\mathbb{R}} y \, d\mathbb{\zeta}(y) = E_{\mathbb{\Pi}}[W], \tag{262}$$

$$\Lambda''_{\mathbb{\zeta}}(0) = \int_{\mathbb{R}} \left(y - \int_{\mathbb{R}} \widetilde{y} \, d\mathbb{\zeta}(\widetilde{y}) \right)^2 d\mathbb{\zeta}(y) = E_{\mathbb{\Pi}}[W^2] - (E_{\mathbb{\Pi}}[W])^2 = Var_{\mathbb{\Pi}}[W]; \tag{263}$$

(C6) under the assumption $\lambda_- \neq \lambda_+$ there holds: $\Lambda_{\mathbb{\zeta}}$ is strictly convex on $]\lambda_-, \lambda_+[$ if and only if $\mathbb{\zeta}$ is not a one-point distribution (Dirac mass) if and only if W is not a.s. constant; otherwise, $\Lambda_{\mathbb{\zeta}}$ is linear;
(C7) under the assumption that $\mathbb{\zeta}$ is not a one-point distribution (Dirac mass) — with the notations $a := \inf supp(\mathbb{\zeta}) = \inf supp(W)$, $b := \sup supp(\mathbb{\zeta}) = \sup supp(W)$, $t_-^{sc} := \inf\{\Lambda'_{\mathbb{\zeta}}(z) \, : \, z \in]\lambda_-, \lambda_+[\} = \lim_{z \downarrow \lambda_-} \Lambda'_{\mathbb{\zeta}}(z)$ and $t_+^{sc} := \sup\{\Lambda'_{\mathbb{\zeta}}(z) \, : \, z \in]\lambda_-, \lambda_+[\} = \lim_{z \uparrow \lambda_+} \Lambda'_{\mathbb{\zeta}}(z)$ — one gets the following assertions:
  (C7i) $]t_-^{sc}, t_+^{sc}[\subseteq]a, b[$;
  (C7ii) if $a > -\infty$, then
  - $\lambda_- = -\infty$,
  - $t_-^{sc} = \lim_{z \to -\infty} \Lambda'_{\mathbb{\zeta}}(z) = \lim_{z \to -\infty} \frac{\Lambda_{\mathbb{\zeta}}(z)}{z} = a$;

  (C7iii) if $b < \infty$, then
  - $\lambda_+ = \infty$,
  - $t_+^{sc} = \lim_{z \to \infty} \Lambda'_{\mathbb{\zeta}}(z) = \lim_{z \to \infty} \frac{\Lambda_{\mathbb{\zeta}}(z)}{z} = b$;

  (C7iv) if $a = -\infty$ and $\lambda_- = -\infty$, then $t_-^{sc} = \lim_{z \to -\infty} \Lambda'_{\mathbb{\zeta}}(z) = -\infty = a$;
  (C7v) if $b = \infty$ and $\lambda_+ = \infty$, then $t_+^{sc} = \lim_{z \to \infty} \Lambda'_{\mathbb{\zeta}}(z) = \infty = b$;
  (C7vi) if $\lambda_- \in]-\infty, 0[$ and $t_-^{sc} > -\infty$, then
  - $a = -\infty$,
  - $\Lambda_{\mathbb{\zeta}}(\lambda_-) \in]-\infty, \infty[$,
  - $\Lambda_{\mathbb{\zeta}}(z) = \infty$ for all $z < \lambda_-$,
  - $\Lambda'_{\mathbb{\zeta}}(\lambda_-) \in]-\infty, \infty[$;

  (C7vii) if $\lambda_+ \in]0, \infty[$ and $t_+^{sc} < \infty$, then
  - $b = \infty$,
  - $\Lambda_{\mathbb{\zeta}}(\lambda_+) \in]-\infty, \infty[$,
  - $\Lambda_{\mathbb{\zeta}}(z) = \infty$ for all $z > \lambda_+$,
  - $\Lambda'_{\mathbb{\zeta}}(\lambda_+) \in]-\infty, \infty[$;

  (C7viii) if $\lambda_- \in]-\infty, 0[$ and $t_-^{sc} = -\infty$, then $a = -\infty$;
  (C7ix) if $\lambda_+ \in]0, \infty[$ and $t_+^{sc} = \infty$, then $b = \infty$.

Notice that (C7ii) to (C7ix) cover all possible constellations. For a proof of (C7ii) to (C7vii) as well as further details, see e.g. Section 9.1 in Borovkov [56]. By contradiction, (C7viii) follows from (C7ii) and (C7ix) follows from (C7iii). Moreover, (C7i) is a consequence (C7ii) to (C7ix). As a side remark, notice that (C6) refines (M6).

According to the representability requirement (6), one has

$$\varphi(t) = \sup_{z \in \mathbb{R}} (z \cdot t - \Lambda_{\mathbb{\zeta}}(z)) =: \Lambda^*_{\mathbb{\zeta}}(t), \qquad t \in \mathbb{R}, \tag{264}$$

(i.e. the divergence generator φ must be equal to the Fenchel-Legendre transform $\Lambda^*_{\mathbb{\zeta}}$ of a cumulant generating function $\Lambda_{\mathbb{\zeta}}$) of some probability distribution $\mathbb{\zeta}$, such that $\lambda_- < 0 < \lambda_+$ holds. Moreover, φ should satisfy $\varphi(1) = 0$, and should be finite as well as strictly convex in a non-empty neighborhood $]t_-^{sc}, t_+^{sc}[$ of 1 (cf. the definition of $\widetilde{\Upsilon}(]a, b[)$). The latter *rules out* that $\mathbb{\zeta}$ is any one-point distribution (Dirac distribution), say $\mathbb{\zeta} = \delta_{y_0}$ for some $y_0 \in \mathbb{R}$, since in such a situation one gets $\Lambda_{\mathbb{\zeta}}(z) = z \cdot y_0$, and thus $\varphi(t) = \Lambda^*_{\mathbb{\zeta}}(t) = 0$ for $t = y_0$ and $\varphi(t) = \Lambda^*_{\mathbb{\zeta}}(t) = \infty$ for all $t \in \mathbb{R} \backslash \{y_0\}$ (even in the case $y_0 = 1$ for which $\varphi(1) = 0$ is satisfied). Consequently, $\Lambda_{\mathbb{\zeta}}$ is strictly convex on $]\lambda_-, \lambda_+[= int(dom(\Lambda_{\mathbb{\zeta}}))$ (cf. (C6)) and (C7) applies. Clearly, by continuity one gets

$$\Lambda^*_{\mathbb{\zeta}}(t) = \sup_{z \in]\lambda_-, \lambda_+[} (t \cdot z - \Lambda_{\mathbb{\zeta}}(z)), \qquad t \in \mathbb{R}. \tag{265}$$

For $t \in]t_-^{sc}, t_+^{sc}[$, the optimization problem (265) can be solved explicitly by the well-known "pure/original" Legendre transform, namely

$$\Lambda_{\mathfrak{C}}^*(t) = t \cdot \Lambda_{\mathfrak{C}}'^{-1}(t) - \Lambda_{\mathfrak{C}}\Big(\Lambda_{\mathfrak{C}}'^{-1}(t)\Big), \qquad t \in]t_-^{sc}, t_+^{sc}[. \tag{266}$$

Let us inspect the further cases $t \leq t_-^{sc}$. In the contexts of (C7iv) and (C7viii), this is obsolete since $t_-^{sc} = a = -\infty$. For (C7ii), where $t_-^{sc} = a > -\infty$, one can show $\Lambda_{\mathfrak{C}}^*(a) = -\log \mathfrak{C}[\{a\}] = -\log \Pi[W = a]$ which together with (264) proves (G10ii); moreover, $\Lambda_{\mathfrak{C}}^*(t) = \infty$ for all $t < a$ (see e.g. Section 9.1 of Borovkov [56]). In the setup (C7vi), where $t_-^{sc} > a = -\infty$ it is clear that $\Lambda_{\mathfrak{C}}^*(t_-^{sc}) = t_-^{sc} \cdot \Lambda_{\mathfrak{C}}'^{-1}(t_-^{sc}) - \Lambda_{\mathfrak{C}}\Big(\Lambda_{\mathfrak{C}}'^{-1}(t_-^{sc})\Big) = t_-^{sc} \cdot \lambda_- - \Lambda_{\mathfrak{C}}(\lambda_-)$ and

$$\Lambda_{\mathfrak{C}}^*(t) = t \cdot \lambda_- - \Lambda_{\mathfrak{C}}(\lambda_-) = \Lambda_{\mathfrak{C}}^*(t_-^{sc}) + \lambda_- \cdot (t - t_-^{sc}) \quad \text{for all } t \in]-\infty, t_-^{sc}[. \tag{267}$$

As far as the cases $t \geq t_+^{sc}$ is concerned, in the situations of (C7v) and (C7ix), this is obsolete since $t_+^{sc} = b = \infty$. For (C7iii), where $t_+^{sc} = b < \infty$, one can show $\Lambda_{\mathfrak{C}}^*(b) = -\log \mathfrak{C}[\{b\}] = -\log \Pi[W = b]$ which together with (264) proves (G10iii); moreover, $\Lambda_{\mathfrak{C}}^*(t) = \infty$ for all $t > b$ (see e.g. Section 9.1 of Borovkov [56]). In the setup (C7vii), where $t_+^{sc} < b = \infty$ it is clear that $\Lambda_{\mathfrak{C}}^*(t_+^{sc}) = t_+^{sc} \cdot \Lambda_{\mathfrak{C}}'^{-1}(t_+^{sc}) - \Lambda_{\mathfrak{C}}\Big(\Lambda_{\mathfrak{C}}'^{-1}(t_+^{sc})\Big) = t_+^{sc} \cdot \lambda_+ - \Lambda_{\mathfrak{C}}(\lambda_+)$ and

$$\Lambda_{\mathfrak{C}}^*(t) = t \cdot \lambda_+ - \Lambda_{\mathfrak{C}}(\lambda_+) = \Lambda_{\mathfrak{C}}^*(t_+^{sc}) + \lambda_+ \cdot (t - t_+^{sc}) \quad \text{for all } t \in]t_+^{sc}, \infty[. \tag{268}$$

As a side effect, we have thus also proved (G10i) and (G3) (notice that in (G3) we have started with a, b to be the endpoints of the support of $\mathfrak{C}$ respectively W, in contrast to Definition 3 where a, b are defined as the endpoints of the effective domain of φ).

To proceed, from (264) and (266) we obtain

$$\varphi'(t) = (\Lambda_{\mathfrak{C}}^*)'(t) = \Lambda_{\mathfrak{C}}'^{-1}(t), \qquad \varphi''(t) = (\Lambda_{\mathfrak{C}}^*)''(t) = \frac{1}{\Lambda_{\mathfrak{C}}''\big(\Lambda_{\mathfrak{C}}'^{-1}(t)\big)} > 0, \qquad t \in]t_-^{sc}, t_+^{sc}[, \tag{269}$$

which — together with the investigations below (266) — provides (G4) and (G5); moreover, (G6) is immediate since the infinite differentiability is straightforward and $\varphi'(1) = 0$ because we have required both the nonnegativity of φ and (G2) (cf. the definition of $\widetilde{\Upsilon}(]a, b[)$). The property (G7) follows from (C7ii), (C7iv), (C7viii), (264), (267) and $\varphi'(t_-^{sc}) = \Lambda_{\mathfrak{C}}'^{-1}(t_-^{sc}) = \lambda_-$. Analogously, we get (G8) from (C7iii), (C7v), (C7ix), (264), (268) and $\varphi'(t_+^{sc}) = \Lambda_{\mathfrak{C}}'^{-1}(t_+^{sc}) = \lambda_+$.

Let us continue with (G9). By applying the general theory of double Fenchel-Legendre transforms (bi-conjugates), (156) turns into

$$\varphi^*(z) = \Lambda_{\mathfrak{C}}(z), \qquad z \in \mathbb{R}, \tag{270}$$

which deduces (G9i). The properties (G9ii), (G9iii) and (G9iv) follow from Theorem 30 (cf. the discussion thereafter). Finally, we obtain (G11i) and (G11ii) from (269), (262) and (263). ■

Proof of Proposition 31. The assertions follow immediately from (157), (158), (159), Theorem 30, (269) (and the discussion thereafter) as well as (M5). ■

Appendix E
Proofs — Part 5

Proof of Proposition 34. The assertion follows straightforwardly from the following two facts:
(i) a moment generating function MGF is infinitely divisible if and only if MGF^c is a moment generating function for all $c > 0$ (cf. e.g. (the MGF-version of) Prop. IV.2.5 of Steutel & van Harn [341]).
(ii) $z \mapsto MGF(z)$ is a moment generating function if and only if $z \mapsto MGF(\breve{c} \cdot z) =: MGF_{\breve{c}}(z)$ is a moment generating function for all $\breve{c} > 0$.
Notice that for each $c > 0$, $\breve{c} > 0$ one has $int(dom(MGF)) = int(dom(MGF^c))$ and $int(dom(MGF_{\breve{c}})) = \frac{1}{\breve{c}} \cdot int(dom(MGF))$, and hence the light-tailedness remains unchanged: $0 \in int(dom(MGF))$ if and only if $0 \in int(dom(MGF^c))$ if and only if $0 \in int(dom(MGF_{\breve{c}}))$. Since $\varphi \in \Upsilon(]a, b[)$, we have

$$\varphi(t) = \sup_{z \in]\lambda_-, \lambda_+[} \left(z \cdot t - \log\Big(\int_{\mathbb{R}} e^{z \cdot y} \, d\mathfrak{C}(y) \Big) \right), \quad t \in]a, b[\,, \tag{271}$$

and thus for the exponential of its Fenchel-Legendre transform

$$\int_{\mathbb{R}} e^{z\cdot y} d\mathfrak{C}(y), \qquad z \in]\lambda_-, \lambda_+[. \tag{272}$$

Now, let $\widetilde{\varphi} := \widetilde{c} \cdot \varphi \in \Upsilon(]a,b[)$ for arbitrarily fixed $\widetilde{c} > 0$. From the application of (6) to $\widetilde{\varphi}$ we obtain

$$\widetilde{\varphi}(t) = \sup_{\widetilde{z} \in]\widetilde{\lambda}_-, \widetilde{\lambda}_+[} \left(\widetilde{z} \cdot t - \log \int_{\mathbb{R}} e^{\widetilde{z}\cdot\widetilde{y}} d\widetilde{\mathfrak{C}}_{\widetilde{c}}(\widetilde{y}) \right), \qquad t \in]a,b[\,, \tag{273}$$

for some unique probability distribution $\widetilde{\mathfrak{C}}_{\widetilde{c}}$ on $\mathbb{R}$. Here, according to (G9i) for $\widetilde{\varphi}$ we have used $\widetilde{\lambda}_- := \inf_{t \in]a,b[} \widetilde{\varphi}'(t) = \widetilde{c} \cdot \lambda_-$ and $\widetilde{\lambda}_+ := \sup_{t \in]a,b[} \widetilde{\varphi}'(t) = \widetilde{c} \cdot \lambda_+$. Dividing (273) by $\widetilde{c}$, we arrive at

$$\begin{aligned} \varphi(t) = \frac{\widetilde{\varphi}(t)}{\widetilde{c}} &= \sup_{\widetilde{z} \in]\widetilde{c}\cdot\lambda_-, \widetilde{c}\cdot\lambda_+[} \left(\frac{\widetilde{z}}{\widetilde{c}} \cdot t - \log \Big(\int_{\mathbb{R}} e^{\frac{\widetilde{z}}{\widetilde{c}}\cdot\widetilde{y}\cdot\widetilde{c}} d\widetilde{\mathfrak{C}}_{\widetilde{c}}(\widetilde{y}) \Big)^{1/\widetilde{c}} \right), \\ &= \sup_{z \in]\lambda_-, \lambda_+[} \left(z \cdot t - \log \Big(\int_{\mathbb{R}} e^{z\cdot\widetilde{y}\cdot\widetilde{c}} d\widetilde{\mathfrak{C}}_{\widetilde{c}}(\widetilde{y}) \Big)^{1/\widetilde{c}} \right), \qquad t \in]a,b[\,, \end{aligned} \tag{274}$$

and hence for the exponential of its Fenchel-Legendre transform

$$e^{\varphi_*(z)} = \Big(\int_{\mathbb{R}} e^{z\cdot\widetilde{y}\cdot\widetilde{c}} d\widetilde{\mathfrak{C}}_{\widetilde{c}}(\widetilde{y}) \Big)^{1/\widetilde{c}}, \qquad z \in]\lambda_-, \lambda_+[. \tag{275}$$

Here, according to (G9i) for $\widetilde{\varphi}$ we have used $\widetilde{\lambda}_- := \inf_{t \in]a,b[} \widetilde{\varphi}'(t) = \widetilde{c} \cdot \lambda_-$ and $\widetilde{\lambda}_+ := \sup_{t \in]a,b[} \widetilde{\varphi}'(t) = \widetilde{c} \cdot \lambda_+$.
From (272) and (275) we deduce for $\widetilde{c} := \frac{1}{n}$ the relation $MGF_{\mathfrak{C}}(z) = (MGF_{\widetilde{\mathfrak{C}}_{1/n}}(\frac{z}{n}))^n$ for all $n \in \mathbb{N}$ which (with the help of (ii)) implies the infinitely divisibility of $\mathfrak{C}$.
For the reverse direction, let us assume that $\varphi \in \Upsilon(]a,b[)$ and that the corresponding $\mathfrak{C}$ is infinitely divisible. Recall that $]a,b[= int(dom(\varphi))$. Moreover, we fix an arbitrary constant $\widetilde{c} > 0$. Of course, there holds $\widetilde{c} \cdot \varphi \in \widetilde{\Upsilon}(]a,b[)$ and $dom(\widetilde{c} \cdot \varphi) = dom(\varphi)$. Furthermore, by multiplying (271) with $\widetilde{c} > 0$ and by employing (i), (ii) we get

$$\begin{aligned} \widetilde{c} \cdot \varphi(t) &= \sup_{z \in]\lambda_-, \lambda_+[} \left(\widetilde{c} \cdot z \cdot t - \log \Big(\int_{\mathbb{R}} e^{\widetilde{c}\cdot z \cdot \frac{y}{\widetilde{c}}} d\mathfrak{C}(y) \Big)^{\widetilde{c}} \right) = \sup_{\widetilde{z} \in]\widetilde{c}\cdot\lambda_-, \widetilde{c}\cdot\lambda_+[} \left(\widetilde{z} \cdot t - \log \Big(\int_{\mathbb{R}} e^{\frac{\widetilde{z}}{\widetilde{c}}\cdot y} \, d\mathfrak{C}(y) \Big)^{\widetilde{c}} \right) \\ &= \sup_{\widetilde{z} \in]\widetilde{c}\cdot\lambda_-, \widetilde{c}\cdot\lambda_+[} \left(\widetilde{z} \cdot t - \log \Big(\int_{\mathbb{R}} e^{\widetilde{z}\cdot y} d\mathfrak{C}_{\widetilde{c}}(y) \Big) \right), \quad t \in]a,b[; \end{aligned} \tag{276}$$

for some probability distribution $\mathfrak{C}_{\widetilde{c}}$ on $\mathbb{R}$. ■

Proof of Proposition 35.
It is well known that a candidate function $M :]-\infty, 0[\mapsto]0, \infty[$ is the moment-generating function of an infinitely divisible probability distribution if and only if $(\log M)'$ is absolutely monotone (see e.g. Theorem 5.11 of Schilling et al. [322]). By applying this to $M(z) := e^{-a\cdot z + \varphi^*(z)}$ respectively $M(z) := e^{b\cdot z + \varphi^*(-z)}$, one gets straightforwardly the assertion (a) respectively (b); notice that the light-tailedness follows then from (G1) to (G8), and $b = \infty$ respectively $a = -\infty$ can be deduced from the fact that the support of an infinitely distribution is always (one-sided or two-sided) unbounded. For the third case $a = -\infty$, $b = \infty$ one can use the assertion (cf. e.g. Morris [267], p.73) that a candidate function $M :]\lambda_-, \lambda_+[\mapsto]0, \infty[$ is the moment-generating function of an infinitely divisible probability distribution if the connected function $z \mapsto (\log M)''(z)/(\log M)''(0)$ is the moment-generating function of some auxiliary probability distribution; but the latter is equivalent to exponentially convexity (cf. Theorem 30(b)). By applying this to $M(z) := e^{\varphi^*(z)}$, one ends up with (c). ■

Appendix F
Proofs — Part 6

Proof of Theorem 36. (i) Clearly, on $]\lambda_-, \lambda_+[$ the function Λ is differentiable with strictly increasing derivative

$$\Lambda'(z) = F^{-1}(z + c) + 1 - F^{-1}(c), \qquad z \in]\lambda_-, \lambda_+[. \tag{277}$$

Hence, Λ is strictly convex and smooth (because of the smoothness of F^{-1}), and satisfies $\Lambda(0) = 0$ as well as $\Lambda'(0) = 1$. Also, the corresponding extensions of Λ to $z = \lambda_-$ and $z = \lambda_+$ are continuous.
(ii) It is straightforward to see that on $]t_-^{sc}, t_+^{sc}[$ the function φ is differentiable with strictly increasing derivative

$$\varphi'(t) = F(t + F^{-1}(c) - 1) - c, \qquad t \in]t_-^{sc}, t_+^{sc}[. \tag{278}$$

Hence, φ is strictly convex and smooth (because of the smoothness of F), and satisfies $\varphi(1) = 0$ as well as $\varphi'(1) = 0$. Also, the corresponding extensions of φ to $t = t_-^{sc}$ and $t = t_-^{sc}$ are continuous. Hence (G1), (G2), (G5) and (G6) hold.

To prove (G3) (and hence (G1)), let us first notice that obviously there holds $a \leq t_-^{sc}$ and $t_+^{sc} \leq b$. Moreover, the validity of $\varphi(t) < \infty$ for all $t \in]t_-^{sc}, t_+^{sc}[$ is clear from (163) since $t + F^{-1}(c) - 1 \in]a_F, b_F[= int(dom(F))$ and the involved integral over the continuous function F^{-1} is taken over a compact interval.

For the subcase $t_-^{sc} = -\infty = a$ we have thus shown $dom(\varphi) \cap] - \infty, 1] \;=] - \infty, 1] \;=]a, 1]$, whereas for the subcase $t_+^{sc} = \infty = b$ we have verified $dom(\varphi) \cap [1, \infty[= [1, \infty[= [1, b[$.

Let us next examine the subcase "$t_-^{sc} > -\infty$ and $\varphi(t_-^{sc}) < \infty$": if $\lambda_- > -\infty$ then $a = -\infty$ and (163) implies $\varphi(t) = \varphi(t_-^{sc}) + \lambda_- \cdot (t - t_-^{sc}) < \infty$ for all $t \in] - \infty, t_-^{sc}] =]a, t_-^{sc}]$, which leads to $dom(\varphi) \cap] - \infty, 1] =] - \infty, 1] =]a, 1]$; in contrast, if $\lambda_- = -\infty$ then $a = t_-^{sc}$ and (163) implies $\varphi(t) = \varphi(t_-^{sc}) + \lambda_- \cdot (t - t_-^{sc}) = \infty$ for all $t \in] - \infty, t_-^{sc}[=] - \infty, a[$, which leads to $dom(\varphi) \cap] - \infty, 1] = [a, 1]$.

In the subcase "$t_-^{sc} > -\infty$ and $\varphi(t_-^{sc}) = \infty$", due to the strict convexity of φ one always has $\lim_{t \downarrow t_-^{sc}} \varphi'(t) = -\infty$; this implies, by the below-mentioned (279), that $\lambda_- = -\infty$ and thus $a = t_-^{sc}$; from (163) we derive $\varphi(t) = \varphi(t_-^{sc}) + \lambda_- \cdot (t - t_-^{sc}) = \infty$ for all $t \in] - \infty, t_-^{sc}[=] - \infty, a[$, which leads to $dom(\varphi) \cap] - \infty, 1] =]a, 1]$.

As a further step, we deal with the subcase "$t_+^{sc} < \infty$ and $\varphi(t_+^{sc}) < \infty$": if $\lambda_+ < \infty$ then $b = \infty$ and (163) implies $\varphi(t) = \varphi(t_+^{sc}) + \lambda_+ \cdot (t - t_+^{sc}) < \infty$ for all $t \in \; [t_+^{sc}, \infty[= [t_+^{sc}, b[$, which leads to $dom(\varphi) \cap [1, \infty[= [1, \infty[= [1, b[$; in contrast, if $\lambda_+ = \infty$ then $b = t_+^{sc}$ and (163) implies $\varphi(t) = \varphi(t_+^{sc}) + \lambda_+ \cdot (t - t_+^{sc}) = \infty$ for all $t \in]t_-^{sc}, \infty[=]b, \infty[$, which leads to $dom(\varphi) \cap [1, \infty[= [1, b]$.

In the subcase "$t_+^{sc} < +\infty$ and $\varphi(t_+^{sc}) = \infty$", due to the strict convexity of φ one always gets $\lim_{t \uparrow t_+^{sc}} \varphi'(t) = \infty$; this implies, by the below-mentioned (280), that $\lambda_+ = \infty$ and thus $b = t_+^{sc}$; from (163) we deduce $\varphi(t) = \varphi(t_+^{sc}) + \lambda_+ \cdot (t - t_+^{sc}) = \infty$ for all $t \in]t_+^{sc}, \infty[=]b, \infty[$, which leads to $dom(\varphi) \cap [1, \infty[= [1, b[$.
Putting things together, we have proved (G3). The property (G4) follows straightforwardly from (278), the continuity of F and from $\lim_{t \downarrow t_-^{sc}} \varphi'(t) = \lambda_-$, $\lim_{t \uparrow t_+^{sc}} \varphi'(t) = \lambda_+$. To see the latter two, from (278) we obtain

$$\lim_{t \downarrow t_-^{sc}} \varphi'(t) = \lim_{t \downarrow t_-^{sc}} F(t + F^{-1}(c) - 1) - c = \lim_{t \downarrow t_-^{sc}} F(t + a_F - t_-^{sc}) - c = \inf\{F(\widetilde{t}) - c : \widetilde{t} \in]a_F, b_F[\} = \lambda_-, \tag{279}$$

$$\lim_{t \uparrow t_+^{sc}} \varphi'(t) = \lim_{t \uparrow t_+^{sc}} F(t + F^{-1}(c) - 1) - c = \lim_{t \uparrow t_+^{sc}} F(t + b_F - t_+^{sc}) - c = \sup\{F(\widetilde{t}) - c : \widetilde{t} \in]a_F, b_F[\} = \lambda_+. \tag{280}$$

The two properties (G7) and (G8) are clear form the above considerations.
(iii) From (277) and (278) one gets easily

$$\Lambda'^{-1}(t) = F\left(t + F^{-1}(c) - 1\right) - c = \varphi'(t), \qquad t \in]t_-^{sc}, t_+^{sc}[, \tag{281}$$

as well as $\Lambda'^{-1}(1) = 0$. From this, we derive

$$\begin{aligned} & t \cdot \Lambda'^{-1}(t) - \Lambda\left(\Lambda'^{-1}(t)\right) \\ & = t \cdot \left[F\left(t + F^{-1}(c) - 1\right) - c\right] + \left[F^{-1}(c) - 1\right] \cdot \left[F\left(t + F^{-1}(c) - 1\right) - c\right] \\ & - \int_0^{F\left(t + F^{-1}(c) - 1\right) - c} F^{-1}(u + c) du \\ & = \varphi(t), \qquad t \in]t_-^{sc}, t_+^{sc}[, \end{aligned} \tag{282}$$

and hence, with the help of (281) in combination with (279), (280)

$$\varphi(t) = \max_{z \in]\lambda_-, \lambda_+[} \left(z \cdot t - \Lambda(z)\right), \qquad t \in]t_-^{sc}, t_+^{sc}[, \tag{283}$$

i.e. on $]t_-^{sc}, t_+^{sc}[$ the divergence generator φ is the classical Legendre transform of the restriction of Λ to $]\lambda_-, \lambda_+[$. If "$\lambda_- > -\infty$, $\Lambda(\lambda_-) \in] - \infty, \infty[$ and $\Lambda'(\lambda_-) \in] - \infty, \infty[$" respectively "$\lambda_+ < -\infty$, $\Lambda(\lambda_+) \in] - \infty, \infty[$ and $\Lambda'(\lambda_+) \in] - \infty, \infty[$", then one can apply classical facts of Fenchel-Legendre transformation to get the corresponding left-hand respectively right-hand linear extensions of φ on the complement of $]t_-^{sc}, t_+^{sc}[$, in order to obtain the desired

$$\varphi(t) = \sup_{z \in] - \infty, \infty[} \left(z \cdot t - \Lambda(z)\right), \qquad t \in \mathbb{R}; \tag{284}$$

notice that $t_-^{sc} = \lim_{z \downarrow \lambda_-} \Lambda'(z)$ and $t_+^{sc} = \lim_{z \uparrow \lambda_+} \Lambda'(z)$.
(iv) This is just the inverse of (iii), by applying standard Fenchel-Legendre-transformation theory. ■

APPENDIX G
FURTHER DETAILS AND PROOFS FOR SUBSECTION VI-A

Proof of Lemma 57. By Assumption (OM), one gets for all $\boldsymbol{\lambda} \in cl(\boldsymbol{\Lambda})$ that $\{\mathbf{x} \in (dom(\widetilde{\varphi})^n : T(\mathbf{x}) = \boldsymbol{\lambda}\} \cap]t_-^{sc}, t_+^{sc}[^n \neq \emptyset$. Moreover, for any $\mathbf{x} = (x_1, .., x_n)$ in $\mathbb{R}^n$, by the independence of the components of $\widetilde{\mathbf{W}}$ as well as (265) and (270), we have

$$I_{\widetilde{\mathbf{W}}}(\mathbf{x}) = \sup_{\mathbf{z}=(z_1,\ldots,z_n)\in\mathbb{R}^n} \left(\langle \mathbf{x}, \mathbf{z} \rangle - \sum_{i=1}^n \Lambda_{\widetilde{\zeta}}(z_i) \right) = \sup_{\mathbf{z}\in]\lambda_-,\lambda_+[^n} \left(\sum_{i=1}^n \Big(x_i \cdot z_i - \Lambda_{\widetilde{\zeta}}(z_i) \Big) \right)$$
$$= \sum_{i=1}^n \left(\sup_{z_i\in]\lambda_-,\lambda_+[} \Big(x_i \cdot z_i - \Lambda_{\widetilde{\zeta}}(z_i) \Big) \right) = \sum_{i=1}^n \widetilde{\varphi}(x_i) = \sum_{k=1}^K \sum_{i\in I_k^{(n)}} \widetilde{\varphi}(x_i) \tag{285}$$

which is finite if and only if $\mathbf{x} \in (dom(\widetilde{\varphi}))^n$ (recall that $\widetilde{\varphi}$ is a nonnegative function). Hence, for each $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ we obtain

$$I(\boldsymbol{\lambda}) \;:=\; \inf_{\mathbf{x}\in\mathbb{R}^n:\,T(\mathbf{x})=\boldsymbol{\lambda}} I_{\widetilde{\mathbf{W}}}(\mathbf{x}) = \inf_{\mathbf{x}\in(dom(\widetilde{\varphi}))^n:\,T(\mathbf{x})=\boldsymbol{\lambda}} I_{\widetilde{\mathbf{W}}}(\mathbf{x}) = \inf_{\mathbf{x}\in(dom(\widetilde{\varphi}))^n:\,T(\mathbf{x})=\boldsymbol{\lambda}} \sum_{k=1}^K \sum_{i\in I_k^{(n)}} \widetilde{\varphi}(x_i) \tag{286}$$

$$= \sum_{k=1}^K n_k \cdot \widetilde{\varphi}(\lambda_k) \;=\; n \cdot \sum_{k=1}^K \widetilde{p}_k \cdot \widetilde{\varphi}(\lambda_k) = \inf_{\mathbf{x}\in]t_-^{sc},t_+^{sc}[^n:\,T(\mathbf{x})=\boldsymbol{\lambda}} I_{\widetilde{\mathbf{W}}}(\mathbf{x})\,; \tag{287}$$

here, we have employed the following facts: (i) the right-most infimum in (286) is achieved by minimizing each of the K terms $\sum_{i\in I_k^{(n)}} \widetilde{\varphi}(x_i)$ under the linear constraint $\frac{1}{n_k} \cdot \sum_{i\in I_k^{(n)}} x_i = \lambda_k$, and by the strict convexity of $\widetilde{\varphi}$ on $]t_-^{sc}, t_+^{sc}[$ (cf. (G5)) the minimum of this generic term is attained when all components x_i are equal to λ_k, and (ii) the outcoming minimum does not depend on the particular (generally non-unique) choice of the x_i's. Notice that we have used the relation $n_k = n \cdot \widetilde{p}_k$ as well. To proceed, let $\underline{\boldsymbol{\lambda}}$ be a minimal rate point of $\boldsymbol{\Lambda}$, which means that $\underline{\boldsymbol{\lambda}} \in \partial\boldsymbol{\Lambda}$ and $I(\underline{\boldsymbol{\lambda}}) \leq I(\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$. By Assumption (OM) one can run all the steps in (286) and (287) with $\underline{\boldsymbol{\lambda}}$ instead of $\boldsymbol{\lambda}$, and hence

$$I(\underline{\boldsymbol{\lambda}}) \;=\; \inf_{\mathbf{x}\in\mathbb{R}^n:\,T(\mathbf{x})=\underline{\boldsymbol{\lambda}}} I_{\widetilde{\mathbf{W}}}(\mathbf{x}) \;=\; \inf_{\mathbf{x}\in]t_-^{sc},t_+^{sc}[^n:\,T(\mathbf{x})=\underline{\boldsymbol{\lambda}}} I_{\widetilde{\mathbf{W}}}(\mathbf{x}) \;=\; n \cdot \sum_{k=1}^K \widetilde{p}_k \cdot \widetilde{\varphi}(\underline{\lambda}_k) \;=\; n \cdot \sum_{k=1}^K \widetilde{p}_k \cdot \widetilde{\varphi}(\underline{\widetilde{q}}_k/\widetilde{p}_k) \tag{288}$$

where for the last equality we have employed the vector $\underline{\widetilde{\mathbf{Q}}} := \mathfrak{D}^{-1}\underline{\boldsymbol{\lambda}}$ which we have called the "dominating point of $\widetilde{\boldsymbol{\Omega}}$". Also we have proved

$$I(\underline{\boldsymbol{\lambda}}) \;=\; n \cdot \inf_{\widetilde{\mathbf{Q}}\in\widetilde{\boldsymbol{\Omega}}} \sum_{k=1}^K \widetilde{p}_k \cdot \widetilde{\varphi}(\widetilde{q}_k/\widetilde{p}_k). \qquad \blacksquare \tag{289}$$

On the obtainment of proxies of minimal rate points by proxy method 2:
For the rest of this section, besides (OM) we assume that $dom(\widetilde{\varphi}) =]a, b[=]t_-^{sc}, t_+^{sc}[$, and that in case of $a = -\infty$ or $b = +\infty$ the divergence generator $\widetilde{\varphi}$ is regularly varying at a or b accordingly, with positive index β, i.e. (with a slight abuse of notation)

- if $a = -\infty$, then for all $\lambda > 0$ there holds
$$\lim_{u\to-\infty} \frac{\widetilde{\varphi}\,(\lambda u)}{\widetilde{\varphi}\,(u)} = \lambda^\beta,$$
- if $b = +\infty$, then for all $\lambda > 0$ there holds
$$\lim_{u\to+\infty} \frac{\widetilde{\varphi}\,(\lambda u)}{\widetilde{\varphi}\,(u)} = \lambda^\beta;$$

this assumption is denoted by (H$\widetilde{\varphi}$).
A proxy of $\underline{\widetilde{\mathbf{Q}}}$ can be obtained by sampling from a distribution on $\mathbb{R}^K$ defined through

$$f(\widetilde{\mathbf{Q}}) := C \cdot \exp\left(-\sum_{k=1}^K \widetilde{p}_k \cdot \widetilde{\varphi}(\widetilde{q}_k/\widetilde{p}_k) \right) \;=\; C \cdot \exp\left(-D_{\widetilde{\varphi}}\left(\widetilde{\mathbf{Q}}, \widetilde{\mathbb{P}} \right) \right) \qquad \text{(cf. (220))}$$

where C is a normalizing constant; strict convexity (cf. (G5)) of $\widetilde{\varphi}$ together with (H$\widetilde{\varphi}$) prove that f is a well-defined (Lebesgue-) density for a random variable $\mathbf{T}$ on $\mathbb{R}^K$. We denote by $\mathbb{F}(\cdot) := \Pi[\mathbf{T} \in \cdot]$ the corresponding distribution on $\mathbb{R}^K$ having density f. The distribution of $\mathbf{T}$ given $\left(\mathbf{T} \in \widetilde{\boldsymbol{\Omega}}\right)$ concentrates on the points in $\widetilde{\boldsymbol{\Omega}}$ which minimize $D_{\widetilde{\varphi}}\left(\widetilde{\mathbf{Q}}, \widetilde{\mathbb{P}} \right)$ as $\widetilde{\mathbf{Q}}$ runs in $\widetilde{\boldsymbol{\Omega}}$, when $D_{\widetilde{\varphi}}(\widetilde{\boldsymbol{\Omega}}, \widetilde{\mathbb{P}})$ is large. This can be argued as follows. We will consider the case when $\widetilde{\boldsymbol{\Omega}}$ is a compact subset in $\mathbb{R}^K_{>0}$ and $\widetilde{\varphi}$ satisfies (H$\widetilde{\varphi}$) with $b = +\infty$. For the case when $\widetilde{\boldsymbol{\Omega}}$ is not compact, or belongs to $\mathbb{R}^K / \{\mathbf{0}\}$, see the Remark 67 hereunder.

Consider a compact set $\mathbf{\Gamma}$ in $\widetilde{\mathbf{\Omega}}$ and let $\mathbf{\Gamma}_t$ be defined as deduced from $\mathbf{\Gamma}$ in a way that makes $D_{\widetilde{\varphi}}\left(\mathbf{\Gamma}_t, \widetilde{\mathbb{P}}\right)$ increase with t for sufficiently large t. For instance, set

$$\mathbf{\Gamma}_t := t \cdot \mathbf{\Gamma}. \tag{290}$$

Hence, in case of $b = +\infty$ the divergence

$$D_{\widetilde{\varphi}}\left(\mathbf{\Gamma}_t, \widetilde{\mathbb{P}}\right) = \inf_{\boldsymbol{g}_t \in \mathbf{\Gamma}_t} \sum_{k=1}^{K} \widetilde{p}_k \cdot \widetilde{\varphi}\left(\frac{(\boldsymbol{g}_t)_k}{\widetilde{p}_k}\right) = \inf_{\boldsymbol{g} \in \mathbf{\Gamma}} \sum_{k=1}^{K} \widetilde{p}_k \cdot \widetilde{\varphi}\left(\frac{t \cdot g_k}{\widetilde{p}_k}\right)$$

tends to infinity as $t \to \infty$; the case $a = -\infty$ works analogously with $t \to -\infty$. In case of $b < \infty$ we may consider

$$\mathbf{\Gamma}_t := \{b - \boldsymbol{g}/t; \boldsymbol{g} \in \mathbf{\Gamma}\} \tag{291}$$

and indeed $D_{\widetilde{\varphi}}\left(\mathbf{\Gamma}_t, \widetilde{\mathbb{P}}\right) \to \infty$ as $t \to \infty$, with a similar statement when $a > -\infty$.

Assume that $\mathbf{\Gamma}$ has a dominating point $\underline{\boldsymbol{g}}$. Then $\mathbf{\Gamma}_t$ has dominating point $\underline{\boldsymbol{g}}_t := t \cdot \underline{\boldsymbol{g}}$. We prove that $\mathbf{T}$ with distribution (220) cannot be too far away (depending on t) from $\underline{\boldsymbol{g}}_t$ whenever $\mathbf{T}$ belongs to $\mathbf{\Gamma}_t$. This argument is valid in the present description of some asymptotics which makes $\mathbf{\Gamma}_t$ as a model for $\widetilde{\mathbf{\Omega}}$ for large t; considering the case when $D_{\widetilde{\varphi}}\left(\widetilde{\mathbf{\Omega}}, \widetilde{\mathbb{P}}\right)$ is large is captured through the asymptotic statement

$$\lim_{t \to \infty} D_{\widetilde{\varphi}}\left(\mathbf{\Gamma}_t, \widetilde{\mathbb{P}}\right) = +\infty.$$

There holds the following

Proposition 65: With the above notation and under condition (H$\widetilde{\varphi}$), denote by $\mathbf{B}$ a neighborhood of $\underline{\boldsymbol{g}}$ and $\mathbf{B}_t := t \cdot \mathbf{B}$. Then

$$\mathbb{F}[\mathbf{\Gamma}_t \cap \mathbf{B}_t^c \,|\, \mathbf{\Gamma}_t] = \mathbb{P}\left[\mathbf{T} \in \mathbf{\Gamma}_t \cap \mathbf{B}_t^c \,|\, \mathbf{T} \in \mathbf{\Gamma}_t\right] \to 0$$

as $t \to \infty$, which proves that simulations under (220) produce proxies of the dominating points $\underline{\boldsymbol{g}}_t$ in $\mathbf{\Gamma}_t$.

Before we start with the proof of Proposition 65, we first quote the following

Lemma 66: Let $\widetilde{\varphi}$ satisfy (H$\widetilde{\varphi}$) with $b = +\infty$. Then for all $\mathbf{A}$ in $\mathbb{R}^K$ such that

$$\breve{\alpha} := D_{\widetilde{\varphi}}(\mathbf{A}, \widetilde{\mathbb{P}}) := \inf_{\mathbf{v} \in \mathbf{A}} \sum_{k=1}^{K} \widetilde{p}_k \cdot \widetilde{\varphi}\left(\frac{v_k}{\widetilde{p}_k}\right)$$

is finite there holds

$$\lim_{t \to \infty} \frac{1}{t} \log \int_{\mathbf{A}} \exp\left(-t \sum_{k=1}^{K} \widetilde{p}_k \cdot \widetilde{\varphi}\left(\frac{v_k}{\widetilde{p}_k}\right)\right) dv_1 \ldots dv_k = -D_{\widetilde{\varphi}}\left(\mathbf{A}, \widetilde{\mathbb{P}}\right).$$

Proof of Lemma 66. Let us first remark that according to the geometry of the set $\mathbf{A}$, various combinations for the asymptotics (290) or (291) may occur; for sake of brevity, we only handle the simplest ones, since all turn to be amenable through the same arguments. Denote for positive r

$$\mathbf{B}(r) := \left\{\mathbf{v} \in \mathbb{R}^K : \sum_{k=1}^{K} \widetilde{p}_k \widetilde{\varphi}\left(\frac{v_k}{\widetilde{p}_k}\right) > r\right\}.$$

It holds, by making the change of variable $r = t \cdot \breve{\alpha} + t \cdot s$,

$$\int_{\mathbf{A}} \exp\left(-t \sum_{k=1}^{K} \widetilde{p}_k \cdot \widetilde{\varphi}\left(\frac{v_k}{\widetilde{p}_k}\right)\right) dv_1 \ldots dv_k = \int \cdots \int 1_{\mathbb{R}^+}(r) \cdot 1_{\mathbf{A}}(\mathbf{v}) \cdot 1_{\left]t \sum_{k=1}^{K} \widetilde{p}_k \widetilde{\varphi}\left(\frac{v_k}{\widetilde{p}_k}\right), \infty\right[}(r) \cdot e^{-r} \, dr \, dv_1 \ldots dv_K$$

$$= t e^{-t\breve{\alpha}} \int \cdots \int 1_{]-\breve{\alpha}, \infty[}(s) \cdot 1_{\mathbf{A}}(\mathbf{v}) \cdot 1_{\mathbf{B}^c(\breve{\alpha}+s)}(\mathbf{v}) \cdot e^{-ts} \, ds \, dv_1 \ldots dv_K \;=\; t e^{-t\breve{\alpha}} \int_{-\breve{\alpha}}^{\infty} Vol\left(\mathbf{A} \cap \mathbf{B}^c(\breve{\alpha}+s)\right) \cdot e^{-ts} \, ds.$$

Let $I_t := t \cdot \int_0^{\infty} Vol\left(\mathbf{A} \cap \mathbf{B}^c(\breve{\alpha}+s)\right) e^{-ts} ds$. We prove that

$$\lim_{t \to \infty} \frac{1}{t} \log I_t = 0. \tag{292}$$

When $a = -\infty$ or $b = +\infty$, since $\widetilde{\varphi}$ satisfies (H$\widetilde{\varphi}$) there exists a polynomial P such that

$$Vol\left(\mathbf{A} \cap \mathbf{B}^c(\breve{\alpha}+s)\right) \leq P(s)\,;$$

whence, assuming without loss of generality that $dom(\widetilde{\varphi}) = \mathbb{R}^+$, we obtain

$$\frac{1}{t}\log I_t \leq \frac{1}{t}\log\int_0^\infty P\left(\frac{u}{t}\right) te^{-u}du$$

which yields that for large t

$$\frac{1}{t}\log I_t < 0.$$

When dealing with a context where a or b have finite value and the corresponding sets $\mathbf{\Gamma}_t$ are "far away" from $\mathbf{\Gamma}$ in terms of the distance measure $D_{\widetilde{\varphi}}\left(\cdot, \widetilde{\mathbb{P}}\right)$, then $Vol\left(\mathbf{A}\cap\mathbf{B}^c(\breve{\alpha}+s)\right)$ is bounded. Hence, $\limsup_{t\to\infty}\frac{1}{t}\log I_t \leq 0$. Now fix $\varepsilon > 0$. Then, since $Vol\left(\mathbf{A}\cap\mathbf{B}^c(a+s)\right)$ is increasing in s, we get

$$\begin{aligned} I_t &\geq t\int_\varepsilon^\infty Vol\left(\mathbf{A}\cap\mathbf{B}^c(\breve{\alpha}+s)\right)e^{-ts}ds \\ &\geq Vol\left(\mathbf{A}\cap\mathbf{B}^c(\breve{\alpha}+\varepsilon)\right)e^{-t\varepsilon}. \end{aligned}$$

Hence

$$\frac{1}{t}\log I_t \geq \frac{1}{t}\log Vol\left(\mathbf{A}\cap\mathbf{B}^c(\breve{\alpha}+\varepsilon)\right) - \varepsilon$$

which yields $\liminf_{t\to\infty}\frac{1}{t}\log I_t \geq 0$. Therefore (292) holds, which concludes the proof. ■

We now turn to the

Proof of Proposition 65. Without loss of generality, let $b = +\infty$, $\mathbf{\Gamma}_t$ as in (290) and Condition (H$\widetilde{\varphi}$) hold. Moreover, consider an arbitrary neighborhood $\mathbf{B}$ of $\underline{\boldsymbol{g}}$ and the corresponding neighborhoods $\mathbf{B}_t := t\cdot\mathbf{B}$ of $\underline{\boldsymbol{g}}_t = t\cdot\underline{\boldsymbol{g}}$. There holds

$$\begin{aligned} \frac{1}{\widetilde{\varphi}(t)}\log\mathbb{\Pi}\left[\mathbf{T}\in\mathbf{\Gamma}_t\right] &= \frac{C}{\widetilde{\varphi}(t)}\log\int_{\mathbf{\Gamma}_t}\exp\left(-\sum_{k=1}^K\widetilde{p}_k\cdot\widetilde{\varphi}\left(\frac{w_k}{\widetilde{p}_k}\right)\right)dw_1\dots dw_K \\ &\overset{(1)}{=} \frac{CK}{\widetilde{\varphi}(t)}\log t + \frac{C}{\widetilde{\varphi}(t)}\log\int_{\mathbf{\Gamma}}\exp\left(-t^\beta\cdot\sum_{k=1}^K\widetilde{p}_k\cdot\left(\widetilde{\varphi}\left(\frac{v_k}{\widetilde{p}_k}\right)\cdot(1+o(1))\right)\right)dv_1\dots dv_K \\ &\overset{(2)}{=} \frac{CK}{\widetilde{\varphi}(t)}\log t + \frac{C}{(\widetilde{\varphi}(t)/t^\beta)}\cdot\frac{1}{t^\beta}\log\left((1+o(1))\cdot\int_{\mathbf{\Gamma}}\exp\left(-t^\beta\sum_{k=1}^K\widetilde{p}_k\cdot\widetilde{\varphi}\left(\frac{v_k}{\widetilde{p}_k}\right)\right)dv_1\dots dv_K\right) \\ &\overset{(3)}{=} -\frac{Ct^\beta}{\widetilde{\varphi}(t)}\cdot D_{\widetilde{\varphi}}(\mathbf{\Gamma},\widetilde{\mathbb{P}})\cdot(1+o(1)) \\ &\overset{(4)}{=} -\breve{l}(t)\cdot D_{\widetilde{\varphi}}(\mathbf{\Gamma},\widetilde{\mathbb{P}})\cdot(1+o(1)) \end{aligned}$$

as t tends to infinity. In the above display, (1) follows from $\widetilde{\varphi}(tx) = (tx)^\beta\cdot\ell(tx) = t^\beta\cdot x^\beta\cdot\ell(x)\cdot\frac{\ell(tx)}{\ell(x)} = t^\beta\cdot\widetilde{\varphi}(x)\cdot(1+o(1))$ as t tends to infinity and x lies in a compact subset of $]0,\infty[$, where ℓ is a slowly varying function. The equality (2) follows from compactness of $\mathbf{\Gamma}$ together with the fact that $\widetilde{\varphi}$ is a regularly varying function with index β, so that

$$\lim_{t\to\infty}\frac{\widetilde{\varphi}(tv)}{\widetilde{\varphi}(t)} = v^\beta$$

uniformly upon v on compact sets in $]0,\infty[$. The remaining equalities (3) and (4) follow from classical properties of regularly varying functions, where $\breve{\ell} := 1/\ell$ is a slowly varying function at infinity, together with standard Laplace-Integral approximation.

In the same way we can show

$$\frac{1}{\widetilde{\varphi}(t)}\log\mathbb{\Pi}\left[\,\mathbf{T}\in\mathbf{\Gamma}_t\cap\mathbf{B}_t^c\,\right] = -\breve{l}(t)\cdot D_{\widetilde{\varphi}}(\mathbf{\Gamma}\cap\mathbf{B}^c,\widetilde{\mathbb{P}})\cdot(1+o(1))$$

as t tends to infinity. Since $\mathbf{B}$ is a neighborhood of the unique dominating point $\underline{\boldsymbol{g}}$ of $\mathbf{\Gamma}$, one gets that $D_{\widetilde{\varphi}}(\mathbf{\Gamma}\cap\mathbf{B}^c,\widetilde{\mathbb{P}}) > D_{\widetilde{\varphi}}(\mathbf{\Gamma},\widetilde{\mathbb{P}})$. This implies that

$$\mathbb{\Pi}\left[\,\mathbf{T}\in\mathbf{\Gamma}_t\cap\mathbf{B}_t^c\,\middle|\,\mathbf{T}\in\mathbf{\Gamma}_t\,\right]\to 0 \qquad \text{as } t\to\infty. \qquad ■$$

Remark 67: Firstly, let us quote that the case when $\widetilde{\mathbf{\Omega}}$ is an unbounded subset in $\mathbb{R}^K/\{\mathbf{0}\}$ is somewhat immaterial for applications. Anyhow, if compactness of $\mathbf{\Gamma}$ is lost, then in order to use the same line of arguments as above, it is necessary to strengthen the assumptions (H$\widetilde{\varphi}$) e.g. as follows: when $b=+\infty$ then $\widetilde{\varphi}$ has to be asymptotically homogeneous with degree $\beta>0$, in the sense that $\widetilde{\varphi}(tx)=t^{\beta}\widetilde{\varphi}(x)\cdot(1+o(1))$ as $t\rightarrow\infty$; for the subcase $a=-\infty$ one employs an analogous assumption as $t\rightarrow-\infty$. The case when $\widetilde{\mathbf{\Omega}}$ is a compact set in $\mathbb{R}^K\backslash\{\mathbf{0}\}$ can be treated as above, by combining the asymptotics in t in the neighborhood of a and b accordingly.

Appendix H
Proof for Subsection VI-B

Proof of Proposition 60. Recall the weighted empirical measure

$$\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{V}}:=\left(\frac{1}{n}\sum_{i\in I_1^{(n)}}V_i,\ldots,\frac{1}{n}\sum_{i\in I_K^{(n)}}V_i\right)$$

which satisfies the K linear constraints defined in (235) through

$$E_S[\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{V}}]=\boldsymbol{\xi}_{M,\mathbf{X}}^{\mathbf{W}^*}=\overline{W^*}\cdot\boldsymbol{\xi}_{M,\mathbf{X}}^{w\mathbf{W}^*}$$

where $\mathbf{Q}^*:=(q_1^*,\ldots,q_K^*)=\boldsymbol{\xi}_{M,\mathbf{X}}^{w\mathbf{W}^*}\in int(\mathbf{\Omega})$ and $\overline{W^*}=\frac{1}{M}\sum_{j=1}^M W_j^*$. The probability distribution S defined on $\mathbb{R}^n$ is the Kullback-Leibler projection of $\mathfrak{C}^{\otimes n}$ on the class of all probability distributions on $\mathbb{R}^n$ which satisfy (235). We prove that $\liminf_{n\rightarrow\infty}S\left[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}}\in\mathbf{\Omega}\right]>0$. To start with, we define for strictly positive δ the set

$$A_{n,\delta}:=\left\{\left|\frac{1}{n}\sum_{i=1}^n V_i-\overline{W^*}\right|\leq\delta\right\}$$

and write

$$S\left[\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}}\in\mathbf{\Omega}\right]=S\left[\{\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}}\in\mathbf{\Omega}\}\cap A_{n,\delta}\right]+S\left[\{\boldsymbol{\xi}_{n,\mathbf{X}}^{w\mathbf{V}}\in\mathbf{\Omega}\}\cap A_{n,\delta}^c\right]=:I+II.$$

By the law of large numbers, the second term II tends to 0 as n tends to infinity. Moreover, one can rewrite

$$I=S\left[\bigcup_{m\in\left[\overline{W^*}-\delta,\overline{W^*}+\delta\right]}\left\{\boldsymbol{\xi}_{n,\mathbf{X}}^{\mathbf{V}}\in m\cdot\mathbf{\Omega}\right\}\right]$$

which entails

$$I\ \geq\ S\left[\frac{1}{n_k}\sum_{i\in I_k^{(n)}}V_i\in\mathcal{V}_\eta\left(\overline{W^*}\frac{q_k^*}{p_k}\right)\text{ for all }k\in\{1,\ldots,K\}\right],$$

where $\mathcal{V}_\eta\left(\overline{W^*}\frac{q_k^*}{p_k}\right)$ denotes a neighborhood of $\overline{W^*}\frac{q_k^*}{p_k}$ with radius η being small when δ is small, for large enough n, making use of the a.s. convergence of n_k/n to p_k. Now, for any $k\in\{1,\ldots,K\}$ one has

$$S\left[\frac{1}{n_k}\sum_{i\in I_k^{(n)}}V_i\notin\mathcal{V}_\eta\left(\overline{W^*}\frac{q_k^*}{p_k}\right)\right]\ \leq\ \exp\left(-n_k\cdot\inf_{x\in\mathcal{V}_\eta\left(\overline{W^*}\frac{q_k^*}{p_k}\right)^c}\varphi(x)\right)\tag{293}$$

since any margin of S with index in $I_k^{(n)}$ is a corresponding Kullback-Leibler projection of $\mathfrak{C}$ on the set of all distributions on $\mathbb{R}$ with expectation $\overline{W^*}\cdot\frac{q_k^*}{p_{M,k}^{emp}}$ — where $p_{M,k}^{emp}$ denotes the fraction of the X_i's (within $X_1,\ldots,X_M$) which are equal to d_k (cf. (29)) — and therefore has a moment generating function which is finite in a non-void neighborhood of 0, which yields (293) by the Markov Inequality. Note that the event $\left\{\boldsymbol{\xi}_{M,\mathbf{X}}^{w\mathbf{W}^*}\in int(\mathbf{\Omega})\right\}$ is regenerative, so that M can be chosen large enough to make $p_{M,k}^{emp}$ close to p_k for all $k\in\{1,\ldots,K\}$. This proves the claim. ■

Acknowledgment

W. Stummer is grateful to the Sorbonne Université Paris for its multiple partial financial support and especially the LPSM for its multiple great hospitality. M. Broniatowski thanks very much the FAU Erlangen-Nürnberg for its partial financial support and hospitality. Moreover, W. Stummer would like to thank Rene Schilling for an interesting discussion on complex-valued foundations of the Bernstein-Widder theorem.

REFERENCES

[1] O.O. Aalen, “Modeling the heterogeneity in survival analysis by the compound Poisson distribution,” *Ann. Appl. Probab.*, Vol. 2, no. 4, pp. 951–972, 1992.

[2] J. Abante, Y. Fang, A.P. Feinberg and J. Goutsias, “Detection of haplotype-dependent allele-specific DNA methylation in WGBS data,” *Nature Commun.*, Vol. 11, No. 5238, 2020; doi:10.1038/s41467-020-19077-1.

[3] P. Acera Mateos, R.F. Balboa, S. Easteal, E. Eyras and H.R. Patel, “PACIFIC: a lightweight deep–learning classifier of SARS–CoV–2 and co–infecting RNA viruses,” *Sci. Rep.*, Vol. 11, 3209, 2021; doi:10.1038/s41598-021-82043-4.

[4] A.-M. Acu, G. Bascanbaz-Tunca and I. Rasa, “Information potential for some probability density functions,” *Appl. Math. Comput.*, Vol. 389, 125578, 2021; doi:10.1016/j.amc.2020.125578.

[5] A. Afek, H. Shi, A. Rangadurai, H. Sahay, A. Senitzki, S. Xhani, M. Fang, R. Salinas, Z. Mielko, M.A. Pufall, G.M.K. Poon, T.E. Haran, M.A. Schumacher, H.M. Al-Hashimi and R. Gordan, “DNA mismatches reveal conformational penalties in protein–DNA recognition,” *Nature*, Vol. 587, pp. 291–296, 2020.

[6] C. Aghamohammadi and J.P. Crutchfield, “Thermodynamics of random number generation,” *Phys. Rev. E*, Vol. 95, pp. 062139-1–062139-11, 2017.

[7] R. Ahlswede, “Identification entropy,” in: R. Ahlswede et al. (eds.), *General Theory of Information Transfer and Combinatorics*, Lecture Notes in Computer Science, vol. 4123, pp. 595—613. Berlin, Germany: Springer, 2006.

[8] R. Ahlswede and N. Cai, “An interpretation of identification entropy,” *IEEE Trans. Inf. Theory*, Vol. 52, no. 9, pp. 4198–4207, 2006.

[9] N.I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis*. Edinburgh, UK: Oliver & Boyd, 1965.

[10] R. Alaiz-Rodriguez and A.C. Parnell, “An information theoretic approach to quantify the stability of feature selection and ranking algorithms,” *Knowl.-Based Syst.*, Vol. 195, 105745, 2020; doi:10.106/j.knosys.2020.105745.

[11] M.S. Ali and D. Silvey, “A general class of coefficients of divergence of one distribution from another,” *J. Roy. Statist. Soc. Series B (Methodological)*, Vol. 28, no. 1, pp. 131—140, 1966.

[12] S.-I. Amari, *Differential-Geometrical Methods in Statistics*. Berlin, Germany: Springer, 1985.

[13] J.P. Amezquita-Sanchez, “Entropy algorithms for detecting incipient damage in high-rise buildings subjected to dynamic vibrations,” *J. Vib. Cotrol*, Vol. 27, no. 3–4, pp. 426—436, 2021.

[14] D.L. Applegate, R.E. Bixby, V. Chvatal and W.J. Cook, *The Traveling Salesman Problem*. Princeton, NJ, USA: Princeton University Press, 2006.

[15] E. Arikan, “An inequality on guessing and its application to sequential decoding,” *IEEE Trans. Inf. Theory*, Vol. 42, no. 1, pp. 99–105, 1996.

[16] S. Arimoto, “Information-theoretical considerations on estimation problems,” *Information and Control*, Vol. 19, pp. 181–194, 1971.

[17] C. Arndt, *Information Measures*. Berlin, Germany: Springer, 2001.

[18] M. Arrigoni and G.K.H. Madsen, “Evolutionary computing and machine learning for discovering of low-energy defect configurations,” *npj Comput. Mater.*, Vol. 7, No. 71, 2021; doi:10.1038/s41524-021-00537-1

[19] Ö. Arslan, “Statistical coverage control of mobile sensor networks,” *IEEE Trans. Robot.*, Vol. 35, no. 4, pp. 889–908, 2019.

[20] K.T. Atanassov, “Intuitionistic fuzzy sets,” *Fuzzy Sets Syst.*, Vol. 20, pp. 87–96, 1986.

[21] S. Aviyente, L.A.W. Brakel, R.K. Kushwaha, M. Snodgrass, H. Shevrin, and W.J. Williams, “Characterization of event related potentials using information theoretic distance measures,” *IEEE Trans. Bio-Med. Eng.*, Vol. 51, no. 5, pp. 737–743, 2004.

[22] G. Avlogiaris, A. Micheas and K. Zografos, “On local divergences between two probability measures,” *Metrika*, Vol. 79, pp. 303—333, 2016.

[23] G. Avlogiaris, A. Micheas and K. Zografos, “On testing local hypotheses via local divergence,” *Statist. Methodol.*, Vol. 31, pp. 20—42, 2016.

[24] Z. Avsec, M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, R. Fropf, C. McAnany, J. Gagneur, A. Kundaje and J. Zeitlinger, “Base-resolution models of transcription-factor binding reveal soft motif syntax,” *Nature Genetics*, Vol. 53, pp. 354—366, 2021.

[25] N. Ay, J. Jost, H.V. Le, and L. Schwachhöfer, *Information Geometry*. Cham, Switzerland: Springer International, 2017.

[26] I.B. Ayed, K. Punithakumar and S. Li, “Distribution matching with the Bhattacharyya similarity: a bound optimization framework,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 37, no. 9, pp. 1777–1791, 2015.

[27] A. Baddeley, “A statistical commentary on mineral prospectivity analysis,” in: B.S.D. Sagar, Q. Cheng, F. Agterberg (eds.), *Handbook of Mathematical Geosciences*, pp. 25—65. Cham, Switzerland: Springer International, 2018.

[28] J.P. Bagrow and E.M. Bollt, “An information-theoretic, all-scales approach to comparing networks,” *Appl. Network Sci.*, Vol. 4, 45, 2018; doi:10.1007/s41109-019-0156-x

[29] L. Bai, A. Velichko and B.W. Drinkwater, “Grain scattering noise modeling and its use in the detection and characterization of defects using ultrasonic arrays,” *IEEE Trans. Ultrason. Ferr. Frequ. Control*, Vol. 66, no. 11, pp. 1798–1813, 2019.

[30] S. Bai, T. Lepoint, A. Roux-Langlois, A. Sakzad, D. Stehle and R. Steinfeld, “Improved security proofs in lattice-based cryptography: using the Renyi divergence rather than the statistical distance,” *J. Cryptol.*, Vol. 31, pp. 610–640, 2018.

[31] K.A. Balygin, V.I. Zaitsev, A.N. Klimov, S.P. Kulik and S.N. Molotkov, “A quantum random number generator based on the 100-Mbit/s Poisson photocount statistics,” *J. Exp. Theor. Phys.*, Vol. 126, no. 6, pp. 728–740, 2018.

[32] P. Barennes, V. Quiniou, M. Shugay, E.S. Egorov, A.N. Davydov, D.M. Chudakov, I. Uddin, M. Ismail, T. Oakes, B. Chain, A. Eugster, K. Kashofer, P.P. Rainer, S. Darko, A. Ransier, D.C. Douek, D. Klatzmann and E. Mariotti-Ferrandiz, “Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases,” *Nature Biotechn.*, Vol. 39, pp. 236—245, 2021.

[33] R.J.M. Bashford-Rogers, L. Bergamaschi, E.F. McKinney, D.C. Pombal1, F. Mescia, J.C. Lee, D.C. Thomas, S.M. Flint, P. Kellam, D.R.W. Jayne, P.A. Lyons and K.G.C. Smith, “Analysis of the B cell receptor repertoire in six immune-mediated diseases,” *Nature*, Vol. 574, pp. 122–126, 2019.

[34] M. Basseville, “Divergence measures for statistical data processing —- an annotated bibliography,” *Signal Process.*, Vol. 93, pp. 621—633, 2013.

[35] A. Basu and B.G. Lindsay, “Minimum disparity estimation for continuous models: efficiency, distributions and robustness,” *Ann. Inst. Statist. Math.*, Vol. 46, no. 4, pp. 683–705, 1994.

[36] A. Basu, H. Shioya and C. Park, *Statistical Inference: The Minimum Distance Approach*. Boca Raton, USA: CRC Press, 2011.

[37] S. Bekhet and A. Ahmed, “Evaluation of similarity measures for video retrieval,” *Multimed. Tools Appl.*, Vol. 79, pp. 6265—6278, 2020.

[38] M. Ben-Bassat, “f-entropies, probability of error, and feature selection,” *Information and Control*, Vol. 39, pp. 227–242, 1978.

[39] C.H. Bennett, G. Brassard, C. Crepeau and U.M. Maurer, “Generalized privacy amplification,” *IEEE Trans. Inf. Theory*, Vol. 41, no. 6, pp. 1915–1923, 1995.

[40] A. Ben-Tal and M. Teboulle, “Rate-distortion theory with generalized information measures via convex programming duality,” *IEEE Trans. Inf. Theory*, Vol. 32, no. 5, pp. 630–641, 1986.

[41] D. Berend, P. Harremoes and A. Kontorovich, “Minimum KL-divergence on complements of L_1 balls,” *IEEE Trans. Inf. Theory*, Vol. 60, no. 6, pp. 3172–3177, 2014.

[42] S. Bernstein, “Sur les fonctions absolument monotones,” *Acta Math.*, Vol. 52, pp. 1–66, 1929.

[43] P. Bertrand, M. Broniatowski, and J.-F. Marcotorchino, *Logical indetermination coupling: a method to minimize drawing mathces and its applications*, arXiv:2012.14674v1, 28 pages, 2020.

[44] P. Bertrand, M. Broniatowski, and J.-F. Marcotorchino, *Independence versus indetermination: basis of two canonical clustering criteria*, preprint, 24 pages, 2020.

[45] D. Bertsimas and R. Weismantel, *Optimization over Integers*. Belmont, MA, USA: Dynamic Ideas, 2005.

[46] A. K. Bhandari, “A novel beta differential evolution algorithm-based fast multilevel thresholding for color image segmentation,” *Neural Comput. Appl.*, Vol. 32, pp. 4583—4613, 2020.

[47] D. Bhandari and N.R. Pal, “Some new information measures for fuzzy sets,” *Inf. Sci.*, Vol. 67, pp. 209–228, 1993.

[48] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, Vol. 35, pp. 99–109, 1943.

[49] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Sankhya*, Vol. 7, no. 4, pp. 401–406, 1946.

[50] A. Bhattacharyya, "On some analogues of the amount of information and their use in statistical estimation (contd.)," *Sankhya*, Vol. 8, no. 3, pp. 201–218, 1947.

[51] S. Bi, M. Broggi and M. Beer, "The role of the Bhattacharyya distance in stochastic model updating," *Mech. Syst. Signal Proc.*, Vol. 129, pp. 437–452, 2019.

[52] S. Bi, M. Broggi, P. Wei and M. Beer, "The Bhattacharyya distance: enriching the P–box in stochastic sensitivity analysis," *Mech. Syst. Signal Proc.*, Vol. 129, pp. 265–281, 2019.

[53] G. Biau, B. Cadre, M. Sangnier and U. Tanielian, "Some theoretical properties of GANs," *Ann. Stat.*, Vol. 48, no. 3, pp. 1539—1566, 2020.

[54] I. Bilik and P. Khomchuk, "Minimum divergence approaches for robust classification of ground moving targets," *IEEE Trans. Aero. Elec. Sys.*, Vol. 48, no. 1, pp. 581–603, 2012.

[55] J. Blanchet, F. He and K. Murthy, "On distributionally robust extreme value analysis," *Extremes*, Vol. 23, pp. 317—347, 2020.

[56] A.A. Borovkov, *Probability Theory*. London, UK: Springer, 2013.

[57] P. Boskoski and D. Juricic, "Fault detection of mechanical drives under variable operating conditions based on wavelet packet Renyi entropy signatures," *Mech. Syst. Signal Process.*, Vol. 31, pp. 369–381, 2012.

[58] M. Broniatowski, "A weighted bootstrap procedure for divergence minimization problems," in: J. Antoch et al. (eds.), *Analytical Methods in Statistics*, Springer Proc. Math. Stat. 193, pp. 1–22. Cham, Switzerland: Springer International, 2017.

[59] M. Broniatowski and A. Decurninge, "Estimation for models defined by conditions on their L-moments," *IEEE Trans. Inf. Theory*, Vol. 62, no. 9, pp. 5181–5198, 2016.

[60] M. Broniatowski and A. Keziou, "Minimization of ϕ-divergences on sets of signed measures," *Stud. Scient. Math. Hungar.*, Vol. 43, pp. 403–442, 2006.

[61] M. Broniatowski and A. Keziou, "Parametric estimation and tests through divergences and the duality technique," *J. Multiv. Anal.*, Vol. 100, no. 1, pp. 16–36, 2009.

[62] M. Broniatowski and A. Keziou, "Divergences and duality for estimation and test under moment condition models," *J. Statistical Planning and Inference*, Vol. 142, pp. 2554—2573, 2012.

[63] M. Broniatowski, E. Miranda and W. Stummer, "Testing the number and the nature of the components in a mixture distribution," in: F. Nielsen and F. Barbaresco (eds.), *Geometric Science of Information GSI 2019*, Lecture Notes in Computer Science, vol. 11712, pp. 309—318. Cham, Switzerland: Springer Nature Switzerland, 2019.

[64] M. Broniatowski and W. Stummer, "Some universal insights on divergences for statistics, machine learning and artificial intelligence," in: F. Nielsen (ed.), *Geometric Structures of Information*, Ser. Signals and Communications Technology, pp. 149—211. Cham, Switzerland: Springer Nature Switzerland, 2019.

[65] M. Broniatowski and I. Vajda, "Several applications of divergence criteria in continuous families," *Kybernetika* Vol. 48, no. 4, pp. 600–636, 2012.

[66] C. Brukner and A. Zeilinger, "Operationally invariant information in quantum measurements," *Phys. Rev. Lett.*, Vol. 83, no. 17, pp. 3354–3357, 1999.

[67] C. Bunte and A. Lapidoth, "Encoding tasks and Renyi entropy," *IEEE Trans. Inf. Theory*, Vol. 60, no. 9, pp. 5065–5076, 2014.

[68] C. Burbea and C.R. Rao, "On the convexity of some divergence measures based on entropy functions," *IEEE Trans. Inf. Theory*, Vol. 28, no. 3, pp. 489–495, 1982.

[69] R. Burkard, M. Dell'Amico and S. Martello, *Assignment Problems*, rev. repr. Philadelphia, PA, USA: SIAM, 2009.

[70] H. Cai, Y. Yang, S. Gehly, C. He, and M. Jah, "Sensor tasking for search and catalog maintenance of geosynchronous space objects," *Acta Astronaut.*, Vol. 175, pp. 234-248, 2020.

[71] L. Cai, Y. Chen, N. Cai, W. Cheng and H. Wang, "Utilizing Amari-alpha divergence to stabilize the training of generative adversarial networks," *Entropy*, Vol. 22, No. 410, 2020; doi:10.3390/e22040410.

[72] L.L. Campbell, "Exponential entropy as a measure of extent of a distribution," *Z. Wahrscheinlichkeit verw. Geb./Probab. Theory Rel. Fields*, Vol. 5, pp. 217–225, 1966.

[73] P. Carravilla, J. Chojnacki, E. Rujas, S. Insausti, E. Largo, D. Waithe, B. Apellaniz, T. Sicard, J.-P. Julien, C. Eggeling and J.L. Nieva, "Molecular recognition of the native HIV-1 MPER revealed by STED microscopy of single virions," *Nat. Commun.*, Vol. 10, no. 78, 2019; doi:10.1038/s41467-018-07962-9.

[74] A. Carre, G. Klausner, M. Edjlali, M. Lerousseau, J. Briend–Diop, R. Sun, S. Ammari, S. Reuze, E. Alvarez Andres, T. Estienne, S. Niyoteka, E. Battistella, M. Vakalopoulou, F. Dhermain, N. Paragios, E. Deutsch, C. Oppenheim, J. Pallud and C. Rober, "Standardization of brain MR images across machines and protocols: bridging the gap for MRI–based radiomics," *Sci. Rep.*, Vol. 10, 12340, 2020; doi:10.1038/s41598-020-69298-z.

[75] A. Chakraborty, S. Easwaran and S. Sinha, "Uncovering hierarchical structure of international FOREX market by using similarity metric between fluctuation distributions of currencies," *Acta Phys. Pol. A*, Vol. 138, no. 1, pp. 105–115, 2020.

[76] R.H. Chan, T.H. Chan, H.M. Yeung and R.W. Wang, "Composition vector method based on maximum entropy principle for sequence comparison," *IEEE/ACM Comput. Biol. Bioinformat.*, Vol. 9, no. 1, pp. 79–87, 2012.

[77] S.T. Chandrasekaran, V.E.G. Karnam and A. Sanya, "0.36-mW, 52-Mbps true random number generator based on a stochastic delta–sigma modulator," *IEEE Solid-State Lett.*, Vol. 3, pp. 190–193, 2020.

[78] B. Chen, J. Hu, L. Pu, and Z. Sun, "Stochastic gradient algorithm under $(h, \phi)-$entropy," *Circuits Syst. Signal Process.*, Vol. 26, pp. 941–960, 2007.

[79] B. Chen, X. He, B. Pan, X. Zou and N. You, "Comparison of beta diversity measures in clustering the high-dimensional microbial data," *PLoS ONE*, Vol. 16, no. 2, e0246893, 2021. doi:10.1371/journal.pone.0246893.

[80] C. Chen, C. Zhou, P. Liu and D. Zhang, "Iterative reweighted Tikhonov-regularized multihypothesis prediction scheme for distributed compressive video sensing," *IEEE Trans. Circ. Syst. Video Techn.*, Vol. 30, no. 1, pp. 1–10, 2020.

[81] D. Chen, D.-D. Shi, M. Qin, S.-M. Xu and G.-J. Pan, "Complex network comparison based on communicability sequence entropy," *Phys. Rev. E*, Vol. 98, No. 012319, 2018; doi:10.1103/PhysRevE.98.012319.

[82] D. Chen, J. Spencer, J.-M. Mirebeau, K. Chen, M. Shu and L.D. Cohen, "A generalized asymmetric dual-front model for active contours and image segmentation," *IEEE Trans. Image Process.*, Vol. 30, pp. 5056–5071, 2021.

[83] D.-S. Chen, R.G. Batson and Y. Dang, *Applied Integer Programming*. Hoboken, NJ, USA: Wiley, 2010.

[84] J. Chen, J.E. Markowitz, V. Lilascharoen, S. Taylor, P. Sheurpukdi, J.A. Keller, J.R. Jensen, B.K. Lim, S.R. Datta and L. Stowers, "Flexible scaling and persistence of social vocal communication," *Nature*, Vol. 593, pp. 108–113, 2021.

[85] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statistics*, Vol. 23 pp. 493–507, 1952.

[86] P.-H. Chiu, P.-H. Tseng and K.-T. Feng, "Interactive mobile augmented reality system for image and hand motion racking," *IEEE Trans. Veh. Technol.*, Vol. 67, no. 10, pp. 9995–10009, 2018.

[87] J. Chong, P. Liu, G. Zhou and J. Xia, "Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data," *Nature Protocols*, Vol. 15, pp. 799—821, 2020.

[88] Y.S. Chow and H. Teicher, *Probability Theory*, 3rd ed. New York, NY, USA: Springer, 1997.

[89] O. Ciftci, M. Mehrtash, and A. Kargarian, "Data-driven nonparametric chance-constrained optimization for microgrid energy management," *IEEE Trans. Ind. Inform.*, Vol. 16, no. 4, pp. 2447–2457, 2020.

[90] D.E. Clark, "Local entropy statistics for point processes," *IEEE Trans. Inf. Theory*, Vol. 66, no. 2, pp. 1155–1163, 2020.

[91] A. Cohen, D. Malak, V.B. Bracha and M. Medard, "Adaptive Causal Network Coding With Feedback," *IEEE Trans. Commun.*, Vol. 68, no. 7, pp. 4325–4341, 2020.

[92] W.J. Cook, *In Pursuit of the Traveling Salesman*. Princeton, NJ, USA: Princeton University Press, 2012.

[93] N. Cressie and T.R.C. Read, "Multinomial goodness-of-fit tests," *J. R. Statist. Soc. B*, Vol. 46, no. 3, pp. 440–464, 1984.

[94] I. Csiszár, "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci.*, Vol. 8, pp. 85—108, 1963.

[95] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathem. Hungar.*, Vol. 2, no. 1-4, pp. 191–213, 1972.

[96] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, Vol. 3, no. 1, pp. 146–158, 1975.

[97] I. Csiszár. "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, Vol. 12, no. 3, pp. 768–793, 1984.

[98] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Trans. Inf. Theory*, Vol. 49, no. 6, pp. 1474–1490, 2003.

[99] I. Csiszár and P.C. Shields, *Information Theory and Statistics: a Tutorial*. Hanover, MA, USA: now Publishers, 2004.

[100] I. Csiszár and T. Breuer, "Expected value minimization in information theoretic multiple priors models," *IEEE Trans. Inf. Theory*, Vol. 64, no. 6, pp. 3957–3974, 2018.

[101] L. Cui, J. Wu, D. Pi, P. Zhang and P. Kennedy, "Dual implicit mining-based latent friend recommendation," *IEEE Trans. Syst. Man Cyb.: Syst.*, Vol. 50, no. 5, pp. 1663–1678, 2020.

[102] D. Dacunha-Castelle and F. Gamboa, "Maximum d'entropie et probleme des moments," *Ann. Inst. Henri Poincare*, Vol. 26, no. 4, pp. 567–596, 1990.

[103] B. Dang, J. Sun, T. Zhang, S. Wang, M. Zhao, K. Liu, L. Xu, J. Zhu , C. Cheng, L. Bao, Y. Yang, H. Wang, Y. Hao and R. Huang, "Physically transient true random number generators based on paired threshold switches enabling Monte Carlo method applications," *IEEE Electron. Device Lett.*, Vol. 40, no. 7, pp. 1096–1099, 2019.

[104] G. Dantzig, R. Fulkerson, and S. Johnson, "Solution of a large-scale traveling-salesman problem," *Oper. Res.*, Vol. 2, no. 4, pp. 393–410, 1954.

[105] E.C.L. De Oliveira, K. Santana, L. Josino, A.H. Lima e Lima and C. de Souza de Sales Junior, "Predicting cell-penetrating peptides using machine learning algorithms and navigating in their chemical space," *Sci. Rep.*, Vol. 11, 7628, 2021; doi:10.1038/s41598-021-87134-w.

[106] A. De Santis, A.G. Gaggia, and U. Vaccaro, "Bounds on entropy in a guessing game," *IEEE Trans. Inf. Theory*, Vol. 47, no. 1, pp. 468–473, 2001.

[107] T. De Wet, F. Österreicher and M. Thaler, "Tsallis'entropies – axiomatics, associated $f-$divergences and Fisher's information," *South Afr. Stat. J.*, Vol. 54, no. 2, pp. 163–175, 2020.

[108] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed (corr. print.). New York, USA: Springer, 2009.

[109] A.P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Stat.*, Vol. 38, no. 2, pp. 325–339, 1967.

[110] Y. Deng, "Deng entropy," *Chaos Solitons Fract.*, Vol. 91, pp. 549–553, 2016.

[111] L. Devroye, "Random variate generation for exponentially and polynomially tilted stable distributions," *ACM Trans. Model. Comput. Simul.*, Vol. 19, no. 4, Article 18, 20 pages, 2009. doi:10.1145/1596519.1596523.

[112] L. Devroye and L. James, "On simulation and properties of the stable law," *Stat. Methods Appl.*, Vol. 23, no. 3, 307–343, 2014. doi:10.1007/s10260-014-0260-0.

[113] M.M. Deza and E. Deza, *Encyclopedia of Distances*, 4th ed. Berlin, Germany: Springer, 2016.

[114] A.B. Dharmawan, S. Mariana, G. Scholz, P. Hörmann, T. Schulze, K. Triyana, Mayra Garces-Schröder, I. Rustenbeck, K. Hiller, H.S. Wasisto and A. Waag, "Nonmechanical parfocal and autofocus features based on wave propagation distribution in lensfree holographic microscopy," *Sci. Rep.*, Vol. 11, 3213, 2021; doi:10.1038/s41598-021-81098-7.

[115] V.S. Dixit and P. Jain, "Proposed similarity measure using Bhattacharyya coefficient for context aware recommender system," *J. Intell. Fuzzy Syst.*, Vol. 36, pp. 3105—3117, 2019.

[116] S.S. Dragomir, "Upper bounds for the Kullback-Leibler distance and applications," *Bull. Math. Soc. Sc. Math. Roumanie*, Vol. 43(91), no. 1, pp. 25–37, 2000.

[117] D. Drahi, N. Walk, M.J. Hoban, A.K. Fedorov, R. Shakhovoy, A. Feimov, Y. Kurochkin, W.S. Kolthammer, J. Nunn, J. Barrett and I.A. Walmsley, "Certified quantum random numbers from untrusted light," *Phys. Rev. X*, Vol. 10, pp. 041048-1 – 041048-32, 2020.

[118] G.A. Einicke, H.A. Sabti, D.V. Thiel and M. Fernandez, "Maximum-entropy-rate selection of features for classifying changes in knee and ankle dynamics during running," *IEEE J. Biomed. Health Inform.*, Vol. 22, no. 4, pp. 1097–1103, 2018.

[119] Y. El Merabet, Y. Ruichek, S. Ghaffarian, Z. Samir, T. Boujiha, R. Messoussi, R. Touahni and A. Sbihi, "Maximal similarity based region classification method through local image region descriptors and Bhattacharyya coefficient-based distance: application to horizon line detection using wide–angle camera," *Neurocomputing*, Vol. 265, pp. 28–41, 2017.

[120] D.M. Endres and J.E. Schindelin, "A new metric for probability distributions," *IEEE Trans. Inf. Theory*, Vol. 40, no. 7, pp. 1858–1860, 1993.

[121] T.T. Erguzel, C. Uyulan, B. Unsalver, A. Evrensel, M. Cebi, C.O. Noyan, B. Metin, G. Eryilmaz, G.H. Sayar and N. Tarhan, "Entropy: a promising EEG biomarker dichotomizing subjects with opioid use disorder and healthy controls," *Clin. EEG Neurosci.*, Vol. 51, no. 6, pp. 373–381, 2020.

[122] A. Eshaghi, A.L. Young, P.A. Wijeratne, F. Prados, D.L. Arnold, S. Narayanan5, C.R.G. Guttmann, F. Barkhof, D.C. Alexander, A.J. Thompson, D. Chard and O. Ciccarelli, "Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data," *Nature Commun.*, Vol. 12, No. 2078, 2021; doi:10.1038/s41467-021-22265-2.

[123] S. Fan, Y. Sun and P. Shui, "Region-merging method with texture pattern attention for SAR image segmentation," *IEEE Trans. Geosci. Remote Sens. Lett.*, Vol. 18, no. 1, pp. 112–116, 2021.

[124] M. Feixas, A. Bardera, J. Rigau, Q. Xu and M. Sbert, *Information Theory Tolls for Image Processing*. San Rafael, CA, USA: Morgan & Claypool, 2014.

[125] C. Feng, H. Zhao, Y. Li, Z. Cheng and J. Wena, "Improved detection of focal cortical dysplasia in normal-appearing FLAIR images using a Bayesian classifier," *Med. Phys.*, Vol. 48, no. 2, pp. 912–925, 2021.

[126] I. Fischer and D.J. Gauthier, "High-speed harvesting of random numbers," *Science*, Vol. 371, 26 February 2021, pp. 889–890, 2021.

[127] J.L. Folks, "Inverse distributions," in: S. Kotz et al. (eds.), *Encyclopedia of Statistical Sciences*, 2nd ed., vol. 6, pp. 3673—3675. Hoboken, USA: John Wiley & Sons Inc., 2006.

[128] Q. Fu, S.B. Villas-Boas and G. Judge, "Entropy-based China income distributions and inequality measures," *China Econ. J.*, Vol. 12, no. 3, pp. 352—368, 2019.

[129] T. Fu, C. Xiao, L.M. Glass and J. Sun, "$\alpha-$MOP: molecule optimization with $\alpha-$divergence," in: *Proc. 2020 IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM)*, pp. 240–244, 2020; doi:10.1109/BIBM49941.2020.9313409.

[130] Y. Fu and Z. He, "Bhattacharyya distance criterion based multibit quantizer design for cooperative spectrum sensing in cognitive radio networks," *Wirel. Netw.*, Vol. 25, pp. 2665—2674, 2019.

[131] B. Gabriel, C. Medin, J. Alves, R. Nduati, R.K. Bosire, D. Wamalwa, C. Farquhar, G. John-Stewart and B.L. Lohman-Payne, "Analysis of the TCR repertoire in HIV-exposed but uninfected infants," *Sci. Rep.*, Vol. 9, 11954, 2019; doi:10.1038/s41598-019-48434-4.

[132] F. Gamboa and E. Gassiat, "Asymptotic distribution of $(h, \phi)-$entropies," *Ann. Stat.*, Vol. 25, no. 1, pp. 328–350, 1997.

[133] Z. German–Sallo, "Entropy indices based fault detection," *Proced. Manufact.*, Vol. 46, pp. 549—554, 2020

[134] R. Gholami and G.A. Hodtani, "A more general information theoretic study of wireless location verification system," *IEEE Trans. Veh. Technol.*, Vol. 69, no. 9, pp. 9938–9950, 2020.

[135] A. Ghosh and A. Basu, "A new family of divergences originating from model adequacy tests and applications to robust statistical inference," *IEEE Trans. Inf. Theory*, Vol. 64, no. 8, pp. 5581–5591, 2018.

[136] A. Ghosh and A. Basu, “A scale-invariant generalization of the Renyi entropy, associated divergences and their optimizations under Tsallis’ nonextensive framework,” *IEEE Trans. Inf. Theory*, Vol. 67, no. 4, pp. 2141–2161, 2021.

[137] C. Gietl and and F.P. Reffel, “Continuity of $f-$projections and applications to the iterative proportional fitting procedure,” *Statistics*, Vol. 51, no. 3, pp. 668–684, 2017.

[138] L. Gong, J. Zhang, H. Liu, L. Sang and Y. Wang, “True random number generators using electrical noise,” *IEEE Access*, Vol. 7, pp. 125796–125805, 2019.

[139] J. Görtler, T. Spinner, D. Streeb, D. Weiskopf and O. Deussen, “Uncertainty-aware principal component analysis,” *IEEE Trans. Visual. Comput. Graph.*, Vol. 26, no. 1, pp. 822–833, 2020.

[140] S.W. Golomb, “The information generating function of a probability distribution,” *IEEE Trans. Inf. Theory*, Vol. 12, no. 1, pp. 75–77, 1966.

[141] M. Greenacre, “Weighted metric multidimensional scaling,” in: M. Vichi, P. Monari, S. Mignani, and A. Montanari (eds.), *New Developments in Classification and Data Analysis*, pp. 141—149. Berlin, Germany: Springer, 2005.

[142] V. Greiff, P. Bhat, S.C. Cook, U. Menzel, W. Kang and S.T. Reddy, “A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status,” *Genome Med.*, Vol. 7, 49, 2015; doi:10.1186/s13073-015-0169-8.

[143] W. Guan, Z. Liu, S. Wen, H. Xie and X. Zhang, “Visible light dynamic positioning method using improved camshift-Kalman algorithm,” *IEEE Photonics J.*, Vol. 11, no. 6, 7906922, 2019; doi:10.1109/jphot.2019.2944080.

[144] S. Guiasu and C. Reischer, “The relative information generating function,” *Inf. Sci.*, Vol. 35, no. 3, pp. 235—241, 1985.

[145] S. Guo, W. Huang and Y. Qiao, “Improving scale invariant feature transform with local color contrastive descriptor for image classification,” *J. Electron. Imaging*, Vol. 26, No. 1, 013015, 2017; doi:10.1117/1.JEI.26.1.013015.

[146] X. Guo and Y. Yuan, “Semi-supervised WCE image classification with adaptive aggregated attention,” *Med. Image Anal.*, Vol. 64, 101733, 2020. doi:10.1016/j.media.2020.101733.

[147] G. Gutin and A.P. Punnen (eds.), *The Traveling Salesman Problem and Its Variations*. New York, NY, USA: Springer, 2007.

[148] L. Györfy and I. Vajda, “A class of modified Pearson and Neyman statistics,” *Statistics & Decisions*, Vol. 19, pp. 239–251, 2001.

[149] A. György and T. Linder, “Optimal entropy-constrained scalar quantization of a uniform source,” *IEEE Trans. Inf. Theory*, Vol. 46, no. 7, pp. 2704–2711, 2000.

[150] H. Gzyl, S. Mayoral and E. Gomes-Goncalves, *Loss Data Analysis. The Maximum Entropy Approach*. Berlin, Germany: de Gruyter, 2018.

[151] W. Ha, E.Y. Sidky, R.F. Barber, T.G. Schmidt and X. Pan, “Estimating the spectrum in computed tomography via Kullback–Leibler divergence constrained optimization,” *Med. Phys.*, Vol. 46, no. 1, pp. 81–92, 2019.

[152] H. Han, Y. Wang, Y. Zou, J. Liao and Y. Xu, “Three-dimensional substructure imaging of blood cells using maximum entropy tomography based on two non-orthogonal phase images,” *Opt. Laser Technol.*, Vol. 136, 106799, 2021; doi:10.1016/j.optlastec.2020.106799.

[153] W. Hanli, C. Hao, W. Lei, H. Jingguang and L. Zhenxing, “A novel pilot protection scheme for transmission lines based on current distribution histograms and their Bhattacharyya coefficient,” *Electr. Pow. Syst. Res.*, Vol. 194, 107056, 2021; doi:10.1016/j.epsr.2021.107056.

[154] L. Hannah and J.A. Kay, *Concentration in Modern Industry*. London, UK: The Macmillan Press, 1977.

[155] P. Harremoes and F. Topsoe, “Inequalities between entropy and index of coincidence derived from information diagrams,” *IEEE Trans. Inf. Theory*, Vol. 47, no. 7, pp. 2944–2960, 2001.

[156] P. Harremoes and I. Vajda, “On the Bahadur-efficient testing of uniformity by means of the entropy,” *IEEE Trans. Inf. Theory*, Vol. 54, no. 1, pp. 321–331, 2008.

[157] J. Havrda and F. Charvat, “Quantification method of classification process,” *Kybernetika*, Vol. 3, pp. 30–34, 1967.

[158] Y. He, C. Song, C. Jing and X. Lei, “Robot active olfaction search in turbulent flow and infotaxis search based on Renyi divergence,” in: *Proc. OCEANS 2017 - Aberdeen*, 2017, 9 pages. doi: 10.1109/OCEANSE.2017.8084882.

[159] M. Herrero-Collantes and J.C. Garcia-Escartin, “Quantum random number generators,” *Rev. Mod. Phys.*, Vol. 89, no. 1, pp. 015004-1 – 015004-48, 2017.

[160] M.O. Hill, “Diversity and evenness: a unifying notation and its consequences,” *Ecology*, Vol. 54, no. 2, pp. 427–431, 1973.

[161] A. Iqbal and A.-K. Seghouane, “An $\alpha-$divergence-based approach for robust dictionary learning,” *IEEE Trans. Image Process.*, Vol. 28, no. 11, pp. 5729–5739, 2019.

[162] A. Jain, S. Nagar, P.K. Singh and J. Dhar, “EMUCF: enhanced multistage user-based collaborative filtering through non–linear similarity for recommendation systems,” *Expert Syst. Appl.*, Vol. 161, 113724, 2020; doi:10.1016/j.eswa.2020.113724.

[163] J. Jaksetic and J. Pecaric, “Exponential convexity method,” *J. Convex Anal.*, Vol. 20, no. 1, pp. 181–197, 2013.

[164] W. Jasinska, M. Manhart, J. Lerner, L. Gauthier, A.W.R. Serohijos and S. Bershtein, “Chromosomal barcoding of E. coli populations reveals lineage diversity dynamics at high resolution,” *Nature Ecol. & Evol.*, Vol. 4, pp. 437—452, 2020.

[165] E.T. Jaynes, “Information theory and statistical mechanics I,” *Phys. Rev.*, Vol. 106, pp. 620–630, 1957.

[166] E.T. Jaynes, “Information theory and statistical mechanics II,” *Phys. Rev.*, Vol. 108, pp. 171–190, 1957.

[167] B. Jena, M.K. Naik, R. Panda and A. Abraham, “Maximum 3D Tsallis entropy based multilevel thresholding of brain MR image using attacking Manta Ray foraging optimization,” *Eng. Appl. Artif. Intell.*, Vol. 103, 104293, 2021; doi:10.1016/j.engappai.2021.104293.

[168] G. Jenkinson, E. Pujadas, J. Goutsias and A.P. Feinberg, “Potential energy landscapes identify the information-theoretic nature of the epigenome,” *Nature Genetics*, Vol. 49, no. 5, pp. 719–729, 2017.

[169] J. Jiang, M. Chen and J.A. Fan, “Deep neural networks for the evaluation and design of photonic devices,” *Nat. Rev. Materials*, 2020; doi:10.1038/s41578-020-00260-1.

[170] R. Jiang, Q. Wang, S. Shi, X. Mou and S. Chen, “Flow-assisted visual tracking using event camera,” *CAAI Trans. Intell. Technol.*, Vol. 6, pp. 192—202, 2021.

[171] J. Jiao, J. Yue, D. Pei and Z. Hu, “Application of feature fusion using coaxial vibration signal for diagnosis of rolling element bearings,” *Shock Vibrat.*, Vol. 2020, 8831723, 2020; doi:10.1155/2020/8831723.

[172] T. Joel and R. Sivakumar, “Nonsubsampled contourlet transform with cross-guided bilateral filter for despeckling of medical ultrasound images,” *Int. J. Imaging Syst. Technol.*, Vol. 31, pp. 763—777, 2021.

[173] M. Johansson and M. Sternad, “Resource allocation under uncertainty using the maximum entropy principle,” *IEEE Trans. Inf. Theory*, Vol. 51, no. 12, pp. 4103–4117, 2005.

[174] S. Jolad, A. Roman, M.C. Shastry, M. Gadgil, and A. Basu, “A new family of bounded divergence measures and application to signal detection,” in: M. De Marsico, G. Sanniti di Baja, and A. Fred (eds.), *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pp. 72–83. Setubal, Portugal: Scitepress, 2016.

[175] K. Joshi, M.R. de Massy, M. Ismail, J.L. Reading, I. Uddin, A. Woolston, E. Hatipoglu, T. Oakes, R. Rosenthal, T. Peacock, T. Ronel, M. Noursadeghi, V. Turati, A.J.S. Furness, A. Georgiou, Y.N.S. Wong, A.B. Aissa, M.W. Sunderland, M. Jamal-Hanjani, S. Veeriah, N.J. Birkbak, G.A. Wilson, C.T. Hiley, E. Ghorani, J.A. Guerra-Assuncao, J. Herrero, T. Enver, S.R. Hadrup, A. Hackshaw, K.S. Peggs, N. McGranahan, C. Swanton, TRACERx consortium, S.A. Quezada and B. Chain, “Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer,” *Nature Med.*, Vol. 25, 1549–1559, 2019.

[176] L. Jost, “Entropy and diversity,” *OIKOS*, Vol. 113, no. 2, pp. 363–375, 2006.

[177] F. Juhasz, “The Hellinger Distance as Used for the Representation of Serological ABO Distances Among Earlier Human Populations,” *Hum. Genet.*, Vol. 63, pp. 228–231, 1983.

[178] T. Kailath, “The divergence and Bhattacharyya distance measures in signal selection,” *IEEE Trans. Commun. Techn.*, Vol. 15, no. 1, pp. 52–60, 1967.

[179] N.B. Kammerer and W. Stummer, “Some dissimilarity measures of branching processes and optimal decision making in the presence of potential pandemics,” *Entropy*, Vol. 22, No. 874, 2020; doi:10.3390/e22080874.

[180] D.T. Kanapram, F. Patrone, P. Marin-Plaza, M. Marchese, E.L. Bodanese, L. Marcenaro, D.M. Gomez and C. Regazzoni, “Collective awareness for abnormality detection in connected autonomous vehicles,” *IEEE Internet Things J.*, Vol. 7, no. 5, pp. 3774–3789, 2020.

[181] B. Kang and Y. Deng, “The maximum Deng entropy,” *IEEE Access*, Vol. 7, pp. 120758–120765, 2020.

[182] M.-S. Kang and K.-T. Kim, “Automatic SAR image registration via Tsallis entropy and iterative search process,” *IEEE Sensors J.*, Vol. 20, no. 14, pp. 7711–7720, 2020.

[183] J.N. Kapur, “Four families of measures of entropy,” *Indian J. Pure Appl. Math.*, Vol. 17, pp. 429–449, 1986.

[184] J.N. Kapur, *Maximum-entropy models in science and engineering*. New York, NY, USA: John Wiley & Sons, 1989.

[185] J.N. Kapur and H.K. Kesavan, *Entropy Optimization Principles With Applications*. San Diego, CA, USA: Academic Press, 1992.

[186] Ö. Kartal, M.W. Schmid and U. Grossniklaus, “Cell type-specific genome scans of DNA methylation divergence indicate an important role for transposable elements,” *Genome Biol.*, Vol. 21, No. 172, 2020; doi:10.1186/s13059-020-02068-2.

[187] H.K. Kesavan and J.N. Kapur, “The generalized maximum entropy principle,” *IEEE Trans. Syst. Man Cyb.*, Vol. 19, no. 5, pp. 1042–1052, 1989.

[188] K. Kim, S. Bittner, Y. Zeng, S. Guazzotti, O. Hess, Q.J. Wang and Hui Cao, “Massively parallel ultrafast random bit generation with a chip-scale laser,” *Science*, Vol. 371, 26 February 2021, pp. 948–952, 2021.

[189] A.-L. Kißlinger and W. Stummer, “New model search for nonlinear recursive models, regressions and autoregressions,” in: F. Nielsen and F. Barbaresco (eds.), *Geometric Science of Information GSI 2015*, Lecture Notes in Computer Science, vol. 9389, pp. 693—701. Berlin, Germany: Springer, 2015.

[190] A.-L. Kißlinger and W. Stummer, “Robust statistical engineering by means of scaled Bregman distances,” in: C. Agostinelli, A. Basu, P. Filzmoser and D. Mukherjee (eds.), *Recent Advances in Robust Statistics – Theory and Applications*, pp. 81–113. New Delhi, India: Springer, 2016.

[191] A.-L. Kißlinger and W. Stummer, “A new toolkit for robust distributional change detection,” *Appl. Stochastic Models Bus. Ind.*, Vol. 34, pp. 682–699, 2018.

[192] Klar, B. (2015), “A note on gamma difference distributions,” *Journal of Statistical Computation and Simulation*, vol. 85, no. 18, pp. 3708-3715.

[193] M.A. Koldobskiy, G. Jenkinson, J. Abante, V.A. Rodriguez DiBlasi, W. Zhou, E. Pujada, A. Idrizi, R. Tryggvadottir, C. Callahan, C.L. Bonifant, K.R. Rabin, P.A. Brown, H. Ji, J. Goutsias and A.P. Feinberg, “Converging genetic and epigenetic drivers of paediatric acute lymphoblastic leukaemia identified by an information-theoretic analysis,” *Nature Biomed. Eng.*, Vol. 5, no. 4, pp. 360–376, 2021.

[194] C. Kollmitzer, S. Schauer, S. Rass and B. Rainer (eds.), *Quantum Random Number Generation*. Cham, Switzerland: Springer Nature, 2020.

[195] B. Korte and J. Vygen, *Combinatorial Optimization*, 6th ed. Berlin, Germany: Springer, 2018.

[196] N.O. Kotelina and A.B. Pevny, “Exponential convexity and total positivity,” *Siberian Electr. Math. Rep.*, Vol. 17, pp. 802-806, 2020; doi:10.33048/semi.2020.17.057.

[197] S. Kotz, T.J. Kozubowski and K. Podgorski, *The Laplace Distribution and Generalizations*. Boston, MA, USA: Birkhäuser, 2001.

[198] S. Krömer and W. Stummer, “A new toolkit for mortality data analytics,” in: A. Steland, E. Rafajlowicz, and O. Okhrin (eds.), *Stochastic Models, Statistics and Their Applications*, pp. 393–407. Cham, Switzerland: Springer Nature Switzerland, 2019.

[199] C.-J. Ku and T.L. Fine, “Testing for stochastic independence: application to blind source separation,” *IEEE Trans. Signal Proces.*, Vol. 53, no. 5, pp. 1815–1826, 2005.

[200] S. Kullback and R.A. Leibler, *On information and sufficiency*. “Asymptotic distribution of $(h, \phi)-$entropies,” *Ann. Math. Statistics*, Vol. 22, pp. 79–86, 1951.

[201] M. Kumar, A.K. Bhandari, N. Singh and A. Ghosh, “A new multilevel histogram thresholding approach using variational mode decomposition,” *Multimed. Tools and Appl.*, Vol. 80, pp. 11331—11363, 2021.

[202] M.A. Kumar and I. Sason, “Projection theorems for the Renyi divergence on $\alpha-$convex sets,” *IEEE Trans. Inf. Theory*, Vol. 62, no. 9, pp. 4924–4935, 2016.

[203] M.A. Kumar and R. Sundaresan, “Minimization problems based on relative α-entropy I: forward projection,” *IEEE Trans. Inf. Theory*, Vol. 61, no. 9, pp. 5063–5080, 2015.

[204] M.A. Kumar and R. Sundaresan, “Minimization problems based on relative α-entropy II: reverse projection,” *IEEE Trans. Inf. Theory*, Vol. 61, no. 9, pp. 5081–5095, 2015.

[205] M. A. Kumar, A. Sunny, A. Thakre, and A. Kumar, *A unified framework for problems on guessing, source coding and task partitioning*, arXiv:1907.06889v1, 22 pages, 2019.

[206] M. Kumbhakar, “Streamwise velocity profile in open-channel flow based on Tsallis relative entropy,” *Chaos*, Vol. 30, 073136, 2020; doi:10.1063/1.5144867.

[207] M. Kumbhakar, R.K. Ray, S.K. Chakraborty, K. Ghoshal and V.P. Singh, “Mathematical modelling of streamwise velocity profile in open channels using Tsallis entropy,” *Commun. Nonlinear Sc. Numer. Simulat.*, Vol. 94, 105581, 2021; doi:10.1016/j.cnsns.105581.

[208] D. Kvitsiani, S. Ranade, B. Hangya, H. Taniguchi, J.Z. Huang and A. Kepecs, “Distinct behavioural and network correlates of two interneuron types in prefrontal cortex,” *Nature*, Vol. 498, pp. 363–366, 2013.

[209] N. Lassance and F. Vrins, “Minimum Renyi entropy portfolios,” *Ann. Oper. Res.*, Vol. 299, pp. 23—46, 2021.

[210] T. Laszlovszky, D. Schlingloff, P. Hegedüs, T.F. Freund, A. Gulyas, A. Kepecs and B. Hangya, “Distinct synchronization, cortical coupling and behavioral function of two basal forebrain cholinergic neuron types,” *Nature Neurosci.*, Vol. 23, pp. 992—1003, 2020.

[211] K.A. Lawson et al., “Functional genomic landscape of cancer-intrinsic evasion of killing by T cells,” *Nature*, Vol. 586, pp. 120—126, 2020.

[212] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. New York, NY, USA: Springer, 1986.

[213] C. Léonard, “Minimization of entropy functionals,” *J. Math. Anal. Appl.*, Vol. 346, pp. 183–204, 2008.

[214] G. Letac and M. Mora, “Natural real exponential families with cubic variance functions,” *Ann. Stat.*, Vol. 18, no. 1, pp. 1–37, 1990.

[215] C.H. Li, E.L. Coffey, A. Dall’Agnese, N.M. Hannett, X. Tang, J.E. Henninger, J.M. Platt, O. Oksuz, A.V. Zamudio, L.K. Afeyan, J. Schuijers, X.S. Liu, S. Markoulaki, T. Lungjangwa, G. LeRoy, D.S. Svoboda, E. Wogram, T.I. Lee, R. Jaenisch and R.A. Young, “MeCP2 links heterochromatin condensates and neurodevelopmental disease,” *Nature*, Vol. 586, pp. 440—444, 2020.

[216] F. Liese and K.J. Miescke, *Statistical Decision Theory: Estimation, Testing, and Selection*. New York, USA: Springer, 2008.

[217] F. Liese and I. Vajda, *Convex Statistical Distances*. Leipzig, Germany: Teubner, 1987.

[218] F. Liese and I. Vajda, “On divergences and informations in statistics and information theory,” *IEEE Trans. Inf. Theory*, Vol. 52, no. 10, pp. 4394—4412, 2006.

[219] J. Lin, “Divergence measures based on the Shannon entropy,” *IEEE Trans. Inf. Theory*, Vol. 37, no. 1, pp. 145–151, 1991.

[220] Z. Lin, H. Qin and S. C. Chan, “A new probabilistic representation of color image pixels and its applications,” *IEEE Trans. Image Process.*, Vol. 28, no. 4, pp. 2037–2050, 2019.

[221] B.G. Lindsay, “Efficiency versus robustness: the case for minimum Hellinger distance and related methods,” *Ann. Statist.*, Vol. 22, no. 2, pp. 1081–1114, 1994.

[222] B.G. Lindsay, “Statistical distances as loss functions in assessing model adequacy,” in: M.P. Taper and S.R. Lele (eds.), *The Nature of Scientific Evidence*, pp. 439—487. Chicago, IL, USA: The University of Chicago Press, 2004. This includes comments by D.R. Cox and S.P. Ellner as well as a rejoinder by the author.

[223] B.G. Lindsay, M. Markatou, S. Ray, K. Yang, and S.-C. Chen, “Quadratic distances on probabilities: a unified foundation,” *Ann. Statist.*, Vol. 36, no. 2, pp. 983–1006, 2008.

[224] J. Li, B. Xie, Y. Jin, Z. Hu, and L. Zhou, "Weighted Conflict Evidence Combination Method Based on Hellinger Distance and the Belief Entropy," *IEEE Access*, Vol. 8, pp. 225507–225521, 2020.

[225] J. Liu and B.G. Lindsay, "Building and using semiparametric tolerance regions for parametric multinomial models," *Ann. Statist.*, Vol. 37, no. 6A, pp. 3644–3659, 2009.

[226] S.-B. Liu, C.-Y. Rong, Z.-M. Wu and T. Lu, "Renyi entropy, Tsallis entropy and Onicescu information energy in density functional reactivity theory," *Acta Phys.-Chim. Sin.*, Vol. 31, no. 11, pp. 2057–2063, 2015.

[227] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to Detect a Salient Object," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 33, no. 2, pp. 353–367, 2011.

[228] Y. Liu, C. Chen, D.D. Yang, Q. Li and X. Li, "Fast true random number generator based on chaotic oscillation in self-feedback weakly coupled superlattices," *IEEE Access*, Vol. 8, pp. 182693–182703, 2020.

[229] X. Liu and S. Sun, "Alpha-divergence minimization with mixed variational posterior for Bayesian neural networks and its robustness against adversarial examples," *Neurocomputing*, Vol. 423, pp. 427–434, 2021.

[230] V. Loia and A. Vaccaro, "Decentralized economic dispatch in smart grids by self-organizing dynamic agents," *IEEE Trans. Syst. Man Cybern.*, Vol. 44, no. 4, pp. 397–408, 2014.

[231] S. Lopez-Rosa, A.L. Martin, J. Antolin and J.C. Angulo, "Electron-pair entropic and complexity measures in atomic systems," *Int. J. Quantum Chem.*, Vol. 119, No. e25861, 2019; doi:10.1002/qua.25861.

[232] X. Luo, Q. Xu, M. Sbert and Klaus Schoeffmann, "F-divergences driven video key frame extraction," in: *Proc. 2014 IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2014, 6 pages.

[233] L.T. Luppino, F.M. Bianchi, G. Moser and S.N. Anfinsen, "Unsupervised image regression for heterogeneous change detection," *IEEE Trans. Geosci. Remote Sens.*, Vol. 57, no. 12, pp. 9960–9975, 2019.

[234] E. Lutwak, D. Yang, and G. Zhang, "Cramer-Rao and moment-entropy inequalities for Renyi entropy and generalized Fisher information," *IEEE Trans. Inf. Theory*, Vol. 51, no. 2, pp. 473–478, 2005.

[235] A. Lysiak and M. Szmajda, "Empirical comparison of the feature evaluation methods based on statistical measures," *IEEE Access*, Vol. 9, pp. 27868–27883, 2021.

[236] A. Lyubushin, "Seismic noise wavelet-based entropy in Southern California," *J. Seismol.*, Vol. 25, pp. 25–39, 2021.

[237] Z. Ma and L. Li, "Measuring metagenome diversity and similarity with Hill numbers," *Mol. Ecol. Resour.*, Vol. 18, pp. 1339–1355, 2018.

[238] Z. Ma, L. Li and Y.-P. Zhang, "Defining individual-level genetic diversity and similarity profiles," *Sci. Rep.*, Vol. 10, 5805, 2020; 10.1038/s41598-020-62362-8.

[239] J.A. T. Machado, J.M. Rocha-Neves and J.P. Andrade, "Computational analysis of the SARS-CoV-2 and other viruses based on the Kolmogorov's complexity and Shannon's information theories," *Nonlinear. Dyn.*, Vol. 101, pp. 1731—1750, 2020.

[240] D.B. Madan and E. Seneta, "The variance gamma (V.G.) model for share market returns," *J. Business*, Vol. 63, no. 4, pp. 511–524, 1990.

[241] Z. Mahboubi and M.J. Kochenderfer, "Learning traffic patterns at small airports from flight tracks," *IEEE Trans. Intell. Transp.*, Vol. 18, no. 4, pp. 917–926, 2017.

[242] M.A. Mahfouz, A. Shoukry and M.A.Ismail, "EKNN: ensemble classifier incorporating connectivity and density into kNN with application to cancer diagnosis," *Artif. Intell. Med.*, Vol. 111, 101985, 2021; doi:10.1016/j.artmed.2020.101985.

[243] K. Makkawi, N. Ait–Tmazirte M. El Badaoui El Najjar and N. Moubayed, "Adaptive diagnosis for fault tolerant data fusion based on $\alpha-$Renyi divergence strategy for vehicle localization," *Entropy*, Vol. 23, No. 463, 2021; doi:10.3390/e23040463.

[244] R.P. Mann and R. Garnett R, "The entropic basis of collective behaviour," *J. R. Soc. Interface*, Vol. 12, No. 20150037, 2015; http://dx.doi.org/10.1098/rsif.2015.0037.

[245] Y. Mao, W. Hong, H. Wang, Q. Li, and S. Zhong, "Privacy-preserving computation offloading for parallel deep neural networks training," *IEEE Trans. Parall. Distr. Syst.*, Vol. 32, no. 7, pp. 1777–1788, 2021.

[246] S. Marano and M. Franceschetti, "Ray propagation in a random lattice: a maximum entropy, anomalous diffusion process," *IEEE Trans. Antenn. Propag.*, Vol. 53, no. 6, pp. 1888–1896, 2005.

[247] Y. Marhuenda, D. Morales, J.A. Pardo and M.C. Pardo, "Choosing the best Rukhin goodness-of-fit statistics," *Comp. Statist. & Data Anal.*, Vol. 49, pp. 643–662, 2005.

[248] M. Markatou and E. Sofikitou, "Non-quadratic distances in model assessment," *Entropy*, Vol. 20, No. 464, 2018; doi:10.3390/e20060464.

[249] M. Markatou and Y. Chen, "Statistical distances and the construction of evidence functions for model adequacy," *Front. Ecol. Evol.*, Vol. 7, No. 447, 2019; doi: 10.3389/fevo.2019.00447.

[250] F. Martin, J. Carballeira, L. Moreno, S. Garrido, and P. Gonzalez, "Using the Jensen-Shannon, density power, and Itakura-Saito divergences to implement an evolutionary-based global localization filter for mobile robots," *IEEE Access*, Vol. 5, pp. 13922–13940, 2017.

[251] J.A. Martin-Fernandez, C. Barcelo-Vidal, and V. Pawlowsky-Glahn, "Measures of difference for compositional data and hierarchical clustering methods," in: A. Buccianti, G. Nardi and R. Potenza (eds.), *Proceedings in IAMG'98, The Fourth Annual Conference of the International Association for Mathematical Geology*, pp. 526–531. Naples, Italy: De Frede, 1998.

[252] K.T. Matchev and P. Shyamsundar, "ThickBrick: optimal event selection and categorization in high energy physics. Part I. Signal discovery," *J. High Energy Phys.*, Vol. 2021, No.3, 291, 2021; doi:10.1007/JHEP03(2021)291.

[253] Y. Matsuyama, "The α-EM algorithm: surrogate likelihood maximization using α-logarithmic information measures," *IEEE Trans. Inf. Theory*, Vol. 49, no. 3, pp. 692–706, 2003.

[254] Y. Matsuyama, "The Alpha-HMM estimation algorithm: prior cycle guides fast paths," *IEEE Trans. Signal Process.*, Vol. 65, no. 13, pp. 3446–3461, 2017.

[255] Y. Matsuyama, "Divergence family attains blockchain applications via α-EM algorithm," in: *Proc. 2019 IEEE Int. Symp. Inf. Theory (ISIT 2019), Paris, France*, p. 727–731.

[256] K. Matusita, "On the theory of statistical functions," *Ann. Inst. Statist. Math. Tokyo*, Vol. 3, pp. 17–35, 1951; corrections in *Ann. Inst. Statist. Math. Tokyo*, Vol. 4, pp. 51–53, 1952.

[257] M.M. Mayoral, "Renyi's entropy as an index of diversity in sample-stage cluster sampling," *J. Information Sciences*, Vol. 105, pp. 101–114, 1998.

[258] C.S. McGinnis, D.A. Siegel, G. Xie, G. Hartoularos, M. Stone, C.J. Ye, Z.J. Gartner, N.R. Roan and S.A. Lee, "No detectable alloreactive transcriptional responses under standard sample preparation conditions during donormultiplexed single-cell RNA sequencing of peripheral blood mononuclear cells," *BMC Biology*, Vol. 19, 10, 2021; doi:10.1186/s12915-020-00941-x.

[259] D. McLeish, "Simulating random variables using moment-generating functions and the saddlepoint approximation," *J. Stat. Comput. Simul.*, Vol. 84, no. 2, pp. 324–334, 2014.

[260] L. Miao, H. Qi and H. Szu, "A maximum entropy approach to unsupervised mixed-pixel decomposition," *IEEE Trans. Image Process.*, Vol. 16, no. 4, pp. 1008–1021, 2007.

[261] S. Mohammadi, J. Davila-Velderrain and M. Kellis, "A multiresolution framework to characterize single-cell state landscapes," *Nature Commun.*, Vol. 11, 5399, 2020; doi:10.1038/s41467-020-18416-6.

[262] A. Mohanty, Q. Li, M.A. Tadayon, S.P. Roberts, G.R. Bhatt, E. Shim, X. Ji, J. Cardenas, S.A. Miller, A. Kepecs and M. Lipson, "Reconfigurable nanophotonic silicon probes for sub-millisecond deep-brain optical stimulation," *Nature Biomed. Eng.*, Vol. 4, no. 2, pp. 223–231, 2020.

[263] A. Momeni, K. Rouhi, H. Rajabalipanah and A. Abdolali, "An information theory-inspired strategy for design of reprogrammable encrypted graphene-based coding metasurfaces at terahertz frequencies," *Sci. Rep.*, Vol. 8, 6200, 2018; doi:10.1038/s41598-018-24553-2.

[264] D. Morales, L. Pardo, M. Salicru and M.L. Menendez, "Asymptotic properties of divergence statistics in a stratified random sampling and its applications to test statistical hypotheses," *J. Statist. Plan. Inf.*, Vol. 38, pp. 201–222, 1994.

[265] D. Morales, L. Pardo and I. Vajda, "Uncertainty of discrete stochastic systems: general theory and statistical inference," *IEEE Trans. Syst. Man Cyb. - Part A: Syst. Humans*, Vol. 26, no. 6, pp. 681–697, 1996.

[266] T. Morimoto, "Markov processes and the H-theorem," *J. Phys. Soc. Jpn.*, Vol. 18, no. 3, pp. 328—331, 1963.

[267] C.N. Morris, "Natural exponential families with quadratic variance functions," *Ann. Stat.*, Vol. 10, no. 1, pp. 65–80, 1982.

[268] C. Mühlroth and M. Grottke, "Artificial intelligence in innovation: how to spot emerging trends and technologies," *IEEE Trans. Eng. Manag.*, early access (to appear), 2021.

[269] A. Namdari and Z.S. Li, "An entropy-based approach for modeling Lithium-Ion battery capacity fade," in: *Proc. 2020 Annual Reliability and Maintainability Symposium (RAMS)*, 2020, 7 pages.

[270] P. Nath, "On a coding theorem connected with Renyi's entropy," *Information and Control*, Vol. 29, pp. 234–242, 1975.

[271] T.K. Nayak, "On diversity measures based on entropy functions," *Commun. Statist. - Theory Meth.*, Vol. 14, no. 1, pp. 203–215, 1985.

[272] M. Necci, D. Piovesan, CAID Predictors, DisProt Curators and S.C.E. Tosatto, "Critical assessment of protein intrinsic disorder prediction," *Nature Meth.*, Vol. 18, pp. 472—481, 2021.

[273] X. Nguyen, M.J. Wainwright,and M.I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inf. Theory*, Vol. 56, no. 11, pp. 5847–5861, 2010.

[274] Y.-K. Noh, J. Hamm, F.C. Park, B.-T. Zhang and D.D. Lee, "Fluid dynamic models for Bhattacharyya-based discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 40, no. 1, pp. 92–105, 2018.

[275] Y.-K. Noh, B.-T. Zhang and D.D. Lee, "Generative local metric learning for nearest neighbor classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 40, no. 1, pp. 106–118, 2018.

[276] A. Nunes, T. Trappenberg and M. Alda, "The definition and measurement of heterogeneity," *Transl. Psychiat.*, Vol. 10, No. 299, 2020; doi:10.1038/s41398-020-00986-0.

[277] D. Okada, N. Nakamura, K. Setoh, T. Kawaguchi, K. Higasa, Y. Tabara, F. Matsuda and R. Yamada, "Genome-wide association study of individual differences of human lymphocyte profiles using large-scale cytometry data," *J. Hum. Genet.*, Vol. 66, pp. 557—567, 2021.

[278] O. Onicescu, "Energie informationnelle," *C. R. Acad. Sci. Paris Ser. A.*, Vol. 263, no. 3, pp. 841–842, 1966.

[279] S. Onn, *Nonlinear Discrete Optimization*. Zürich, Switzerland: European Math. Society Publ. House, 2010.

[280] B. Pandey, "Renyi entropy as a measure of cosmic homogeneity," *J. Cosmol. Astropart. Phys.*, Vol. 2021, No. 2, 023, 2021; doi:10.1088/1475-7516/2021/02/023.

[281] L. Pardo, "Order$-\alpha$ weighted information energy," *Inf. Science*, Vol. 40, pp. 155–164, 1986.

[282] L. Pardo, *Statistical Inference Based on Divergence Measures*. Boca Raton, USA: Chapman & Hall/CRC, 2006.

[283] L. Pardo and I.J. Taneja, "Information energy and its applications," *Adv. Electron. El. Phys.*, Vol. 80, pp. 165–241, 1991.

[284] M.C. Pardo and I. Vajda, "About distances of discrete distributions satisfying the data processing theorem of information theory," *IEEE Trans. Inf. Theory*, Vol. 43, no. 4, pp. 1288–1293, 1997.

[285] R. Pascuzzo, N.P. Oxtoby, A.L. Young, J. Blevins, G. Castelli, S. Garbarino, M.L. Cohen, L.B. Schonberger, P. Gambetti, B.S. Appleby, D.C. Alexander and A. Bizzi, "Prion propagation estimated from brain diffusion MRI is subtype dependent in sporadic Creutzfeldt–Jakob disease," *Acta Neuropathologica*, Vol. 140, pp. 169—181, 2020.

[286] G.P. Patil and C. Taillie, "Diversity as a concept and its measurement," *J. Amer. Statist. Assoc.*, Vol. 77, no. 379, pp. 548–561, 1982.

[287] B.K. Patra, R. Launonen, V. Ollikainen and S. Nandi, "A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data," *Knowl.-Based Syst.*, Vol. 82, pp. 163—177, 2015.

[288] B. Pelletier, "Inference in $\varphi-$-families of distributions," *Statistics*, Vol. 45, no. 3, pp. 223–236, 2011.

[289] X.-B. Peng and Z.-J. Li, "Target scale adaptive control based on comparing Bhattacharyya coefficient," *Adv. Mat. Res.*, Vol. 971–973, pp. 1772–1777, 2014.

[290] S. Perera, D. Kasthurirathna and M. Bliemer, "Toplogical rationality of supply chain networks," *Int. J. Prod. Res.*, Vol. 58, no. 10, pp. 3126–3149, 2020.

[291] A.M. Peter and A. Rangarajan, "Information geometry for landmark shape analysis: unifying shape representation and deformation," *IEEE Trans. IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 31, no. 2, pp. 337–350, 1993.

[292] D.-T. Pham, F. Vrins, and M. Verleysen, "On the risk of using Renyi's entropy for blind source separation," *IEEE Trans. Signal Process.*, Vol. 56, no. 10, pp. 4611–4620, 2008.

[293] E.C. Pielou, "The measurement of diversity in different types of biological Collections," *J. Theoret. Biol.*, Vol. 13, pp. 131–143, 1966.

[294] F. Pierri, C. Piccardi and S. Ceri, "Topology comparison of Twitter diffusion networks effectively reveals misleading information," *Sci. Rep.*, Vol. 10, 1372, 2020; doi:10.1038/s41598-020-58166-5.

[295] T.D. Popescu and D. Aiordachioaie, "New procedure for change detection operating on Renyi entropy with application in seismic signals Processing," *Circuits Syst. Signal Process.*, Vol. 36, pp. 3778—3798, 2017.

[296] T. Prest, "Sharper bounds in lattice-based cryptography using the Renyi divergence," in: T. Takagi and T. Peyrin (eds.), *ASIACRYPT 2017*, Part I, Lecture Notes in Computer Science, vol. 10624, pp. 347—374. Cham, Switzerland: Springer Nature, 2017.

[297] J.C. Principe, *Information Theoretic Learning*. New York, NY, USA: Springer, 2010.

[298] Y. Qiao and N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Trans. Signal Process.*, Vol. 58, no. 7, pp. 3884–3890, 2010.

[299] R. Rabadan, Y. Mohamedi, U. Rubin, T. Chu, A.N. Alghalith, O. Elliott, L. Arnes, S. Cal, A.J. Obaya, A.J. Levine and P.G. Camara, "Identification of relevant genetic alterations in cancer using topological data analysis," *Nature Commun.*, Vol. 11, 3808, 2020; doi:10.1038/s41467-020-17659-7.

[300] Z. Ramezani and A. Pourdarvish, "Transfer learning using Tsallis entropy: an application to Gravity Spy," *Physica A*, Vol. 561, 125273, 2020; doi.org/10.1016/j.physa:2020.125273.

[301] C.R. Rao, "Cluster analysis applied to a study of race mixture in human populations," in: J. van Ryzin (ed.), *Classification and clustering*, pp. 175—197. New York, USA: Academic Press, 1977.

[302] P.N. Rathie and P. Kannappan, "A directed-divergence function of type β," *Information and Control*, Vol. 20, pp. 38–45, 1972.

[303] T.R.C. Read and N.A.C. Cressie, *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York, USA: Springer, 1988.

[304] M.D. Reid and R.C. Williamson, "Information, divergence and risk for binary experiments," *J. Machine Learn. Res.*, Vol. 12, pp. 731–817, 2011.

[305] D. Reising, J. Cancelleri, T.D. Loveless, F. Kandah and A. Skjellum, "Radio identity verification-based IoT security using RF-DNA fingerprints and SVM," *IEEE Internet Things J.*, Vol. 8, no. 10, pp. 8356–8371, 2021.

[306] J.G. Reiter, W.-T. Hung, I-H. Lee, S. Nagpal1, P. Giunta, S. Degner, G. Liu, E.C.E. Wassenaar, W.R. Jeck, M.S. Taylor, A.A. Farahani, H.D. Marble, S. Knott, O. Kranenburg, J.K. Lennerz and K. Naxerova, "Lymph node metastases develop through a wider evolutionary bottleneck than distant metastases," *Nature Genetics*, Vol. 52, pp. 692—700, 2020.

[307] A.M. Rekavandi, A.-K. Seghouane and R.J. Evans, "Robust subspace detectors based on $\alpha-$divergence with application to detection in imaging," *IEEE Trans. Image Process.*, Vol. 30, pp. 5017-5031, 2021.

[308] M. Ren, J. Zhang, M. Jiang, M. Yu, and J. Xu, "Minimum $(h,\phi)-$entropy control for non-Gaussian stochastic networked control systems and its application to a networked DC motor control system," *IEEE Trans. Control Syst. Technol.*, Vol. 23, no. 1, pp. 406–411, 2015.

[309] A. Renyi, "On measures of entropy and information," in: J. Neyman (ed.), *Proc. 4th Berkeley Symp. Math. Stat. Probab. Vol.1*, pp. 547–561. Berkeley, CA, USA: Univ. of California Press, 1961.

[310] E.S.C. Rodrigues, F.A. Rodrigues, R.L.A. Rocha and P.L.P. Correa, "Adaptive approach for a maximum entropy algorithm in ecological niche modeling," *IEEE Lat. Amer. Trans.*, Vol. 9, no. 3, pp. 331–338, 2011.

[311] C. Rong, B. Wang, D. Zhao and S. Liu, "Information-theoretic approach in density functional theory and its recent applications to chemical problems," *WIREs Comput Mol Sci.*, Vol. 10, No. e1461, 2020; doi:10.1002/wcms.1461.

[312] A.L. Rukhin, "Optimal estimator for the mixture parameter by the method of moments and information affinity," in: *Trans. 12th Prague Conf. Information Theory, Statistical Decision Functions and Random Processes*, pp. 214–219. Prague, Czech Republic: Czech Acad. Sci., 1994.

[313] J.S. Sadowsky and J.A. Bucklew, "On Large Deviations Theory and Asymptotically Efficient Monte Carlo Estimation," *IEEE Trans. Inf. Theory*, Vol. 36, no. 3, pp. 579–588, 1990.

[314] M. Salicru, M.L. Menendez, D. Morales and L. Pardo, "Asymptotic distribution of $(h, \phi)-$entropies," *Commun. Statist. - Theory Meth.*, Vol. 22, no. 7, pp. 2015–2031, 1993.

[315] O. Salem, A. Serhrouchni, A. Mehaoua, and R. Boutaba, "Event detection in wireless body area networks using Kalman filter and power divergence," *IEEE Trans. Netw. Serv. Manag.*, Vol. 15, no. 3, pp. 1018–1034, 2018.

[316] L.D. Sanghvi, "Comparison of genetical and morphological methods for a study of biological differences," *Am. J. Phys. Anthropol.*, Vol. 11, no. 3, pp. 385–404, 1953.

[317] I. Sason, "Tight bounds for symmetric divergence measures and a new inequality relating $f-$divergences," in: *Proc. 2015 IEEE Information Theory Workshop (ITW)*, 2014, 5 pages.

[318] I. Sason, "On the Renyi divergence, joint range of relative entropies, and a channel coding theorem," *IEEE Trans. Inf. Theory*, Vol. 62, no. 1, pp. 23–34, 2016.

[319] I. Sason, "Tight bounds on the Renyi entropy via majorization with applications to guessing and compression," *Entropy*, Vol. 20, No. 896, 2018; doi:10.3390/e20120896.

[320] I. Sason, "On data-processing and majorization inequalities for f-Divergences with applications," *Entropy*, Vol. 21, No. 1022, 2019; doi:10.3390/e21101022.

[321] I. Sason and S. Verdu, "Improving bounds on lossless source coding and guessing moments via Renyi measures," *IEEE Trans. Inf. Theory*, Vol. 64, no. 6, pp. 4323–4346, 2018.

[322] R.L. Schilling, R. Song and Z. Vondracek, *Bernstein Functions*, 2nd ed. Berlin, Germany: de Gruyter, 2012.

[323] K. Schober, F. Voit, S. Grassmann, T.R. Müller, J. Eggert, S. Jarosch, Bianca Weißbrich, P. Hoffmann, L. Borkner, E. Nio, L. Fanchi, C.R. Clouser, A. Radhakrishnan, L. Mihatsch, P. Lückemeier, J. Leube, G. Dössinger1, L. Klein, M. Neuenhahn, J.D. Oduro, L. Cicin-Sain, V.R. Buchholz and D.H. Busch, "Reverse TCR repertoire evolution toward dominant low-affinity clones during chronic CMV infection," *Nature Immunol.*, Vol. 21, pp. 434–441, 2020.

[324] A. Schrijver, *Combinatorial Optimization*, Vol.A,B,C. Heidelberg, Germany: Springer, 2003.

[325] A.-K. Seghouane and N. Shokouhi, "Adaptive learning for robust radial basis function networks," *IEEE Trans. Cybernetics*, Vol. 51, No. 5, pp. 2847–2856, 2021.

[326] M.T. Seweryn, M. Pietrzak and Q. Mab, "Application of information theoretical approaches to assess diversity and similarity in single-cell transcriptomics," *Comput. Struct. Biotechn. J.*, Vol. 18, pp. 1830—1837, 2020.

[327] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, USA: Princeton University Press, 1976.

[328] C.E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, Vol. 27, no. 3, pp. 379–423, 1948.

[329] B.D. Sharma and D.P. Mittal, "New nonadditive measures of entropy for discrete probability distributions," *J. Math. Sci. (Delhi)*, Vol. 10, pp. 28–40, 1975.

[330] B.D. Sharma and D.P. Mittal, "New nonadditive measures of entropy for discrete probability distributions," *J. Combin. Inform. System Sci.*, Vol. 2, no. 4, pp. 122–132, 1977.

[331] M. Shiino, "H-Theorem with generalized relative entropies and the Tsallis statistics," *J. Phys. Soc. Japan*, Vol. 67, no. 11, pp. 3658–3660, 1998.

[332] D.-D. Shi, D. Chen and G.-J. Pan, "Characterization of network complexity by communicability sequence entropy and associated Jensen-Shannon divergence," *Phys. Rev. E*, Vol. 101, No. 042305, 2020; doi:10.1103/PhysRevE.101.042305.

[333] D.S. Shucker, "Extensions and generalizations of a theorem of Widder and of the theory of symmetric local semigroups," *J. Funct. Anal.*, Vol. 58, pp. 291–309, 1984.

[334] R. Sibson, "Information radius," *Probab. Theory Rel. Fields*, Vol. 14, pp. 149–160, 1969.

[335] J. S. Sigmon et al., "Content and Performance of the MiniMUGA Genotyping Array: A New Tool To Improve Rigor and Reproducibility in Mouse Research," *Genetics*, Vol. 216, pp. 905—930, 2020.

[336] A.K. Singh, H.P. Singh and Karmeshu, "Analysis of finite buffer queue: maximum entropy probability distribution with shifted fractional geometric and arithmetic means," *IEEE Commun. Lett.*, Vol. 19, no. 2, pp. 163–166, 2015.

[337] M.A. Skinnider, C.W. Johnston, M. Gunabalasingam, N.J. Merwin, A.M. Kieliszek, R.J. MacLellan, H. Li, M.R.M. Ranieri, A.L.H. Webster, M.P.T. Cao, A. Pfeifle, N. Spencer, Q. H. To, D.P. Wallace, C.A. Dejong 3 and N.A. Magarvey, "Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences," *Nature Commun.*, Vol. 11, 6058, 2020; doi:10.1038/s41467-020-19986-1.

[338] T. Skrbic, A. Maritan, A. Giacometti and J.R. Banavar, "Local sequence-structure relationships in proteins," *Protein Sci.*, Vol. 30, pp. 818—829, 2021.

[339] Y. Song and Y. Deng, "Divergence measure of belief function and its application in data fusion," *IEEE Access*, Vol. 7, pp. 107465–107472, 2020.

[340] A.D. Staszowska, P. Fox-Roberts, L.M. Hirvonen, C.J. Peddie, L.M. Collinson, G.E. Jones and S. Cox, "The Renyi divergence enables accurate and precise cluster analysis for localization microscopy," *Bioinformatics*, Vol. 34, no. 23, pp. 4102—4111, 2018.

[341] F.W. Steutel and K. van Harn, *Infinite Divisibility of Probability Distributions On The Real Line*. New York, USA: Marcel Dekker Inc., 2004.

[342] S. Stoller and K.A. Campbell, "Demonstration of three true random number generator circuits using memristor created entropy and commercial off-the-shelf components," *Entropy*, Vol. 23, No. 371, 2021; doi:10.3390/e23030371.

[343] D.W. Stroock, *Probability Theory: An Analytic View*, 2nd ed. New York, USA: Cambridge University Press, 2011.

[344] W. Stummer, *Exponentials, Diffusions, Finance, Entropy and Information*. Aachen, Germany: Shaker, 2004.

[345] W. Stummer, "Optimal transport with some directed distances," in: F. Nielsen and F. Barbaresco (eds.), *Geometric Science of Information GSI 2021*, Lecture Notes in Computer Science, vol. 12829, pp. xxx—xxx. Cham, Switzerland: Springer Nature Switzerland, 2021.

[346] W. Stummer and A.-L. Kißlinger, "Some new flexibilizations of Bregman divergences and their asymptotics," In: F. Nielsen and F. Barbaresco (eds.), *Geometric Science of Information GSI 2017*, Lecture Notes in Computer Science, vol. 10589, pp. 514—522. Cham, Switzerland: Springer International Publishing, 2017.

[347] W. Stummer and W. Lao, "Limits of Bayesian decision related quantities of binomial asset price models," *Kybernetika*, Vol. 48, no.4, pp. 750–767, 2012.

[348] W. Stummer and I. Vajda, "Optimal statistical decisions about some alternative financial models," *J. Econometrics*, Vol. 137, no.2, pp. 441–471, 2007.

[349] W. Stummer and I. Vajda, "On divergences of finite measures and their applicability in statistics and information theory," *Statistics*, Vol. 44, no. 2, pp. 169—187, 2010.

[350] W. Stummer and I. Vajda, "On Bregman distances and divergences of probability measures," *IEEE Trans. Inf. Theory*, Vol. 58, no. 3, pp. 1277–1288, 2012.

[351] W. Sun, J. Wang and F. Jin, “An automatic coordinate unification method of multitemporal point clouds based on virtual reference datum detection,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, Vol. 13, pp. 3942–3950, 2020.
[352] R. Sundaresan, “A measure of discrimination and its geometric properties,” in: *Proc. 2002 IEEE Int. Symp. Inf. Theory (ISIT 2002), Lausanne, Switzerland*, p. 264.
[353] R. Sundaresan, “Guessing Under Source Uncertainty,” *IEEE Trans. Inf. Theory*, Vol. 53, no. 1, pp. 269–287, 2007.
[354] S. Suo, Q. Zhu, A. Saadatpour, L. Fei, G. Guo and G.-C. Yuan, “Revealing the critical regulators of cell identity in the mouse cell atlas,” *Cell Rep.*, Vol. 25, pp. 1436—1445, 2018.
[355] B. Tao, L. Zhu, H. Ding and Y. Xiong, “An alternative time-domain index for condition monitoring of rolling element bearings – a comparison study,” *Reliab. Eng. Syst. Saf.*, Vol. 92, pp. 660–670, 2007.
[356] A. Tarighati and J. Jalden, “Optimality of rate balancing in wireless sensor networks,” *IEEE Trans. Signal Process.*, Vol. 64, no. 14, pp. 3735–3749, 2016.
[357] M. Teboulle and I. Vajda, “Convergence of best $\phi-$entropy estimates,” *IEEE Trans. Inf. Theory*, Vol. 39, no. 1, pp. 297–301, 1993.
[358] J.S. Teh, A. Samsudin, M. Al-Mazrooie and A. Akhavan, “GPUs and chaos: a new true random number generator,” *Nonlinear Dyn.*, Vol. 82, pp. 1913–1922, 2015.
[359] A. Theodorescu, “Energie informationelle et notions apparentees,” *Trabajos de Estadist. Investigacion Oper.*, Vol. XXVIII, no. 2-3, pp. 183–206, 1977.
[360] F. Topsoe, “Some inequalities for information divergence and related measures of discrimination,” *IEEE Trans. Inf. Theory*, Vol. 46, no. 4, pp. 1602–1609, 2000.
[361] E. Torgersen, *Comparison of Statistical Experiments*. Cambridge, UK: Cambridge University Press, 1991.
[362] G.T. Toussaint, “Probability of error, expected divergence and the affinity of several distributions,” *IEEE Trans. Syst. Man Cyb.*, Vol. 8, no. 6, pp. 482–485, 1978.
[363] C. Tsallis, “Possible generalization of Boltzmann-Gibbs Statistics,” *Journal of Statistical Physics*, Vol. 52, no. 1/2, pp. 479–487, 1988.
[364] C. Tsallis, “Generalized entropy-based criterion for consistent testing,” *Phys. Rev. E*, Vol. 58, No. 2, pp. 1442–1445, 1998;
[365] S. Tsiapoki, O. Bahrami, M.W. Häckell, J.P. Lynch and R. Rolfes, “Combination of damage feature decisions with adaptive boosting for improving the detection performance of a structural health monitoring framework: validation on an operating wind turbine,” *Struct. Health. Monit.*, Vol. 20, no. 2, pp. 637—660, 2021.
[366] R.R. Tucci, “Method for sampling probability distributions using a quantum computer,” *United States Patent*, Patent No. US 8543627 B1, 24th Sep. 2013.
[367] S. Tuo, H. Liu and H. Chen, “Multipopulation harmony search algorithm for the detection of high-order SNP interactions,” *Bioinformatics*, Vol. 36, no. 16, pp. 4389—4398, 2020.
[368] M.C.K. Tweedie, “Inverse statistical variates,” *Nature*, Vol. 155, no. 3937, pp. 453–453, 1945.
[369] M.C.K. Tweedie, “Functions of a statistical variate with given means, with special reference to Laplacian distributions,” *Proc. Camb. Philos. Soc.*, Vol. 43, pp. 41–49, 1947.
[370] S. Uttam, A.M. Stern, C.J. Sevinsky, S. Furman, F. Pullara, D. Spagnolo, L. Nguyen, A. Gough, F. Ginty, D.L. Taylor and S.C. Chennubhotla, “Spatial domain analysis predicts risk of colorectal cancer recurrence and infers associated tumor microenvironment networks,” *Nature Commun.*, Vol. 11, 3515, 2020; 10.1038/s41467-020-17083-x.
[371] I. Vajda, *Theory of Statistical Inference and Information*. Dordrecht, NL: Kluwer, 1989.
[372] I. Vajda, “About perceptron realizations of Bayesian decisions,” in: *IEEE International Conference on Neural Networks, Washington 1996*, pp. 253—257, 1996.
[373] I. Vajda, “On metric divergences of probability measures,” *Kybernetika*, Vol. 45, no. 6, pp. 885–900, 2009.
[374] I. Vajda and E.C. van der Meulen, “Goodness-of-fit criteria based on observations quantized by hypothetical and empirical percentiles,” in: Z.A. Karian ZA and E.J. Dudewicz (eds.), *Handbook of Fitting Statistical Distributions with R*, pp. 917 – 994. Heidelberg, Germany: CRC, 2010.
[375] I. Vajda and K. Vasek, “Majorizations, concave entropies, and comparison of experiments,” *Problems of Control and Information Theory*, Vol. 14, no. 2, pp. 105–115, 1985.
[376] I. Vajda and J. Zvarova, “On generalized entropies, Bayesian decisions and statistical diversity,” *Kybernetika*, Vol. 43, no. 5, pp. 675–696, 2007.
[377] S. Valpione, E. Galvani, J. Tweedy, P.A. Mundra1, A. Banyard, P. Middlehurst, J. Barry, S. Mills, Z. Salih, J. Weightman, A. Gupta, G. Gremel, F. Baenke, N. Dhomen, P.C. Lorigan and R. Marais, “Immune-awakening revealed by peripheral T cell dynamics after one cycle of immunotherapy,” *Nature Cancer*, Vol. 1, No. 2, pp. 210–221, 2021.
[378] B. Van de Sande, C. Flerin, K. Davie, M. De Waegeneer, G. Hulselmans, S. Aibar, R. Seurinck, W. Saelens, R. Cannoodt, Q. Rouchon, T. Verbeiren, D. De Maeyer, J. Reumers, Y. Saeys and S. Aerts, “A scalable SCENIC workflow for single-cell gene regulatory network analysis,” *Nature Protocols*, Vol. 15, pp. 2247–2276, 2020.
[379] J.C.A. van der Lubbe, “An axiomatic theory of heterogeneity and homogeneity,” *Metrika*, Vol. 33, pp. 223–245, 1986.
[380] T. van Erven and P. Harremoes, “Renyi divergence and Kullback-Leibler divergence,” *IEEE Trans. Inf. Theory*, Vol. 60, no. 7, pp. 3797–3820, 2014.
[381] P. van Molle, T. Verbelen, B. Vankeirsbilck, J. De Vylder, B. Diricx, T. Kimpe, P. Simoens and B. Dhoedt, “Leveraging the Bhattacharyya coefficient for uncertainty quantification in deep neural networks,” *Neural Comput. Appl.*, early access (to appear), 2021; doi.org/10.1007/s00521-021-05789-y.
[382] R. Verma, “Multiple attribute group decision-making based on order-α divergence and entropy measures under q-rung orthopair fuzzy environment,” *Int. J. Intell. Syst.*, Vol. 35, pp. 718–750, 2020.
[383] J. Verrelst, J.P. Rivera, G. Leonenko, L. Alonso, and J. Moreno, “Optimizing LUT-based RTM inversion for semiautomatic mapping of crop biophysical parameters from Sentinel-2 and -3 Data: role of cost functions,” *IEEE Trans. Geosci. Remote Sens.*, Vol. 52, no. 1, pp. 257–269, 2014.
[384] I. Vincze, “On the concept and measure of information contained in an observation,” in: J. Gani and V.K. Rohatgi (eds.), *Contributions to Probability*, pp. 207—214. New York, NY, USA: Academic Press, 1981.
[385] I.K. Vlachos and G.D. Sergiadis, “Intuitionistic fuzzy information – applications to pattern recognition,” *Pattern Recogn. Lett.*, Vol. 28, pp. 197–206, 2007.
[386] R. Wang, M. Dang, K. Harada, G. Han, F. Wang, M.P. Pizzi, M. Zhao, G. Tatlonghari, S. Zhang, D. Hao, Y. Lu, S. Zhao, B.D. Badgwell, M. B. Murphy, N. Shanbhag, J.S. Estrella, S. Roy-Chowdhuri, A.A.F. Abdelhakeem, Y. Wang, G. Peng, S. Hanash, G.A. Calin, X. Song, Y. Chu, J. Zhang, M. Li, K. Chen, A.J. Lazar, A. Futreal, S. Song, J.A. Ajani and L. Wang, “Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma,” *Nature Med.*, Vol. 27, 141–151, 2021.
[387] X. Wang, J. Wang, C. Miao and K. Zeng, “Forewarning method of downburst based on feature recognition and extrapolation,” *Natural Hazards*, Vol. 103, pp. 903—921, 2020.
[388] Y. Wang, P. Wang, Z. Liu and L.Y. Zhang, “A new item similarity based on $\alpha-$divergence for collaborative filtering in sparse data,” *Expert Syst. Appl.*, Vol. 166, 114074, 2021; doi:10.1016/j.eswa.2020.114074.
[389] A.J. Webster, K. Gaitskell, I. Turnbull, B.J. Cairns and R. Clarke, “Characterisation, identification, clustering, and classification of disease,” *Sci. Rep.*, Vol. 11, 5405, 2021; doi:10.1038/s41598-021-84860-z.
[390] T. Wen and W. Jiang, “Measuring the complexity of complex networks by Tsallis entropy,” *Physica A*, Vol. 526, 121054, 2019; doi:10.1016/j.physa.2019.121054.
[391] D.V. Widder, “Necessary and sufficient conditions for the representation of a function by a doubly infinite Laplace integral,” *Bull. Amer. Math. Soc.*, Vol. 40, no. 4, pp. 321–326, 1934.
[392] D.V. Widder, *The Laplace Transform*. Princeton, NJ, USA: Princeton University Press, 1941.

[393] L.A. Wolsey, *Integer Programming*, 2nd ed. Hoboken, NJ, USA: Wiley, 2021.
[394] A.J. Wood, B.F. Wollenberg and G.B. Sheble, *Power Generation, Operation, and Control*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2014.
[395] W.K. Wootters, "Statistical distance and Hilbert space," *Phys. Rev. D*, Vol. 23, no. 2, pp. 357–362, 1981.
[396] T. Xiahou, Z. Zeng and Y. Liu, "Remaining useful life prediction by fusing expert knowledge and condition monitoring information," *IEEE Trans. Ind. Inform.*, Vol. 17, no. 4, pp. 2653–2663, 2021.
[397] F. Xiao, "Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy," *Inf. Fusion*, Vol. 46, pp. 23–32, 2019.
[398] F. Xiao, "A new divergence measure for belief functions in D-S evidence theory for multisensor data fusion," *Inf. Sci.*, Vol. 514, pp. 462–483, 2020.
[399] F. Xiao and W. Ding, "Divergence measure of Pythagorean fuzzy sets and its application in medical diagnosis," *Appl. Soft. Comput. J.*, Vol. 79, pp. 254–267, 2019.
[400] D. Xiong, M. Zhang and H. Li, "A maximum-entropy segmentation model for statistical machine translation," *IEEE Trans. Audio Speech Lang. Process.*, Vol. 19, no. 8, pp. 2494–2505, 2011.
[401] C. Xu, K. Wang, P. Li, R. Xia, S. Guo, and M. Guo, "Renewable energy-aware big data analytics in geo-distributed data centers with reinforcement learning," *IEEE Trans. Netw. Sci. Eng.*, Vol. 7, no. 1, pp. 205–215, 2020.
[402] M. Xu, V. Jog and P.-L. Loh, "Optimal rates for community estimation in the weighted stochastic block model," *Ann. Stat.*, Vol. 48, no. 1, pp. 183–204, 2020.
[403] Y. Xu, Y. Xue, G. Hua and J. Cheng, "An adaptive distributed compressed video sensing algorithm based on normalized Bhattacharyya coefficient for coal mine monitoring video," *IEEE Access*, Vol. 8, pp. 158369–158379, 2020.
[404] Y. Xu, W. Zhang, W. Liu and W. Yu, *Distributed Energy Management of Electrical Power Systems*. Hoboken, NJ, USA: John Wiley & Sons, 2021.
[405] R.R. Yager, "Pythagorean fuzzy subsets," in: *Proc. of the 2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*, 2013, pp. 57–61.
[406] R.R. Yager, "Generalized orthopair fuzzy sets," *IEEE Trans. Fuzzy Syst.*, Vol. 25, no. 5, pp. 1222–1230, 2017.
[407] R.R. Yager and A.M. Abbasov, "Pythagorean membership grades, complex numbers, and decision making," *Int. J. Intell. Syst.*, Vol. 28, pp. 436–452, 2013.
[408] S. Yagli, Y. Altug and S. Verdu, "Minimax Renyi redundancy," *IEEE Trans. Inf. Theory*, Vol. 64, no. 5, pp. 3715–3733, 2018.
[409] G.C. Yalcin and C. Beck, "Generalized statistical mechanics of cosmic rays: application to positron-electron spectral indices," *Sci. Rep.*, Vol. 8, 1764, 2018; doi:10.1038/s41598-018-20036-6.
[410] T. Yamano, "Some bounds for skewed $\alpha-$Jensen-Shannon divergence," *Results in Appl. Math.*, Vol. 3, No. 100064, 2019; doi:10.1016/j.rinam.2019.100064.
[411] J. Yanase and E. Triantaphyllou, "A systematic survey of computer-aided diagnosis in medicine: past and present developments," *Expert Syst. Appl.*, Vol. 138, 112821, 2019; doi:10.1016/j.eswa.2019.112821.
[412] S. Yang, S. Tan and J.-X. Xu, "Consensus based approach for economic dispatch problem in a smart grid," *IEEE Trans. Power Syst.*, Vol. 28, no. 4, pp. 4416–4426, 2013.
[413] Z. Yang, W. Yan, F. Li, Z. Yu and L. Guo, "Evaluating onset times of acoustic emission signals using histogram distances," *IEEE Trans. Ind. Electron.*, Vol. 68, no. 6, pp. 5237–5247, 2021.
[414] L. Yu and V.Y.F. Tan, "Wyner's common information under Renyi divergence measures," *IEEE Trans. Inf. Theory*, Vol. 64, no. 5, pp. 3616–3632, 2018; Comments and Corrections: Vol. 66, no. 4, pp. 2599–2508, 2020.
[415] L. Yu and V.Y.F. Tan, "Simulation of random variables under Renyi divergence measures of all orders," *IEEE Trans. Inf. Theory*, Vol. 65, no. 6, pp. 3349–3383, 2019.
[416] X. Yu, H. Zhang, C. Luo, H. Qi and P. Ren, "Oil spill segmentation via adversarial f-divergence learning," *IEEE Trans. Geosci. Remote Sens.*, Vol. 56, no. 9, pp. 4973–4988, 2018.
[417] L.A. Zadeh, "Fuzzy sets," *Information and Control*, Vol. 8, pp. 338–353, 1965.
[418] G. Zhang, X. Su and V.P. Singh, "Modelling groundwater-dependent vegetation index using entropy theory," *Ecol. Model.*, Vol. 416, 108916, 2020; doi:10.1016/j.ecolmodel.2019.108916.
[419] G. Zhang, F. Lian, C. Han, H. Chen and N. Fu, "Two novel sensor control schemes for multitarget tracking via delta generalised labelled multi-Bernoulli filtering," *IET Signal Process.*, Vol. 12, no. 9, pp. 1131–1139, 2018.
[420] T. Zhang, J. Cheng, H. Fu, Z. Gu, Y. Xiao, K. Zhou, S. Gao, R. Zheng and J. Liu, "Noise adaptation generative adversarial network for medical image analysis," *IEEE Trans. Med. Imag.*, Vol. 39, no. 4, pp. 1149–1159, 2020.
[421] X. Zhang, C. Delpha, and D. Diallo, "Jensen-Shannon divergence for non-destructive incipient crack detection and estimation," *IEEE Access*, Vol. 8, pp. 116148–116162, 2020.
[422] Z. Zhang, J. Ding, J. Xu, J. Tang and F. Guo, "Multi-scale time-series kernel-based learning method for brain disease diagnosis," *IEEE J. Biomed. Health Inform.*, Vol. 25, no. 1, pp. 209–217, 2021.
[423] D. Zhao, S. Liu, C. Rong, A. Zhong, and S. Liu, "Toward understanding the isomeric stability of fullerenes with density functional theory and the information-theoretic approach," *ACS Omega*, Vol. 3, pp. 17986–17990, 2018.
[424] S. Zhao, G. Ding, Y. Gao, X. Zhao, Y. Tang, J. Han, H. Yao and Q. Huang, "Discrete probability distribution prediction of image emotions with shared sparse learning," *IEEE Trans. Affect. Comput.*, Vol. 11, no. 4, pp. 574–587, 2020.
[425] T. Zhi, Y. Liu, J. Wang and H. Zhang, "Resist interest flooding attacks via Entropy—SVM and Jensen—Shannon divergence in information–centric networking," *IEEE Syst. J.*, Vol. 14, no. 2, pp. 297–301, 2020.
[426] L. Zhou, "Multiple private key generation for continuous memoryless sources with a helper," *IEEE Trans. Foren. Sec.*, Vol. 15, pp. 2629–2640, 2020.
[427] Y. Zhou and Y. Yu, "Human visual search follows a suboptimal Bayesian strategy revealed by a spatiotemporal computational model and experiment," *Commun. Biol.*, Vol. 4, No. 34, 2021; doi:10.1038/s42003-020-01485-0.
[428] V.M. Zolotarev, *One-dimensional Stable Distributions*. Providence, USA: American Mathematical Society, 1986.