Some Universal Insights on Divergences for Statistics, Machine Learning and Artificial Intelligence

Michel Broniatowski and Wolfgang Stummer

Abstract Dissimilarity quantifiers such as divergences (e.g. Kullback–Leibler information, relative entropy) and distances between *probability distributions* are widely used in statistics, machine learning, information theory and adjacent artificial intelligence (AI). Within these fields, in contrast, some applications deal with divergences between *other-type* real-valued functions and vectors. For a broad readership, we present a correspondingly unifying framework which – by its nature as a "structure on structures" – also qualifies as a basis for similarity-based multistage AI and more humanlike (robustly generalizing) machine learning. Furthermore, we discuss some specificalities, subtleties as well as pitfalls when e.g. one "moves away" from the probability context. Several subcases and examples are given, including a new approach to obtain parameter estimators in continuous models which is based on noisy divergence minimization.

1 Outline

The goals formulated in the abstract are achieved in the following way and order: to address a wide audience, throughout the paper (with a few connection-indicative exceptions) we entirely formulate and investigate divergences and distances between

M. Broniatowski
Sorbonne Universite Pierre et Marie Curie, LPSM, 4 place Jussieu, 75252 Paris, France
e-mail: michel.broniatowski@sorbonne-universite.fr

W. Stummer (✉)
Department of Mathematics, University of Erlangen–Nürnberg, Cauerstrasse 11, 91058 Erlangen, Germany
e-mail: stummer@math.fau.de

W. Stummer
Affiliated Faculty Member of the School of Business and Economics, University of Erlangen–Nürnberg, Lange Gasse 20, 90403 Nürnberg, Germany

F. Nielsen (ed.), *Geometric Structures of Information*, Signals and Communication Technology, https://doi.org/10.1007/978-3-030-02520-5_8

functions, even for the probability context. In Sect. 2, we provide some non-technical background and overview of some of their principally possible usabilities for tasks in data analytics such as statistics, machine learning, and artificial intelligence (AI). Furthermore, we indicate some connections with geometry and information. Thereafter, in Sect. 3 we introduce a new *structured* framework (toolkit) of divergences between functions, and discuss their building-blocks, boundary behaviour, as well as their identifiability properties. Several subcases, running examples, technical subtleties of practical importance, etc. are illuminated, too. Finally, we study divergences between "entirely different functions" which e.g. appear in the frequent situation when for data-derived *discrete* functions one wants to find a closest possible continuous-function model (cf. Sect. 4); several corresponding noisy minimum-divergence procedures are compared – for the first time within a unifying framework – and new methods are derived too.

2 Some General Motivations and Uses of Divergences

2.1 Quantification of Proximity

As a starting motivation, it is basic knowledge that there are numerous ways of evaluating the proximity $d(p,q)$ of two real numbers p and q of primary interest. For instance, to quantify that p and q nearly coincide one could use the difference $d^{(1)}(p,q) := p-q \approx 0$ or the fraction $d^{(2)}(p,q) := \frac{p}{q} \approx 1$, scaled (e.g. magnifying, zooming-in) versions $d^{(3)}_m(p,q) := m \cdot (p-q) \approx 0$ or $d^{(4)}_m(p,q) := m \cdot \frac{p}{q} \approx 1$ with "scale" m of secondary (auxiliary) interest, as well as more flexible hybrids $d^{(5)}_{m_1,m_2,m_3}(p,q) := m_3 \cdot \left(\frac{p}{m_1} - \frac{q}{m_2}\right) \approx 0$ where m_i may also take one of the values p, q. All these "dissimilarities" $d^{(j)}(\cdot,\cdot)$ can principally take any sign and they are asymmetric which is consistent with the – in many applications required – desire that one of the two primary-interest numbers (say p) plays a distinct role; moreover, the involved divisions cause technical care if one principally allows for (convergence to) zero-valued numbers. A more sophisticated, nonlinear alternative to $d^{(1)}(\cdot,\cdot)$ is given by the dissimilarity $d^{(6)}_\phi(p,q) := \phi(p) - (\phi(q) + \phi'(q)\cdot(p-q))$ where $\phi(\cdot)$ is a strictly convex, differentiable function and thus $d^{(6)}_\phi(p,q)$ quantifies the difference between $\phi(p)$ and the value at p of the tangent line taken at $\phi(q)$. Notice that $d^{(6)}_\phi(\cdot,\cdot)$ is generally still asymmetric but always stays nonnegative independently of the possible signs of the "generator" ϕ and the signs of p,q. In contrast, as a nonlinear alternative to $d^{(4)}_m(\cdot,\cdot)$ one can construct from ϕ the dissimilarity $d^{(7)}_\phi(p,q) := q \cdot \phi\left(\frac{p}{q}\right)$ (where $m = q$) which is also asymmetric but can become negative depending on the signs of p, q, ϕ. More generally, one often wants to work with dissimilarities $d(\cdot,\cdot)$ having the properties

(D1) $d(p,q) \geqslant 0$ for all p, q (nonnegativity),
(D2) $d(p,q) = 0$ if and only if $p = q$ (reflexivity; identity of indiscernibles[1]),

and such $d(\cdot,\cdot)$ is then called a *divergence* (or disparity, contrast function). Loosely speaking, the divergence $d(p,q)$ of p and q can be interpreted as a kind of "directed distance from p to q".[2] As already indicated above, the underlying directness turns out to be especially useful in contexts where the first component (point), say p, is always/principally of "more importance" or of "higher attention" than the second component, say q; this is nothing unusual, since after all, one of our most fundamental daily-life constituents – namely time – is directed (and therefore also time-dependent quantities)! Moreover, as a further analogue consider the "way/path-length" $d(p,q)$ a taxi would travel from point p to point q in parts of a city with at least one one-way street. Along the latter, there automatically exist points $p \neq q$ such that $d(p,q) \neq d(q,p)$; this non-equality may even hold for all $p \neq q$ if the street pattern is irregular enough; the same holds on similar systems of connected "one-way loops", directed graphs, etc. However, sometimes the application context demands for the usage of a dissimilarity $d(\cdot,\cdot)$ satisfying (D1), (D2) and

(D3) $d(p,q) = d(q,p)$ for all p, q (symmetry),

and such $d(\cdot,\cdot)$ is denoted as a *distance*; notice that we don't assume that the triangle inequality holds. Hence, we regard a distance as a symmetric divergence. Moreover, a distance $d(\cdot,\cdot)$ can be constructed from a divergence $\widetilde{d}(\cdot,\cdot)$ e.g. by means of either the three "symmetrizing operations" $d(p,q) := \widetilde{d}(p,q) + \widetilde{d}(q,p)$, $d(p,q) := \min\{\widetilde{d}(p,q), \widetilde{d}(q,p)\}$, $d(p,q) := \max\{\widetilde{d}(p,q), \widetilde{d}(q,p)\}$ for all p and q.

In many real-life applications, the numbers p, q of primary interest as well as the scaling numbers m_i of secondary interest are typically replaced by real-valued functions $x \to p(x)$, $x \to q(x)$, $x \to m_i(x)$, where $x \in \mathscr{X}$ is taken from some underlying set $\mathscr{X}$. To address the entire functions as objects we use the abbreviations $P := \{p(x)\}_{x\in\mathscr{X}}$, $Q := \{q(x)\}_{x\in\mathscr{X}}$, $M_i := \{m_i(x)\}_{x\in\mathscr{X}}$, and alternatively sometimes also $p(\cdot)$, $q(\cdot)$, $m_i(\cdot)$. This is conform with the high-level data processing paradigms in "functional programming" and "symbolic computation", where functions are basically treated as whole entities, too.

Depending on the nature of the data-analytical task, the function P of primary interest may stem either from a hypothetical model, or its analogue derived from observed/measured data, or its analogue derived from artificial computer-generated (simulated) data; the same holds for Q where "cross-over constellations" (w.r.t. to the origin of P) are possible.

The basic underlying set (space) $\mathscr{X}$ respectively the function argument x can play different roles, depending on the application context. For instance, if $\mathscr{X} \subset \mathbb{N}$ is a subset of the integers $\mathbb{N}$ then $x \in \mathscr{X}$ may be an index and $p(x)$ may describe the xth real-valued data-point. Accordingly, P is then a s-dimensional vector where s is the total number of elements in $\mathscr{X}$ with eventually allowing for $s = \infty$. In other

[1] See e.g. Weller-Fahy et al. [93].

[2] Alternatively, one can think of $d(p,q)$ as degree of proximity from p to q.

situations, x itself may be a data point of arbitrary nature (i.e. $\mathscr{X}$ can be any set) and $p(x)$ a real value attributed to x; this $p(x)$ may be of direct or of indirect use. The latter holds for instance in cases where $p(\cdot)$ is a density function (on $\mathscr{X}$) which roughly serves as a "basis" for the operationalized calculation of the "local aggregations over all[3] $A \subset \mathscr{X}$" in the sense of $A \to \sum_{x \in A} p(x)$ or $A \to \int_A p(x)\, \mathrm{d}\lambda(x)$ subject to some "integrator" $\widetilde{\lambda}(\cdot)$ (including classical Riemann integrals $\mathrm{d}\widetilde{\lambda}(x) = \mathrm{d}x$); as examples for nonnegative densities $p(\cdot) \geqslant 0$ one can take "classical" (volumetric, weights-concerning) inertial-mass densities, population densities, probability densities, whereas densities $p(\cdot)$ with possible negative values can occur in electromagnetism (charge densities, polarization densities), in other fields of contemporary physics (negative inertial-mass respectively gravitational-mass densities) as well as in the field of acoustic metamaterials (effective density), to name but a few.

Especially when used as a set of possible states/data configurations (rather than indices), $\mathscr{X}$ can be of arbitrary complexity. For instance, each x itself may be a real-valued continuous function on a time interval $[0, T]$ (i.e. $x : [0, T] \to]-\infty, \infty[$) which describes the scenario of the overall time-evolution of a quantity of interest (e.g. of a time-varying quantity in an deterministic production process of one machine, of the return on a stock, of a neural spike train). Accordingly, one can take e.g. $\mathscr{X} = C\big([0, T],]-\infty, \infty[\big)$ to be the set of all such continuous functions, and e.g. $p(\cdot)$ a density thereupon (which is then a function on functions). Other kinds of functional data analytics can be covered in an analogous fashion.

To proceed with the proximity-quantification of the primary-interest functions $P := \big\{p(x)\big\}_{x \in \mathscr{X}}, Q := \big\{q(x)\big\}_{x \in \mathscr{X}}$, in accordance with the above-mentioned investigations one can deal with the pointwise dissimilarities/divergences $d_\phi^{(j)}(p(x), q(x))$, $d^{(5)}_{m_1(x), m_2(x), m_3(x)}(p(x), q(x))$ for fixed $x \in \mathscr{X}$, but in many contexts it is crucial to take "summarizing" dissimilarities/divergences

$$D_\phi^{(j)}(P, Q) := \sum_{x \in \mathscr{X}} d_\phi^{(j)}(p(x), q(x)) \cdot \lambda(x) \quad \text{or} \quad D_\phi^{(j)}(P, Q) := \int_{\mathscr{X}} d_\phi^{(j)}(p(x), q(x))\, \mathrm{d}\lambda(x)$$

subject to some weight-type "summator"/"integrator" $\lambda(\cdot)$ (including classical Riemann integrals); analogously, one can deal with $D^{(5)}_{\phi, M_1, M_2, M_3}(P, Q) := \sum_{x \in \mathscr{X}} d^{(5)}_{m_1(x), m_2(x), m_3(x)}(p(x), q(x)) \cdot \lambda(x)$ or $D^{(5)}_{\phi, M_1, M_2, M_3}(P, Q) := \int_{\mathscr{X}} d^{(5)}_{m_1(x), m_2(x), m_3(x)}(p(x), q(x))\, \mathrm{d}\lambda(x)$. Notice that the requirements (D1), (D2) respectively (D3) carry principally over in a straightforward manner also to these pointwise and aggregated dissimilarities between the functions (rather than real points), and accordingly one calls them (pointwise/aggregated) divergences respectively distances, too.

[3] Measurable.

2.2 Divergences and Geometry

There are several ways how pointwise dissimilarities $d(\cdot,\cdot)$ respectively aggregated dissimilarities $D(\cdot,\cdot)$ between two functions $P := \{p(x)\}_{x\in\mathscr{X}}$ and $Q := \{q(x)\}_{x\in\mathscr{X}}$ can be connected with geometric issues. To start with an "all-encompassing view", following the lines of e.g. Birkhoff [14] and Millmann and Parker [50], one can build from any set $\mathscr{S}$, whose elements can be interpreted as "points", together with a collection $\mathscr{L}$ of non-empty subsets of $\mathscr{S}$, interpreted as "lines" (as a manifestation of a principle sort of structural connectivity between points), and an arbitrary *distance* $\mathfrak{d}(\cdot,\cdot)$ on $\mathscr{S}\times\mathscr{S}$, an axiomatic constructive framework of geometry which can be of far-reaching nature; therein, $\mathfrak{d}(\cdot,\cdot)$ plays basically the role of a marked ruler. Accordingly, each triplet $(\mathscr{S},\mathscr{L},\mathfrak{d}(\cdot,\cdot))$ forms a distinct "quantitative geometric system"; the most prominent classical case is certainly $\mathscr{S}=\mathbb{R}^2$ with $\mathscr{L}$ as the collection of all vertical and non-vertical lines, equipped with the Euclidean distance $\mathfrak{d}(\cdot,\cdot)$, hence generating the usual Euclidean geometry in the two-dimensional space. In the case that $\mathfrak{d}(\cdot,\cdot)$ is only an *asymmetric divergence* but not a distance anymore, we propose that some of the outcoming geometric building blocks have to be interpreted in a direction-based way (e.g. the use of $\mathfrak{d}(\cdot,\cdot)$ as a marked directed ruler, the construction of points of equal divergence from a center viewed as distorted directed spheres, etc.). For $d(\cdot,\cdot)$ one takes $\mathscr{S}\subset\mathbb{R}$ whereas for $D(\cdot,\cdot)$ one has to work with $\mathscr{S}$ being a family of real-valued functions on $\mathscr{X}$.

Secondly, from any *distance* $\mathfrak{d}(\cdot,\cdot)$ on a "sufficiently rich" set $\mathscr{S}$ and a finite number of (fixed or adaptively flexible) distinct "reference points" s_i $(i=1,\ldots,n)$ one can construct the corresponding Voronoi cells $V(s_i)$ by

$$V(s_i) := \{z\in\mathscr{S}:\ \mathfrak{d}(z,s_i)\leqslant\mathfrak{d}(z,s_j)\text{ for all } j=1,\ldots,n\}.$$

This produces a tesselation (tiling) of $\mathscr{S}$ which is very useful for classification purposes. Of course, the geometric shape of these tesselations is of fundamental importance. In the case that $\mathfrak{d}(\cdot,\cdot)$ is only an *asymmetric divergence* but not a distance anymore, then $V(s_i)$ has to be interpreted as a directed Voronoi cell and then there is also the "reversely directed" alternative

$$\widetilde{V}(s_i) := \{z\in\mathscr{S}:\ \mathfrak{d}(s_i,z)\leqslant\mathfrak{d}(s_j,z)\text{ for all } j=1,\ldots,n\}.$$

Recent applications where $\mathscr{S}\subset\mathbb{R}^d$ and $\mathfrak{d}(\cdot,\cdot)$ is a Bregman divergence or a more general conformal divergence, can be found e.g. in Boissonnat et al. [15], Nock et al. [64] (and the references therein), where they also deal with the corresponding adaption of k-nearest neighbour classification methods.

Thirdly, consider a "specific framework" where the functions $P := \widetilde{P}_{\theta_1} := \{\widetilde{p}_{\theta_1}(x)\}_{x\in\mathscr{X}}$ and $Q := \widetilde{P}_{\theta_2} := \{\widetilde{p}_{\theta_2}(x)\}_{x\in\mathscr{X}}$ depend on some parameters $\theta_1\in\Theta$, $\theta_2\in\Theta$, which reflect the strive for a complexity-reducing representation of "otherwise intrinsically complicated" functions P, Q. The way of dependence of the function (say) $\widetilde{p}_\theta(\cdot)$ on the underlying parameter θ from an appropriate space Θ

of e.g. manifold type, may show up directly e.g. via its operation/functioning as a relevant system-indicator, or it may be manifested implicitly e.g. such that $\widetilde{p}_\theta(\cdot)$ is the solution of an optimization problem with θ-involving constraints. In such a framework, one can induce divergences $D(\widetilde{P}_{\theta_1}, \widetilde{P}_{\theta_2}) =: f(\theta_1, \theta_2)$ and – under sufficiently smooth dependence – study their corresponding differential-geometric behaviour of $f(\cdot, \cdot)$ on Θ. An example is provided by the Kullback–Leibler divergence between two distributions of the same exponential family of distributions, which defines a Bregman divergence on the parameter space. This and related issues are subsumed in the research field of "information geometry"; for comprehensive overviews see e.g. Amari [3], Amari [1], Ay et al. [8]. Moreover, for recent connections between divergence-based information geometry and optimal transport the reader is e.g. referred to Pal and Wong [66, 67], Karakida and Amari [34], Amari et al. [2], Peyre and Cuturi [71], and the literature therein.

Further relations of divergences with other approaches to geometry can be overviewed e.g. from the wide-range-covering research-article collections in Nielsen and Bhatia [58], Nielsen and Barbaresco [55–57]. Finally, geometry also enters as a tool for visualizing quantitative effects on divergences.

2.3 Divergences and Uncertainty in Data

In general, data-uncertainty (including "deficiencies" like data incompleteness, fakery, unreliability, faultiness, vagueness, etc.) can enter the framework in various different ways. For instance, in situations where $x \in \mathscr{X}$ plays the role of an index (e.g. $\mathscr{X} = \{1, 2, \ldots, s\}$) and $p(x)$ describes the xth real-valued data-point, the uncertainty is typically[4] incorporated by adding a random argument $\omega \in \Omega$ to end up with the "vectors" $P(\omega) := \left\{p(x, \omega)\right\}_{x \in \mathscr{X}}$, $Q(\omega) := \left\{q(x, \omega)\right\}_{x \in \mathscr{X}}$ of random data points. Accordingly, one ends up with random-variable-type pointwise divergences $\omega \to d_\phi^{(j)}(p(x, \omega), q(x, \omega))$, $\omega \to d^{(5)}_{m_1(x), m_2(x), m_3(x)}(p(x, \omega), q(x, \omega))\,(x \in \mathscr{X})$ as well as with the random-variable-type "summarizing" divergences
$\omega \to D_\phi^{(j)}(P(\omega), Q(\omega)) := \sum_{x \in \mathscr{X}} d_\phi^{(j)}(p(x, \omega), q(x, \omega)) \cdot \lambda(x)$ respectively
$\omega \to D_\phi^{(j)}(P(\omega), Q(\omega)) := \int_{\mathscr{X}} d_\phi^{(j)}(p(x, \omega), q(x, \omega))\, \mathrm{d}\lambda(x)$, as well as with
$\omega \to D^{(5)}_{\phi, M_1, M_2, M_3}(P(\omega), Q(\omega)) := \sum_{x \in \mathscr{X}} d^{(5)}_{m_1(x), m_2(x), m_3(x)}(p(x, \omega), q(x, \omega)) \cdot \lambda(x)$, resp. $\omega \to D^{(5)}_{\phi, M_1, M_2, M_3}(P(\omega), Q(\omega)) := \int_{\mathscr{X}} d^{(5)}_{m_1(x), m_2(x), m_3(x)}(p(x, \omega), q(x, \omega))\mathrm{d}\lambda(x)$. More generally, one can allow for random scales $m_1(x, \omega), m_2(x, \omega), m_3(x, \omega)$.

In other situations with finitely-many-elements carrying $\mathscr{X}$, the state x may e.g. describe a possible outcome $Y(\omega)$ of an uncertainty-prone observation of a quantity Y of interest and $p(x), q(x)$ represent the corresponding probability mass functions ("discrete density functions") at x under two alternative probability mechanisms Pr, $\widetilde{Pr}$ (i.e. $p(x) = Pr[\{\omega \in \Omega : Y(\omega) = x\}]$, $q(x) = \widetilde{Pr}[\{\omega \in \Omega : Y(\omega) = x\}]$); as

[4] In a probabilistic approach rather than a chaos-theoretic approach.

already indicated above, $P := \left\{p(x)\right\}_{x\in\mathscr{X}}$ respectively $Q := \left\{q(x)\right\}_{x\in\mathscr{X}}$ serve then as a kind of "basis" for the computation of the probabilities $\sum_{x\in A} p(x)$ respectively $\sum_{x\in A} q(x)$ that an arbitrary event $\{\omega \in \Omega : Y(\omega) \in A\}$ $(A \subset \mathscr{X})$ occurs. Accordingly, the pointwise divergences $d_\phi^{(j)}(p(x), q(x))$, $d^{(5)}_{m_1(x),m_2(x),m_3(x)}(p(x), q(x))$ $(x \in \mathscr{X})$, and the aggregated divergences $D_\phi^{(j)}(P, Q) := \sum_{x\in\mathscr{X}} d_\phi^{(j)}(p(x), q(x))$, $D^{(5)}_{\phi,M_1,M_2,M_3}(P, Q) := \sum_{x\in\mathscr{X}} d^{(5)}_{m_1(x),m_2(x),m_3(x)}(p(x), q(x))$, $D^{(5)}_{\phi,M_1,M_2,M_3}(P, Q) := \int_{\mathscr{X}} d^{(5)}_{m_1(x),m_2(x),m_3(x)}(p(x), q(x))\,\mathrm{d}\lambda(x)$ can then be regarded as (nonnegative, reflexive) dissimilarities between the two alternative uncertainty-quantification-bases P and Q. Analogously, when e.g. $\mathscr{X} = \mathbb{R}^n$ is the n-dimensional Euclidean space and P, Q are classical probability density functions interpreted roughly via $p(x)\mathrm{d}x = Pr[\{\omega \in \Omega : Y(\omega) \in [x, x + \mathrm{d}x[\}$, $q(x)\mathrm{d}x = \widetilde{Pr}[\{\omega \in \Omega : Y(\omega) \in [x, x + \mathrm{d}x[\}$, then $d_\phi^{(j)}(p(x), q(x))$, $d^{(5)}_{m_1(x),m_2(x),m_3(x)}(p(x), q(x))$ $(x \in \mathscr{X})$, $D_\phi^{(j)}(P, Q) := \int_{\mathscr{X}} d_\phi^{(j)}(p(x), q(x))\,\mathrm{d}x$, $D^{(5)}_{\phi,M_1,M_2,M_3}(P, Q) := \int_{\mathscr{X}} d^{(5)}_{m_1(x),m_2(x),m_3(x)}(p(x), q(x))\,\mathrm{d}x$ serve as dissimilarities between the two alternative uncertainty-quantification-bases P, Q.

Let us finally mention that in concrete applications, the "degree" of intrinsic data-uncertainty may be zero (deterministic), low (e.g. small random data contamination and small random deviations from a "basically" deterministic system, slightly noisy data, measurement errors) or high (forecast of the price of a stock in one year from now). Furthermore, the data may contain "high unusualnesses" ("surprising observations") such as outliers and inliers. All this should be taken into account when choosing or even designing the right type of divergence which have different sensitivity to such issues (see e.g. Kißlinger and Stummer [37] and the references therein).

2.4 Divergences, Information and Model Uncertainty

In the main spirit of this book on geometric structures of information, let us also connect the latter with dissimilarities in a wide sense which is appropriate enough for our ambitions of universal modeling. In correspondingly adapting some conception e.g. of Buckland [20] to our above-mentioned investigations, in the following we regard a density function (say) $p(\cdot)$ as a fundamental basis of information understood as quantified real – respectively hypothetical – knowledge which can be communicated about some particular (family of) subjects or (family of) events; according to this information-as-knowledge point of view, pointwise dissimilarities/divergences/distances $d(p(x), q(x))$ $(x \in \mathscr{X})$ respectively aggregated dissimilarities/divergences/distances $D(P, Q)$ quantify the proximity between the two information-bases $P := \left\{p(x)\right\}_{x\in\mathscr{X}}$ and $Q := \left\{q(x)\right\}_{x\in\mathscr{X}}$ in a directed/nonnegative directed/nonnegative symmetric way. Hence, $d(\cdot, \cdot)$ respectively $D(\cdot, \cdot)$ themselves can be seen as a higher-level information on pairs of information bases.

Divergences can be used for the quantification of information-concerning issues for model uncertainty (model risk) and exploratory model search in various different ways. For instance, suppose that we search for (respectively learn to understand) a true unknown density function $Q^{true} := \left\{q^{true}(x)\right\}_{x\in\mathscr{X}}$ of an underlying data-generating mechanism of interest, which is often supposed to be a member of a prefixed class $\mathscr{P}$ of "hypothetical model-candidate density functions"; frequently, this task is (e.g. for the sake of fast tractability) simplified to a setup of finding the true unknown parameter $\theta = \theta_0$ – and hence $Q^{true} = Q_{\theta_0}$ – within a parametric family $\mathscr{P} := \{Q_\theta\}_{\theta\in\Theta}$. Let us first consider the case where the data-generating mechanism of interest Q^{true} is purely deterministic and hence also all the candidates $Q \in \mathscr{P}$ are (taken to be) *not* of probability-density-function type. Although one has no intrinsic data-uncertainty, one faces another type of knowledge-lack called model-uncertainty. Then, one standard goal is to "track down" (respectively learn to understand) this true unknown Q^{true} respectively Q_{θ_0} by collecting and purpose-appropriately postprocessing some corresponding data observations. Accordingly, one attempts to design a density-function-construction rule (mechanism, algorithm) $data \to P^{data} := \left\{p^{data}(x)\right\}_{x\in\mathscr{X}}$ to produce data-derived information-basis-type replica of a "comparable principal form" as the anticipated Q^{true}. This rule should theoretically guarantee that P^{data} converges – with reasonable "operational" speed – to Q^{true} as the number N^{data} of data grows, which particularly implies that (say) $D(P^{data}, Q^{true})$ for some prefixed aggregated divergence $D(\cdot,\cdot)$ becomes close to zero "fast enough". On these grounds, one reasonable strategy to down-narrow the true unknown data-generating mechanism Q^{true} is to take a prefixed class $\mathscr{P}^{hyp}$ of hypothetical density-function models and compute $infodeg := \inf_{Q\in\mathscr{P}^{hyp}} D(P^{data}, Q)$ which in the light of the previous discussions can be interpreted as an "unnormalized degree of informative evidence of Q^{true} being a member of $\mathscr{P}^{hyp}$", or from a reversed point of view, as an "unnormalized degree of goodness of approximation (respectively fit) of the data-derived density function P^{data} through/by means of $\mathscr{P}^{hyp}$". Within this current paradigm, if $infodeg$ is too large (to be specified in a context-dependent, appropriately quantified sense by taking into account the size of N^{data}), then one has to repeat the same procedure with a different class $\widetilde{\mathscr{P}^{hyp}}$; on the other hand, if (and roughly only if) $infodeg$ is small enough then $\widehat{Q^{data}} := \arg\inf_{Q\in\mathscr{P}^{hyp}} D(P^{data}, Q)$ (which may not be unique) is "the most reasonable" approximation. This procedure is repeated recursively as soon as new data points are observed.

In contrast to the last paragraph, let us now cope with the case where the true unknown data-generating mechanism of interest is prone to uncertainties (i.e. is random, noisy, risk-prone) and hence Q^{true} as well as all the candidates $Q \in \mathscr{P}$ *are* of probability-density-function type. Even more, the data-derived information-basis-type replica $\omega \to data(\omega) \to P^{data(\omega)} := \left\{p^{data(\omega)}(x)\right\}_{x\in\mathscr{X}}$ of Q^{true} is now a density-function-valued (!) random variable; notice that in an above-mentioned "full-scenario" time-evolutionary context, this becomes a density-function-on-functions-valued random variable. Correspondingly, the above-mentioned procedure for the deterministic case has to be adapted and the notions

of convergence and smallness have to be stochastified, which leads to the need of considerably more advanced techniques.

Another field of applying divergences to a context of synchronous model and data uncertainty is Bayesian sequential updating. In such a "doubly uncertain" framework, one deals with a parametric context of probability density functions $Q^{true} = Q_{\theta_0}$, $\mathscr{P} := \{Q_\theta\}_{\theta\in\Theta}$ where the uncertain knowledge about the parameter θ (to be learnt) is operationalized by replacing it with a random variable ϑ on Θ. Based on both (i) an initial prior distribution $Prior_1[\cdot] := Pr[\vartheta \in \cdot\,]$ of ϑ (with probability density function pdf $\theta \to prior_1(\theta)$) and (ii) observed data $data_1(\omega), \ldots, data_{N^{data}}(\omega)$ of number N^{data}, a posterior distribution $Post_1[\cdot, \omega] := Pr[\vartheta \in \cdot \,|\, data_1(\omega), \ldots, data_{N^{data}}(\omega);\ prior[\cdot]\,]$ of ϑ (with pdf $\theta \to post_1(\theta, \omega)$) is determined with (amongst other things) the help of the well-known Bayes formula. This procedure is repeated recursively with new incoming data input (block) $data_{N^{data}+1}$, where the new prior distribution $Prior_2[\cdot, \omega] := Post_1[\cdot, \omega]$ is chosen as the old posterior and the new posterior distribution is $Post_2[\cdot, \omega] := Pr[\vartheta \in \cdot \,|\, data_1, \ldots, data_{N^{data}}, data_{N^{data}+1};\ Prior_2[\cdot, \omega]\,]$ (with pdf $\theta \to post_2(\theta, \omega)$), etc. The corresponding (say) aggregated divergence $D(P(\omega), Q(\omega))$ between the probability-density-valued random variables $\omega \to P(\omega) := \left\{prior_2(\theta, \omega)\right\}_{\theta\in\Theta}$, and $\omega \to Q(\omega) := \left\{post_2(\theta, \omega)\right\}_{\theta\in\Theta}$ serves as "degree of informativity of the new data-point observation on the learning of the true unknown θ_0".

As another application in a "doubly uncertain" framework, divergences $D(P, Q)$ appear also in a dichotomous Bayesian testing problem between the two alternative probability densities functions P and Q, where $D(P, Q)$ represents an appropriate average (over prior probabilities) of the corresponding difference between the prior Bayes risk (prior minimal mean decision loss) and the posterior Bayes risk (posterior minimal mean decision loss). This, together with non-averaging versions and an interpretation of $D(P, Q)$ as a (weighted-average) statistical information measure in the sense of De Groot [29] can be found e.g. in Österreicher and Vajda [65]; see also Stummer [78–80], Liese and Vajda [42], Reid and Williamson [73]. In contrast of this employment of $D(P, Q)$ as quantifier of "decision risk reduction" respectively "model risk reduction" respectively "information gain", a different use of divergences $D(P, Q)$ in a "double uncertain" general Bayesian context of dichotomous loss-dependent decisions between arbitrary probability density functions P and Q can be found in Stummer and Vajda [81], where they achieve $D_{\phi_\alpha}(P, Q)$ (for some power functions ϕ_α cf. (5)) as upper and lower bound of the Bayes risk (minimal mean decision loss) itself and also give applications to decision making of time-continuous, non-stationary financial stochastic processes.

Divergences can be also employed to detect distributional changes in streams (respectively clouds) $(data_j)_{j\in\tau}$ of uncertain (random, noisy, risk-prone) data indexed by j from an arbitrary countable set τ (e.g. the integers, an undirected graph); a survey together with some general framework can be found in Kißlinger and Stummer [38]: the basic idea is to pick out two[5] non-identical, purpose-appropriately chosen subcollections respectively sample patterns (e.g. windows)

[5] Where one of them may e.g. stem from training data.

$data_{one}(\omega) := (data_{s_1}(\omega), \ldots, data_{s_{N_1}}(\omega))$, $data_{two}(\omega) := (data_{t_1}(\omega), \ldots, data_{t_{N_2}}(\omega))$, and to build from them data-derived probability-density functions $\omega \to data_{one}(\omega) \to P^{data_{one}(\omega)} := \left\{p^{data_{one}(\omega)}(x)\right\}_{x\in\mathscr{X}}$, $\omega \to data_{two}(\omega) \to P^{data_{two}(\omega)} := \left\{p^{data_{two}(\omega)}(x)\right\}_{x\in\mathscr{X}}$. If a correspondingly chosen (say) aggregated divergence $D\left(P^{data_{one}(\omega)}, P^{data_{two}(\omega)}\right)$ – which plays the role of a condensed change-score – is "significantly large" in the sense that it is large enough – compared to some sound threshold which within the model reflects the desired "degree of confidential plausibility" – then there is strong indication of a distributional change which we then "believe in". Notice that both components of the divergence $D(\cdot,\cdot)$ are now probability-density-function-valued random variables. The sound threshold can e.g. be derived from advanced random asymptotic theory.

From the above discussion it is clear that divergence-based model-uncertainty methods are useful tools in concrete applications for machine learning and artificial intelligence, see e.g. Collins et al. [25], Murata et al. [54], Banerjee et al. [9], Tsuda et al. [87], Cesa-Bianchi and Lugosi [21], Nock and Nielsen [63], Sugiyama et al. [85], Wu et al. [94], Nock et al. [62], Nielsen et al. [60], respectively Minka [51], Cooper et al. [26], Lizier [46], Zhang et al. [96], Chhogyal [22], Cliff et al. [23, 24].

3 General Framework

For the rest of this paper, we shall use the following

Main (i.e. non-locally used) Notation and Symbols

$\mathbb{R}, \mathbb{N}, \mathbb{R}^d$	Set of real respectively integer numbers respectively d-dimensional vectors
Θ, θ	Set of parameters, see p. 188
$\mathbb{1}$	Function with constant value 1
$\mathbf{1}_A(z) = \delta_z[A]$	Indicator function on the set A evaluated at data point z, which is equal to Dirac's one-point distribution on z evaluated at A
$\#A$	Number of elements in set A
$\mathscr{X}$; $\mathscr{X}_\#$	Space/set where data can take values in; space/set of countable size
$\mathscr{F}$	System of admissible events/data-collections (σ-algebra) on $\mathscr{X}$
λ	Reference measure/integrator/summator, see p. 160 & Sect. 3.1 on p. 165
λ-a.a.	λ-almost all, see p. 160
λ_L	Lebesgue measure ("Riemann-type" integrator), see p. 160, & Sect. 3.1
$\lambda_\#$	Counting measure ("classical summator"), see p. 160 & Sect. 3.1 on p. 165
$P := \left\{p(x)\right\}_{x\in\mathscr{X}}$	Function from which the divergence/dissimilarity is measured from, see p. 160
$Q := \left\{q(x)\right\}_{x\in\mathscr{X}}$	Function to which the divergence/dissimilarity is measured to, see p. 160
$M_i := \left\{m_i(x)\right\}_{x\in\mathscr{X}}$	Scaling function ($i = 1, 2$) respec. aggregation function ($i = 3$), see p. 161, (1) and paragraph (I1) thereafter, as well as Sect. 3.3 on p. 170
$p(\cdot), q(\cdot), m_i(\cdot)$, $R := \left\{r(x)\right\}_{x\in\mathscr{X}}$	Alternative representations of P, Q, M_i Function used for the aggregation function $m_3(\cdot)$, see Sect. 3.3.1 on p. 171
W_i	Connector function of the form $W_i := \left\{w_i(x, y, z)\right\}_{x,y,z\in\ldots}$, for adaptive scaling and aggregation functions $m_i(x) = w_i(x, p(x), q(x))$ $(i = 1, 2, 3)$, see e.g. Assumption 2 on p. 163 and Sect. 3.3.1.3 on p. 181

$\mathbb{P}, \mathbb{Q}, \mathbb{M}_i, \mathbb{W}_i$	Functions with $\mathbb{p}(x) \geqslant 0$, $\mathbb{q}(x) \geqslant 0$, $\mathbb{m}_i(x) \geqslant 0$, $\mathbb{w}_i(x) \geqslant 0$ for λ-a.a. $x \in \mathscr{X}$
$\mathbb{Q}^\chi := \{\mathbb{q}^\chi(x)\}_{x\in\mathscr{X}}$	Function for the aggregation function $m_3(\cdot)$, see Sect. 4.2 on p. 184, (73)
$\mathbb{P}, \mathbb{Q}$	λ-probability density functions (incl. probability mass functions for $\lambda = \lambda_\#$), i.e. for which $\mathbb{p}(x) \geqslant 0$, $\mathbb{q}(x) \geqslant 0$ for λ-a.a. $x \in \mathscr{X}$ and $\int_{\mathscr{X}} \mathbb{p}(x)\,\mathrm{d}\lambda(x) = 1$, see Remark 2 on p. 172
$\vec{\mathbb{Q}}_\theta := \{\vec{\mathbb{q}}_\theta(x)\}_{x\in\mathscr{X}}$	λ-probab. density function which depends on a parameter $\theta \in \Theta$, see p. 188
$\mathscr{R}\left(\frac{P}{M_1}\right)$	Range (image) of the function $\left\{\frac{p(x)}{m_1(x)}\right\}_{x\in\mathscr{X}}$, see paragraph (I2) on p. 161
$\mathscr{R}(Y_1, \ldots, Y_N)$	Range (image) of the random variables $Y_1, \ldots, Y_N$, see p. 182
$\widetilde{\mathbb{Q}}_\theta := \{\widetilde{\mathbb{q}}_\theta(x)\}_{x\in\mathscr{X}}$	λ-probab. density function (modification of $\vec{\mathbb{Q}}_\theta$) defined by $\widetilde{\mathbb{q}}_\theta(x) := \vec{\mathbb{q}}_\theta(x) \cdot (1 - \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x))$, see p. 191
$\phi := \{\phi(t)\}_{t\in]a,b[}$	Divergence generator, a convex real-valued function on $]a, b[$, see p. 161, (1) and paragraph (I2), as well as Sect. 3.2 on p. 165
$\boldsymbol{\Phi}(]a, b[)$;	Class of all such ϕ, see paragraph (I2) on p. 161
$\overline{\phi} := \{\phi(t)\}_{t\in[a,b]}$	Continuous extension of ϕ on $[a, b]$, with $\overline{\phi}(t) = \phi(t)$ for all $t \in]a, b[$, see (I2)
$\phi'_{+,c}(t)$	c-weighted mixture of left-hand and right-hand derivative of ϕ at t, see (I2)
$\boldsymbol{\Phi}_{C_1}(]a, b[)$	Subclass of everywhere continuously differentiable ϕ, with derivative $\phi'(t)$ (being equal to $\phi'_{+,c}(t)$ for all $c \in [0, 1]$), see (I2) on p. 161
ϕ_α	α-power-function type divergence generator, see (5) on p. 166, (14), (18), (19)
ϕ_{TV}	Generator of total variation distance, see (31) on p. 169
ϕ_{ie}	Divergence generator with interesting effects, see (35) on p. 170
$\underline{\psi_{\phi,c}}$	Function given by $\psi_{\phi,c}(s, t) := \phi(s) - \phi(t) - \phi'_{+,c}(t) \cdot (s - t) \geqslant 0$, see (I2)
$\overline{\psi_{\phi,c}}$	Bivariate extension of $\psi_{\phi,c}$, see (I2) on p. 161
$\overline{\int}_{\mathscr{X}} \ldots, \overline{\sum}_{\mathscr{X}} \ldots$	Integral/sum over extension of integrand/summand . . ., see (I2) & (2) on p. 165
$D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P, Q)$	Divergence between two functions P (scaled by M_1) and Q (scaled by M_2), generated by ϕ and weight c, and aggregated by $\mathbb{M}_3$ and λ, see (1) on p. 161
$D_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P, Q)$	As above, but with $\phi \in \boldsymbol{\Phi}_{C_1}(]a, b[)$ and obsolete c, see Sect. 3.2 on p. 165
$D_\lambda(P, Q)$	General λ-aggregated divergence, see p. 189, respectively pseudo-divergence, see Definition 2 on p. 195
$D_{\mathbb{M},\lambda}(\vec{\mathbb{P}}, \vec{\mathbb{Q}})$	Pointwise decomposable pseudo-divergence, scaled by $\mathbb{M}$ and aggregated by $\mathbb{M}$ and λ, see Sect. 4.6 on p. 200
NN0, NN1	Nonnegativity setup 0 respectively 1, see p. 166 resp. p. 171
$\mathfrak{P}^{\mathbb{R}\cdot\lambda}, \mathfrak{Q}^{\mathbb{R}\cdot\lambda}, \mathfrak{M}^{\mathbb{R}\cdot\lambda}$	Measures with λ-densities $\mathbb{p}(\cdot) \cdot \mathbb{r}(\cdot)$, $\mathbb{q}(\cdot) \cdot \mathbb{r}(\cdot)$, $\mathbb{m}(\cdot) \cdot \mathbb{r}(\cdot)$, see Remark 2 on p. 171
$\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda}, \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda}$	Probability measures (distributions) with λ-densities $\vec{\mathbb{p}}(\cdot)$, $\vec{\mathbb{q}}(\cdot)$, see Remark 2
$\mathscr{Q}^{\lambda_2}_\Theta, \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_2}_\theta$	Class of probability measures with λ_2-densities $\vec{\mathbb{q}}_\theta(\cdot)$ with parameter $\theta \in \Theta$, see p. 188
$\vec{\mathfrak{P}}^{emp}_N, \vec{\mathbb{P}}^{emp}_N, \vec{\mathbb{p}}^{emp}_N(\cdot)$	Data-derived empirical (probability) distribution, and probability mass function ($\lambda_\#$-density) thereof, see Remark 2 on p. 172
$\vec{\mathfrak{P}}^{\overline{emp}(\omega)}_N, \vec{\mathbb{P}}^{\overline{emp}(\omega)}_N$	Data-derived "extended" empirical (probability) distribution, and probability mass function thereof, see (85) on p. 190 and thereafter
DPD, CASD	Density-power divergences (see p. 174), Csiszar–Ali–Silvey divergences (see p. 177)
$\ell i_1, \phi^*(0), \ell i_2, \ell i_3$	Certain limits, see (50), (71), (72)
$\mathbb{P} \perp \mathbb{Q}$	The functions $\mathbb{P}, \mathbb{Q}$ are "essentially different", see (64) to (66) and thereafter
$\mathbb{P} \not\perp \mathbb{Q}$	Negation of $\mathbb{P} \perp \mathbb{Q}$, see p. 192
$\mathbb{P} \sim \mathbb{Q}$	The functions $\mathbb{P}, \mathbb{Q}$ are "equivalent" (concerning zeros), see (80)
$\mathbb{P} \nsim \mathbb{Q}$	Negation of $\mathbb{P} \sim \mathbb{Q}$, see p. 195
$\widehat{\theta}_{N,D_{\lambda_2}}$	Minimum-divergence estimator ("approximator") of the true unknown parameter θ_0, based on N data observations, see (82) on p. 189

$\widehat{\theta}_{N,D_{\lambda_\#}}, \widehat{\theta}_{N,D_\lambda}$	Certain minimum-divergence estimators, see (83), (86)
$\widehat{\theta}_{N,decD_\lambda}, \widehat{\theta}_{N,decD_{\underrightarrow{\mathbb{Q}}_\tau,\lambda}}$	Certain minimum-divergence estimators, see (107), (123)
$\widehat{\theta}_{N,sup\mathscr{D}_{\phi,\lambda}}$	Certain minimum-divergence estimator, see (135)
$\mathscr{P}^\lambda$	Certain class of nonnegative, mutually equivalent functions, see p. 194
$\mathscr{P}^{\lambda\sim}, \widetilde{\mathscr{P}}^\lambda$	Certain classes of nonnegative functions, see p. 194
$\mathscr{P}^\lambda_\Theta, \mathscr{P}^{\lambda\perp}_{emp}, \mathscr{P}^\lambda_{\Theta,emp}$	Certain classes of nonnegative functions, see p. 195
$\mathfrak{D}^0, \mathfrak{D}^1, \rho_{\mathbb{Q}}$	Functionals and mapping for decomposable pseudo-divergences, see Definition 3 on p. 195
$\psi^{dec}, \psi^0, \psi^1, \rho$	Mappings for pointwise decomposable pseudo-divergences, see Definition 3 on p. 196
h_0, h_1, h_2	Mappings for pointwise decomposable pseudo-divergences, see Definition 3 on p. 196
ψ^{dec}_m	Perspective function of ψ^{dec}, see (120)

New Divergence Toolkit

In the above Sect. 2, we have motivated that for many different tasks within a broad spectrum of situations, it is useful to employ divergences as "directed distances", including distances as their symmetric special case. For the rest of the paper, we shall only deal with aggregated forms of divergences, and thus drop the attribute "aggregated" from now on. In the following, we present a fairly universal, flexible, multi-component system of divergences by adapting and widening the concept of scaled Bregman divergences of Stummer [81] and Stummer and Vajda [84] to the current context of arbitrary (measurable) functions. To begin with, let us assume that the modeled respectively observed (random) data take values in a state space $\mathscr{X}$ (with at least two distinct values), equipped with a system $\mathscr{F}$ of admissible events (σ-algebra) and a σ-finite measure λ (e.g. the Lebesgue measure, the counting measure, etc.). Furthermore, we suppose that $x \to p(x) \in [-\infty, \infty]$ and $x \to q(x) \in [-\infty, \infty]$ are (correspondingly measurable) functions on $\mathscr{X}$ which satisfy $p(x) \in]-\infty, \infty[$, $q(x) \in]-\infty, \infty[$ for λ-almost all (abbreviated as λ-a.a.) $x \in \mathscr{X}$.[6] To address the entire functions as objects we write $P := \{p(x)\}_{x\in\mathscr{X}}$, $Q := \{q(x)\}_{x\in\mathscr{X}}$ and alternatively sometimes also $p(\cdot), q(\cdot)$. To better highlight the very important special case of λ-*probability density functions* – where $p(x) \geqslant 0, q(x) \geqslant 0$ for λ-a.a. $x \in \mathscr{X}$ and $\int_{\mathscr{X}} p(x)\,\mathrm{d}\lambda(x) = 1, \int_{\mathscr{X}} q(x)\,\mathrm{d}\lambda(x) = 1$ – we use the notation $\vec{\mathbb{P}}, \vec{\mathbb{p}}, \vec{\mathbb{Q}}, \vec{\mathbb{q}}$ instead of P, p, Q, q (where $\vec{\cdot\cdot}$ symbolizes a lying 1). For instance, if $\lambda = \lambda_L$ is the Lebesgue measure on the s-dimensional Euclidean space $\mathscr{X} = \mathbb{R}^s$, then $\vec{\mathbb{P}}, \vec{\mathbb{Q}}$ are "classical" (e.g. Gaussian) probability density functions. In contrast, in the *discrete setup* where the state space (i.e. the set of all possible data points) $\mathscr{X} = \mathscr{X}_\#$ has countably many elements and $\lambda := \lambda_\#$ is the counting measure (i.e., $\lambda_\#[\{x\}] = 1$ for all $x \in \mathscr{X}_\#$), then $\vec{\mathbb{P}}, \vec{\mathbb{Q}}$ are probability mass functions and (say) $\vec{\mathbb{p}}(x)$ can be interpreted as probability that the data point x is taken by the underlying random (uncertainty-prone) mechanism. If $p(x) \geqslant 0$, $q(x) \geqslant 0$ for λ-a.a. $x \in \mathscr{X}$ (but not necessarily with the restrictions $\int_{\mathscr{X}} p(x)\,\mathrm{d}\lambda(x) = 1 = \int_{\mathscr{X}} q(x)\,\mathrm{d}\lambda(x)$) then we write $\mathbb{P}, \mathbb{Q}, \mathbb{p}, \mathbb{q}$ instead of P, p, Q, q.

[6] This means that there exists a $N \in \mathscr{F}$ with $\lambda[N] = 0$ (where the empty set $N = \emptyset$ is allowed) such that for all $x \in \mathscr{X}\backslash\{N\}$ (say) $p(x) \in]-\infty, \infty[$ holds.

Back to generality, we quantify the dissimilarity between the two functions P, Q in terms of divergences $D^c_\beta(P, Q)$ with $\beta = (\phi, M_1, M_2, \mathbb{M}_3, \lambda)$, defined by

$$0 \leqslant D^c_{\phi, M_1, M_2, \mathbb{M}_3, \lambda}(P, Q)$$
$$:= \int_{\mathscr{X}} \left[\phi\left(\tfrac{p(x)}{m_1(x)}\right) - \phi\left(\tfrac{q(x)}{m_2(x)}\right) - \phi'_{+,c}\left(\tfrac{q(x)}{m_2(x)}\right) \cdot \left(\tfrac{p(x)}{m_1(x)} - \tfrac{q(x)}{m_2(x)}\right)\right] \cdot \mathbb{m}_3(x)\, \mathrm{d}\lambda(x) \quad (1)$$

(see Stummer [81], Stummer and Vajda [84] for the case $c = 1, m_1(x) = m_2(x) = \mathbb{m}_3(x)$). Here, we use:

(I1) (measurable) *scaling functions* $m_1 : \mathscr{X} \to [-\infty, \infty]$ and $m_2 : \mathscr{X} \to [-\infty, \infty]$ as well as a nonnegative (measurable) *aggregating function* $\mathbb{m}_3 : \mathscr{X} \to [0, \infty]$ such that $m_1(x) \in]-\infty, \infty[$, $m_2(x) \in]-\infty, \infty[$, $\mathbb{m}_3(x) \in [0, \infty[$ for λ-a.a. $x \in \mathscr{X}$.[7] In accordance with the above notation, we use the symbols $M_i := \{m_i(x)\}_{x \in \mathscr{X}}$ respectively $m_i(\cdot)$ to refer to the entire functions, and $\mathbb{M}_i$, $\mathbb{m}_i(\cdot)$ when they are nonnegative as well as $\overset{\rightarrow}{\mathbb{M}}_i$, $\overset{\rightarrow}{\mathbb{m}}_i(\cdot)$ when they manifest λ-probability density functions. Furthermore, let us emphasize that we allow for / cover adaptive situations in the sense that all three functions $m_1(x), m_2(x), \mathbb{m}_3(x)$ (evaluated at x) may also depend on $p(x)$ and $q(x)$.

(I2) the so-called "divergence-generator" ϕ which is a continuous, convex (finite) function $\phi : E \to]-\infty, \infty[$ on some appropriately chosen open interval $E =]a, b[$ such that $[a, b]$ covers (at least) the union $\mathscr{R}\left(\frac{P}{M_1}\right) \cup \mathscr{R}\left(\frac{Q}{M_2}\right)$ of both ranges $\mathscr{R}\left(\frac{P}{M_1}\right)$ of $\left\{\frac{p(x)}{m_1(x)}\right\}_{x \in \mathscr{X}}$ and $\mathscr{R}\left(\frac{Q}{M_2}\right)$ of $\left\{\frac{q(x)}{m_2(x)}\right\}_{x \in \mathscr{X}}$; for instance, $E =]0, 1[$, $E =]0, \infty[$ or $E =]-\infty, \infty[$; the class of all such functions will be denoted by $\Phi(]a, b[)$. Furthermore, we assume that ϕ is continuously extended to $\overline{\phi} : [a, b] \to [-\infty, \infty]$ by setting $\overline{\phi}(t) := \phi(t)$ for $t \in]a, b[$ as well as $\overline{\phi}(a) := \lim_{t \downarrow a} \phi(t)$, $\overline{\phi}(b) := \lim_{t \uparrow b} \phi(t)$ on the two boundary points $t = a$ and $t = b$. The latter two are the only points at which infinite values may appear. Moreover, for any fixed $c \in [0, 1]$ the (finite) function $\phi'_{+,c} :]a, b[\to]-\infty, \infty[$ is well-defined by $\phi'_{+,c}(t) := c \cdot \phi'_+(t) + (1 - c) \cdot \phi'_-(t)$, where $\phi'_+(t)$ denotes the (always finite) right-hand derivative of ϕ at the point $t \in]a, b[$ and $\phi'_-(t)$ the (always finite) left-hand derivative of ϕ at $t \in]a, b[$. If $\phi \in \Phi(]a, b[)$ is also continuously differentiable – which we denote by $\phi \in \Phi_{C_1}(]a, b[)$ – then for all $c \in [0, 1]$ one gets $\phi'_{+,c}(t) = \phi'(t)$ $(t \in]a, b[)$ and in such a situation we always suppress the obsolete indices $c, +$ in the corresponding expressions. We also employ the continuous continuation $\overline{\phi'_{+,c}} : [a, b] \to [-\infty, \infty]$ given by $\overline{\phi'_{+,c}}(t) := \phi'_{+,c}(t)$ $(t \in]a, b[)$, $\overline{\phi'_{+,c}}(a) := \lim_{t \downarrow a} \phi'_{+,c}(t)$, $\overline{\phi'_{+,c}}(b) := \lim_{t \uparrow b} \phi'_{+,c}(t)$. To explain the precise meaning of (1), we also make use of the (finite, nonnegative) function $\psi_{\phi,c} :]a, b[\times]a, b[\to [0, \infty[$ given by $\psi_{\phi,c}(s, t) := \phi(s) - \phi(t) - \phi'_{+,c}(t) \cdot (s - t) \geqslant 0$ $(s, t \in]a, b[)$. To extend this to a lower semi-continuous

[7] As an example, let $\mathscr{X} = \mathbb{R}$, $\lambda = \lambda_L$ be the Lebesgue measure (and hence, except for rare cases, the integral turns into a Riemann integral) and $\mathbb{m}_1(x) := \frac{1}{2} \cdot x^{-1/2} \cdot \mathbf{1}_{[0,1]}(x) \geqslant 0$; since $\int_{\mathscr{X}} \mathbb{m}_1(x)\, \mathrm{d}\lambda(x) = 1$ this qualifies as a probability density and thus is a possible candidate for $\overset{\rightarrow}{\mathbb{m}}_1(x) = \vec{q}_1(x)$ in Sect. 3.3.1.2 below.

function $\overline{\psi_{\phi,c}} : [a,b] \times [a,b] \to [0,\infty]$ we proceed as follows: firstly, we set $\overline{\psi_{\phi,c}}(s,t) := \psi_{\phi,c}(s,t)$ for all $s,t \in]a,b[$. Moreover, since for fixed $t \in]a,b[$, the function $s \to \psi_{\phi,c}(s,t)$ is convex and continuous, the limit $\overline{\psi_{\phi,c}}(a,t) := \lim_{s \to a} \psi_{\phi,c}(s,t)$ always exists and (in order to avoid overlines in (1)) will be interpreted/abbreviated as $\phi(a) - \phi(t) - \phi'_{+,c}(t) \cdot (a-t)$. Analogously, for fixed $t \in]a,b[$ we set $\overline{\psi_{\phi,c}}(b,t) := \lim_{s \to b} \psi_{\phi,c}(s,t)$ with corresponding shorthand notation $\phi(b) - \phi(t) - \phi'_{+,c}(t) \cdot (b-t)$. Furthermore, for fixed $s \in]a,b[$ we interpret $\phi(s) - \phi(a) - \phi'_{+,c}(a) \cdot (s-a)$ as

$$\overline{\psi_{\phi,c}}(s,a) := \Big\{\phi(s) - \overline{\phi'_{+,c}}(a) \cdot s + \lim_{t \to a} \big(t \cdot \overline{\phi'_{+,c}}(a) - \phi(t)\big)\Big\} \cdot \mathbf{1}_{]-\infty,\infty[}\big(\overline{\phi'_{+,c}}(a)\big) + \infty \cdot \mathbf{1}_{\{-\infty\}}\big(\overline{\phi'_{+,c}}(a)\big),$$

where the involved limit always exists but may be infinite. Analogously, for fixed $s \in]a,b[$ we interpret $\phi(s) - \phi(b) - \phi'_{+,c}(b) \cdot (s-b)$ as

$$\overline{\psi_{\phi,c}}(s,b) := \Big\{\phi(s) - \overline{\phi'_{+,c}}(b) \cdot s + \lim_{t \to b} \Big(t \cdot \overline{\phi'_{+,c}}(b) - \phi(t)\Big)\Big\} \cdot \mathbf{1}_{]-\infty,\infty[}\big(\overline{\phi'_{+,c}}(b)\big) + \infty \cdot \mathbf{1}_{\{+\infty\}}\big(\overline{\phi'_{+,c}}(b)\big),$$

where again the involved limit always exists but may be infinite. Finally, we always set $\overline{\psi_{\phi,c}}(a,a) := 0$, $\overline{\psi_{\phi,c}}(b,b) := 0$, and $\overline{\psi_{\phi,c}}(a,b) := \lim_{s \to a} \overline{\psi_{\phi,c}}(s,b)$, $\overline{\psi_{\phi,c}}(b,a) := \lim_{s \to b} \overline{\psi_{\phi,c}}(s,a)$. Notice that $\overline{\psi_{\phi,c}}(\cdot,\cdot)$ is lower semicontinuous but not necessarily continuous. Since ratios are ultimately involved, we also consistently take $\overline{\psi_{\phi,c}}\big(\frac{0}{0},\frac{0}{0}\big) := 0$. Taking all this into account, we interpret $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)$ as $\int_{\mathscr{X}} \overline{\psi_{\phi,c}}\big(\frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big) \mathbb{m}_3(x)\, \mathrm{d}\lambda(x)$ at first glance (see further investigations in Assumption 2 below), and use the (in lengthy examples) less clumsy notation $\overline{\int}_{\mathscr{X}} \psi_{\phi,c}\big(\frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big) \mathbb{m}_3(x)\, \mathrm{d}\lambda(x)$ as a shortcut for the implicitly involved boundary behaviour. □

Notice that despite of the "difference-structure" in the integrand of (1), the splitting of the integral into differences of several "autonomous" integrals may not always be feasible due to the possible appearance of differences between infinite integral values. Furthermore, there is non-uniqueness in the construction (1); for instance, one (formally) gets $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) = D^c_{\tilde{\phi},M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)$ for any $\tilde{\phi}(t) := \phi(t) + c_1 + c_2 \cdot t \quad (t \in E)$ with $c_1, c_2 \in \mathbb{R}$. Moreover, there exist "essentially different" pairs $(\phi, \mathbb{M})$ and $(\breve{\phi}, \breve{\mathbb{M}})$ (where $\phi(t) - \breve{\phi}(t)$ is nonlinear in t) for which $D^c_{\phi,\mathbb{M},\mathbb{M},\mathbb{M},\lambda}(P,Q) = D^c_{\breve{\phi},\breve{\mathbb{M}},\breve{\mathbb{M}},\breve{\mathbb{M}},\lambda}(P,Q)$ (see e.g. [37]). Let us also mention that we could further generalize (1) by adapting the divergence concept of Stummer and Kißlinger [82] who also deal even with non-convex non-concave divergence generators ϕ; for the sake of brevity, this is omitted here.

Notice that by construction we obtain the following important assertion:

Theorem 1 *Let $\phi \in \Phi(]a,b[)$ and $c \in [0,1]$. Then there holds $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) \geqslant 0$ with equality if $\frac{p(x)}{m_1(x)} = \frac{q(x)}{m_2(x)}$ for λ-almost all $x \in \mathscr{X}$. Depending on the concrete situation, $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)$ may take infinite value.*

To get "sharp identifiability" (i.e. reflexivity) one needs further assumptions on $\phi \in \Phi(]a,b[)$, $c \in [0,1]$. As a motivation, consider the case where $\mathbb{m}_3(x) \equiv 1$ and $\phi \in \Phi(]a,b[)$ is affine linear on the whole interval $]a,b[$, and hence its extension $\overline{\phi}$ is affine-linear on $[a,b]$. Accordingly, one gets for the integrand-builder $\overline{\psi_{\phi,c}}(s,t) \equiv 0$ and hence $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) = \int_{\mathscr{X}} \overline{\psi_{\phi,c}}\big(\frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big)\,\mathrm{d}\lambda(x) = 0$ even in cases where $\frac{p(x)}{m_1(x)} \neq \frac{q(x)}{m_2(x)}$ for λ-a.a. $x \in \mathscr{X}$. In order to avoid such and similar phenomena, we use the following set of requirements:

Assumption 2 Let $c \in [0,1]$, $\phi \in \Phi(]a,b[)$ and $\mathscr{R}\big(\frac{P}{M_1}\big) \cup \mathscr{R}\big(\frac{Q}{M_2}\big) \subset [a,b]$. The aggregation function is supposed to be of the form $\mathbb{m}_3(x) = \mathbb{w}_3\big(x, \frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big)$ for some (measur.) function $\mathbb{w}_3 : \mathscr{X} \times [a,b] \times [a,b] \to [0,\infty]$. Moreover, for all $s \in \mathscr{R}\big(\frac{P}{M_1}\big)$, all $t \in \mathscr{R}\big(\frac{Q}{M_2}\big)$ and λ-a.a. $x \in \mathscr{X}$, let the following conditions hold:

(a) ϕ is strictly convex at t;
(b) if ϕ is differentiable at t and $s \neq t$, then ϕ is not affine-linear on the interval $[\min(s,t), \max(s,t)]$ (i.e. between t and s);
(c) if ϕ is not differentiable at t, $s > t$ and ϕ is affine linear on $[t,s]$, then we exclude $c = 1$ for the ("globally/universally chosen") subderivative $\phi'_{+,c}(\cdot) = c \cdot \phi'_+(\cdot) + (1-c) \cdot \phi'_-(\cdot)$;
(d) if ϕ is not differentiable at t, $s < t$ and ϕ is affine linear on $[s,t]$, then we exclude $c = 0$ for $\phi'_{+,c}(\cdot)$;
(e) $\mathbb{w}_3(x,s,t) < \infty$;
(f) $\mathbb{w}_3(x,s,t) > 0$ if $s \neq t$;
(g) $\mathbb{w}_3(x,a,a) \cdot \psi_{\phi,c}(a,a) := 0$ by convention (even in cases where the function $\mathbb{w}_3(x,\cdot,\cdot) \cdot \psi_{\phi,c}(\cdot,\cdot)$ is not continuous on the boundary point (a,a));
(h) $\mathbb{w}_3(x,b,b) \cdot \psi_{\phi,c}(b,b) := 0$ by convention (even in cases where the function $\mathbb{w}_3(x,\cdot,\cdot) \cdot \psi_{\phi,c}(\cdot,\cdot)$ is not continuous on the boundary point (b,b));
(i) $\mathbb{w}_3(x,a,t) \cdot \psi_{\phi,c}(a,t) > 0$, where $\mathbb{w}_3(x,a,t) \cdot \psi_{\phi,c}(a,t) := \lim_{s\to a} \mathbb{w}_3(x,s,t) \cdot \psi_{\phi,c}(s,t)$ if this limit exists, and otherwise we set by convention $\mathbb{w}_3(x,a,t) \cdot \psi_{\phi,c}(a,t) := 1$ (or any other strictly positive constant);
(j) $\mathbb{w}_3(x,b,t) \cdot \psi_{\phi,c}(b,t) > 0$, where $\mathbb{w}_3(x,b,t) \cdot \psi_{\phi,c}(b,t)$ is analogous to (i);
(k) $\mathbb{w}_3(x,s,a) \cdot \psi_{\phi,c}(s,a) > 0$, where $\mathbb{w}_3(x,s,a) \cdot \psi_{\phi,c}(s,a) := \lim_{t\to a} \mathbb{w}_3(x,s,t) \cdot \psi_{\phi,c}(s,t)$ if this limit exists, and otherwise we set by convention $\mathbb{w}_3(x,s,a) \cdot \psi_{\phi,c}(s,a) := 1$ (or any other strictly positive constant);
(l) $\mathbb{w}_3(x,s,b) \cdot \psi_{\phi,c}(s,b) > 0$, where $\mathbb{w}_3(x,s,b) \cdot \psi_{\phi,c}(s,b)$ is analogous to (k);
(m) $\mathbb{w}_3(x,a,b) \cdot \psi_{\phi,c}(a,b) > 0$, where $\mathbb{w}_3(x,a,b) \cdot \psi_{\phi,c}(a,b) := \lim_{s\to a} \mathbb{w}_3(x,s,b) \cdot \psi_{\phi,c}(s,b)$ if this limit exists, and otherwise we set by convention $\mathbb{w}_3(x,a,b) \cdot \psi_{\phi,c}(a,b) := 1$ (or any other strictly positive constant);
(n) $\mathbb{w}_3(x,b,a) \cdot \psi_{\phi,c}(b,a) > 0$, where $\mathbb{w}_3(x,b,a) \cdot \psi_{\phi,c}(b,a) := \lim_{s\to b} \mathbb{w}_3(x,s,a) \cdot \psi_{\phi,c}(s,a)$ if this limit exists, and otherwise we set by convention $\mathbb{w}_3(x,b,a) \cdot \psi_{\phi,c}(b,a) := 1$ (or any other strictly positive constant). □

Under Assumption 2, we always interpret the corresponding divergence

$$D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) := D^c_{\phi,M_1,M_2,\mathbb{W}_3,\lambda}(P,Q) :=$$
$$:= \int_{\mathscr{X}} \overline{\mathbb{w}_3}\Big(x, \frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\Big) \cdot \Big[\phi\Big(\frac{p(x)}{m_1(x)}\Big) - \phi\Big(\frac{q(x)}{m_2(x)}\Big)$$
$$-\phi'_{+,c}\Big(\frac{q(x)}{m_2(x)}\Big) \cdot \Big(\frac{p(x)}{m_1(x)} - \frac{q(x)}{m_2(x)}\Big)\Big] \mathrm{d}\lambda(x)$$

as $\int_{\mathscr{X}} \overline{\mathbb{w}_3 \cdot \psi_{\phi,c}}\big(x, \frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big)\, \mathrm{d}\lambda(x)$, where $\overline{\mathbb{w}_3 \cdot \psi_{\phi,c}}(x,s,t)$ denotes the extension of the function $\mathscr{X} \times]a,b[\times]a,b[\ni (x,s,t) \to \mathbb{w}_3(x,s,t) \cdot \psi_{\phi,c}(s,t)$ on $\mathscr{X} \times [a,b] \times [a,b]$ according to the conditions (g) to (n) above.

Remark 1 (a) We could even work with a weaker assumption obtained by replacing s with $\frac{p(x)}{m_1(x)}$ as well as t with $\frac{q(x)}{m_2(x)}$ and by requiring that then the correspondingly plugged-in conditions (a) to (n) hold for λ-a.a. $x \in \mathscr{X}$.
(b) Notice that our above context subsumes aggregation functions of the form $\mathbb{m}_3(x) = \tilde{\mathbb{w}}_3(x, p(x), q(x), m_1(x), m_2(x))$ with $\tilde{\mathbb{w}}_3(x, z_1, z_2, z_3, z_4)$ having appropriately imbeddable behaviour in its arguments x, z_1, z_2, z_3, z_4, the outcoming ratios $\frac{z_1}{z_3}, \frac{z_2}{z_4}$ and possible boundary values thereof. □

The following requirement is stronger than the "model-individual/dependent" Assumption 2 but is more "universally applicable" (amongst *all* models such that $\mathscr{R}\big(\frac{P}{M_1}\big) \cup \mathscr{R}\big(\frac{Q}{M_2}\big) \subset [a,b]$, take e.g. $E =]a,b[$ as $E =]0,\infty[$ or $E =]-\infty,\infty[$):

Assumption 3 Let $c \in [0,1]$, $\phi \in \Phi(]a,b[)$ on some fixed $]a,b[\in]-\infty,+\infty[$ such that $]a,b[\supset \mathscr{R}\big(\frac{P}{M_1}\big) \cup \mathscr{R}\big(\frac{Q}{M_2}\big)$. The aggregation function is of the form $\mathbb{m}_3(x) = \mathbb{w}_3\big(x, \frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big)$ for some (measurable) function $\mathbb{w}_3 : \mathscr{X} \times [a,b] \times [a,b] \to [0,\infty]$. Furthermore, for all $s \in]a,b[$, $t \in]a,b[$ and λ-a.a. $x \in \mathscr{X}$, the conditions (a) to (n) of Assumption 2 hold.

Important examples in connection with the Assumptions 2, 3 will be given in Sect. 3.2 (for ϕ) and Sect. 3.3 (for m_1, m_2, $\mathbb{w}_3$) below. With these assumptions at hand, we obtain the following non-negativity and reflexivity assertions:

Theorem 4 *Let the Assumption 2 be satisfied. Then there holds:*
(1) $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) \geqslant 0$. *Depending on the concrete situation,* $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)$ *may take infinite value.*

(2) $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q) = 0$ *if and only if* $\dfrac{p(x)}{m_1(x)} = \dfrac{q(x)}{m_2(x)}$ *for* λ*-a.a.* $x \in \mathscr{X}$.

Theorem 4 – whose proof will be given in the appendix – says that $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)$ is indeed a "proper" divergence under the Assumption 2. Hence, the latter will be assumed for the rest of the paper, unless stated otherwise: for instance, we shall sometimes work with the stronger Assumption 3; thus, for more comfortable reference, we state explicitly

Corollary 1 *Under the more universally applicable Assumption 3, the Assertions (1) and (2) of Theorem 4 hold.*

Under some non-obvious additional constraints on the functions P, Q it may be possible to show the Assertions (1), (2) of Theorem 4 by even dropping the purely generator-concerning Assumptions 2(b) to (d); see e.g. Sect. 3.3.1.2 below. In the following, we discuss several important features and special cases of $\beta = (\phi, M_1, M_2, \mathbb{M}_3, \lambda)$ in a well-structured way. Let us start with the latter.

3.1 The Reference Measure λ

In (1), λ can be interpreted as a "governer" upon the *principle* aggregation structure, whereas the "aggregation function" $\mathbb{m}_3$ tunes the *fine* aggregation details. For instance, if one chooses $\lambda = \lambda_L$ as the Lebesgue measure on $\mathscr{X} \subset \mathbb{R}$, then the integral in (1) turns out to be of Lebesgue-type and (with some rare exceptions) consequently of Riemann-type. In contrast, in the *discrete setup* where $\mathscr{X} := \mathscr{X}_\#$ has countably many elements and is equipped with the counting measure $\lambda := \lambda_\# := \sum_{z \in \mathscr{X}_\#} \delta_z$ (where δ_z is Dirac's one-point distribution $\delta_z[A] := \mathbf{1}_A(z)$, and thus $\lambda_\#[\{z\}] = 1$ for all $z \in \mathscr{X}_\#$) then (1) simplifies to

$$
\begin{aligned}
&0 \leqslant D^c_{\phi, M_1, M_2, \mathbb{M}_3, \lambda_\#}(P, Q) \\
&:= \overline{\sum}_{z \in \mathscr{X}} \Big[\phi\big(\tfrac{p(z)}{m_1(z)}\big) - \phi\big(\tfrac{q(z)}{m_2(z)}\big) - \phi'_{+,c}\big(\tfrac{q(z)}{m_2(z)}\big) \cdot \big(\tfrac{p(z)}{m_1(z)} - \tfrac{q(z)}{m_2(z)}\big) \Big] \cdot \mathbb{m}_3(z)\,, \qquad (2)
\end{aligned}
$$

which we interpret as $\sum_{z \in \mathscr{X}} \overline{\psi_{\phi,c}}\big(\frac{p(z)}{m_1(z)}, \frac{q(z)}{m_2(z)}\big) \cdot \mathbb{m}_3(z)$ with the same conventions and limits as in the paragraph right after (1); if $\mathscr{X}_\# = \{z_0\}$ for arbitrary $z_0 \in \widetilde{X}$, we obtain the corresponding one-point divergence over any space $\widetilde{X}$.

3.2 The Divergence Generator ϕ

We continue with the inspection of interesting special cases of $\beta = (\phi, M_1, M_2, \mathbb{M}_3, \lambda)$ by dealing with the first component. For this, let $\Phi_{C_1}(]a, b[)$ be the class of all functions $\phi \in \Phi(]a, b[)$ which are also continuously differentiable on $E =]a, b[$. For divergence generator $\phi \in \Phi_{C_1}(]a, b[)$, the formula (1) becomes (recall that we suppress the obsolete c and subderivative index $+$)

$$
\begin{aligned}
&0 \leqslant D_{\phi, M_1, M_2, \mathbb{M}_3, \lambda}(P, Q) \\
&:= \overline{\int}_{\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{m_1(x)}\big) - \phi\big(\tfrac{q(x)}{m_2(x)}\big) - \phi'\big(\tfrac{q(x)}{m_2(x)}\big) \cdot \big(\tfrac{p(x)}{m_1(x)} - \tfrac{q(x)}{m_2(x)}\big) \Big] \cdot \mathbb{m}_3(x)\, \mathrm{d}\lambda(x)\,, \qquad (3)
\end{aligned}
$$

whereas (2) turns into

$$0 \leqslant D_{\phi,M_1,M_2,\mathbb{M}_3,\lambda_\#}(P,Q)$$
$$:= \overline{\sum}_{x\in\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{m_1(x)}\big) - \phi\big(\tfrac{q(x)}{m_2(x)}\big) - \phi'\big(\tfrac{q(x)}{m_2(x)}\big)\cdot\big(\tfrac{p(x)}{m_1(x)} - \tfrac{q(x)}{m_2(x)}\big)\Big]\cdot \mathbb{m}_3(x).$$

Formally, by defining the integral functional $g_{\phi,\mathbb{M}_3,\lambda}(\xi) := \int_{\mathscr{X}} \phi(\xi(x))\cdot \mathbb{m}_3(x)\,\mathrm{d}\lambda(x)$ and plugging in e.g. $g_{\phi,\mathbb{M}_3,\lambda}\big(\frac{P}{M_1}\big) = \int_{\mathscr{X}} \phi\big(\frac{p(x)}{m_1(x)}\big)\cdot \mathbb{m}_3(x)\,\mathrm{d}\lambda(x)$, the divergence in (3) can be interpreted as

$$\begin{aligned} &0 \leqslant D_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P,Q)\\ &= g_{\phi,\mathbb{M}_3,\lambda}\big(\tfrac{P}{M_1}\big) - g_{\phi,\mathbb{M}_3,\lambda}\big(\tfrac{Q}{M_2}\big) - g'_{\phi,\mathbb{M}_3,\lambda}\big(\tfrac{Q}{M_2},\, \tfrac{P}{M_1} - \tfrac{Q}{M_2}\big) \end{aligned} \tag{4}$$

where $g'_{\phi,\mathbb{M}_3,\lambda}\big(\eta,\,\cdot\big)$ denotes the corresponding directional derivate at $\eta = \frac{Q}{M_2}$. If one has a "nonnegativity-setup" (NN0) in the sense that for all $x \in \mathscr{X}$ there holds $\frac{p(x)}{m_1(x)} \geqslant 0$ and $\frac{q(x)}{m_2(x)} \geqslant 0$ (but not necessarily $p(x) \geqslant 0$, $q(x) \geqslant 0$, $m_1(x) \geqslant 0$, $m_2(x) \geqslant 0$) then one can take $a = 0$, $b = \infty$, i.e. $E =]0,\infty[$, and employ the strictly convex power functions

$$\begin{aligned} &\tilde\phi(t) := \tilde\phi_\alpha(t) := \tfrac{t^\alpha - 1}{\alpha(\alpha-1)} \in]-\infty,\infty[, \qquad t \in]0,\infty[,\ \alpha \in \mathbb{R}\backslash\{0,1\}\,,\\ &\phi(t) := \phi_\alpha(t) := \tilde\phi_\alpha(t) - \tilde\phi'_\alpha(1)\cdot(t-1) = \tfrac{t^\alpha-1}{\alpha(\alpha-1)} - \tfrac{t-1}{\alpha-1} \in [0,\infty[, \quad t \in]0,\infty[,\\ &\qquad\qquad \alpha \in \mathbb{R}\backslash\{0,1\}\,, \end{aligned} \tag{5}$$

which satisfy (with the notations introduced in the paragraph right after (1))

$$\phi_\alpha(1) = 0, \quad \phi'_\alpha(t) = \tfrac{t^{\alpha-1}-1}{\alpha-1}, \quad \phi'_\alpha(1) = 0, \quad \phi''_\alpha(t) = t^{\alpha-2} > 0, \quad t \in]0,\infty[, \tag{6}$$

$$\phi_\alpha(0) := \lim_{t\downarrow 0}\phi_\alpha(t) = \tfrac{1}{\alpha}\cdot \mathbf{1}_{]0,1]\cup]1,\infty[}(\alpha) + \infty\cdot\mathbf{1}_{]-\infty,0[}(\alpha),$$
$$\phi_\alpha(\infty) := \lim_{t\uparrow\infty}\phi_\alpha(t) = \infty, \tag{7}$$

$$\begin{aligned} &\phi'_\alpha(0) := \lim_{t\downarrow 0}\phi'_\alpha(t) = \tfrac{1}{1-\alpha}\cdot\mathbf{1}_{]1,\infty[}(\alpha) - \infty\cdot\mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha),\\ &\phi'_\alpha(\infty) := \lim_{t\uparrow\infty}\phi'_\alpha(t) = \infty\cdot\mathbf{1}_{]1,\infty[}(\alpha) + \tfrac{1}{1-\alpha}\cdot\mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha) = \lim_{t\uparrow\infty}\tfrac{\phi_\alpha(t)}{t}, \end{aligned} \tag{8}$$

$$\psi_{\phi_\alpha}(s,t) = \tfrac{1}{\alpha\cdot(\alpha-1)}\cdot\Big[s^\alpha + (\alpha-1)\cdot t^\alpha - \alpha\cdot s\cdot t^{\alpha-1}\Big], \quad s,t \in]0,\infty[, \tag{9}$$

$$\psi_{\phi_\alpha}(0,t) = \tfrac{t^\alpha}{\alpha}\cdot\mathbf{1}_{]0,1[\cup]1,\infty[}(\alpha) + \infty\cdot\mathbf{1}_{]-\infty,0[}(\alpha), \quad t \in]0,\infty[, \tag{10}$$

$$\begin{aligned} &\psi_{\phi_\alpha}(\infty,t) = \infty, \quad t \in]0,\infty[,\\ &\lim_{s\to\infty}\tfrac{1}{s}\cdot\psi_{\phi_\alpha}(s,1) = \tfrac{1}{1-\alpha}\cdot\mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha) + \infty\cdot\mathbf{1}_{]1,\infty[}(\alpha),\\ &\psi_{\phi_\alpha}(s,0) = \tfrac{s^\alpha}{\alpha\cdot(\alpha-1)}\cdot\mathbf{1}_{]1,\infty[}(\alpha) + \infty\cdot\mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha), \quad s \in]0,\infty[, \end{aligned} \tag{11}$$

$$\begin{aligned} &\psi_{\phi_\alpha}(s,\infty) = \tfrac{s^\alpha}{\alpha\cdot(\alpha-1)}\cdot\mathbf{1}_{]-\infty,0[}(\alpha) + \infty\cdot\mathbf{1}_{]0,1[\cup]1,\infty[}(\alpha), \quad s \in]0,\infty[,\\ &\psi_{\phi_\alpha}(0,0) := 0 \text{ (which is unequal to } \lim_{t\to 0}\lim_{s\to 0}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha < 0\\ &\qquad\qquad \text{and which is unequal to } \lim_{s\to 0}\lim_{t\to 0}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha > 1),\\ &\psi_{\phi_\alpha}(\infty,\infty) := 0 \text{ (which is unequal to } \lim_{t\to\infty}\lim_{s\to\infty}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha \in \mathbb{R}\backslash\{0,1\}\\ &\qquad\qquad \text{and which is unequal to } \lim_{s\to\infty}\lim_{t\to\infty}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha \in]0,1[\cup]1,\infty[),\\ &\psi_{\phi_\alpha}(0,\infty) := \lim_{s\to 0}\lim_{t\to\infty}\psi_{\phi_\alpha}(s,t) = \infty \end{aligned} \tag{12}$$

$$
\begin{aligned}
&\text{(which coincides with } \lim_{t\to\infty}\lim_{s\to 0}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha\in\mathbb{R}\backslash\{0,1\}),\\
&\psi_{\phi_\alpha}(\infty,0) := \lim_{s\to\infty}\lim_{t\to 0}\psi_{\phi_\alpha}(s,t) = \infty \qquad (13)\\
&\text{(which coincides with } \lim_{t\to 0}\lim_{s\to\infty}\psi_{\phi_\alpha}(s,t) \text{ for } \alpha\in\mathbb{R}\backslash\{0,1\}).
\end{aligned}
$$

The perhaps most important special case is $\alpha = 2$, for which (5) turns into

$$\phi_2(t) := \tfrac{(t-1)^2}{2}, \quad t \in]0,\infty[= E, \tag{14}$$

having for $s, t \in]0,\infty[$ the properties (cf. (7)–(13))

$$
\begin{aligned}
&\phi_2(1) = 0, \quad \phi_2'(1) = 0, \quad \phi_2(0) = \tfrac{1}{2}, \quad \phi_2(\infty) = \infty, \quad \phi_2'(0) = -\tfrac{1}{2},\\
&\phi_2'(\infty) = \infty = \lim_{t\uparrow\infty}\tfrac{\phi_2(t)}{t}, \psi_{\phi_2}(s,t) = \tfrac{(s-t)^2}{2}, \qquad (15)\\
&\psi_{\phi_2}(0,t) = \tfrac{t^2}{2}, \quad \psi_{\phi_2}(\infty,t) = \infty, \quad \lim_{s\to\infty}\tfrac{1}{s}\cdot\psi_{\phi_2}(s,1) = \infty,\\
&\psi_{\phi_2}(s,0) = \tfrac{s^2}{2}, \quad \psi_{\phi_2}(s,\infty) = \infty, \quad \psi_{\phi_2}(0,0) := 0, \qquad (16)\\
&\psi_{\phi_2}(\infty,\infty) := 0, \quad \psi_{\phi_2}(0,\infty) = \infty, \quad \psi_{\phi_2}(\infty,0) = \infty.
\end{aligned}
$$

Also notice that the divergence-generator ϕ_2 of (14) can be trivially extended to

$$\bar{\phi}_2(t) := \tfrac{(t-1)^2}{2}, \quad t \in]-\infty,\infty[= \bar{E}, \tag{17}$$

which is useful in a general setup (GS) where for all $x \in \mathscr{X}$ one has $\frac{p(x)}{m_1(x)} \in [-\infty,\infty]$ and $\frac{q(x)}{m_2(x)} \in [-\infty,\infty]$. Convex extensions to $]a,\infty[$ with $a \in]-\infty,0[$ can be easily done by the shift $\bar{\phi}_\alpha(t) := \phi_\alpha(t-a)$.

Further examples of everywhere strictly convex differentiable divergence generators $\phi \in \Phi_{C_1}(]a,b[)$ for the "nonnegativity-setup" (NN0) (i.e. $a = 0$, $b = \infty$, $E =]0,\infty[$) can be obtained by taking the α-limits

$$
\begin{aligned}
&\tilde{\phi}_1(t) := \lim_{\alpha\to 1}\phi_\alpha(t) = t\cdot\log t \ \in [-e^{-1},\infty[, \qquad t \in]0,\infty[,\\
&\phi_1(t) := \lim_{\alpha\to 1}\phi_\alpha(t) = \tilde{\phi}_1(t) - \tilde{\phi}_1'(1)\cdot(t-1) = t\cdot\log t + 1 - t \in [0,\infty[, \ t \in]0,\infty[, \qquad (18)\\
&\tilde{\phi}_0(t) := \lim_{\alpha\to 0}\phi_\alpha(t) = -\log t \ \in]-\infty,\infty[, \qquad t \in]0,\infty[,\\
&\phi_0(t) := \lim_{\alpha\to 0}\phi_\alpha(t) = \tilde{\phi}_0(t) - \tilde{\phi}_0'(1)\cdot(t-1) = -\log t + t - 1 \in [0,\infty[, \ t \in]0,\infty[, \qquad (19)
\end{aligned}
$$

which satisfy

$$\phi_1(1)=0,\quad \phi_1'(t)=\log t,\quad \phi_1'(1)=0,\quad \phi_1''(t)=t^{-1}>0,\quad t\in]0,\infty[,$$

$$\phi_1(0):=\lim_{t\downarrow 0}\phi_1(t)=1,\quad \phi_1(\infty):=\lim_{t\uparrow\infty}\phi_1(t)=\infty, \tag{20}$$

$$\phi_1'(0):=\lim_{t\downarrow 0}\phi_1'(t)=-\infty,\quad \phi_1'(\infty):=\lim_{t\uparrow\infty}\phi_1'(t)=+\infty=\lim_{t\uparrow\infty}\tfrac{\phi_1(t)}{t}, \tag{21}$$

$$\psi_{\phi_1}(s,t)=s\cdot\log\left(\tfrac{s}{t}\right)+t-s,\quad s,t\in]0,\infty[, \tag{22}$$

$$\psi_{\phi_1}(0,t)=t,\quad \psi_{\phi_1}(\infty,t)=\infty,\quad \lim_{s\to\infty}\tfrac{1}{s}\cdot\psi_{\phi_1}(s,1)=\infty,\quad t\in]0,\infty[, \tag{23}$$

$$\psi_{\phi_1}(s,0)=\infty,\quad \psi_{\phi_1}(s,\infty)=\infty,\quad s\in]0,\infty[, \tag{24}$$

$$\psi_{\phi_1}(0,0):=0 \text{ (which coincides with } \lim_{t\to 0}\lim_{s\to 0}\psi_{\phi_1}(s,t)$$
$$\text{but which does not coincide with } \lim_{s\to 0}\lim_{t\to 0}\psi_{\phi_1}(s,t)=\infty),$$

$$\psi_{\phi_1}(\infty,\infty):=0 \text{ (which does not coincide with}$$
$$\lim_{t\to\infty}\lim_{s\to\infty}\psi_{\phi_1}(s,t)=\lim_{s\to\infty}\lim_{t\to\infty}\psi_{\phi_1}(s,t)=\infty,$$

$$\psi_{\phi_1}(0,\infty):=\lim_{s\to 0}\lim_{t\to\infty}\psi_{\phi_1}(s,t)=\infty$$
$$\text{(which coincides with } \lim_{t\to\infty}\lim_{s\to 0}\psi_{\phi_1}(s,t)),$$

$$\psi_{\phi_1}(\infty,0):=\lim_{s\to\infty}\lim_{t\to 0}\psi_{\phi_1}(s,t)=\infty$$
$$\text{(which coincides with } \lim_{t\to 0}\lim_{s\to\infty}\psi_{\phi_1}(s,t)),$$

as well as

$$\phi_0(1)=0,\quad \phi_0'(t)=1-\tfrac{1}{t},\quad \phi_0'(1)=0,\quad \phi_0''(t)=t^{-2}>0,\quad t\in]0,\infty[, \tag{25}$$

$$\phi_0(0):=\lim_{t\downarrow 0}\phi_0(t)=\infty,\quad \phi_0(\infty):=\lim_{t\uparrow\infty}\phi_0(t)=\infty, \tag{26}$$

$$\phi_0'(0):=\lim_{t\downarrow 0}\phi_0'(t)=-\infty,\quad \phi_0'(\infty):=\lim_{t\uparrow\infty}\phi_0'(t)=1=\lim_{t\uparrow\infty}\tfrac{\phi_0(t)}{t}, \tag{27}$$

$$\psi_{\phi_0}(s,t)=-\log\left(\tfrac{s}{t}\right)+\tfrac{s}{t}-1,\quad s,t\in]0,\infty[, \tag{28}$$

$$\psi_{\phi_0}(0,t)=\infty,\quad \psi_{\phi_0}(\infty,t)=\infty,\quad \lim_{s\to\infty}\tfrac{1}{s}\cdot\psi_{\phi_0}(s,1)=1,\quad t\in]0,\infty[, \tag{29}$$

$$\psi_{\phi_0}(s,0)=\infty,\quad \psi_{\phi_0}(s,\infty)=\infty,\quad s\in]0,\infty[, \tag{30}$$

$$\psi_{\phi_0}(0,0):=0 \text{ (which does not coincide with}$$
$$\lim_{t\to 0}\lim_{s\to 0}\psi_{\phi_0}(s,t)=\lim_{s\to 0}\lim_{t\to 0}\psi_{\phi_0}(s,t)=\infty),$$

$$\psi_{\phi_0}(\infty,\infty):=0 \text{ (which does not coincide with}$$
$$\lim_{t\to\infty}\lim_{s\to\infty}\psi_{\phi_0}(s,t)=\lim_{s\to\infty}\lim_{t\to\infty}\psi_{\phi_0}(s,t)=\infty),$$

$$\psi_{\phi_0}(0,\infty):=\lim_{s\to 0}\lim_{t\to\infty}\psi_{\phi_0}(s,t)=\infty$$
$$\text{(which coincides with } \lim_{t\to\infty}\lim_{s\to 0}\psi_{\phi_0}(s,t)),$$

$$\psi_{\phi_0}(\infty,0):=\lim_{s\to\infty}\lim_{t\to 0}\psi_{\phi_0}(s,t)=\infty$$
$$\text{(which coincides with } \lim_{t\to 0}\lim_{s\to\infty}\psi_{\phi_0}(s,t)).$$

An important, but (in our context) technically delicate, convex divergence generator is $\phi_{TV}(t):=|t-1|$ which is non-differentiable at $t=1$; the latter is also the only point of strict convexity. Further properties are for arbitrarily fixed $s,t\in]0,\infty[$, $c\in[0,1]$ (if not stated otherwise)

$$\phi_{TV}(1)=0,\quad \phi_{TV}(0)=1,\quad \phi_{TV}(\infty)=\infty, \tag{31}$$

$$\phi'_{TV,+,c}(t)=\mathbf{1}_{]1,\infty[}(t)+(2c-1)\cdot\mathbf{1}_{\{1\}}(t)-\mathbf{1}_{]0,1[}(t),$$

$$\phi'_{TV,+,1}(t)=\mathbf{1}_{[1,\infty[}(t)-\mathbf{1}_{]0,1[}(t),$$

$$\phi'_{TV,+,\frac{1}{2}}(t)=\mathbf{1}_{]1,\infty[}(t)-\mathbf{1}_{]0,1[}(t)=\mathrm{sgn}(t-1)\cdot\mathbf{1}_{]0,\infty[}(t),$$

$$\phi'_{TV,+,c}(1)=2c-1,\qquad \phi'_{TV,+,1}(1)=1,\quad \phi'_{TV,+,\frac{1}{2}}(1)=0, \tag{32}$$

$$\phi'_{TV,+,c}(0)=\lim_{t\to 0}\phi'_{TV,+,c}(t)=-1,\quad \phi'_{TV,+,c}(\infty)=\lim_{t\to\infty}\phi'_{TV,+,c}(t)=1,$$

$$\psi_{\phi_{TV},c}(s,t)=\mathbf{1}_{]0,1[}(t)\cdot 2(s-1)\cdot\mathbf{1}_{]1,\infty[}(s)+\mathbf{1}_{]1,\infty[}(t)\cdot 2(1-s)\cdot\mathbf{1}_{]0,1]}(s)$$
$$+\mathbf{1}_{\{1\}}(t)\cdot\Big[2(1-c)\cdot(s-1)\cdot\mathbf{1}_{]1,\infty[}(s)+2c\cdot(1-s)\cdot\mathbf{1}_{]0,1]}(s)\Big],$$

$$\psi_{\phi_{TV},\frac{1}{2}}(s,1)=|s-1|, \tag{33}$$

$$\psi_{\phi_{TV},c}(0,t)=\lim_{s\to 0}\psi_{\phi_{TV},c}(s,t)=2\cdot\mathbf{1}_{]1,\infty[}(t)+2c\cdot\mathbf{1}_{\{1\}}(t),$$

$$\psi_{\phi_{TV},c}(\infty,t)=\lim_{s\to\infty}\psi_{\phi_{TV},c}(s,t)=\infty\cdot\mathbf{1}_{]0,1[}(t)+\ \infty\cdot\mathbf{1}_{\{1\}}(t)\cdot\mathbf{1}_{[0,1[}(c),$$

$$\lim_{s\to\infty}\tfrac{1}{s}\cdot\psi_{\phi_{TV},c}(s,1)=2(1-c), \tag{34}$$

$$\psi_{\phi_{TV},c}(s,0)=\lim_{t\to 0}\psi_{\phi_{TV},c}(s,t)=2(s-1)\cdot\mathbf{1}_{]1,\infty[}(s),$$

$$\psi_{\phi_{TV},c}(s,\infty)=\lim_{t\to\infty}\psi_{\phi_{TV},c}(s,t)=2(1-s)\cdot\mathbf{1}_{]0,1]}(s),$$

$$\psi_{\phi_{TV},c}(0,0):=0\ (\text{which coincides with both }\lim_{t\to 0}\lim_{s\to 0}\psi_{\phi_{TV},c}(s,t)$$
$$\text{and }\lim_{s\to 0}\lim_{t\to 0}\psi_{\phi_{TV},c}(s,t)),$$

$$\psi_{\phi_{TV},c}(\infty,\infty):=0\ (\text{which coincides with both }\lim_{t\to\infty}\lim_{s\to\infty}\psi_{\phi_{TV},c}(s,t)$$
$$\text{and }\lim_{s\to\infty}\lim_{t\to\infty}\psi_{\phi_{TV},c}(s,t)),$$

$$\psi_{\phi_{TV},c}(0,\infty):=\lim_{s\to 0}\lim_{t\to\infty}\psi_{\phi_{TV},c}(s,t)=2$$
$$(\text{which coincides with }\lim_{t\to\infty}\lim_{s\to 0}\psi_{\phi_{TV},c}(s,t)),$$

$$\psi_{\phi_{TV},c}(\infty,0):=\lim_{s\to\infty}\lim_{t\to 0}\psi_{\phi_{TV},c}(s,t)=\infty$$
$$(\text{which coincides with }\lim_{t\to 0}\lim_{s\to\infty}\psi_{\phi_{TV},c}(s,t)).$$

In particular, one sees from Assumption 2(a) that – in our context – ϕ_{TV} can only be potentially applied if $\frac{q(x)}{m_2(x)}=1$ for λ-a.a. $x\in\mathscr{X}$ and from Assumption 2(c), (d) that we *generally* have to exclude $c=1$ and $c=0$ for $\phi'_{+,c}(\cdot)$ (i.e. we choose $c\in]0,1[$); as already mentioned above, under some non-obvious additional constraints on the functions P, Q it may be possible to drop the Assumptions 2(c), (d), see for instance Sect. 3.3.1.2 below.

Another interesting and technically delicate example is the divergence generator $\phi_{ie}(t):=t-1+\frac{(1-t)^3}{3}\cdot\mathbf{1}_{[0,1]}(t)$ which is convex, twice continuously differentiable, strictly convex at any point $t\in]0,1]$ and affine-linear on $[1,\infty[$. More detailed, one obtains for arbitrarily fixed $s,t\in]0,\infty[$ (if not stated otherwise):

$$
\begin{aligned}
&\phi_{ie}(1)=0,\quad \phi_{ie}(0)=-\tfrac{2}{3},\quad \phi_{ie}(\infty)=\infty, \qquad (35)\\
&\phi'_{ie}(t)=1-(1-t)^2\cdot \mathbf{1}_{]0,1[}(t),\\
&\phi'_{ie}(1)=1,\quad \phi'_{ie}(0)=\lim_{t\to 0}\phi'_{ie}(t)=0,\quad \phi'_{ie}(\infty)=\lim_{t\to\infty}\phi'_{ie}(t)=1,\\
&\phi''_{ie}(t)=2(1-t)\cdot \mathbf{1}_{]0,1[}(t),\quad \phi''_{ie}(1)=0,\\
&\psi_{\phi_{ie}}(s,t)=\tfrac{(1-s)^3}{3}\cdot \mathbf{1}_{]0,1[}(s)+\ (1-t)^2\cdot\Big[\tfrac{2}{3}\cdot(1-t)+(s-1)\Big]\cdot \mathbf{1}_{]0,1[}(t),\\
&\psi_{\phi_{ie}}(s,1)=\tfrac{(1-s)^3}{3}\cdot \mathbf{1}_{]0,1[}(s),\\
&\psi_{\phi_{ie}}(0,t)\ =\ \lim_{s\to 0}\psi_{\phi_{ie}}(s,t)\ =\ \tfrac{1}{3}\cdot \mathbf{1}_{[1,\infty[}(t)+\tfrac{1}{3}\ \cdot\Big[1-(1-t)^2\cdot(1-2t)\Big]\cdot \mathbf{1}_{]0,1[}(t),\\
&\psi_{\phi_{ie}}(\infty,t)=\lim_{s\to\infty}\psi_{\phi_{ie}}(s,t)=\infty\cdot \mathbf{1}_{]0,1[}(t),\\
&\lim_{s\to\infty}\tfrac{1}{s}\cdot\psi_{\phi_{ie}}(s,1)=0,\\
&\psi_{\phi_{ie}}(s,0)=\lim_{t\to 0}\psi_{\phi_{ie}}(s,t)=\big(s-\tfrac{1}{3}\big)\cdot \mathbf{1}_{[1,\infty[}(s)+s^2\cdot\big(1-\tfrac{s}{3}\big)\cdot \mathbf{1}_{]0,1[}(s),\\
&\psi_{\phi_{ie}}(s,\infty)=\lim_{t\to\infty}\psi_{\phi_{ie}}(s,t)=\tfrac{(1-s)^3}{3}\cdot \mathbf{1}_{]0,1[}(s),\\
&\psi_{\phi_{ie}}(0,0):=0\ (\text{which coincides with both } \lim_{t\to 0}\lim_{s\to 0}\psi_{\phi_{ie}}(s,t)\\
&\qquad\qquad \text{and } \lim_{s\to 0}\lim_{t\to 0}\psi_{\phi_{ie}}(s,t)),\\
&\psi_{\phi_{ie}}(\infty,\infty):=0\ (\text{which coincides with both } \lim_{t\to\infty}\lim_{s\to\infty}\psi_{\phi_{ie}}(s,t)\\
&\qquad\qquad \text{and } \lim_{s\to\infty}\lim_{t\to\infty}\psi_{\phi_{ie}}(s,t)),\\
&\psi_{\phi_{ie}}(0,\infty):=\lim_{s\to 0}\lim_{t\to\infty}\psi_{\phi_{ie}}(s,t)=\tfrac{1}{3}\\
&\qquad\qquad (\text{which coincides with } \lim_{t\to\infty}\lim_{s\to 0}\psi_{\phi_{ie}}(s,t)),\\
&\psi_{\phi_{ie}}(\infty,0):=\lim_{s\to\infty}\lim_{t\to 0}\psi_{\phi_{ie}}(s,t)=\infty\\
&\qquad\qquad (\text{which coincides with } \lim_{t\to 0}\lim_{s\to\infty}\psi_{\phi_{ie}}(s,t)).
\end{aligned}
$$

In particular, one sees from the Assumptions 2(a), (b) that – in our context – ϕ_{ie} can only be potentially applied in the following two disjoint situations:

(i) $\frac{q(x)}{m_2(x)}<1$ for λ-a.a. $x\in\mathscr{X}$;
(ii) $\frac{q(x)}{m_2(x)}=1$ and $\frac{p(x)}{m_1(x)}\leqslant 1$ for λ-a.a. $x\in\mathscr{X}$.

As already mentioned above, under some non-obvious additional constraints on the functions P, Q it may be possible to drop Assumption 2(b) and consequently (ii) can then be replaced by
$\widetilde{(ii)}$ $\frac{q(x)}{m_2(x)}=1$ for λ-a.a. $x\in\mathscr{X}$;
see for instance Sect. 3.3.1.2 below.

3.3 *The Scaling and the Aggregation Functions m_1, m_2, m_3*

In the above two Sects. 3.1 and 3.2, we have illuminated details of the choices of the first and the last component of $\beta=(\phi, M_1, M_2, \mathbb{M}_3, \lambda)$. Let us now discuss the *principal* roles as well as examples of m_1, m_2, m_3, which widen considerably the

divergence-modeling flexibility and thus bring in a broad spectrum of goal-oriented situation-based applicability. To start with, recall that in accordance with (1), the aggregation function m_3 tunes the fine aggregation details (whereas λ can be interpreted as a "governer" upon the basic/principle aggregation structure); furthermore, the function $m_1(\cdot)$ scales the function $p(\cdot)$ and $m_2(\cdot)$ the function $q(\cdot)$. From a modeling perspective, these two scaling functions can e.g. be "purely direct" in the sense that $m_1(x), m_2(x)$ are chosen to directly reflect some dependence on the data-reflecting state $x \in \mathscr{X}$ (independent of the choice of P,Q), or "purely adaptive" in the sense that $m_1(x) = w_1(p(x), q(x))$, $m_2(x) = w_2(p(x), q(x))$ for some appropriate (measurable) "connector functions" w_1, w_2 on the product $\mathscr{R}(P) \times \mathscr{R}(Q)$ of the ranges of $\{p(x)\}_{x\in\mathscr{X}}$ and $\{q(x)\}_{x\in\mathscr{X}}$, or "hybrids" $m_1(x) = w_1(x, p(x), q(x))$ $m_2(x) = w_2(x, p(x), q(x))$. Also recall that in consistency with Assumption 2 we always assume $\mathrm{m}_3(x) = \mathrm{w}_3\big(x, \frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big)$ for some (measurable) function $\mathrm{w}_3 : \mathscr{X} \times [a,b] \times [a,b] \to [0,\infty]$. Whenever applicable and insightfulness-enhancing, we use the notation $D^c_{\phi,W_1,W_2,\mathrm{W}_3,\lambda}(P,Q)$ instead of $D^c_{\phi,M_1,M_2,\mathrm{M}_3,\lambda}(P,Q)$.

Let us start with the following important sub-setup:

3.3.1 $\mathbf{m_1(x) = m_2(x) := m(x)}$, $\mathbf{\mathrm{m}_3(x) = r(x)\cdot m(x) \in [0,\infty]}$ for Some (meas.) Function $\mathbf{r : \mathscr{X} \to \mathbb{R}}$ Satisfying $\mathbf{r(x) \in]-\infty, 0[\cup]0,\infty[}$ for $\boldsymbol{\lambda}$−a.a. $\mathbf{x \in \mathscr{X}}$

As an interpretation, here the scaling functions are strongly coupled with the aggregation function; in order to avoid "case-overlapping", we assume that the function $r(\cdot)$ does not (explicitly) depend on the functions $m(\cdot)$, $p(\cdot)$ and $q(\cdot)$ (i.e. it is not of the form $r(\cdot) = h(\cdot, m(\cdot), p(\cdot), q(\cdot))$). From (1) one can deduce

$$\begin{aligned} &0 \leqslant D^c_{\phi,M,M,R\cdot M,\lambda}(P,Q) \\ &:= \overline{\int}_{\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{m(x)}\big) - \phi\big(\tfrac{q(x)}{m(x)}\big) - \phi'_{+,c}\big(\tfrac{q(x)}{m(x)}\big)\cdot\big(\tfrac{p(x)}{m(x)} - \tfrac{q(x)}{m(x)}\big)\Big]\cdot m(x)\cdot r(x)\,\mathrm{d}\lambda(x)\,, \end{aligned} \tag{36}$$

which for the discrete setup $(\mathscr{X},\lambda) = (\mathscr{X}_\#,\lambda_\#)$ (recall $\lambda_\#[\{x\}] = 1$ for all $x \in \mathscr{X}_\#$) simplifies to

$$\begin{aligned} &0 \leqslant D^c_{\phi,M,M,R\cdot M,\lambda_\#}(P,Q) \\ &= \overline{\sum}_{x\in\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{m(x)}\big) - \phi\big(\tfrac{q(x)}{m(x)}\big) - \phi'_{+,c}\big(\tfrac{q(x)}{m(x)}\big)\cdot\big(\tfrac{p(x)}{m(x)} - \tfrac{q(x)}{m(x)}\big)\Big]\cdot m(x)\cdot r(x)\,. \end{aligned} \tag{37}$$

Remark 2 (a) If one has a "nonnegativity-setup" (NN1) in the sense that for λ-almost all $x \in \mathscr{X}$ there holds $\mathrm{m}(x) \geqslant 0$, $\mathrm{r}(x) \geqslant 0$, $\mathrm{p}(x) \geqslant 0$, $\mathrm{q}(x) \geqslant 0$, then (36) (and hence also (37)) can be interpreted as scaled Bregman divergence $B_\phi\big(\mathfrak{P},\mathfrak{Q} \,|\, \mathfrak{M}\big)$ between the two nonnegative measures $\mathfrak{P},\mathfrak{Q}$ (on $(\mathscr{X},\mathscr{F})$) defined by $\mathfrak{P}[\bullet] := \mathfrak{P}^{\mathbb{R}\cdot\lambda}[\bullet] := \int_\bullet \mathrm{p}(x)\cdot\mathrm{r}(x)\,\mathrm{d}\lambda(x)$ and $\mathfrak{Q}[\bullet] := \mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\bullet] := \int_\bullet \mathrm{q}(x)\cdot\mathrm{r}(x)\,\mathrm{d}\lambda(x)$, with scaling by the nonnegative measure $\mathfrak{M}[\bullet] := \mathfrak{M}^{\mathbb{R}\cdot\lambda}[\bullet] := \int_\bullet \mathrm{m}(x)\cdot\mathrm{r}(x)\,\mathrm{d}\lambda(x)$.

(b) In a context of $\mathbb{r}(x) \equiv 1$ and "λ-probability-densities" $\vec{\mathbb{p}}$, $\vec{\mathbb{q}}$ on general state space $\mathscr{X}$, then $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda}[\bullet] := \int_\bullet \vec{\mathbb{p}}(x)\,\mathrm{d}\lambda(x)$ and $\vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda}[\bullet] := \int_\bullet \vec{\mathbb{q}}(x)\,\mathrm{d}\lambda(x)$ are probability measures (where $\mathbb{1}$ stands for the function with constant value 1). Accordingly, (36) (and hence also (37)) can be interpreted as scaled Bregman divergence $B_\phi\big(\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda}, \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda} \,|\, \mathfrak{M}^{\mathbb{1}\cdot\lambda}\big)$ which has been first defined in Stummer [81], Stummer and Vajda [84], see also Kisslinger and Stummer [35–37] for the "purely adaptive" case $\mathbb{m}(x) = \mathbb{w}\big(\vec{\mathbb{p}}(x), \vec{\mathbb{q}}(x)\big)$ and indications on non-probability measures. For instance, if Y is a random variable taking values in the discrete space $\mathscr{X}_\#$, then (with a slight abuse of notation[8]) $\vec{\mathbb{q}}(x) = \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_\#}[Y = x]$ may be its probability mass function under a hypothetical/candidate law $\vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_\#}$, and $\vec{\mathbb{p}}(x) = \frac{1}{N}\cdot\#\{i \in \{1,\ldots,N\} : Y_i = x\} =: \vec{\mathbb{p}}_N^{emp}(x)$ is the probability mass function of the corresponding data-derived "empirical distribution" $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda_\#}[\bullet] := \vec{\mathfrak{P}}_N^{emp}[\bullet] := \frac{1}{N}\cdot\sum_{i=1}^N \delta_{Y_i}[\bullet]$ of an N-size independent and identically distributed (i.i.d.) sample $Y_1,\ldots,Y_N$ of Y which is nothing but the probability distribution reflecting the underlying (normalized) histogram. Typically, for small respectively medium sample size N one gets $\vec{\mathbb{p}}_N^{emp}(x) = 0$ for some states $x \in \mathscr{X}$ which are feasible but "not yet" observed; amongst other things, this explains why density-zeros play an important role especially in statistics and information theory. This concludes the current Remark 2. □

In the following, we illuminate two important special cases of the scaling (and aggregation-part) function $m(\cdot)$, namely $\mathbb{m}(x) := 1$ and $m(x) := q(x)$:

3.3.1.1 $\mathbb{m}_1(\mathbf{x}) = \mathbb{m}_2(\mathbf{x}) := \mathbf{1}$, $\mathbb{m}_3(\mathbf{x}) = \mathbb{r}(\mathbf{x})$ for Some (Measurable) Function $\mathbb{r} : \mathscr{X} \to [\mathbf{0}, \infty]$ Satisfying $\mathbb{r}(\mathbf{x}) \in]\mathbf{0}, \infty[$ for $\boldsymbol{\lambda}$−a.a. $\mathbf{x} \in \mathscr{X}$

Accordingly, (36) turns into

$$\begin{aligned} &0 \leqslant D^c_{\phi,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(P,Q) \\ &:= \overline{\int}_{\mathscr{X}} \Big[\phi\big(p(x)\big) - \phi\big(q(x)\big) - \phi'_{+,c}\big(q(x)\big)\cdot\big(p(x)-q(x)\big)\Big]\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)\,, \end{aligned} \tag{38}$$

which for the discrete setup $(\mathscr{X},\lambda) = (\mathscr{X}_\#,\lambda_\#)$ becomes[9]

$$\begin{aligned} &0 \leqslant D^c_{\phi,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda_\#}(P,Q) \\ &:= \overline{\sum}_{x\in\mathscr{X}} \Big[\phi\big(p(x)\big) - \phi\big(q(x)\big) - \phi'_{+,c}\big(q(x)\big)\cdot\big(p(x)-q(x)\big)\Big]\cdot\mathbb{r}(x) \end{aligned} \tag{39}$$

[8] Respectively working with canonical space representation and $Y := id$.

[9] As a side remark, let us mention here that in the special case of continuously differentiable *strictly log-convex* divergence generator ϕ, one can construct divergences which are tighter than (38) respectively (39), see Stummer and Kißlinger [82]; in a finite discrete space and for differentiable *exponentially concave* divergence generator ϕ, a similar tightening (called L-divergence) can be found in Pal and Wong [66, 67].

Notice that for $\mathbb{r}(x) \equiv 1$, the divergences (38) and (39) are "consistent extensions" of the motivating pointwise dissimilarity $d_\phi^{(6)}(\cdot,\cdot)$ from Sect. 2. A special case of (38) is e.g. the rho-tau divergence (cf. Lemma 1 of Zhang and Naudts [95]).
Let us exemplarily illuminate the special case $\phi = \phi_\alpha$ together with $\mathbb{p}(x) \geqslant 0$, $\mathbb{q}(x) \geqslant 0$, for λ-almost all $x \in \mathscr{X}$ which by means of (9), (22), (28) turns (38) into the "explicit-boundary" version (of the corresponding "implicit-boundary-describing" $\overline{f}\ldots$)[10]

$$
\begin{aligned}
&0 \leqslant D_{\phi_\alpha,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \overline{f}_{\mathscr{X}} \tfrac{\mathbb{r}(x)}{\alpha\cdot(\alpha-1)} \cdot \big[\mathbb{p}(x)^\alpha + (\alpha-1)\cdot\mathbb{q}(x)^\alpha - \alpha\cdot\mathbb{p}(x)\cdot\mathbb{q}(x)^{\alpha-1}\big]\,\mathrm{d}\lambda(x) \qquad (40)\\
&= \int_{\mathscr{X}} \tfrac{\mathbb{r}(x)}{\alpha\cdot(\alpha-1)} \cdot \big[\mathbb{p}(x)^\alpha + (\alpha-1)\cdot\mathbb{q}(x)^\alpha - \alpha\cdot\mathbb{p}(x)\cdot\mathbb{q}(x)^{\alpha-1}\big] \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+ \int_{\mathscr{X}} \mathbb{r}(x) \cdot \big[\tfrac{\mathbb{p}(x)^\alpha}{\alpha\cdot(\alpha-1)} \cdot \mathbf{1}_{]1,\infty[}(\alpha) + \infty\cdot\mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha)\big] \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+ \int_{\mathscr{X}} \mathbb{r}(x) \cdot \big[\tfrac{\mathbb{q}(x)^\alpha}{\alpha} \cdot \mathbf{1}_{]0,1[\cup]1,\infty[}(\alpha) + \infty\cdot\mathbf{1}_{]-\infty,0[}(\alpha)\big] \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\,, \\
&\qquad\qquad \text{for } \alpha \in \mathbb{R}\backslash\{0,1\}, \qquad (41)\\
&0 \leqslant D_{\phi_1,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \int_{\mathscr{X}} \mathbb{r}(x)\cdot\big[\mathbb{p}(x)\cdot\log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) + \mathbb{q}(x) - \mathbb{p}(x)\big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+ \int_{\mathscr{X}} \mathbb{r}(x)\cdot\infty\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+ \int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x) \qquad (42)\\
&0 \leqslant D_{\phi_0,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \int_{\mathscr{X}} \mathbb{r}(x)\cdot\Big[-\log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) + \tfrac{\mathbb{p}(x)}{\mathbb{q}(x)} - 1\Big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+ \int_{\mathscr{X}} \mathbb{r}(x)\cdot\infty\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+ \int_{\mathscr{X}} \mathbb{r}(x)\cdot\infty\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\,, \qquad (43)
\end{aligned}
$$

where we have employed (10), (11) (23), (24), (29), (30); notice that $D_{\phi_1,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q})$ is a generalized version of the Kullback–Leibler information divergence (resp. of the relative entropy). According to the above calculations, one should exclude $\alpha \leqslant 0$ whenever $\mathbb{p}(x) = 0$ for all x in some A with $\lambda[A] > 0$, respectively $\alpha \leqslant 1$ whenever $\mathbb{q}(x) = 0$ for all x in some $\tilde{A}$ with $\lambda[\tilde{A}] > 0$ (a refined alternative for $\alpha = 1$ is given in Sect. 3.3.1.2 below). As far as splitting of the first integral e.g. in (42) resp. (43) is concerned, notice that the integral $(\mathfrak{P}^{\mathbb{R}\cdot\lambda} - \mathfrak{Q}^{\mathbb{R}\cdot\lambda})[\mathscr{X}] := \int_{\mathscr{X}} \big[\mathbb{q}(x) - \mathbb{p}(x)\big]\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)$ resp. $\int_{\mathscr{X}} \Big[\frac{\mathbb{p}(x)}{\mathbb{q}(x)} - 1\Big]\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)$ may be finite even in cases where $\mathfrak{P}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] = \int_{\mathscr{X}} \mathbb{p}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) = \infty$ and $\mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] = \int_{\mathscr{X}} \mathbb{q}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) = \infty$ (especially in case of unbounded data space (e.g. $\mathscr{X} = \mathbb{R}$) when an additive constant is involved and $\mathbb{r}(\cdot)$ is bounded from above); furthermore, there are situations where $\mathfrak{P}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] = \mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] < \infty$ and thus $(\mathfrak{P}^{\mathbb{R}\cdot\lambda} - \mathfrak{Q}^{\mathbb{R}\cdot\lambda})[\mathscr{X}] = 0$ but $\int_{\mathscr{X}} \Big[\frac{\mathbb{p}(x)}{\mathbb{q}(x)} - 1\Big]\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) = \infty$. For $\alpha = 2$, we obtain from (41) and (15) to (16)

[10] The first resp. second resp. third integral in (41) can be interpreted as divergence-contribution of the function-(support-)overlap resp. of one part of the function-nonoverlap (e.g. describing "extreme outliers") resp. of the other part of the function-nonoverlap (e.g. describing "extreme inliers").

$$0 \leqslant D_{\phi_2,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) = \int_{\mathscr{X}} \tfrac{\mathbb{r}(x)}{2} \cdot \big[\mathbb{p}(x) - \mathbb{q}(x)\big]^2 \, \mathrm{d}\lambda(x) \,, \tag{44}$$

where we can exceptionally drop the non-negativity constraints $\mathbb{p}(x) \geqslant 0, \mathbb{q}(x) \geqslant 0$. As for interpretation, (44) is nothing but half of the $\mathbb{r}(\cdot)$-weighted squared $L^2(\lambda)$-distance between $\mathbb{p}(\cdot)$ and $\mathbb{q}(\cdot)$.

In the special sub-setup of $\mathbb{r}(x) \equiv 1$ and "λ-probability-densities" $\vec{\mathbb{p}}$ $\vec{\mathbb{q}}$ on data space $\mathscr{X}$ (cf. Remark 2(b)), we can deduce from (41)–(43) the divergences

$$D_{\phi_\alpha,\mathbb{1},\mathbb{1},\mathbb{1}\cdot\mathbb{1},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}}) \tag{45}$$

which for the choice $\alpha > 0$ can be interpreted as "order$-\alpha$" density-power divergences DPD of Basu et al. [10] between the two corresponding probability measures $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda}$ and $\vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda}$; for their statistical applications see e.g. Basu et al. [12], Ghosh and Basu [30, 31] and the references therein, and for general $\alpha \in \mathbb{R}$ see e.g. Stummer and Vajda [84]. In particular, the case $\alpha = 1$ corresponding divergence in (45) is called "Kullback–Leibler information divergence" between $\vec{\mathbb{P}}$ and $\vec{\mathbb{Q}}$, and is also known under the name "relative entropy". For $\alpha = 2$, we derive $D_{\phi_2,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}})$ from (44) with $\mathbb{r}(x) = 1$ which is nothing but half of the squared L^2-distance between the two "λ-probability-densities" $\vec{\mathbb{p}}$ and $\vec{\mathbb{q}}$.

For the special discrete setup $(\mathscr{X}, \lambda) = (\mathscr{X}_\#, \lambda_\#)$ (recall $\lambda_\#[\{x\}] = 1$ for all $x \in \mathscr{X}_\#$), the divergences (41)–(44) simplify to

$$\begin{aligned}
&0 \leqslant D_{\phi_\alpha,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \textstyle\sum_{x\in\mathscr{X}} \tfrac{\mathbb{r}(x)}{\alpha\cdot(\alpha-1)} \cdot \big[\big(\mathbb{p}(x)\big)^\alpha + (\alpha-1)\cdot\big(\mathbb{q}(x)\big)^\alpha - \alpha\cdot\mathbb{p}(x)\cdot\big(\mathbb{q}(x)\big)^{\alpha-1}\big] \\
&\qquad\qquad \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \big[\tfrac{\mathbb{p}(x)^\alpha}{\alpha\cdot(\alpha-1)} \cdot \mathbf{1}_{]1,\infty[}(\alpha) + \infty \cdot \mathbf{1}_{]-\infty,0[\cup]0,1[}(\alpha)\big] \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \big[\tfrac{\mathbb{q}(x)^\alpha}{\alpha} \cdot \mathbf{1}_{]0,1[\cup]1,\infty[}(\alpha) + \infty \cdot \mathbf{1}_{]-\infty,0[}(\alpha)\big] \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) \,, \\
&\qquad\qquad \text{for } \alpha \in \mathbb{R}\backslash\{0,1\}, \qquad (46) \\
&0 \leqslant D_{\phi_1,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \big[\mathbb{p}(x)\cdot\log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) + \mathbb{q}(x) - \mathbb{p}(x)\big] \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x)\cdot\infty\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) \,, \\
&0 \leqslant D_{\phi_0,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \big[-\log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) + \tfrac{\mathbb{p}(x)}{\mathbb{q}(x)} - 1\big] \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x)\cdot\infty\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) \\
&+ \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x)\cdot\infty\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) \,, \\
&0 \leqslant D_{\phi_2,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda_\#}(\mathbb{P},\mathbb{Q}) = \textstyle\sum_{x\in\mathscr{X}} \tfrac{\mathbb{r}(x)}{2} \cdot \big[\mathbb{p}(x) - \mathbb{q}(x)\big]^2 \,.
\end{aligned}$$

Hence, as above, one should exclude $\alpha \leqslant 0$ whenever $\mathbb{p}(x) = 0$ for all x in some A with $\lambda[A] > 0$, respectively $\alpha \leqslant 1$ whenever $\mathbb{q}(x) = 0$ for all x in some $\tilde{A}$ with $\lambda[\tilde{A}] > 0$ (a refined alternative for $\alpha = 1$ is given in Sect. 3.3.1.2 below).

In particular, take the probability context of Remark 2(b), with discrete random variable Y, hypothetical probability mass function $\mathbb{q}(x) := \vec{\mathfrak{q}}(x) = \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_\#}[Y = x]$, and data-derived probability mass function (relative frequency) $\mathbb{p}(x) := \vec{\mathbb{P}}_N^{emp}(x) = \frac{1}{N} \cdot \#\{i \in \{1, \ldots, N\} : Y_i = x\}$ with sample size N. For $\mathbb{r}(x) \equiv 1$, the corresponding sample-size-weighted divergences $2N \cdot D_{\phi_\alpha, \mathbb{1}, \mathbb{1}, \mathbb{1}, \lambda_\#}(\vec{\mathbb{P}}_N^{emp}, \vec{\mathbb{Q}})$ (for $\alpha \in \mathbb{R}$) can be used as goodness-of-fit test statistics; see e.g. Kisslinger and Stummer [37] for their limit behaviour as the sample size N tends to infinity.

3.3.1.2 $\mathbf{m_1(x) = m_2(x) := q(x)}$, $\mathbb{m}_3(\mathbf{x}) = \mathbf{r(x) \cdot q(x)} \in [\mathbf{0}, \infty]$ for Some (meas.) Function $\mathbf{r} : \mathscr{X} \to \mathbb{R}$ Satisfying $\mathbf{r(x)} \in]-\infty, \mathbf{0}[\cup]\mathbf{0}, \infty[$ for $\boldsymbol{\lambda}$−a.a. $\mathbf{x} \in \mathscr{X}$

In such a set-up, the divergence (36) becomes

$$
\begin{aligned}
&0 \leqslant D^c_{\phi, Q, Q, R\cdot Q, \lambda}(P, Q) \\
&= \overline{\int}_{\mathscr{X}} \big[\phi\big(\tfrac{p(x)}{q(x)}\big) - \phi\big(1\big) - \phi'_{+,c}\big(1\big) \cdot \big(\tfrac{p(x)}{q(x)} - 1\big)\big] \cdot q(x) \cdot r(x)\, \mathrm{d}\lambda(x) \qquad (47) \\
&= \overline{\int}_{\mathscr{X}} \big[q(x) \cdot \phi\big(\tfrac{p(x)}{q(x)}\big) - q(x) \cdot \phi\big(1\big) - \phi'_{+,c}\big(1\big) \cdot \big(p(x) - q(x)\big)\big] \cdot r(x)\, \mathrm{d}\lambda(x)\,, \qquad (48)
\end{aligned}
$$

where in accordance with the descriptions right after (1) we require that $\phi :]a, b[\to \mathbb{R}$ is convex and strictly convex at $1 \in]a, b[$ and incorporate the zeros of $p(\cdot), q(\cdot), r(\cdot)$ by the appropriate limits and conventions. In the following, we demonstrate this in a non-negativity set-up where for λ-almost all $x \in \mathscr{X}$ one has $\mathbb{r}(x) \in]0, \infty[$ as well as $\mathbb{p}(x) \in [0, \infty[$, $\mathbb{q}(x) \in [0, \infty[$, and hence $E =]a, b[=]0, \infty[$. In order to achieve a reflexivity result in the spirit of Theorem 4, we have to check for – respectively analogously adapt most of – the points in Assumption 2: to begin with, the weight $w(x, s, t)$ evaluated at $s := \mathbb{p}(x), t := \mathbb{q}(x)$ has to be substituted/replaced by $\widetilde{w}(x, \widetilde{t}) := \mathbb{r}(x) \cdot \widetilde{t}$ evaluated at $\widetilde{t} = \mathbb{q}(x)$, and the dissimilarity $\psi_{\phi,c}(s, t)$ has to be substituted/replaced by $\widetilde{\widetilde{\psi}}_{\phi,c}(\widetilde{s}, \widetilde{t}) := \psi_{\phi,c}\big(\frac{\widetilde{s}}{\widetilde{t}}, 1\big)$ with the plug-in $\widetilde{s} = \mathbb{p}(x)$. Putting things together, instead of the integrand-generating term $w(x, s, t) \cdot \psi_{\phi,c}(s, t)$ we have to inspect the boundary behaviour of $\widetilde{w}(x, \widetilde{t}) \cdot \widetilde{\widetilde{\psi}}_{\phi,c}(\widetilde{s}, \widetilde{t})$ being explicitly given (with a slight abuse of notation) by the function $\widetilde{\psi}_{\phi,c} :]0, \infty[^3 \to [0, \infty[$ in

$$
\begin{aligned}
&\widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big) := r \cdot \widetilde{t} \cdot \psi_{\phi,c}\big(\tfrac{\widetilde{s}}{\widetilde{t}}, 1\big) = r \cdot \widetilde{t} \cdot \big[\phi\big(\tfrac{\widetilde{s}}{\widetilde{t}}\big) - \phi(1) - \phi'_{+,c}(1) \cdot \big(\tfrac{\widetilde{s}}{\widetilde{t}} - 1\big)\big] \\
&= r \cdot \widetilde{t} \cdot \big[\phi\big(\tfrac{\widetilde{s} \cdot r}{\widetilde{t} \cdot r}\big) - \phi(1) - \phi'_{+,c}(1) \cdot \big(\tfrac{\widetilde{s} \cdot r}{\widetilde{t} \cdot r} - 1\big)\big] = r \cdot \widetilde{t} \cdot \psi_{\phi,c}\big(\tfrac{\widetilde{s} \cdot r}{\widetilde{t} \cdot r}, 1\big)\,. \qquad (49)
\end{aligned}
$$

Since the general right-hand-derivative concerning assumption $t \in \mathscr{R}\big(\frac{Q}{M_2}\big)$ has $\frac{\widetilde{s}}{\widetilde{t}} = 1$ as its analogue, we require that the convex function $\phi :]0, \infty[\to]-\infty, \infty[$ is strictly convex (only) at 1 in conformity with Assumption 2(a) (which is also employed in

Assumption 3); for the sake of brevity we use the short-hand notation 2(a) etc. in the following discussion. We shall not need 2(b) to 2(d) in the prevailing context, so that the above-mentioned generator $\phi_{TV}(t) := |t-1|$ is allowed for achieving reflexivity (for reasons which will become clear in the proof of Theorem 5 in the appendix). The analogue of 2(e) is $\mathbb{r}(x) \cdot \widetilde{t} < \infty$ which is always (almost surely) automatically satisfied (a.a.sat.), whereas 2(f) converts to "$\mathbb{r}(x) \cdot \widetilde{t} > 0$ for all $\widetilde{s} \neq \widetilde{t}$" which is also a.a.sat. except for the case $\widetilde{t} = 0$ which will be below incorporated in combination with $\psi_{\phi,c}$-multiplication (cf. (50)). For the derivation of the analogue of 2(k) we observe that for fixed $r > 0, \widetilde{s} > 0$ the function $\widetilde{t} \to \widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big)$ is (the r-fold of) the perspective function (at $\widetilde{s}$) of the convex function $\psi_{\phi,c}\big(\cdot, 1\big)$ and thus convex with existing limit

$$\ell i_1 := r \cdot 0 \cdot \psi_{\phi,c}\big(\tfrac{\widetilde{s}}{0}, 1\big) := \lim_{t \to 0} \widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big) =$$
$$= -r \cdot \widetilde{s} \cdot \phi'_{+,c}(1) + r \cdot \widetilde{s} \cdot \lim_{\widetilde{t} \to 0} \big[\tfrac{\widetilde{t}}{\widetilde{s}} \cdot \phi\big(\tfrac{\widetilde{s}}{\widetilde{t}}\big)\big] = r \cdot \widetilde{s} \cdot (\phi^*(0) - \phi'_{+,c}(1)) \geqslant 0\,, \quad (50)$$

where $\phi^*(0) := \lim_{u \to 0} u \cdot \phi\big(\frac{1}{u}\big) = \lim_{v \to \infty} \frac{\phi(v)}{v}$ exists but may be infinite (recall that $\phi'_{+,c}(1)$ is finite). Notice that in contrast to 2(k) we need not assume $\ell i_1 > 0$ (and thus do not exclude ϕ_{TV}). To convert 2(i), we employ the fact that for fixed $r > 0, \widetilde{t} > 0$ the function $\widetilde{s} \to \widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big)$ is convex with existing limit

$$r \cdot \widetilde{t} \cdot \psi_{\phi,c}\big(\frac{0}{\widetilde{t}}, 1\big) := \lim_{s \to 0} \widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big) = r \cdot \widetilde{t} \cdot (\phi(0) + \phi'_{+,c}(1) - \phi(1)) > 0\,,$$

where $\phi(0) := \lim_{u \to 0} \phi(u)$ exists but may be infinite. To achieve the analogue of 2(g), let us first remark that for fixed $r > 0$ the function $(\widetilde{s}, \widetilde{t}) \to \widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\big)$ may not be continuous at $(\widetilde{s}, \widetilde{t}) = (0, 0)$, but due to the very nature of a divergence we make the 2(g)-conform convention of setting

$$r \cdot 0 \cdot \psi_{\phi,c}\big(\tfrac{0}{0}, 1\big) := \widetilde{\psi}_{\phi,c}\big(r, 0, 0\big) := 0$$

(notice that e.g. the power function ϕ_{-1} of (5) with index $\alpha = -1$ obeys $\lim_{\widetilde{t} \to 0} \widetilde{\psi}_{\phi_{-1}}\big(r, \widetilde{t}, \widetilde{t}\big) = 0 \neq \frac{r}{2} = \lim_{\widetilde{t} \to 0} \widetilde{\psi}_{\phi_{-1}}\big(r, \widetilde{t}^2, \widetilde{t}\big)$). The analogues of the remaining Assumptions 2(h),(j),(ℓ),(m),(n) are (almost surely) obsolete because of our basic (almost surely) finiteness requirements. Summing up, with the above-mentioned limits and conventions we write (47) explicitly as

$$
\begin{aligned}
&0 \leqslant D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \int_{\mathscr{X}} \mathbb{r}(x)\cdot\Big[\mathbb{q}(x)\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)\cdot\phi\big(1\big) - \phi'_{+,c}\big(1\big)\cdot\big(\mathbb{p}(x)-\mathbb{q}(x)\big)\Big] \\
&\qquad\qquad \cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+\big[\phi^*(0) - \phi'_{+,c}(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+\big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x) \\
&= \int_{\mathscr{X}} \mathbb{r}(x)\cdot\Big[\mathbb{q}(x)\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)\cdot\phi\big(1\big) - \phi'_{+,c}\big(1\big)\cdot\big(\mathbb{p}(x)-\mathbb{q}(x)\big)\Big] \\
&\qquad\qquad \cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+\big[\phi^*(0) - \phi'_{+,c}(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+\big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\,. \qquad (51)
\end{aligned}
$$

In case of $\mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] := \int_{\mathscr{X}} \mathbb{q}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$, the divergence (51) becomes

$$
\begin{aligned}
&0 \leqslant D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \int_{\mathscr{X}} \mathbb{r}(x)\cdot\Big[\mathbb{q}(x)\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \phi'_{+,c}\big(1\big)\cdot\big(\mathbb{p}(x)-\mathbb{q}(x)\big)\Big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+\big[\phi^*(0) - \phi'_{+,c}(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+\big[\phi(0) + \phi'_{+,c}(1)\big]\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x) - \phi(1)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\,\mathrm{d}\lambda(x)\,. \qquad (52)
\end{aligned}
$$

Moreover, in case of $\phi\big(1\big) = 0$ and $(\mathfrak{P}^{\mathbb{R}\cdot\lambda} - \mathfrak{Q}^{\mathbb{R}\cdot\lambda})[\mathscr{X}] = \int_{\mathscr{X}} \big(\mathbb{p}(x) - \mathbb{q}(x)\big)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) \in]-\infty,\infty[$ (but not necessarily $\mathfrak{P}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] = \int_{\mathscr{X}} \mathbb{p}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$, $\mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\mathscr{X}] = \int_{\mathscr{X}} \mathbb{q}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$), the divergence (51) turns into

$$
\begin{aligned}
&0 \leqslant D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = \int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big)\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+\phi^*(0)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) + \phi(0)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x) \\
&-\phi'_{+,c}\big(1\big)\cdot\int_{\mathscr{X}} \mathbb{r}(x)\cdot\big(\mathbb{p}(x)-\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\,. \qquad (53)
\end{aligned}
$$

Let us remark that (53) can be interpreted as ϕ-divergence $D^c_\phi\big(\mu,\nu\big)$ between the two nonnegative measures μ,ν (on $(\mathscr{X},\mathscr{F})$) (cf. Stummer and Vajda [83]), where $\mu[\bullet] := \mathfrak{P}^{\mathbb{R}\cdot\lambda}[\bullet]$ and $\nu[\bullet] := \mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\bullet]$. In the following, we briefly discuss two important sub-cases. First, in the "λ-probability-densities" context of Remark 2(b) one has for general $\mathscr{X}$ the manifestation $\mathbb{p}(x) := \vec{\mathbb{p}}(x) \geqslant 0$, $\mathbb{q}(x) := \vec{\mathbb{q}}(x) \geqslant 0$, and under the constraint $\phi(1) = 0$ the corresponding divergence $D^c_{\phi,\vec{\mathbb{Q}},\vec{\mathbb{Q}},\mathbb{R}\cdot\vec{\mathbb{Q}},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}})$ turns out to be the ($\mathbb{r}$-)"local ϕ-divergence of Avlogiaris et al. [6, 7]; in case of $\mathbb{r}(x) \equiv 1$ this reduces – due to the fact $\int_{\mathscr{X}} \big(\vec{\mathbb{p}}(x) - \vec{\mathbb{q}}(x)\big)\,\mathrm{d}\lambda(x) = 0$ – to the classical Csiszar-Ali-Silvey ϕ-divergence CASD ([4, 27], see also e.g. Liese and Vajda [41], Vajda [89])

$$
\begin{aligned}
0 \leqslant D^c_{\phi,\vec{\mathbb{Q}},\vec{\mathbb{Q}},\mathbb{1}\cdot\vec{\mathbb{Q}},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}}) &= \int_{\mathscr{X}} \vec{\mathfrak{q}}(x)\cdot\phi\Big(\tfrac{\vec{\mathbb{p}}(x)}{\vec{\mathfrak{q}}(x)}\Big)\cdot\mathbf{1}_{]0,\infty[}\big(\vec{\mathbb{p}}(x)\cdot\vec{\mathfrak{q}}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^*(0)\cdot\int_{\mathscr{X}}\vec{\mathbb{p}}(x)\cdot\mathbf{1}_{\{0\}}\big(\vec{\mathfrak{q}}(x)\big)\,\mathrm{d}\lambda(x)+\phi(0)\cdot\int_{\mathscr{X}}\vec{\mathfrak{q}}(x)\cdot\mathbf{1}_{\{0\}}\big(\vec{\mathbb{p}}(x)\big)\,\mathrm{d}\lambda(x)\\
&-\phi'_{+,c}\big(1\big)\cdot\int_{\mathscr{X}}\big(\vec{\mathbb{p}}(x)-\vec{\mathfrak{q}}(x)\big)\,\mathrm{d}\lambda(x)\\
&=\int_{\mathscr{X}}\vec{\mathfrak{q}}(x)\cdot\phi\Big(\tfrac{\vec{\mathbb{p}}(x)}{\vec{\mathfrak{q}}(x)}\Big)\cdot\mathbf{1}_{]0,\infty[}\big(\vec{\mathbb{p}}(x)\cdot\vec{\mathfrak{q}}(x)\big)\,\mathrm{d}\lambda(x)\\
&\quad+\phi^*(0)\cdot\mathfrak{P}^{\mathbb{1}\cdot\lambda}[\vec{\mathfrak{q}}(x)=0]+\phi(0)\cdot\mathfrak{Q}^{\mathbb{1}\cdot\lambda}[\vec{\mathbb{p}}(x)=0]\,;
\end{aligned}
\tag{54}
$$

if $\phi(1)\neq 0$ then one has to additionally subtract $\phi(1)$ (cf. the corresponding special case of (52)). In particular, for the special sub-setup where for λ-almost all $x\in\mathscr{X}$ there holds $\mathbb{p}(x):=\vec{\mathbb{p}}(x)>0$, $\mathfrak{q}(x):=\vec{\mathfrak{q}}(x)>0$, $\mathbb{r}(x)\equiv 1$, $\phi(1)=0$, one ends up with the reduced Csiszar-Ali-Silvey divergence

$$
0\leqslant D^c_{\phi,\vec{\mathbb{Q}},\vec{\mathbb{Q}},\mathbb{1}\cdot\vec{\mathbb{Q}},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}})=\int_{\mathscr{X}}\vec{\mathfrak{q}}(x)\cdot\phi\Big(\tfrac{\vec{\mathbb{p}}(x)}{\vec{\mathfrak{q}}(x)}\Big)\,\mathrm{d}\lambda(x)
$$

which can be interpreted as a "consistent extension" of the motivating pointwise dissimilarity $d^{(7)}_\phi(\cdot,\cdot)$ from the introductory Sect. 2; notice the fundamental structural difference to the divergence (38) which reflects $d^{(6)}_\phi(\cdot,\cdot)$. For comprehensive treatments of statistical applications of CASD, the reader is referred to Liese and Vajda [41], Read and Cressie [72], Vajda [89], Pardo [68], Liese and Miescke [40], Basu et al. [13].

Returning to the general divergence setup (51), we derive the reflexivity result (to be proved in the appendix):

Theorem 5 *Let $c\in[0,1]$, $\mathbb{r}(x)\in]0,\infty[$ for λ-a.a. $x\in\mathscr{X}$, $\mathscr{R}\big(\frac{\mathbb{P}}{\mathbb{Q}}\big)\cup\{1\}\subset[a,b]$, and $\phi\in\Phi(]a,b[)$ be strictly convex at $t=1$. Moreover, suppose that*

$$
\int_{\mathscr{X}}\big(\mathbb{p}(x)-\mathfrak{q}(x)\big)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)\;=\;0 \tag{55}
$$

(but not necessarily $\int_{\mathscr{X}}\mathbb{p}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)<\infty$, $\int_{\mathscr{X}}\mathfrak{q}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)<\infty$). Then:
(1) $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q})\geqslant 0$. Depending on the concrete situation, $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q})$ may take infinite value.

$$
\text{(2)}\;\; D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q})=0 \;\text{ if and only if }\; \mathbb{p}(x)=\mathfrak{q}(x)\text{ for }\lambda\text{-a.a. }x\in\mathscr{X}\,. \tag{56}
$$

Remark 3 (a) In the context of non-negative measures, the special case $c=1$ – together with $\int_{\mathscr{X}}\mathbb{p}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)<\infty$, $\int_{\mathscr{X}}\mathfrak{q}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)<\infty$ – of Theorem 5 was first achieved by Stummer and Vajda [83].
(b) Assumption (55) is always automatically satisfied if one has coincidence of finite total masses in the sense of $\mathfrak{P}^{\mathbb{R}\cdot\lambda}[\mathscr{X}]=\int_{\mathscr{X}}\mathbb{p}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)=\int_{\mathscr{X}}\mathfrak{q}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)=\mathfrak{Q}^{\mathbb{R}\cdot\lambda}[\mathscr{X}]<\infty$. For $\mathbb{r}(x)\equiv 1$ this is always satisfied for λ-probability densities $\mathbb{p}(x):=\vec{\mathbb{p}}(x)$, $\mathfrak{q}(x):=\vec{\mathfrak{q}}(x)$, since $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda}[\mathscr{X}]=\vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda}[\mathscr{X}]=1$.

(c) Notice that in contrast to Theorem 4, the generator-concerning Assumptions 2(b)–(d) are replaced by the "model-concerning" constraint (55). This opens the gate for the use of the generators ϕ_{ie} and ϕ_{TV} for cases where (55) is satisfied. For the latter, we obtain with $c = \frac{1}{2}$ explicitly from (49) and (33)

$$\widetilde{\psi}_{\phi_{TV},\frac{1}{2}}\big(r,\widetilde{s},\widetilde{t}\big) := r\cdot\widetilde{t}\cdot\psi_{\phi_{TV},\frac{1}{2}}\big(\frac{\widetilde{s}}{\widetilde{t}},1\big) = r\cdot\widetilde{t}\cdot\big|\frac{\widetilde{s}}{\widetilde{t}}-1\big| = r\cdot\big|\widetilde{s}-\widetilde{t}\big|,$$

and hence from (51) together with $\phi_{TV}(1)=0$, $\phi_{TV}(0)=1$ (cf. (31)), $\phi'_{TV,+,\frac{1}{2}}(1) = 0$ (cf. (32)), $\phi^*_{TV}(0) = \lim_{s\to\infty}\frac{1}{s}\cdot\psi_{\phi_{TV},\frac{1}{2}}(s,1) = 1$ (cf. (34)) we get

$$0 \leqslant D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = \int_{\mathscr{X}} \mathbb{r}(x)\cdot\big|\mathbb{p}(x)-\mathbb{q}(x)\big|\,\mathrm{d}\lambda(x) \tag{57}$$

which is nothing but the (possibly infinite) $\mathbb{r}(\cdot)$-weighted L_1-distance between the functions $x\to\mathbb{p}(x)$ and $x\to\mathbb{q}(x)$.

(d) In the light of (52), Theorem 4 (adapted to the current context) and Theorem 5, let us indicate that if one wants to use $\mathit{\Xi} := \int_{\mathscr{X}} \mathbb{q}(x)\cdot\phi\big(\frac{\mathbb{p}(x)}{\mathbb{q}(x)}\big)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)$ (with appropriate zero-conventions) as a divergence, then one should either employ generators ϕ satisfying $\phi(1)=\phi'_{+,c}(1)=0$, or employ models fulfilling the assumption (56) together with generators ϕ satisfying $\phi(1)=0$. On the other hand, if this integral $\mathit{\Xi}$ appears in your application context "naturally", then one should be aware that $\mathit{\Xi}$ may become negative depending on the involved set-up; for a counter-example, see Stummer and Vajda [83]. This concludes Remark 3.

As an important example, we illuminate the special case $\phi=\phi_\alpha$ with $\alpha\in\mathbb{R}\backslash\{0,1\}$ (cf. (5)) under the constraint $(\mathfrak{P}^{\mathbb{R}\cdot\lambda}-\mathfrak{Q}^{\mathbb{R}\cdot\lambda})[\mathscr{X}] = \int_{\mathscr{X}}\big(\mathbb{p}(x)-\mathbb{q}(x)\big)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) \in]-\infty,\infty[$. Accordingly, the "implicit-boundary-describing" divergence (48) resp. the corresponding "explicit-boundary" version (53) turn into the generalized power divergences of order α (cf. Stummer and Vajda [83] for $\mathbb{r}(x)\equiv 1$)[11]

$$\begin{aligned}
&0\leqslant D_{\phi_\alpha,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q})\\
&=\overline{\int}_{\mathscr{X}}\tfrac{1}{\alpha\cdot(\alpha-1)}\cdot\Big[\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big)^\alpha-\alpha\cdot\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}+\alpha-1\Big]\cdot\mathbb{q}(x)\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x)\\
&=\tfrac{1}{\alpha\cdot(\alpha-1)}\cdot\int_{\mathscr{X}}\mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\Big[\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big)^\alpha-\alpha\cdot\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}+\alpha-1\Big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^*_\alpha(0)\cdot\int_{\mathscr{X}}\mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)+\phi_\alpha(0)\cdot\int_{\mathscr{X}}\mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\\
&=\tfrac{1}{\alpha\cdot(\alpha-1)}\int_{\mathscr{X}}\mathbb{r}(x)\cdot\Big[\mathbb{p}(x)^\alpha\cdot\mathbb{q}(x)^{1-\alpha}-\mathbb{q}(x)\Big]\cdot\mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\tfrac{1}{1-\alpha}\cdot\int_{\mathscr{X}}\mathbb{r}(x)\cdot\big(\mathbb{p}(x)-\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)+\infty\cdot\mathbf{1}_{]1,\infty[}(\alpha)\cdot\int_{\mathscr{X}}\mathbb{r}(x)\cdot\mathbb{p}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big(\tfrac{1}{\alpha\cdot(1-\alpha)}\cdot\mathbf{1}_{]0,1]\cup]1,\infty[}(\alpha)+\infty\cdot\mathbf{1}_{]-\infty,0[}(\alpha)\big)\cdot\int_{\mathscr{X}}\mathbb{r}(x)\cdot\mathbb{q}(x)\cdot\mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x),
\end{aligned}\tag{58}$$

[11] This can be interpreted analogously as in footnote 10.

where we have employed (8) and (7); especially, one gets for $\alpha = 2$

$$\begin{aligned}
&0 \leqslant D_{\phi_2,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = \overline{\int}_{\mathscr{X}} \tfrac{1}{2} \cdot \tfrac{(\mathbb{p}(x)-\mathbb{q}(x))^2}{\mathbb{q}(x)} \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) \\
&= \tfrac{1}{2}\int_{\mathscr{X}} \mathbb{r}(x) \cdot \tfrac{(\mathbb{p}(x)-\mathbb{q}(x))^2}{\mathbb{q}(x)} \cdot \mathbf{1}_{[0,\infty[}(\mathbb{p}(x)) \cdot \mathbf{1}_{]0,\infty[}(\mathbb{q}(x))\,\mathrm{d}\lambda(x) \\
&+\infty \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)
\end{aligned}$$

which is called Pearsons's chisquare divergence. Under the same constraint $(\mathfrak{P}^{\mathbb{R}\cdot\lambda} - \mathfrak{Q}^{\mathbb{R}\cdot\lambda})[\mathscr{X}] \in]-\infty,\infty[$, the case $\alpha = 1$ leads by (18)–(22) to the generalized Kullback–Leibler divergence (generalized relative entropy)

$$\begin{aligned}
&0 \leqslant D_{\phi_1,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = \overline{\int}_{\mathscr{X}} \Big[\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)} \cdot \log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) + 1 - \tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\Big] \cdot \mathbb{q}(x) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) \\
&= \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x) \cdot \log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+ \int_{\mathscr{X}} \mathbb{r}(x) \cdot (\mathbb{q}(x) - \mathbb{p}(x))\,\mathrm{d}\lambda(x) + \infty \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)
\end{aligned}$$

(which equals (42)), and for $\alpha = 0$ one gets from (19), (25)–(27) the generalized reverse Kullback–Leibler divergence (generalized reverse relative entropy)

$$\begin{aligned}
&0 \leqslant D_{\phi_0,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = \overline{\int}_{\mathscr{X}} \big[-\log\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) + \tfrac{\mathbb{p}(x)}{\mathbb{q}(x)} - 1\big] \cdot \mathbb{q}(x) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) \\
&= \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \log\big(\tfrac{\mathbb{q}(x)}{\mathbb{p}(x)}\big) \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) \\
&+ \int_{\mathscr{X}} \mathbb{r}(x) \cdot (\mathbb{p}(x) - \mathbb{q}(x))\,\mathrm{d}\lambda(x) + \infty \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x).
\end{aligned}$$

Notice that instead of the limit in (50) one could also use the convention $r \cdot 0 \cdot \psi_\phi\big(\frac{s}{0}, 1\big) := \widetilde{\psi}_\phi\big(r, s, 0\big) := 0$; in the context of λ-probability densities, one then ends up with divergence by Rüschendorf [75].

For the discrete setup $(\mathscr{X}, \lambda) = (\mathscr{X}_\#, \lambda_\#)$, the divergence in (51) simplifies to

$$\begin{aligned}
&0 \leqslant D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda_\#}(\mathbb{P},\mathbb{Q}) \\
&= \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \Big[\mathbb{q}(x) \cdot \phi\big(\frac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x) \cdot \phi\big(1\big) - \phi'_{+,c}\big(1\big) \cdot \big(\mathbb{p}(x) - \mathbb{q}(x)\big)\Big] \\
&\qquad\qquad \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big) \\
&+\big[\phi^*(0) - \phi'_{+,c}(1)\big] \cdot \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) \\
&+\big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big] \cdot \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)
\end{aligned} \tag{59}$$

which in case of $\phi(1) = \phi'_{+,c}(1) = 0$ – respectively $\phi(1) = 0$ and (55) – turns into

$$\begin{aligned}
&0 \leqslant D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda_\#}(\mathbb{P},\mathbb{Q}) = \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \phi\big(\frac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\cdot\mathbb{q}(x)\big) \\
&+\phi^*(0) \cdot \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) + \phi(0) \cdot \textstyle\sum_{x\in\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big).
\end{aligned} \tag{60}$$

3.3.1.3 $\mathbf{m_1(x) = m_2(x) := w(p(x), q(x))}$, $\mathbf{m_3(x) = r(x) \cdot w(p(x), q(x)) \in [0, \infty[}$ for Some (Measurable) Functions $\mathbf{w : \mathscr{R}(P) \times \mathscr{R}(Q) \to \mathbb{R}}$ and $\mathbf{r : \mathscr{X} \to \mathbb{R}}$

Such a choice extends the context of the previous Sect. 3.3.1.2 where the "connector function" w took the simple form $w(u, v) = v$, as well as the setup of Sect. 3.3.1.1 dealing with constant $w(u, v) \equiv 1$. This introduces a wide flexibility with divergences of the form

$$
\begin{aligned}
&0 \leqslant D^c_{\phi, W(P,Q), W(P,Q), R \cdot W(P,Q), \lambda}(P, Q) \\
&:= \overline{\int}_{\mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{w(p(x),q(x))}\big) - \phi\big(\tfrac{q(x)}{w(p(x),q(x))}\big) \\
&\quad - \phi'_{+,c}\big(\tfrac{q(x)}{w(p(x),q(x))}\big) \cdot \big(\tfrac{p(x)}{w(p(x),q(x))} - \tfrac{q(x)}{w(p(x),q(x))}\big) \Big] \cdot w(p(x), q(x)) \cdot r(x) \, \mathrm{d}\lambda(x), \quad (61)
\end{aligned}
$$

which for the discrete setup $(\mathscr{X}, \lambda) = (\mathscr{X}_\#, \lambda_\#)$ (recall $\lambda_\#[\{x\}] = 1$ for all $x \in \mathscr{X}_\#$) simplifies to

$$
\begin{aligned}
0 \leqslant D^c_{\phi, W(P,Q), W(P,Q), R \cdot W(P,Q), \lambda_\#}(P, Q) &= \overline{\sum}_{x \in \mathscr{X}} \Big[\phi\big(\tfrac{p(x)}{w(p(x),q(x))}\big) - \phi\big(\tfrac{q(x)}{w(p(x),q(x))}\big) \\
- \phi'_{+,c}\big(\tfrac{q(x)}{w(p(x),q(x))}\big) \cdot \big(\tfrac{p(x)}{w(p(x),q(x))} - \tfrac{q(x)}{w(p(x),q(x))}\big) \Big] &\cdot w(p(x), q(x)) \cdot r(x) \, . \quad (62)
\end{aligned}
$$

A detailed discussion of this wide class of divergences (61),(62) is beyond the scope of this paper. For the λ-probability density context (and an indication for more general functions), see the comprehensive paper of Kisslinger and Stummer [37] and the references therein. Finally, by appropriate choices of $w(\cdot, \cdot)$ we can even derive divergences of the form (60) but with non-convex non-concave ϕ: see e.g. the "perturbed" power divergences of Roensch and Stummer [74].

3.3.2 Global Scaling and Aggregation, and Other Paradigms

Our universal framework also contains, as special cases, scaling and aggregation functions of the form $m_i(x) := m_{\ell,i}(x) \cdot H_i\big((m_{g,i}(z))_{z \in \mathscr{X}}\big)$ for some (meas., possibly nonnegative) functions $m_{l,i} : \mathscr{X} \mapsto \mathbb{R}$, $m_{g,i} : \mathscr{X} \mapsto \mathbb{R}$ and some nonzero scalar functionals H_i thereupon ($i = 1, 2, 3$, $x \in \mathscr{X}$). Accordingly, the components $H_i\big(\dots\big)$ can be viewed as "global tunings", and may depend adaptively on the primary-interest functions P and Q, i.e. $m_{g,i}(z) = w_{g,i}(x, p(x), q(x))$. For instance, in a finite discrete setup $(\mathscr{X}_\#, \lambda_\#)$ with strictly convex and differentiable ϕ, $m_1(x) \equiv m_2(x) \equiv 1$, $m_3(x) = H_i\big((w_{g,3}(q(x)))_{z \in \mathscr{X}}\big)$ this reduces to the conformal divergences of Nock et al. [64] (they also indicate the extension to equal non-unity scaling $m_1(x) \equiv m_2(x)$), for which the subcase $w_{g,3}(q(x)) := \big(\phi'(q(x))\big)^2$, $H_3\big(\big(h(x)\big)_{x \in \mathscr{X}}\big) := \big(1 + \sum_{x \in \mathscr{X}} h(x)\big)^{-1/2}$ leads to the total Bregman divergences of Liu et al. [44, 45], Vemuri et al. [91]. In contrast, Nock et al. [62] use $m_1(x) \equiv m_1 = H_1\big((p(x))_{z \in \mathscr{X}}\big)$, $m_2(x) \equiv m_1 = H_1\big((q(x))_{z \in \mathscr{X}}\big)$, $m_3(x) \equiv 1$. A more detailed discussion can be found in Stummer and Kißlinger [82] and Roensch and Stummer [74],

where they also give versions for nonconvex nonconcave divergence generators. Let us finally mention that for the construction of divergence families, there are other recent paradigms which are essentially different to (1), e.g. by means of measuring the tightness of inequalities (cf. Nielsen et al. [60, 61]), respectively of comparative convexity (cf. Nielsen et al. [59]).

4 Divergences for Essentially Different Functions

4.1 Motivation

Especially in divergence-based statistics, one is often faced with the situation where the functions $p(\cdot)$ and $q(\cdot)$ are of "essentially different nature". For instance, consider the situation where the uncertainty-prone data-generating mechanism is a random variable Y taking values in $\mathscr{X} = \mathbb{R}$ having a "classical" (e.g. Gaussian) probability density $\mathfrak{q}(\cdot)$ with respect to the one-dimensional Lebesque measure λ_L, i.e. $Pr[Y \in \bullet\,] := \vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_L}[\bullet] := \int_\bullet \vec{\mathfrak{q}}(x)\,\mathrm{d}\lambda_L(x)$ where the latter is almost always a Riemann integral (i.e. $\mathrm{d}\lambda_L(x) = \mathrm{d}x$); notice that we have set $\mathfrak{r}(x) \equiv 1$ $(x \in \mathbb{R})$. As already indicated above, under independent and identically distributed (i.i.d.) data observations $Y_1, \ldots, Y_N$ of Y one often builds the corresponding "empirical distribution" $\vec{\mathfrak{P}}_N^{emp}[\bullet] := \frac{1}{N} \cdot \sum_{i=1}^N \delta_{Y_i}[\bullet]$ which is nothing but the probability distribution reflecting the underlying (normalized) histogram. By rewriting $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda_\#}[\bullet] := \vec{\mathfrak{P}}_N^{emp}[\bullet] = \int_\bullet \vec{\mathfrak{p}}(x)\,\mathrm{d}\lambda_\#(x)$ with empirical probability mass function $\vec{\mathfrak{p}}(x) := \frac{1}{N} \cdot \#\{i \in \{1, \ldots, N\} : Y_i = x\} =: \vec{\mathfrak{p}}_N^{emp}(x)$ one encounters some basic problems for a straightforward application of divergence concepts: the two aggregating measures λ_L and $\lambda_\#$ do not coincide and actually they are of "essentially different" nature; moreover, $\vec{\mathfrak{p}}(\cdot)$ is nonzero only on the range $\mathscr{R}(Y_1, \ldots, Y_N) = \{z_1, \ldots, z_s\}$ of distinguishable points $z_1, \ldots, z_s$ $(s \leqslant N)$ occupied by $Y_1, \ldots, Y_N$. In particular, one has $\lambda_L[\{z_1, \ldots, z_s\}] = 0$. Accordingly, building a "non-coarsely discriminating" dissimilarity/divergence $D(\vec{\mathbb{P}}, \vec{\mathbb{Q}})$ between such type of functions $\vec{\mathbb{P}} := \left\{\vec{\mathfrak{p}}(x)\right\}_{x\in\mathscr{X}}$ and $\vec{\mathbb{Q}} := \left\{\vec{\mathfrak{q}}(x)\right\}_{x\in\mathscr{X}}$, is a task like "comparing apples with pears". There are several solutions to tackle this. To begin with, in the following we take the "encompassing" approach of quantifying their dissimilarity by means of their common superordinate characteristics as "fruits". Put in mathematical terms, we choose e.g. $\mathscr{X} = \mathbb{R}$, $\lambda = \lambda_L + \lambda_\#$ and work with the particular representations $\vec{\mathfrak{p}}(x) := \widetilde{\mathfrak{p}}(x) \cdot \mathbf{1}_{\{z_1,\ldots,z_s\}}(x)$ with $\widetilde{\mathfrak{p}}(x) > 0$ for λ-almost all $x \in \{z_1, \ldots, z_s\}$ as well as $\vec{\mathfrak{q}}(x) := \widetilde{\mathfrak{q}}(x) \cdot \mathbf{1}_{\widetilde{A}\setminus\{z_1,\ldots,z_s\}}(x)$ with $\widetilde{\mathfrak{q}}(x) > 0$ for λ-almost all $x \in \widetilde{A}\setminus\{z_1, \ldots, z_s\}$ with some large enough (measurable) subset $\widetilde{A}$ of $\mathscr{X} = \mathbb{R}$ such that

$$1 = \int_{\mathscr{X}} \vec{\mathfrak{p}}(x)\,\mathrm{d}\lambda_\#(x) = \int_{\mathscr{X}} \vec{\mathfrak{p}}(x)\,\mathrm{d}\lambda(x) \text{ and } 1 = \int_{\mathscr{X}} \vec{\mathfrak{q}}(x)\,\mathrm{d}\lambda_L(x) = \int_{\mathscr{X}} \vec{\mathfrak{q}}(x)\,\mathrm{d}\lambda(x) \qquad (63)$$

hold. In fact, with these choices one gets $Pr[Y \in \bullet] = \int_\bullet \mathbb{q}(x)\,\mathrm{d}\lambda(x)$ and $\vec{\mathfrak{P}}_N^{emp}[\bullet] = \int_\bullet \mathbb{p}(x)\,\mathrm{d}\lambda(x)$, as well as

$$\mathbb{p}(x) \cdot \mathbb{q}(x) = 0 \quad \text{for } \lambda\text{-almost all } x \in \mathscr{X}, \tag{64}$$

$$\mathbb{p}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) = \mathbb{p}(x) \quad \text{for } \lambda\text{-almost all } x \in \mathscr{X}, \tag{65}$$

$$\mathbb{q}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) = \mathbb{q}(x) \quad \text{for } \lambda\text{-almost all } x \in \mathscr{X} \tag{66}$$

for the special choices $\mathbb{p}(x) = \mathbb{p}(x)$ and $\mathbb{q}(x) = \mathbb{q}(x)$. By means of these and (63), the divergence (51) simplifies to

$$D^c_{\phi,\vec{\mathbb{Q}},\vec{\mathbb{Q}},\mathbb{1}\cdot\vec{\mathbb{Q}},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}}) = \phi^*(0) + \phi(0) - \phi(1) \;>\; 0. \tag{67}$$

Since for arbitrary space $\mathscr{X}$ (and not only $\mathbb{R}$) and any aggregator λ thereupon, the formula (67) holds for all functions $\vec{\mathbb{P}} := \big\{\mathbb{p}(x)\big\}_{x\in\mathscr{X}}$, $\vec{\mathbb{Q}} := \big\{\mathbb{q}(x)\big\}_{x\in\mathscr{X}}$ which satisfy (63) as well as (64)–(66) for λ-almost all $x \in \mathscr{X}$, and since $\phi^*(0) + \phi(0) - \phi(1)$ is just a constant (which may be infinite), these divergences $D^c_{\phi,\vec{\mathbb{Q}},\vec{\mathbb{Q}},\mathbb{1}\cdot\vec{\mathbb{Q}},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}})$ are not suitable for discriminating between such "essentially different" (basically orthogonal) λ-probability densities $\vec{\mathbb{P}}$ and $\vec{\mathbb{Q}}$. More generally, under the validity of (64)–(66) for λ-almost all $x \in \mathscr{X}$ – which we denote by $\mathbb{P} \perp \mathbb{Q}$ and which basically amounts to pair of functions of the type

$$\mathbb{p}(x) := \widetilde{p}(x) \cdot \mathbf{1}_A(x) \quad \text{with } \widetilde{p}(x) > 0 \text{ for } \lambda\text{-almost all } x \in A, \tag{68}$$

$$\mathbb{q}(x) := \widetilde{q}(x) \cdot \mathbf{1}_{B\setminus A}(x) \quad \text{with } \widetilde{q}(x) > 0 \text{ for } \lambda\text{-almost all } x \in B\setminus A, \tag{69}$$

with some (measurable) subsets $\widetilde{A} \subset B$ of $\mathscr{X}$ – the divergence (51) turns into

$$\begin{aligned} D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q}) = {} & \big[\phi^*(0) - \phi'_{+,c}(1)\big] \cdot \textstyle\int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x)\,\mathrm{d}\lambda(x) \\ & + \big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big] \cdot \textstyle\int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x)\,\mathrm{d}\lambda(x) \;>\; 0 \end{aligned} \tag{70}$$

which now depends on $\mathbb{P}$ and $\mathbb{Q}$, in a rudimentary "weighted-total-mass" way. Inspired by this, we specify a statistically interesting divergence subclass:

Definition 1 We say that a divergence (respectively dissimilarity respectively distance)[12] $D(\cdot,\cdot)$ is encompassing for a class $\widetilde{\mathscr{P}}$ of functions if

- for arbitrarily fixed $Q := \big\{q(x)\big\}_{x\in\mathscr{X}} \in \widetilde{\mathscr{P}}$ the function $P := \big\{p(x)\big\}_{x\in\mathscr{X}} \to D(P,Q)$ is non-constant on the subfamily of all $P \in \widetilde{\mathscr{P}}$ with $P \perp Q$, and
- for arbitrarily fixed $P \in \widetilde{\mathscr{P}}$ the function $Q \to D(P,Q)$ is non-constant on the subfamily of all $Q \in \widetilde{\mathscr{P}}$ with $Q \perp P$.

Accordingly, due to (67) the prominently used divergences $D^c_{\phi,\vec{\mathbb{Q}},\vec{\mathbb{Q}},\mathbb{1}\cdot\vec{\mathbb{Q}},\lambda}(\vec{\mathbb{P}},\vec{\mathbb{Q}})$ are not encompassing for the class of $\widetilde{\mathscr{P}}$ of all λ-probability densities; more gener-

[12] i.e. the properties (D1) and (D2) (respectively (D2) respectively (D1), (D2) and (D3)) are satisfied.

ally, because of (70) the divergences $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P},\mathbb{Q})$ are in general encompassing for the class of $\widetilde{\mathscr{P}}$ of all λ-probability densities, but not for $\widetilde{\mathscr{P}} := \{\widetilde{P} := \{\widetilde{\mathbb{p}}(x)\}_{x\in\mathscr{X}} \mid \int_{\mathscr{X}} \mathbb{r}(x)\cdot\widetilde{\mathbb{p}}(x)\,\mathrm{d}\lambda(x) = \widetilde{c}\}$ for any fixed $\widetilde{c}$.

4.2 $\mathbb{m}_1(x) = \mathbb{m}_2(x) := \mathbb{q}(x)$, $\mathbb{m}_3(x) = \mathbb{r}(x)\cdot\mathbb{q}(x)^\chi \in [0,\infty]$ *for Some* $\chi > 1$ *and Some (Measurable) Function* $\mathbb{r} : \mathscr{X} \to [0,\infty[$

In the following, we propose a new way of repairing the above-mentioned encompassing-concerning deficiency for λ-probability density functions, by introducing a new divergence in terms of choosing a generator $\phi :]0,\infty[\to \mathbb{R}$ which is convex and strictly convex at 1, the scaling function $\mathbb{m}_1(x) = \mathbb{m}_2(x) := \mathbb{q}(x)$ as in the non-negativity set-up of Sect. 3.3.1.2, but the more general aggregation function $\mathbb{m}_3(x) = \mathbb{r}(x)\cdot\mathbb{q}(x)^\chi \in [0,\infty[$ for some power $\chi > 1$ and some (measurable) function $\mathbb{r} : \mathscr{X} \to [0,\infty[$ which satisfies $\mathbb{r}(x) \in]0,\infty[$ for λ-almost all $x \in \mathscr{X}$. To incorporate the zeros of $\mathbb{p}(\cdot)$, $\mathbb{q}(\cdot)$, $\mathbb{r}(\cdot)$ by appropriate limits and conventions, we proceed analogously to Sect. 3.3.1.2. Accordingly, we inspect the boundary behaviour of the function $\widetilde{\psi}_{\phi,c} :]0,\infty[^3 \to [0,\infty[$ given by

$$\widetilde{\psi}_{\phi,c}(r,\widetilde{s},\widetilde{t}) := r\cdot\widetilde{t}^\chi\cdot\psi_{\phi,c}\big(\tfrac{\widetilde{s}}{\widetilde{t}},1\big) = r\cdot\widetilde{t}^\chi\cdot\big[\phi\big(\tfrac{\widetilde{s}}{\widetilde{t}}\big) - \phi(1) - \phi'_{+,c}(1)\cdot\big(\tfrac{\widetilde{s}}{\widetilde{t}}-1\big)\big]$$
$$= r\cdot\widetilde{t}^\chi\cdot\big[\phi\big(\tfrac{\widetilde{s}\cdot r}{\widetilde{t}\cdot r}\big) - \phi(1) - \phi'_{+,c}(1)\cdot\big(\tfrac{\widetilde{s}\cdot r}{\widetilde{t}\cdot r}-1\big)\big] = r\cdot\widetilde{t}^\chi\cdot\psi_{\phi,c}\big(\tfrac{\widetilde{s}\cdot r}{\widetilde{t}\cdot r},1\big).$$

As in Sect. 3.3.1.2, the Assumption 2(a) is conformly satisfied, for which we use the short-hand notation 2(a) etc. in the following discussion. Moreover, we require the validity of 2(b)–2(d) at the point $t = 1$. The analogue of 2(e) is $\mathbb{r}(x)\cdot\widetilde{t}^\chi < \infty$ which is always (almost surely) automatically satisfied (a.a.sat.), whereas 2(f) converts to "$\mathbb{r}(x)\cdot\widetilde{t}^\chi > 0$ for all $\widetilde{s} \neq \widetilde{t}$" which is also a.a.sat. except for the case $\widetilde{t} = 0$ which will be incorporated below. For the derivation of the analogue of 2(k) we observe that for fixed $r > 0$, $\widetilde{s} > 0$

$$\ell i_2 := r\cdot 0^\chi\cdot\psi_{\phi,c}\big(\tfrac{\widetilde{s}}{0},1\big) := \lim_{t\to 0}\widetilde{\psi}_{\phi,c}\big(r,\widetilde{s},\widetilde{t}\big) =$$
$$= r\cdot\widetilde{s}^\chi\cdot\lim_{\widetilde{t}\to 0}\big[\tfrac{\widetilde{t}^\chi}{\widetilde{s}^\chi}\cdot\phi\big(\tfrac{\widetilde{s}}{\widetilde{t}}\big)\big] = r\cdot\widetilde{s}^\chi\cdot\phi^*_\chi(0) \geqslant 0, \tag{71}$$

where $\phi^*_\chi(0) := \lim_{u\to 0} u^{\chi-1}\cdot u\cdot\phi\big(\frac{1}{u}\big) = \lim_{v\to\infty}\frac{\phi(v)}{v^\chi}$ exists but may be infinite. To convert 2(i), we employ the fact that for fixed $r > 0$, $\widetilde{t} > 0$ the function $\widetilde{s} \to \widetilde{\psi}_{\phi,c}\big(r,\widetilde{s},\widetilde{t}\big)$ is convex with existing limit

$$\ell i_3 := r\cdot\widetilde{t}^\chi\cdot\psi_{\phi,c}\big(\tfrac{0}{\widetilde{t}},1\big) := \lim_{s\to 0}\widetilde{\psi}_{\phi,c}\big(r,\widetilde{s},\widetilde{t}\big)$$
$$= r\cdot\widetilde{t}^\chi\cdot(\phi(0) + \phi'_{+,c}(1) - \phi(1)) > 0. \tag{72}$$

To achieve the analogue of 2(g), let us first remark that for fixed $r > 0$ the function $(\widetilde{s}, \widetilde{t}) \to \widetilde{\psi}_{\phi,c}\big(r, \widetilde{s}, \widetilde{t}\,\big)$ may not be continuous at $(\widetilde{s}, \widetilde{t}) = (0, 0)$, but due to the very nature of a divergence we make the 2(g)-conform convention of setting

$$r \cdot 0^{\chi} \cdot \psi_{\phi,c}\big(\tfrac{0}{0}, 1\big) := \widetilde{\psi}_{\phi,c}\big(r, 0, 0\big) := 0\,.$$

The analogues of the Assumptions 2(h), (j), (ℓ), (m), (n) are obsolete because of our basic finiteness requirements. Putting together all the building-blocks, with the above-mentioned limits and conventions we obtain the divergence

$$\begin{aligned}
&0 \leqslant D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^{\chi},\lambda}(\mathbb{P},\mathbb{Q})\\
&:= \overline{\int}_{\mathscr{X}} \mathbb{r}(x) \cdot \Big[\mathbb{q}(x)^{\chi} \cdot \phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)^{\chi} \cdot \phi\big(1\big) - \phi'_{+,c}\big(1\big) \cdot \big(\mathbb{p}(x) \cdot \mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big)\Big]\,\mathrm{d}\lambda(x)\\
&:= \int_{\mathscr{X}} \mathbb{r}(x) \cdot \Big[\mathbb{q}(x)^{\chi} \cdot \phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)^{\chi} \cdot \phi\big(1\big) - \phi'_{+,c}\big(1\big) \cdot \big(\mathbb{p}(x) \cdot \mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big)\Big]\\
&\qquad\qquad \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x) \cdot \mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^{*}_{\chi}(0) \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x)^{\chi} \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big] \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x)^{\chi} \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{q}(x)\big) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\\
&= \int_{\mathscr{X}} \mathbb{r}(x) \cdot \Big[\mathbb{q}(x)^{\chi} \cdot \phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)^{\chi} \cdot \phi\big(1\big) - \phi'_{+,c}\big(1\big) \cdot \big(\mathbb{p}(x) \cdot \mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big)\Big]\\
&\qquad\qquad \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x) \cdot \mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^{*}_{\chi}(0) \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x)^{\chi} \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big] \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x)^{\chi} \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\,. \qquad (73)
\end{aligned}$$

In case of $\mathfrak{Q}^{\mathbb{R}\cdot\lambda}_{\chi}[\mathscr{X}] := \int_{\mathscr{X}} \mathbb{q}(x)^{\chi} \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$, the divergence (73) becomes

$$\begin{aligned}
&0 \leqslant D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^{\chi},\lambda}(\mathbb{P},\mathbb{Q})\\
&= \int_{\mathscr{X}} \mathbb{r}(x) \cdot \Big[\mathbb{q}(x)^{\chi} \cdot \phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \phi'_{+,c}\big(1\big) \cdot \big(\mathbb{p}(x) \cdot \mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big)\Big]\\
&\qquad\qquad \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x) \cdot \mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^{*}_{\chi}(0) \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x)^{\chi} \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\big[\phi(0) + \phi'_{+,c}(1)\big] \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x)^{\chi} \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\\
&-\phi(1) \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x)^{\chi}\,\mathrm{d}\lambda(x)\,. \qquad (74)
\end{aligned}$$

Moreover, in case of $\phi\big(1\big) = 0$ and $\int_{\mathscr{X}} \big(\mathbb{p}(x) \cdot \mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) \in [0, \infty[$ (but not necessarily $\int_{\mathscr{X}} \mathbb{p}(x) \cdot \mathbb{q}(x)^{\chi-1} \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$, $\int_{\mathscr{X}} \mathbb{q}(x)^{\chi} \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x) < \infty$), the divergence (73) turns into

$$\begin{aligned}
&0 \leqslant D^{c}_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^{\chi},\lambda}(\mathbb{P},\mathbb{Q}) = \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x)^{\chi} \cdot \phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x) \cdot \mathbb{q}(x)\big)\,\mathrm{d}\lambda(x)\\
&+\phi^{*}_{\chi}(0) \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x)^{\chi} \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) + \phi(0) \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x)^{\chi} \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big)\,\mathrm{d}\lambda(x)\\
&-\phi'_{+,c}\big(1\big) \cdot \int_{\mathscr{X}} \big(\mathbb{p}(x) \cdot \mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^{\chi}\big) \cdot \mathbb{r}(x)\,\mathrm{d}\lambda(x)\,.
\end{aligned}$$

In contrast to the case $\chi = 1$ where for λ-probability-density functions $\mathbb{p}$, $\mathbb{q}$, the divergence (53) was further simplified due to $\int_{\mathscr{X}} \big(\mathbb{p}(x) - \mathbb{q}(x)\big)\,\mathrm{d}\lambda(x) = 0$, for the current setup $\chi > 1$ the latter has no impact for further simplification. However, in general, for the new divergence defined by (73) one gets for any $\mathbb{P} \perp \mathbb{Q}$ from (68), (69), (64)–(66) the expression

$$
\begin{aligned}
&0 \leqslant D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q}) \\
&= \phi^*_\chi(0) \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{p}(x)^\chi \,\mathrm{d}\lambda(x) + \big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big] \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x)^\chi \,\mathrm{d}\lambda(x) \qquad (75)
\end{aligned}
$$

which is encompassing for the class of λ-probability functions. By inspection of the above calculations, one can even relax the assumptions away from convexity:

Theorem 6 *Let $\chi > 1$, $c \in [0,1]$, $\phi :]0,\infty[\to \mathbb{R}$ such that both $\phi'_{+,c}(1)$ and $\phi(0) := \lim_{s\to 0}\phi(s)$ exist and $\psi_{\phi,c}(s,1) = \phi(s) - \phi(1) - \phi'_{+,c}(1)\cdot(s-1) \geqslant 0$ for all $s > 0$. Moreover, assume that $\psi_{\phi,c}(s,1) = 0$ if and only if $s = 1$. Furthermore, let the limits $\ell i_2 \geqslant 0$ defined by (71) and $\ell i_3 \geqslant 0$ defined by (72) exist and satisfy $\ell i_2 + \ell i_3 > 0$. Then one gets for the divergence defined by (73):*
(1) $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q}) \geqslant 0$. Depending on the concrete situation, $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q})$ may take infinite value.
(2) $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q}) = 0$ if and only if $\mathbb{p}(x) = \mathbb{q}(x)$ for λ-a.a. $x \in \mathscr{X}$.
(3) For $\mathbb{P} \perp \mathbb{Q}$, the representation (75) holds.

Remark 4 (1) As seen above, if the generator ϕ is in $\Phi(]0,\infty[)$ and satisfies the Assumptions 2(a)–(d) for $t = 1$, then the requirements on ϕ in Theorem 6 are automatically satisfied. The case $\chi = 1$ has already been covered by Theorem 5.
(2) For practical purposes, it is sometimes useful to work with a sub-setup of choices $\chi > 1$, $c \in [0,1]$ and ϕ such that $\ell i_2 \in]0,\infty[$ and/or $\ell i_3 \in]0,\infty[$. □

Let us give some examples. To begin with, for $\alpha \in \mathbb{R}\setminus\{0,1\}$ take the power functions $\phi(t) := \phi_\alpha(t) := \frac{t^\alpha - 1}{\alpha(\alpha-1)} - \frac{t-1}{\alpha-1} \in [0,\infty[, \quad t \in]0,\infty[$, with the properties $\phi_\alpha(1) = 0, \phi'_\alpha(1) = 0$ (cf. (6)) and $\phi_\alpha(0) := \lim_{t\downarrow 0}\phi_\alpha(t) = \frac{1}{\alpha}\cdot \mathbf{1}_{]0,1]\cup]1,\infty[}(\alpha) + \infty \cdot \mathbf{1}_{]-\infty,0[}(\alpha)$. Then, for arbitrary $\chi \in \mathbb{R}$ one gets the representation

$$
\begin{aligned}
&0 \leqslant D_{\phi_\alpha,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q}) \\
&:= \overline{\int}_{\mathscr{X}} \mathbb{r}(x) \cdot \Big[\mathbb{q}(x)^\chi \cdot \phi_\alpha\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x)^\chi \cdot \phi_\alpha(1) - \phi'_\alpha(1)\cdot\big(\mathbb{p}(x)\cdot\mathbb{q}(x)^{\chi-1} - \mathbb{q}(x)^\chi\big)\Big]\,\mathrm{d}\lambda(x) \qquad (76) \\
&= \overline{\int}_{\mathscr{X}} \Big[\phi_\alpha\big(\tfrac{\mathbb{p}(x)}{w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x))}\big) - \phi_\alpha\big(\tfrac{\mathbb{q}(x)}{w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x))}\big) \\
&\quad - \phi'_\alpha\big(\tfrac{\mathbb{q}(x)}{w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x))}\big)\cdot\big(\tfrac{\mathbb{p}(x)}{w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x))} - \tfrac{\mathbb{q}(x)}{w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x))}\big)\Big]\cdot w_{\widetilde{\chi}}(\mathbb{p}(x),\mathbb{q}(x))\cdot\mathbb{r}(x)\,\mathrm{d}\lambda(x) \\
&= D_{\phi_\alpha,\mathbb{Q}^{\widetilde{\chi}},\mathbb{Q}^{\widetilde{\chi}},\mathbb{R}\cdot\mathbb{Q}^{\widetilde{\chi}},\lambda}(\mathbb{P},\mathbb{Q}) \qquad (77)
\end{aligned}
$$

with the adaptive scaling/aggregation function $w_{\widetilde{\chi}}(u,v) = v^{\widetilde{\chi}}$ and $\widetilde{\chi} := 1 + \frac{\chi-1}{1-\alpha}$; in other words, the divergence (76) can be seen as a particularly adaptively scaled

Bregman divergence of non-negative functions in the sense of Kißlinger and Stummer [37], from which their robustness and non-singularity-asymptotical-statistics properties can be derived as a special case (for the probability setup $\overrightarrow{\mathbb{P}}, \overrightarrow{\mathbb{Q}}, \mathbb{r}(x) \equiv 1$, and beyond). From (77), it is immediate to see that the case $\chi = 1$ corresponds to the generalized power divergences (58) of order $\alpha \in \mathbb{R}\backslash\{0, 1\}$, whereas $\chi = \alpha$ corresponds to the unscaled divergences (40), i.e.

$$0 \leqslant D_{\phi_\alpha,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\alpha,\lambda}(\mathbb{P},\mathbb{Q}) = D_{\phi_\alpha,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{P},\mathbb{Q}) \tag{78}$$
$$= \overline{\int}_{\mathscr{X}} \tfrac{\mathbb{r}(x)}{\alpha\cdot(\alpha-1)} \cdot \Big[\mathbb{p}(x)^\alpha + (\alpha-1)\cdot \mathbb{q}(x)^\alpha - \alpha\cdot\mathbb{p}(x)\cdot\mathbb{q}(x)^{\alpha-1}\Big]\,\mathrm{d}\lambda(x) \ (cf.\ (40))$$

which for $\alpha > 1, \mathbb{r}(x) \equiv 1, \mathbb{p} = \mathbb{\dot{p}}, \mathbb{q} = \mathbb{\dot{q}}$ is a multiple of the α-order density-power divergences DPD used by Basu et al. [10]; as a side remark, in the latter setup our divergence (77) manifests a smooth interconnection between PD and DPD which differs from that of Patra et al. [70], Ghosh et al. [32].

For (76), let us shortly inspect the corresponding ℓi_2 from (71) as well as ℓi_3 from (72). Only for $\alpha \in]0, 1[\cup]1, \infty[$, one gets finite $\ell i_3 = \frac{r\widetilde{t}^\chi}{\alpha} \in]0, \infty[$ for all $\chi \in \mathbb{R}, r > 0, \widetilde{t} > 0$. Additionally, one obtains finite ℓi_2 only for $\chi = 1, \alpha \in]0, 1[$ where $\ell i_2 = \frac{r\widetilde{s}}{1-\alpha}$ (PD case), respectively for $\chi > 1, \alpha \in]0, 1[\cup]1, \chi[$ where $\ell i_2 = 0$, respectively for $\alpha = \chi > 1$ where $\ell i_2 = \frac{r\widetilde{s}^\alpha}{\alpha\cdot(\alpha-1)}$ (DPD case), for all $r > 0, \widetilde{s} > 0$.

Another interesting example for the divergence $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q})$ in (73) is given for $\alpha \in \mathbb{R}\backslash\{0, 1\}$ by the generators

$$\phi(t) := \overset{\approx}{\phi}_\alpha(t) := \tfrac{(\alpha-1)\cdot t^\alpha - \alpha\cdot t^{\alpha-1}+1}{\alpha\cdot(\alpha-1)}, \quad t > 0, \quad \overset{\approx}{\phi}_\alpha(1) = 0, \ \overset{\approx}{\phi'}_\alpha(1) = 0,$$

for which $t \to \overset{\approx}{\phi}_\alpha(t) = \overset{\approx}{\phi}_\alpha(t) - \overset{\approx}{\phi}_\alpha(0) - \overset{\approx}{\phi'}_\alpha(1)\cdot(t-1) = \psi_{\phi_\alpha}(t, 1)$ is strictly decreasing on $]0, 1[$ and strictly increasing on $]1, \infty[$. Hence, the corresponding assumptions of Theorem 6 are satisfied. Beyond this, notice that $\overset{\approx}{\phi}_\alpha(\cdot)$ is strictly convex on $]0, \infty[$ if $\alpha \in]1, 2]$, respectively strictly convex on $]1 - \frac{1}{\alpha-1}, \infty[$ and strictly concave on $]0, 1 - \frac{1}{\alpha-1}[$ if $\alpha > 2$, respectively strictly convex on $]0, 1 + \frac{1}{1-\alpha}[$ and strictly concave on $]1 + \frac{1}{1-\alpha}, \infty[$ if $\alpha \in]-\infty, 0[\cup]0, 1[$. Furthermore, the corresponding ℓi_3 is finite only for $\alpha > 1$, namely $\ell i_3 = \frac{r\widetilde{t}^\chi}{\alpha\cdot(\alpha-1)} \in]0, \infty[$ for all $\chi \in \mathbb{R}$, $r > 0, \widetilde{t} > 0$. Additionally, if $\alpha > 1$ one gets finite ℓi_2 only for $\chi > \alpha > 1$ where $\ell i_2 = 0$, respectively for $\alpha = \chi > 1$ where $\ell i_2 = \frac{r\widetilde{s}^\alpha}{\alpha}$ for all $r > 0, \widetilde{s} > 0$. Notice that for $\chi = \alpha > 1$, the limits $\ell i_2, \ell i_3$ for the cases ϕ_α and $\overset{\approx}{\phi}_\alpha$ are asymmetric. Indeed, by straightforward calculations one can easily see that

$$0 \leqslant D_{\overset{\approx}{\phi}_\alpha,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\alpha,\lambda}(\mathbb{P},\mathbb{Q}) = D_{\phi_\alpha,\mathbb{1},\mathbb{1},\mathbb{R}\cdot\mathbb{1},\lambda}(\mathbb{Q},\mathbb{P})$$
$$= \overline{\int}_{\mathscr{X}} \tfrac{\mathbb{r}(x)}{\alpha\cdot(\alpha-1)} \cdot \Big[\big(\mathbb{q}(x)\big)^\alpha + (\alpha-1)\cdot\big(\mathbb{p}(x)\big)^\alpha - \alpha\cdot\mathbb{q}(x)\cdot\big(\mathbb{p}(x)\big)^{\alpha-1}\Big]\,\mathrm{d}\lambda(x) \tag{79}$$

which is the "reversion" of the divergence (40).

4.3 Minimum Divergences - The Encompassing Method

So far, we have almost entirely dealt with aggregated divergences between functions $P := \{p(x)\}_{x\in\mathscr{X}}$, $Q := \{q(x)\}_{x\in\mathscr{X}}$ under the *same* aggregator (measure) λ. On the other hand, in Sect. 4.1 we have already encountered an important statistical situation where *two* aggregators λ_1 and λ_2 come into play. Let us now investigate such a context in more detail. To achieve this, for the rest of this paper we confine ourselves to the following probabilistic setup: the modeled respectively observed (random) data take values in a state space $\mathscr{X}$ (with at least two distinct values), equipped with a system $\mathscr{F}$ of admissible events (σ-algebra) and two σ-finite measures λ_1 and λ_2. Furthermore, let $\vec{\mathbb{P}} := \{\vec{\mathbb{p}}\}_{x\in\mathscr{X}}$, $\vec{\mathbb{Q}} := \{\vec{\mathbb{q}}\}_{x\in\mathscr{X}}$ such that $\vec{\mathbb{p}}(x) \geqslant 0$ for λ_1-a.a. $x \in \mathscr{X}$, $\vec{\mathbb{q}}(x) \geqslant 0$ for λ_2-a.a. $x \in \mathscr{X}$, $\int_{\mathscr{X}} \vec{\mathbb{p}}(x)\,\mathrm{d}\lambda_1(x) = 1$, and $\int_{\mathscr{X}} \vec{\mathbb{q}}(x)\,\mathrm{d}\lambda_2(x) = 1$; in other words, $\vec{\mathbb{P}}$ is a λ_1-probability density function and $\vec{\mathbb{Q}}$ is a λ_2-probability density function; the two corresponding probability measures are denoted by $\vec{\mathfrak{P}}^{\mathbb{1}\cdot\lambda_1}[\bullet] := \int_\bullet \vec{\mathbb{p}}(x)\,\mathrm{d}\lambda_1(x)$ and $\vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_2}[\bullet] := \int_\bullet \vec{\mathbb{q}}(x)\,\mathrm{d}\lambda_2(x)$. Notice that we henceforth assume $\mathbb{r}(x) = 1$ for all $x \in \mathscr{X}$.

More specific, we deal with a parametric framework of double uncertainty in the data and in the model (cf. Sect. 2.4). The former is described by a random variable Y taking values in the space $\mathscr{X}$ and by its probability law $\vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2}[\bullet]$ which (as far as model risk is concerned) is supposed to be unknown but belong to a class $\mathscr{Q}_\Theta^{\lambda_2} = \{\vec{\mathfrak{Q}}_\theta^{\mathbb{1}\cdot\lambda_2}[\bullet] : \theta \in \Theta\}$ of probability measures on $(\mathscr{X}, \mathscr{F})$ indexed by a set of parameters $\Theta \subset \mathbb{R}^d$ (the non-parametric case works basically in analogous way, with more sophisticated technicalities). Accordingly, all $Pr[Y \in \bullet \mid \theta] = \vec{\mathfrak{Q}}_\theta^{\mathbb{1}\cdot\lambda_2}[\bullet] = \int_\bullet \vec{\mathbb{q}}_\theta(x)\,\mathrm{d}\lambda_2(x)$ ($\theta \in \Theta$) are principal model-candidate laws, with θ_0 to be found out (approximately and with high confidence) by N concrete data observations described by the independent and identically distributed random variables $Y_1, \ldots Y_N$. Furthermore, we assume that the true unknown parameter θ_0 (to be learnt) is identifiable and that the family $\mathscr{Q}_\Theta^{\lambda_2}$ is (measure-theoretically) equivalent in the sense

$$\vec{\mathfrak{Q}}_\theta^{\mathbb{1}\cdot\lambda_2} \neq \vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2} \quad \text{and} \quad \vec{\mathfrak{Q}}_\theta^{\mathbb{1}\cdot\lambda_2} \sim \vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2} \quad \text{for all } \theta,\ \theta_0 \in \Theta \quad \text{with } \theta \neq \theta_0. \tag{80}$$

As usual, the equivalence $\vec{\mathfrak{Q}}^{\mathbb{1}\cdot\lambda_2} \sim \widehat{\vec{\mathfrak{Q}}}^{\mathbb{1}\cdot\lambda_2}$ means that for λ_2-a.a. $x \in \mathscr{X}$ there holds the density-function-relation: $\vec{\mathbb{q}}(x) = 0$ if and only if $\widehat{\vec{\mathbb{q}}}(x) = 0$; this implies in particular that $\vec{\mathbb{q}}(x) \cdot \mathbf{1}_{\{0\}}\big(\widehat{\vec{\mathbb{q}}}(x)\big) = 0$ and $\widehat{\vec{\mathbb{q}}}(x) \cdot \mathbf{1}_{\{0\}}\big(\vec{\mathbb{q}}(x)\big) = 0$ for λ_2-a.a. $x \in \mathscr{X}$, and by cutting off "datapoints/states of zero contributions" one can then even take $\mathscr{X}$ small enough such that $\vec{\mathbb{q}}(x) \cdot \widehat{\vec{\mathbb{q}}}(x) > 0$ (and hence, $\mathbf{1}_{]0,\infty[}\big(\vec{\mathbb{q}}(x) \cdot \widehat{\vec{\mathbb{q}}}(x)\big) = 1$) for λ_2-a.a. $x \in \mathscr{X}$. Clearly, since any λ_2-aggregated divergence $D_{\lambda_2}(\cdot,\cdot)$ satisfies (the aggregated version of) the axioms (D1) and (D2), and since θ_0 is identifiable, one gets immediately in terms of the corresponding λ_2-probability density functions $\vec{\mathbb{Q}}_\theta := \{\vec{\mathbb{q}}_\theta(x)\}_{x\in\mathscr{X}}$

$$\theta_0 = \mathrm{argmin}_{\theta\in\Theta}\ D_{\lambda_2}\big(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}\big) \quad \text{for every } \theta_0 \in \Theta. \tag{81}$$

Inspired by this, one major idea of tracking down (respectively, learning) the true unknown θ_0 is to replace $\vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2}$ by a data-observation-derived – and thus noisy – probability law $\omega \to \vec{\mathfrak{P}}_N^{obs(\omega);\mathbb{1}\cdot\lambda_1}[\bullet] := \int_\bullet \vec{\mathbb{P}}^{Y_1(\omega),\ldots,Y_N(\omega)}(x)\,\mathrm{d}\lambda_1(x)$ where the λ_1-probability density function $\vec{\mathbb{P}}_N^{obs(\omega)} := \big\{\vec{\mathbb{P}}^{Y_1(\omega),\ldots,Y_N(\omega)}(x)\big\}_{x\in\mathscr{X}}$ depends, as indexed, on the outcome of the observations $Y_1(\omega), \ldots, Y_N(\omega)$. If $\vec{\mathfrak{P}}_N^{obs(\omega);\mathbb{1}\cdot\lambda_1}$ converges in distribution to $\vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2}$ as N tends to infinity, then one *intuitively* expects to obtain the so-called *minimum-divergence estimator* ("approximator")

$$\widehat{\theta}_N(\omega) := \widehat{\theta}_{N,D_{\lambda_2}}(\omega) := \mathrm{arginf}_{\theta\in\Theta}\ D_{\lambda_2}\big(\vec{\mathbb{P}}_N^{obs(\omega)}, \vec{\mathbb{Q}}_{\theta}\big) \tag{82}$$

which estimates θ_0 consistently in the usual sense of the convergence $\theta_n \to \theta_0$ for $n \to \infty$. However, by the nature of our divergence construction, the method (82) makes principal sense only if the two aggregators λ_1 and λ_2 coincide (and if (82) is analytically respectively computationally solvable)! Remark that the minimum distance estimator (82) depends on the choice of the divergence $D_{\lambda_2}(\cdot,\cdot)$.

Subsetup 1. For instance, if by nature the set $\mathscr{X}$ of all possible data points has only countably many elements, say $\mathscr{X} = \mathscr{X}_\# = \{z_1, \ldots z_s\}$ (where s is an integer larger than one or infinity), then a natural model-concerning aggregator is the counting measure $\lambda_2 := \lambda_\#$ (recall $\lambda_\#[\{x\}] = 1$ for all $x \in \mathscr{X}$), and hence $\vec{\mathfrak{Q}}_\theta^{\mathbb{1}\cdot\lambda_2}[\bullet] = \sum_{x\in\bullet} \vec{\mathbb{q}}_\theta(x) = \sum_{x\in\mathscr{X}} \mathbf{1}_\bullet(x)\cdot\vec{\mathbb{q}}_\theta(x)$ (where $\bullet$ stands for any arbitrary subset of $\mathscr{X}$). In such a context, a popular choice for the data-observation-derived probability law is the so-called "empirical distribution" $\omega \to \vec{\mathfrak{P}}_N^{obs(\omega);\mathbb{1}\cdot\lambda_1}[\bullet] = \int_\bullet \vec{\mathbb{P}}^{Y_1(\omega),\ldots,Y_N(\omega)}(x)\,\mathrm{d}\lambda_1(x) := \sum_{x\in\bullet} \vec{\mathbb{P}}_N^{emp(\omega)}(x) =: \vec{\mathfrak{P}}_N^{emp(\omega)}[\bullet]$, where $\lambda_1 := \lambda_\# = \lambda_2$ and $\vec{\mathbb{P}}_N^{emp(\omega)}(x) := \frac{1}{N}\cdot\#\{i \in \{1,\ldots,N\} : Y_i(\omega) = x\}$ is the total number of x-observations divided by the total number N of observations. In other words, $\vec{\mathfrak{P}}_N^{obs(\omega);\mathbb{1}\cdot\lambda_1}[\bullet] := \vec{\mathfrak{P}}_N^{emp(\omega)}[\bullet] := \frac{1}{N}\cdot\sum_{i=1}^N \delta_{Y_i(\omega)}[\bullet]$, where $\delta_z[\bullet]$ is the corresponding Dirac (resp. one-point) distribution given by $\delta_z[A] := \mathbf{1}_A(z)$. Hence, in such a set-up it makes sense to solve the noisy minimization problem

$$\widehat{\theta}_N(\omega) := \widehat{\theta}_{N,D_{\lambda_\#}}(\omega) := \mathrm{arginf}_{\theta\in\Theta}\ D_{\lambda_\#}\big(\vec{\mathbb{P}}_N^{emp(\omega)}, \vec{\mathbb{Q}}_{\theta}\big) \tag{83}$$

where $\vec{\mathbb{P}}_N^{emp(\omega)} := \big\{\vec{\mathbb{P}}_N^{emp}(x)\big\}_{x\in\mathscr{X}}$ and $D_{\lambda_\#}(\cdot,\cdot)$ is the discrete version of any of the divergences above. Notice that – at least for small enough number N of observations – for some $x \in \mathscr{X}$ with $\lambda_\#[\{x\}] > 0$ one has $\vec{\mathbb{P}}_N^{emp}(x) = 0$ but $\vec{\mathbb{q}}_\theta(x) > 0$ (i.e. an "extreme inlier"), and hence, $\vec{\mathbb{q}}_\theta(x)\cdot\mathbf{1}_{\{0\}}\big(\vec{\mathbb{P}}_N^{emp}(x)\big) > 0$; this must be taken into

account in the calculation of the explicit forms of the corresponding divergences.[13] By the assumed convergence, this effect disappears as N becomes large enough. □

Subsetup 2. Consider the "crossover case" where $\mathscr{X}$ is uncountable (e.g. $\mathscr{X} = \mathbb{R}$) and the family $\mathscr{Q}_\Theta^{\lambda_2}$ is assumed to be *continuous (nonatomic)* in the sense

$$0 = \vec{\mathfrak{Q}}_\theta^{\mathbb{1}\cdot\lambda_2}[\{z\}] = Pr[Y \in \{z\} \,|\, \theta] = \int_{\mathscr{X}} \mathbf{1}_{\{z\}}(x) \cdot \vec{\mathfrak{q}}_\theta(x)\, \mathrm{d}\lambda_2(x) \quad \text{for all } z \in \mathscr{X}, \theta \in \Theta \tag{84}$$

(e.g. $\vec{\mathfrak{q}}_\theta(\cdot)$ are Gaussian densities with mean θ and variance 1), and the data-observation-derived probability law is the "extended" empirical distribution

$$\begin{aligned} &\omega \to \vec{\mathfrak{P}}_N^{obs(\omega);\mathbb{1}\cdot\lambda_1}[\bullet] = \int_\bullet \vec{\mathbb{p}}^{Y_1(\omega),\ldots,Y_N(\omega)}(x)\, \mathrm{d}\lambda_1(x) \\ &:= \sum_{x\in\bullet} \vec{\mathbb{p}}_N^{emp(\omega)}(x) \cdot \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x) =: \vec{\mathfrak{P}}_N^{\overline{emp}(\omega)}[\bullet]\,, \end{aligned} \tag{85}$$

where the extension on $\mathscr{X}$ is accomplished by attributing zeros to all x outside the finite range $\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega)) = \{z_1(\omega),\ldots,z_s(\omega)\}$ of distinguishable points $z_1(\omega),\ldots,z_s(\omega)$ $(s \leqslant N)$ occupied by the observations $Y_1(\omega),\ldots,Y_N(\omega)$; notice that the involved counting measure given by $\lambda_1[\bullet] := \sum_{z\in\mathscr{X}} \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(z) \cdot \delta_z[\bullet]$ puts 1 to each data-point z which has been observed. Because λ_1 and λ_2 are now essentially different, the minimum-divergence method (82) can not be applied directly (by taking either $\lambda := \lambda_1$ or $\lambda := \lambda_2$), despite of $\vec{\mathfrak{P}}_N^{\overline{emp}(\omega)}$ converging in distribution to $\vec{\mathfrak{Q}}_{\theta_0}^{\mathbb{1}\cdot\lambda_2}$ as N tends to infinity. □

There are several ways to circumvent the problem in Subsetup 2. In the following, we discuss in more detail our abovementioned new encompassing approach:

(Enc1) take the encompassing aggregator $\lambda := \lambda_1 + \lambda_2$ and the imbedding $\vec{\mathbb{P}}_N^{\overline{emp}(\omega)} := \{\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\}_{x\in\mathscr{X}}$ with $\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x) := \vec{\mathbb{p}}_N^{emp(\omega)}(x) \cdot \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x)$;

(Enc2) choose a "sufficiently discriminating" (e.g. encompassing) divergence $D_\lambda(\cdot,\cdot)$ from above and evaluate them with the density-functions obtained in (Enc1);

(Enc3) solve the corresponding noisy minimization problem

$$\widehat{\theta}_N(\omega) := \widehat{\theta}_{N,D_\lambda}(\omega) := \mathrm{arginf}_{\theta\in\Theta}\, D_\lambda\big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \breve{\vec{\mathbb{Q}}}_\theta\big) \tag{86}$$

for $\breve{\vec{\mathbb{Q}}}_\theta := \vec{\mathbb{Q}}_\theta$ respectively $\breve{\vec{\mathbb{Q}}}_\theta := \widetilde{\vec{\mathbb{Q}}}_\theta$ (to be defined right below);

(Enc4) compute the noisy minimal distance $D_\lambda\big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \breve{\vec{\mathbb{Q}}}_\theta\big) > 0$ as an indicator of "goodness of fit" (goodness of noisy approximation");

---

[13] E.g. applying the divergence (46) for $\alpha \in \mathbb{R}\backslash\{0,1\}$, the sum-entry $\mathbb{r}(x) \cdot \frac{\vec{\mathbb{q}}_\theta(x)^\alpha}{\alpha}$ appears, which can be viewed as penalty for the cell x being empty of data observations ("intrinsic empty-cell-penalty"); for divergence (60), the penalty is $\phi(0) \cdot \mathbb{r}(x) \cdot \vec{\mathfrak{q}}_\theta(x)$.

(Enc5) investigate sound statistical properties of the outcoming estimator $\widehat{\theta}_N(\omega)$, e.g. show probabilistic convergence (as N tends to infinity) to the true unknown parameter θ_0, compute the corresponding convergence speed, analyze its robustness against data-contamination, etc.

Typically, for fixed N the step (Enc3) is not straightforward to solve, and consequently, the tasks described in the unavoidable step (Enc4) become even much more complicated; a detailed discussion of both is – for the sake of brevity – beyond the scope of this paper. As far as (Enc1) is concerned, things are non-trivial due to the generally well-known fact that "continuous" densities are only almost-surely unique. Indeed, consider e.g. the case where the θ-family of functions $\overrightarrow{\mathbb{Q}}_\theta := \left\{\overrightarrow{\mathfrak{q}}_\theta(x)\right\}_{x\in\mathscr{X}}$ satisfies

$$\overrightarrow{\mathfrak{q}}_\theta(x) > 0 \ \text{ for all } x \in \mathscr{X} \text{ and } \overrightarrow{\mathfrak{Q}}_\theta^{\mathbb{1}\cdot\lambda_2}[\mathscr{X}] = \int_{\mathscr{X}} \overrightarrow{\mathfrak{q}}_\theta(x)\,\mathrm{d}\lambda_2(x) = 1 \ \text{ for all } \theta \in \Theta \tag{87}$$

and the alternative θ-family of functions $\widetilde{\overrightarrow{\mathbb{Q}}}_\theta := \left\{\widetilde{\overrightarrow{\mathfrak{q}}}_\theta(x)\right\}_{x\in\mathscr{X}}$ defined by $\widetilde{\overrightarrow{\mathfrak{q}}}_\theta(x) := \overrightarrow{\mathfrak{q}}_\theta(x)\cdot(1-\mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x))$; for the latter, one obtains

$$\widetilde{\overrightarrow{\mathfrak{Q}}}_\theta^{\mathbb{1}\cdot\lambda_2}[\mathscr{X}] = \int_{\mathscr{X}} \widetilde{\overrightarrow{\mathfrak{q}}}_\theta(x)\,\mathrm{d}\lambda_2(x) = \int_{\mathscr{X}} \widetilde{\overrightarrow{\mathfrak{q}}}_\theta(x)\,\mathrm{d}(\lambda_1+\lambda_2)(x) = 1 \ \text{ for all } \theta \in \Theta. \tag{88}$$

Furthermore, due to (85) one has

$$1 = \overrightarrow{\mathfrak{P}}_N^{\overline{emp}(\omega)}[\mathscr{X}] = \int_{\mathscr{X}} \overrightarrow{\mathfrak{p}}_N^{\overline{emp}(\omega)}(x)\,\mathrm{d}\lambda_1(x) = \int_{\mathscr{X}} \overrightarrow{\mathfrak{p}}_N^{\overline{emp}(\omega)}(x)\,\mathrm{d}(\lambda_1+\lambda_2)(x) \tag{89}$$

and the validity of (64)–(66) with $\mathfrak{p}(x) := \overrightarrow{\mathfrak{p}}_N^{\overline{emp}(\omega)}(x)$, $\mathfrak{q}(x) := \widetilde{\overrightarrow{\mathfrak{q}}}_\theta(x)$ and $\lambda = \lambda_1 + \lambda_2$; in other words, there holds the singularity (measure-theoretical orthogonality) $\overrightarrow{\mathbb{P}}_N^{\overline{emp}(\omega)} \perp \widetilde{\overrightarrow{\mathbb{Q}}}_\theta$ for all $\theta \in \Theta$. Accordingly, for the step (Enc2) one can e.g. take directly the (family of) encompassing divergences $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q})$ of (73) for $\mathbb{P} := \overrightarrow{\mathbb{P}}_N^{\overline{emp}(\omega)}$, $\mathbb{Q} := \widetilde{\overrightarrow{\mathbb{Q}}}_\theta$, $\lambda := \lambda_1 + \lambda_2$, $\mathbb{r}(x) \equiv 1$, and apply (75) to get

$$\begin{aligned} 0 \leqslant\ & D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{1}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q}) = \phi^*_\chi(0)\cdot\textstyle\sum_{x\in\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}\left(\overrightarrow{\mathbb{P}}_N^{emp(\omega)}(x)\right)^\chi \\ & + \left[\phi(0)+\phi'_{+,c}(1)-\phi(1)\right]\cdot\textstyle\int_{\mathscr{X}}\left(\overrightarrow{\mathfrak{q}}_\theta(x)\right)^\chi \mathrm{d}\lambda_2(x)\,; \end{aligned} \tag{90}$$

hence, the corresponding solution of (Enc3) does not depend on the data-observations $Y_1(\omega),\ldots,Y_N(\omega)$, and thus is "statistically non-relevant". As an important remark for the rest of this paper, let us mention that – only – in situations where no observations are taken into account, then $\widetilde{\overrightarrow{\mathbb{Q}}}_\theta = \overrightarrow{\mathbb{Q}}_\theta$, $\mathscr{R}(Y_1,\ldots,Y_N) = \emptyset$, and λ_1 collapses to the "zero aggregator" (i.e. $\lambda_1[\bullet] \equiv 0$).

In contrast, let us replace the alternative θ-family $\widetilde{\overrightarrow{\mathbb{Q}}}_\theta$ by the original $\overrightarrow{\mathbb{Q}}_\theta$, on which λ_1 acts differently. In fact, instead of (88) there holds

$$1 = \vec{\Omega}_\theta^{\mathbb{1}\cdot\lambda_2}[\mathscr{X}] = \int_{\mathscr{X}} \vec{\mathbb{q}}_{\mathbb{1}\theta}(x)\,\mathrm{d}\lambda_2(x) < \int_{\mathscr{X}} \vec{\mathbb{q}}_{\mathbb{1}\theta}(x)\,\mathrm{d}(\lambda_1+\lambda_2)(x)$$
$$= 1 + \sum_{x\in\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))} \vec{\mathbb{q}}_{\mathbb{1}\theta}(x) \qquad \text{for all } \theta\in\Theta\,; \tag{91}$$

moreover, one has for all $\theta\in\Theta$ the non-singularity $\vec{\mathbb{P}}_N^{\overline{emp}(\omega)} \not\perp \vec{\mathbb{Q}}_\theta$ but

$$\mathbf{1}_{\{0\}}\big(\vec{\mathbb{q}}_{\mathbb{1}\theta}(x)\big) = 0 \quad \text{for all } x\in\mathscr{X}, \tag{92}$$
$$\mathbf{1}_{\{0\}}\big(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\big) = 1 - \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x) \quad \text{for all } x\in\mathscr{X}, \tag{93}$$
$$\mathbf{1}_{]0,\infty[}\big(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\cdot\vec{\mathbb{q}}_{\mathbb{1}\theta}(x)\big) = \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x) \quad \text{for all } x\in\mathscr{X}\,. \tag{94}$$

Correspondingly, for the step (Enc2) one can e.g. take directly the (family of) encompassing divergences $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q}^\chi,\lambda}(\mathbb{P},\mathbb{Q})$ of (73) for $\mathbb{P} := \vec{\mathbb{P}}_N^{\overline{emp}(\omega)}$, $\mathbb{Q} := \vec{\mathbb{Q}}_\theta$, $\lambda := \lambda_1+\lambda_2$, $\mathbb{r}(x)\equiv 1$; the corresponding solution of the noisy minimization problem (Enc3) generally *does depend* on the data-observations $Y_1(\omega),\ldots,Y_N(\omega)$, as required. Let us demonstrate this exemplarily for the special subsetup where $\phi:[0,\infty[\to[0,\infty[$ is continuous (e.g. strictly convex on $]0,\infty[$), differentiable at 1, $\phi(1)=\phi'(1)=0$, $\phi(t)\in]0,\infty[$ for all $t\in[0,1[\cup]1,\infty[$, $\chi>1$, $\mathbb{r}(x)\equiv 1$, and $\int_{\mathscr{X}} \vec{\mathbb{q}}_{\mathbb{1}\theta}(x)^\chi\,\mathrm{d}\lambda_2(x)\in]0,\infty[$ for all $\theta\in\Theta$. Then, for each fixed $\theta\in\Theta$ we derive from (73) and (92)–(94) the divergence

$$0 < D_{\vec{\phi},\vec{\mathbb{Q}}_\theta,\vec{\mathbb{Q}}_\theta,\mathbb{1}\cdot\vec{\mathbb{Q}}_\theta^{\,\chi},\lambda_1+\lambda_2}\big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)},\vec{\mathbb{Q}}_\theta\big)$$
$$= \int_{\mathscr{X}} \vec{\mathbb{q}}_{\mathbb{1}\theta}(x)^\chi\cdot\phi\Big(\frac{\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)}{\vec{\mathbb{q}}_\theta(x)}\Big)\cdot\mathbf{1}_{]0,\infty[}\big(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\cdot\vec{\mathbb{q}}_{\mathbb{1}\theta}(x)\big)\,\mathrm{d}(\lambda_1+\lambda_2)(x)$$
$$+\phi(0)\cdot\int_{\mathscr{X}} \vec{\mathbb{q}}_{\mathbb{1}\theta}(x)^\chi\cdot\mathbf{1}_{\{0\}}\big(\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\big)\,\mathrm{d}(\lambda_1+\lambda_2)(x)$$
$$= \sum_{x\in\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))} \vec{\mathbb{q}}_{\mathbb{1}\theta}(x)^\chi\cdot\phi\Big(\frac{\vec{\mathbb{p}}_N^{\,emp(\omega)}(x)}{\vec{\mathbb{q}}_{\mathbb{1}\theta}(x)}\Big) + \phi(0)\cdot\int_{\mathscr{X}} \vec{\mathbb{q}}_{\mathbb{1}\theta}(x)^\chi\,\mathrm{d}\lambda_2(x) < \infty\,. \tag{95}$$

When choosing this divergence (95) in step (Enc2), we call the solution $\widehat{\theta}_N(\omega)$ of the corresponding noisy minimization problem (86) of step (Enc3) a *minimum* (ϕ,χ)-*divergence estimator* of the true unknown parameter θ_0; in ML and AI contexts, the pair (ϕ,χ) may be regarded as "hyperparameter". Exemplarily, for the power functions $\phi := \phi_\alpha$ (cf. (5)) with $\alpha=\chi>1$, we obtain from (95) (see also (78), (41)) the divergence

$$]0,\infty[\ni D_{\vec{\phi}_\alpha,\vec{\mathbb{Q}}_\theta,\vec{\mathbb{Q}}_\theta,\mathbb{1}\cdot\vec{\mathbb{Q}}_\theta^{\,\alpha},\lambda_1+\lambda_2}\big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)},\vec{\mathbb{Q}}_\theta\big) = \tfrac{1}{\alpha}\cdot\int_{\mathscr{X}} \vec{\mathbb{q}}_{\mathbb{1}\theta}(x)^\alpha\,\mathrm{d}\lambda_2(x)$$
$$+\sum_{x\in\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}\Big[\tfrac{(\vec{\mathbb{p}}_N^{\,emp(\omega)}(x))^\alpha}{\alpha\cdot(\alpha-1)} - \vec{\mathbb{p}}_N^{\,emp(\omega)}(x)\cdot\tfrac{\vec{\mathbb{q}}_\theta(x)^{\alpha-1}}{\alpha-1} + \tfrac{\vec{\mathbb{q}}_\theta(x)^\alpha}{\alpha}\Big]$$
$$= \tfrac{1}{\alpha}\cdot\int_{\mathscr{X}} \vec{\mathbb{q}}_{\mathbb{1}\theta}(x)^\alpha\,\mathrm{d}\lambda_2(x)$$
$$+\tfrac{1}{N}\sum_{i=1}^N\Big[\tfrac{(\vec{\mathbb{p}}_N^{\,emp(\omega)}(Y_i(\omega)))^{\alpha-1}}{\alpha\cdot(\alpha-1)} - \tfrac{\vec{\mathbb{q}}_\theta(Y_i(\omega))^{\alpha-1}}{\alpha-1} + \tfrac{\vec{\mathbb{q}}_\theta(Y_i(\omega))^\alpha}{\alpha\cdot\vec{\mathbb{p}}_N^{\,emp(\omega)}(Y_i(\omega))}\Big], \tag{96}$$

where for the last equality we have used the representation

$$\sum_{x\in\mathscr{X}} \mathbb{P}_N^{emp(\omega)}(x)\cdot \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x)\cdot \delta_x[\bullet] \;=\; \tfrac{1}{N}\cdot\sum_{i=1}^N \delta_{Y_i(\omega)}[\bullet]\,; \quad (97)$$

notice that $\mathbb{P}_N^{emp(\omega)}(Y_i(\omega)) = \#\{j\in\{1,\ldots,N\} : Y_j(\omega)=Y_i(\omega)\}/N$. Clearly, the outcoming minimum (ϕ,χ)-divergence estimator of (95) (and in particular, the minimum (ϕ_α,α)-divergence estimator of (96)) depends on the data observations $Y_1(\omega),\ldots,Y_N(\omega)$, where for technical reasons as e.g. existence and uniqueness – as well as for the tasks (Enc4), (Enc5) – some further assumptions are generally needed; for the sake of brevity, corresponding details will appear in a forthcoming paper.

4.4 Minimum Divergences - Grouping and Smoothing

Next, we briefly indicate two other ways to circumvent the problem described in Subsetup 2 of Sect. 4.3, with continuous (nonatomic) $\mathscr{Q}_\Theta^{\lambda_2}$ and λ_2 from (84):

(GR) grouping (partitioning, quantization) of data: convert[14] everything into a purely discrete context, by subdividing the data-point-set $\mathscr{X}=\bigcup_{j=1}^s A_j$ into countably many – (say) $s\in\mathbb{N}\cup\{\infty\}\backslash\{1\}$ – (measurable) disjoint classes $A_1,\ldots,A_s$ with the property $\lambda_2[A_j]>0$ ("essential partition"); proceed as in Subsetup 1 of Sect. 4.3, with $\mathscr{X}^{new}:=\{A_1,\ldots,A_s\}$ instead of $\{z_1,\ldots,z_s\}$, and thus the ith data observation $Y_i(\omega)$ and the corresponding running variable x) manifest (only) the corresponding class-membership. For the subcase of Csiszar-Ali-Slivey divergences and adjacently related divergences, thorough statistical investigations (such as efficiency, robustness, types of grouping, grouping-error sensitivity, etc.) of the corresponding minimum-divergence-estimation can be found e.g. in Victoria-Feser and Ronchetti [92], Menendez et al. [47–49], Morales et al. [52, 53], Lin and He [43].

(SM) smoothing of the empirical density function: convert everything to a purely continuous context, by keeping the original data-point-set $\mathscr{X}$ and by "continuously modifying" (e.g. with the help of kernels) the empirical density $\mathbb{P}_N^{emp}(\cdot)$ to a function $\mathbb{P}_N^{emp,smo}(\cdot)\geqslant 0$ such that $\int_{\mathscr{X}}\mathbb{P}_N^{emp,smo}(x)\,\mathrm{d}\lambda_2(x)=1$ and that for all $\theta\in\Theta$ there holds: $\mathbb{P}_N^{emp,smo}(x)=0$ if and only if $\mathbb{q}_\theta(x)=0$ (in addition to (80)). For the subcase of Csiszar-Ali-Slivey divergences, thorough statistical investigations (such as efficiency, robustness, information loss, etc.) of the corresponding minimum-divergence-estimation can be found e.g. in Basu and Lindsay [11], Park and Basu [69], Chapter 3 of Basu et al. [13], Kuchibhotla and Basu [39], Al Mohamad [5], and the references therein. Due to the "curse of dimensionality", such a solution cannot be applied successfully in a large-dimension setting, as required in the

[14] In several situations, such a conversion can appear in a natural way; e.g. an institution may generate/collect data of "continuous value" but mask them for external data analysts to group-frequencies, for reasons of confidentiality (information asymmetry).

so called "big data" paradigm. For instance (in preparation for divergence valuation), take $\mathscr{X} = \mathbb{R}^d$, λ_2 to be the $d-$dimensional Lebesgue measure and $\mathbb{p}_N^{emp,smo}(x) := \frac{1}{N}\sum_{i=1}^N K(x, Y_i, h_n) = \int_{\mathscr{X}} K(x, y, h_n)\, \mathrm{d}\vec{\mathfrak{P}}_N^{emp}(y)$ where $K(\cdot,\cdot,\cdot)$ is an appropriate smooth kernel function with "bandwidth" h_n, e.g. $K(x, y, h_n) := \frac{1}{h_n}\widehat{K}\left(\frac{x-y}{h_n}\right)$ with appropriate nonnegative function $\widehat{K}(\cdot)$ satisfying $\int_{\mathbb{R}^d} \widehat{K}(y)\, \mathrm{d}\lambda_2(y) = 1$. Since such kernel smoothers KS use local averaging, and for large d most neighborhoods tend to be empty of data observations (because data often "live" on lower-dimensional manifolds, sparsity of data), a typical KS technique (choosing concrete kernels and bandwidths, etc.) needs then a huge amount N of data to provide a reasonable accuracy; for $d = 8$ one may need N to be 1 million. For background details, the reader is e.g. referred to DasGupta [28], Scott and Wand [77], Chapter 7 of Scott [76] and the references therein.

For the sake of brevity, a detailed discussion of (GR) and (SM) is beyond the scope of this paper.

4.5 Minimum Divergences - The Decomposability Method

Let us discuss yet another strategy to circumvent the problem described in Subsetup 2 of Sect. 4.3. As a motivation, for a divergence of the form

$$\begin{aligned} 0 \leqslant D_\lambda(\mathbb{P}, \mathbb{Q}) &= \int_{\mathscr{X}} f_1(x) \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x) \cdot \mathbb{q}(x)\big)\, \mathrm{d}\lambda(x) \\ &+ \int_{\mathscr{X}} f_2(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{]0,\infty[}(\mathbb{q}(x))\, \mathrm{d}\lambda(x) \\ &+ \int_{\mathscr{X}} f_3(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) \cdot \mathbf{1}_{]0,\infty[}(\mathbb{p}(x))\, \mathrm{d}\lambda(x) \end{aligned} \tag{98}$$

with $f_1(x) \geqslant 0$, $f_2(x) \geqslant 0$, $f_3(x) \geqslant 0$, and an "adjacent" dissimilarity

$$\begin{aligned} \widetilde{D}_\lambda(\mathbb{P}, \mathbb{Q}) &= \int_{\mathscr{X}} f_1(x) \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x) \cdot \mathbb{q}(x)\big)\, \mathrm{d}\lambda(x) \\ &+ \int_{\mathscr{X}} g_2(x)) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{]0,\infty[}(\mathbb{q}(x))\, \mathrm{d}\lambda(x) \\ &+ \int_{\mathscr{X}} g_3(x)) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{q}(x)\big) \cdot \mathbf{1}_{]0,\infty[}(\mathbb{p}(x))\, \mathrm{d}\lambda(x), \end{aligned} \tag{99}$$

there holds $D_\lambda(\mathbb{P}, \mathbb{Q}) = \widetilde{D}_\lambda(\mathbb{P}, \mathbb{Q})$ for all equivalent $\mathbb{P} \sim \mathbb{Q}$ (where for both, the second and third integral become zero), but (in case that $g_2(\cdot), g_3(\cdot)$ differ sufficiently enough from $f_2(\cdot), f_3(\cdot)$) one gets $D_\lambda(\mathbb{P}, \mathbb{Q}) \neq \widetilde{D}_\lambda(\mathbb{P}, \mathbb{Q})$ for $\mathbb{P} \perp \mathbb{Q}$ and even for $\mathbb{P} \nsim \mathbb{Q}$; in the latter two cases, depending on the signs of $g_2(\cdot)$, $g_3(\cdot)$, $\widetilde{D}_\lambda(\mathbb{P}, \mathbb{Q})$ may even become negative.

Such issues are of importance for our current problem where e.g. $\mathbb{P} := \vec{\mathbb{P}}_N^{\overline{emp}(\omega)} \perp \widetilde{\mathbb{Q}}_\theta =: \mathbb{Q}$. For further illuminations, and for the sake of a compact presentation, we use henceforth the notations $\mathscr{P}^\lambda$ for an arbitrarily fixed class of nonnegative, mutually equivalent functions (i.e. $\mathbb{P}_1 \sim \mathbb{P}_2$ for all $\mathbb{P}_1 \in \mathscr{P}^\lambda$, $\mathbb{P}_2 \in \mathscr{P}^\lambda$), and $\mathscr{P}^{\lambda\sim}$ for a

corresponding class of nonnegative (not necessarily mutually equivalent) functions such that $\mathbb{P}_1 \nsim \mathbb{P}_2$ for all $\mathbb{P}_1^\lambda \in \mathscr{P}^\lambda$, $\mathbb{P}_2 \in \mathscr{P}^{\lambda\nsim}$. Furthermore, we employ $\widetilde{\mathscr{P}}^\lambda := \mathscr{P}^\lambda \cup \mathscr{P}^{\lambda\nsim}$ and specify:

Definition 2 We say that a function $D_\lambda : \widetilde{\mathscr{P}}^\lambda \otimes \mathscr{P}^\lambda \to \mathbb{R}$ is a pseudo-divergence on $\widetilde{\mathscr{P}}^\lambda \times \mathscr{P}^\lambda$, if its restriction to $\mathscr{P}^\lambda \cup \mathscr{P}^\lambda$ is a divergence, i.e.

$$\begin{aligned} &D_\lambda(\mathbb{P}, \mathbb{Q}) \geqslant 0 \text{ for all } \mathbb{P} \in \mathscr{P}^\lambda, \mathbb{Q} \in \mathscr{P}^\lambda, \quad \text{and} \\ &D_\lambda(\mathbb{P}, \mathbb{Q}) = 0 \text{ if and only if } \mathbb{P} = \mathbb{Q} \in \mathscr{P}^\lambda . \end{aligned} \tag{100}$$

If also $D_\lambda(\mathbb{P}, \mathbb{Q}) > 0$ for all $\mathbb{P} \in \mathscr{P}^{\lambda\nsim}$,$\mathbb{Q} \in \mathscr{P}^\lambda$, then $D_\lambda(\cdot, \cdot)$ is a divergence.

As for interpretation, a pseudo-divergence $D_\lambda(\cdot, \cdot)$ acts like a divergence if both arguments are from $\mathscr{P}^\lambda$, but only like a dissimilarity if the first argument is from $\mathscr{P}^{\lambda\nsim}$ and thus is "quite different" from the second argument. In the following, we often use pseudo-divergences for our noisy minimum-distance-estimation problem – cf. (81), (82) – by taking $\lambda = \lambda_1 + \lambda_2$, $\mathscr{P}^\lambda := \mathscr{P}_\Theta^\lambda := \big(\widetilde{\mathbb{Q}}_\theta\big)_{\theta\in\Theta} := \big(\big\{\widetilde{\mathbb{q}}_\theta(x)\big\}_{x\in\mathscr{X}}\big)_{\theta\in\Theta}$ (cf. (87), (88)), and $\mathscr{P}^{\lambda\nsim} := \mathscr{P}_{emp}^{\lambda\perp} := \big(\mathbb{P}_N^{\overline{emp}(\omega)}\big)_{N\in\mathbb{N}} = \big(\big\{\mathbb{p}_N^{\overline{emp}(\omega)}(x)\big\}_{x\in\mathscr{X}}\big)_{N\in\mathbb{N}}$ (cf. (85), (Enc1)) covering all numbers N of data observations (sample sizes), and the according $\widetilde{\mathscr{P}}^\lambda := \mathscr{P}_{\Theta,emp}^\lambda = \mathscr{P}_\Theta^\lambda \cup \mathscr{P}_{emp}^{\lambda\perp}$; notice that by construction we have even the function-class-relationship $\perp$ which is stronger than $\nsim$. In such a setup, we have seen that for the choice $\mathbb{P} := \mathbb{P}_N^{\overline{emp}(\omega)}$, $\mathbb{Q} := \widetilde{\mathbb{Q}}_\theta$ the divergence $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{1}\cdot\mathbb{Q}^\chi,\lambda}$ $(\mathbb{P}, \mathbb{Q}) > 0$ of (90) is unpleasant for (Enc3) since the solution does not depend on the data-observations $Y_1(\omega), \ldots, Y_N(\omega)$; also recall the special case of power functions $\phi := \phi_\alpha$ (cf. (5)) with $\alpha = \chi > 1$ which amounts to the unscaled divergences (78), (40) and thus to (41). In (95), for general ϕ we have repaired this deficiency by replacing $\mathbb{Q} := \widetilde{\mathbb{Q}}_\theta$ with $\mathbb{Q} := \overline{\mathbb{Q}}_\theta$, at the cost of getting total mass larger than 1 but by keeping the strict positivity of the involved divergence; especially for $\phi := \phi_\alpha$, the divergence (41) has then amounted to (96).

In contrast, let us show another method to repair the (Enc3)-deficiency of (41), by sticking to $\mathbb{Q} := \widetilde{\mathbb{Q}}_\theta$ but changing the basically underlying divergence. In fact, we deal with the even more general

Definition 3 (a) We say that a pseudo-divergence $D_\lambda : \widetilde{\mathscr{P}}^\lambda \otimes \mathscr{P}^\lambda \to \mathbb{R}$ is decomposable if there exist functionals $\mathfrak{D}^0 : \widetilde{\mathscr{P}}^\lambda \mapsto \mathbb{R}$, $\mathfrak{D}^1 : \mathscr{Q} \mapsto \mathbb{R}$ and a (measurable) mapping $\rho_\mathbb{Q} : \mathscr{X} \mapsto \mathbb{R}$ (for each $\mathbb{Q} \in \mathscr{P}^\lambda$) such that[15]

[15] In an encompassing way, the part (a) reflects a measure-theoretic "plug-in" version of decomposable pseudo-divergences $D : (\mathscr{P}^{meas,\lambda_1} \cup \mathscr{P}^{meas,\lambda_2}) \otimes \mathscr{P}^{meas,\lambda_1} \mapsto \mathbb{R}$, where $\mathscr{P}^{meas,\lambda_1}$ is a family of mutually equivalent nonnegative measures of the form $\mathfrak{P}[\bullet] := \mathfrak{P}^{\mathbb{1}\cdot\lambda_1}[\bullet] := \int_\bullet \mathbb{p}(x)\,\mathrm{d}\lambda_1(x)$, $\mathscr{P}^{meas,\lambda_2}$ is a family of nonnegative measures of the form $\overline{\mathfrak{P}}[\bullet] := \overline{\mathfrak{P}}^{\mathbb{1}\cdot\lambda_2}[\bullet] := \int_\bullet \mathbb{q}(x)\,\mathrm{d}\lambda_2(x)$ such that any $\mathfrak{P} \in \mathscr{P}^{meas,\lambda_1}$ is not equivalent to any $\overline{\mathfrak{P}} \in \mathscr{P}^{meas,\lambda_2}$, and (101) is replaced with $D(\mathfrak{P}, \mathfrak{Q}) = \mathfrak{D}^0(\mathfrak{P}) + \mathfrak{D}^1(\mathfrak{Q}) + \int_\mathscr{X} \rho_\mathfrak{Q}(x)\,\mathrm{d}\mathfrak{P}(x)$ for all $\mathbb{P} \in \mathfrak{P} \in \mathscr{P}^{meas,\lambda_1} \cup \mathscr{P}^{meas,\lambda_2}$, $\mathfrak{Q} \in \mathscr{P}^{meas,\lambda_2}$; cf. Vajda [90], Broniatowski and Vajda [18], Broniatowski et al. [19]; part (b) is new.

$$D_\lambda(\mathbb{P},\mathbb{Q}) \;=\; \mathfrak{D}^0(\mathbb{P}) + \mathfrak{D}^1(\mathbb{Q}) + \int_{\mathscr{X}} \rho_{\mathbb{Q}}(x)\cdot \mathbb{p}(x)\,\mathrm{d}\lambda(x) \quad \text{for all } \mathbb{P}\in\widetilde{\mathscr{P}}^\lambda,\ \mathbb{Q}\in\mathscr{P}^\lambda. \tag{101}$$

(b) We say that a pseudo-divergence $D_\lambda : \widetilde{\mathscr{P}}^\lambda \otimes \mathscr{P}^\lambda \to \mathbb{R}$ is pointwise decomposable if it is of the form $D_\lambda(\mathbb{P},\mathbb{Q}) = \int_{\mathscr{X}} \psi^{dec}(\mathbb{p}(x), \mathbb{q}(x))\,\mathrm{d}\lambda(x)$ for some (measurable) mapping $\psi^{dec} : [0,\infty[\times[0,\infty[\mapsto \mathbb{R}$ with representation

$$\begin{aligned}
&\psi^{dec}(s,t) := \psi^0\big(s + h_0(x,s)\cdot \mathbf{1}_{\{0\}}(t)\big)\cdot \mathbf{1}_{]\overline{c}_0,\infty[}(s)\cdot \mathbf{1}_{]c_0,\infty[}(t)\\
&+\psi^1\big(t + h_1(x)\cdot \mathbf{1}_{\{0\}}(t)\big)\cdot \mathbf{1}_{]c_1,\infty[}(t)\\
&+\rho\big(t + h_2(x)\cdot \mathbf{1}_{\{0\}}(t)\big)\cdot s \quad \text{for all } (s,t)\in[0,\infty[\times[0,\infty[\backslash\{(0,0)\}\,,\\
&\psi^{dec}(0,0) := 0,
\end{aligned} \tag{102}$$

with constants $c_0, c_1, \overline{c}_0 \in \{0,1\}$, and (measurable) mappings $\psi^0, \psi^1, \rho : [0,\infty[\mapsto \mathbb{R}$, $h_1, h_2 : \mathscr{X} \mapsto [0,\infty[$, $h_0 : \mathscr{X}\times[0,\infty[\mapsto\mathbb{R}$, such that

$$\psi^{dec}(s,t) = \psi^0(s) + \psi^1(t) + \rho(t)\cdot s \geqslant 0 \quad \text{for all } (s,t)\in]0,\infty[\times]0,\infty[\,, \tag{103}$$

$$\psi^{dec}(s,t) = 0 \quad \text{if and only if} \quad s = t\,, \tag{104}$$

$$s + h_0(x,s) \geqslant 0 \quad \text{for all } s\in[0,\infty[\text{ and } \lambda\text{-almost all } x\in\mathscr{X}\,.$$

Remark 5 (a) Any pointwise decomposable pseudo-divergence is decomposable, under the additional assumption that the integral $\int_{\mathscr{X}} \ldots \mathrm{d}\lambda(x)$ can be split into three appropriate parts.
(b) For use in (Enc3), $\mathfrak{D}^1(\cdot)$ and $\rho_{\mathbb{Q}}(\cdot)$ should be non-constant.
(c) In the Definitions 2 and 3 we have put the "extension-role" to the first component $\mathbb{P}$; of course, everything can be worked out analogously for the second component $\mathbb{Q}$ by using (pseudo-)divergences $D_\lambda : \mathscr{P}^\lambda \times \widetilde{\mathscr{P}}^\lambda \to \mathbb{R}$.
(d) We could even extend (102) for bivariate functions $h_1(x,s)$, $h_2(x,s)$. □

Notice that from (102) one obtains the boundary behaviour

$$\mathbb{R} \ni \psi^{dec}(s,0) \;=\; \psi^0(s + h_0(x,s))\;\cdot\; \breve{c}_0 \;+\; \psi^1(h_1(x))\cdot \breve{c}_1 \;+\; \rho(h_2(x))\;\cdot\; s \text{ for all } s>0, \tag{105}$$

$$\mathbb{R} \ni \psi^{dec}(0,t) = \psi^0(0)\cdot \breve{\overline{c}}_0 + \psi^1(t) \quad \text{for all } t>0\,, \tag{106}$$

with $\breve{c}_0 := \mathbf{1}_{]c_0,\infty[}(0)$, $\breve{c}_1 := \mathbf{1}_{]c_1,\infty[}(0)$, $\breve{\overline{c}}_0 := \mathbf{1}_{]\overline{c}_0,\infty[}(0)$. Notice that $\psi^{dec}(s,0)$ of (105) does generally not coincide with the eventually existent "(103)-limit" $\lim_{t\to 0}[\psi^0(s) + \psi^1(t) + \rho(t)\cdot s]$ $(s>0)$, which reflects a possibly "non-smooth boundary behaviour" (also recall (98), (99)). Moreover, when choosing a decomposable pseudo-divergence (101) in step (Enc2), we operationalize the solution $\widehat{\theta}_N(\omega)$ of the corresponding noisy minimization problem (86) of step (Enc3) as follows:

Definition 4 (a) We say that a functional $T_{D_\lambda} : \mathscr{P}^{\lambda}_{\Theta,emp} \mapsto \Theta$ generates a minimum decomposable pseudo-divergence estimator (briefly, min $-decD_\lambda$-estimator)

$$\widehat{\theta}_{N,decD_\lambda}(\omega) := T_{D_\lambda}\big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}\big) \quad \text{for } \vec{\mathbb{P}}_N^{\overline{emp}(\omega)} \in \mathscr{P}^{\lambda\perp}_{emp} \tag{107}$$

of the true unknown parameter θ_0, if $D_\lambda(\cdot,\cdot) : \mathscr{P}^{\lambda}_{\Theta,emp} \otimes \mathscr{P}^{\lambda}_{\Theta} \mapsto \mathbb{R}$ is a decomposable pseudo-divergence and

$$T_{D_\lambda}\big(\vec{\mathbb{P}}\big) = \operatorname{arginf}_{\theta\in\Theta}\big[\mathfrak{D}^1(\vec{\mathbb{Q}}_\theta) + \int_{\mathscr{X}} \rho_{\vec{\mathbb{Q}}_\theta}(x) \cdot \vec{\mathbb{p}}(x)\,\mathrm{d}\lambda(x)\big] \text{ for all } \vec{\mathbb{P}} \in \mathscr{P}^{\lambda}_{\Theta,emp}. \tag{108}$$

(b) If $D_\lambda(\cdot,\cdot)$ is a pointwise decomposable pseudo-divergence we replace (108) by

$$T_{D_\lambda}\big(\vec{\mathbb{P}}\big) = \operatorname{arginf}_{\theta\in\Theta} \int_{\mathscr{X}} \psi^{dec}(\vec{\mathbb{p}}(x), \widetilde{\vec{\mathbb{q}}}_\theta(x))\,\mathrm{d}\lambda(x) \quad \text{for all } \vec{\mathbb{P}} \in \mathscr{P}^{\lambda}_{\Theta,emp}\,,$$

but do not introduce a new notion (also recall that $\lambda = \lambda_2$ and $\widetilde{\vec{\mathbb{q}}}_\theta(\cdot) = \vec{\mathbb{q}}_\theta(\cdot)$ for the case of no observations, e.g. if $\vec{\mathbb{P}} \in \mathscr{P}^{\lambda_2}_{\Theta}$).

To proceed, let us point out that by (107) and (97) every min $-decD_\lambda$-estimator rewrites straightforwardly as

$$\widehat{\theta}_{N,decD_\lambda}(\omega) = \operatorname{arginf}_{\theta\in\Theta}\big[\mathfrak{D}^1(\vec{\mathbb{Q}}_\theta) + \tfrac{1}{N}\textstyle\sum_{i=1}^N \rho_{\vec{\mathbb{Q}}_\theta}(Y_i(\omega))\big] \tag{109}$$

and is Fisher consistent in the sense that

$$T_{\mathfrak{D}}(\vec{\mathbb{Q}}_{\theta_0}) = \operatorname{arginf}_{\theta\in\Theta} \mathfrak{D}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_\theta) = \theta_0 \quad \text{for all } \theta_0 \in \Theta\,. \tag{110}$$

Furthermore, the criterion to be minimized in (109) is of the form

$$\theta \;\mapsto\; \mathfrak{D}^1(\vec{\mathbb{Q}}_\theta) + \tfrac{1}{N}\textstyle\sum_{i=1}^N \rho_{\vec{\mathbb{Q}}_\theta}(Y_i(\omega))$$

which e.g. for the task (Enc5) opens the possibility to apply the methods of the asymptotic theory of so-called M-estimators (cf. e.g. Hampel et al. [33], van der Vaart and Wellner [88], Liese and Mieske [40]). The concept of min $-decD_\lambda$-estimators (101) were introduced in Vajda [90], Broniatowski and Vajda [18] within the probability-law-restriction of the non-encompassing, "plug-in" context of footnote 15.

In the following, we demonstrate that our new concept of pointwise decomposability defined by (102) is very useful and flexible for creating new min $-decD_\lambda$-estimators and imbedding existing ones. In fact, since in our current statistics-ML-AI context we have chosen $\lambda[\bullet] := \lambda_1[\bullet] + \lambda_2[\bullet]$ with $\lambda_1[\bullet] := \sum_{z\in\mathscr{Z}} \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(z) \cdot \delta_z[\bullet]$ and $\lambda_2[\bullet]$ stemming from (87), we have seen that $\mathbb{P} := \vec{\mathbb{P}}_N^{\overline{emp}(\omega)} \perp \vec{\mathbb{Q}}_\theta =: \vec{\mathbb{Q}}$ for all $\theta \in \Theta$. Hence, from (102), (105), (106) we obtain

$$
\begin{aligned}
&D_\lambda\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big) = \int_{\mathscr{X}} \psi^{dec}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x), \widetilde{\vec{\mathbb{q}}}_\theta(x)\Big)\, \mathrm{d}\lambda(x)\\
&= \int_{\mathscr{X}} \Big[\psi^0(0)\cdot \breve{\overline{c}}_0 + \psi^1\Big(\widetilde{\vec{\mathbb{q}}}_\theta(x)\Big)\Big]\cdot \mathbf{1}_{\{0\}}(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x))\, \mathrm{d}(\lambda_1+\lambda_2)(x)\\
&+ \int_{\mathscr{X}} \Big[\psi^0\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x) + h_0\Big(x, \vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x)\Big)\Big)\cdot \breve{c}_0\\
&+\psi^1(h_1(x))\cdot \breve{c}_1 + \rho(h_2(x))\cdot \vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x)\Big]\cdot \mathbf{1}_{\{0\}}(\widetilde{\vec{\mathbb{q}}}_\theta(x))\, \mathrm{d}(\lambda_1+\lambda_2)(x)\\
&= \int_{\mathscr{X}} \Big[\psi^0(0)\cdot \breve{\overline{c}}_0 + \psi^1\Big(\vec{\mathbb{q}}_\theta(x)\Big)\Big]\, \mathrm{d}\lambda_2(x) + \sum_{x\in\mathscr{X}} \Big[\psi^0\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x)\\
&+h_0\Big(x, \vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x)\Big)\Big)\cdot \breve{c}_0\\
&+\psi^1(h_1(x))\cdot \breve{c}_1 + \rho(h_2(x))\cdot \vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x)\Big]\cdot \mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x)\\
&= \int_{\mathscr{X}} \Big[\psi^0(0)\cdot \breve{\overline{c}}_0 + \psi^1\Big(\vec{\mathbb{q}}_\theta(x)\Big)\Big]\, \mathrm{d}\lambda_2(x) + \tfrac{1}{N}\sum_{i=1}^N \rho(h_2(Y_i(\omega)))\\
&+\tfrac{1}{N}\sum_{i=1}^N \frac{\psi^0\Big(\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega)) + h_0\Big(Y_i(\omega), \vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))\Big)\Big)\cdot \breve{c}_0 + \psi^1(h_1(Y_i(\omega)))\cdot \breve{c}_1}{\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))}\,,
\end{aligned}
\tag{111}
$$

where we have employed (97); recall that $\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega)) = \#\{j \in \{1,\ldots,N\} : Y_j(\omega) = Y_i(\omega)\}/N$. Hence, we always choose $\mathfrak{D}^1(\vec{\mathbb{Q}}) = \mathfrak{D}^1(\widetilde{\vec{\mathbb{Q}}}_\theta) = \int_{\mathscr{X}} \Big[\psi^0(0) + \psi^1\Big(\widetilde{\vec{\mathbb{q}}}_\theta(x)\Big)\Big]\, \mathrm{d}\lambda_2(x) = \int_{\mathscr{X}} \Big[\psi^0(0) + \psi^1\Big(\vec{\mathbb{q}}_\theta(x)\Big)\Big]\, \mathrm{d}\lambda_2(x) = \mathfrak{D}^1(\vec{\mathbb{Q}}_\theta)$. Notice that the functions h_0, h_1, h_2 may depend on the parameter θ. Indeed, for $h_0(x,s) \equiv 0$, $h_1(x) \equiv 0$, $h_2(x) = \vec{\mathbb{q}}_\theta(x)$ $(\neq \widetilde{\vec{\mathbb{q}}}_\theta(x))$, the pseudo-divergence (111) turns into

$$
\begin{aligned}
&D_\lambda\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big) = \int_{\mathscr{X}} \Big[\psi^0(0)\cdot \breve{\overline{c}}_0 + \psi^1\Big(\vec{\mathbb{q}}_\theta(x)\Big)\Big]\, \mathrm{d}\lambda_2(x) + \tfrac{1}{N}\sum_{i=1}^N \rho\Big(\vec{\mathbb{q}}_\theta(Y_i(\omega))\Big)\\
&+\tfrac{1}{N}\sum_{i=1}^N \frac{\psi^0\Big(\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))\Big)\cdot \breve{c}_0 + \psi^1(0)\cdot \breve{c}_1}{\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))}\,,
\end{aligned}
\tag{112}
$$

whereas for $h_0(x,s) \equiv 0$, $h_1(x) = \vec{\mathbb{q}}_\theta(x)$, $h_2(x) = \vec{\mathbb{q}}_\theta(x)$, (111) becomes

$$
\begin{aligned}
&D_\lambda\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big) = \int_{\mathscr{X}} \Big[\psi^0(0)\cdot \breve{\overline{c}}_0 + \psi^1\Big(\vec{\mathbb{q}}_\theta(x)\Big)\Big]\, \mathrm{d}\lambda_2(x)\\
&+\tfrac{1}{N}\sum_{i=1}^N \Big[\rho\Big(\vec{\mathbb{q}}_\theta(Y_i(\omega))\Big) + \frac{\breve{c}_1\cdot\psi^1(\vec{\mathbb{q}}_\theta(Y_i(\omega)))}{\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))}\Big] + \tfrac{1}{N}\sum_{i=1}^N \frac{\psi^0\Big(\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))\Big)\cdot \breve{c}_0}{\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))}\,.
\end{aligned}
\tag{113}
$$

The last sum in (112) respectively (113) is the desired $\mathfrak{D}^0(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)})$. As an example, let us take $c_0 = c_1 = \overline{c}_0 = -1$ (and hence, $\breve{c}_0 = \breve{c}_1 = \breve{\overline{c}}_0 = 1$) and for $\alpha > 1$ the power functions $\phi(t) := \phi_\alpha(t) := \frac{t^\alpha - \alpha\cdot t + \alpha - 1}{\alpha\cdot(\alpha-1)}$ $(t \in]0,\infty[)$ of (6), for which by (9)

and (103) one derives immediately the decomposition $\psi^0(t) := \psi^0_\alpha(t) := \frac{t^\alpha}{\alpha(\alpha-1)} > 0, \psi^1(t) := \psi^1_\alpha(t) := \frac{t^\alpha}{\alpha} > 0, \rho(t) := \rho_\alpha(t) := -\frac{t^{\alpha-1}}{\alpha-1} < 0\,(t \in]0,\infty[)$. Accordingly, (111) simplifies to

$$\begin{aligned} &D_\lambda\Big(\overrightarrow{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\overrightarrow{\mathbb{Q}}}_\theta\Big) := D_{\lambda,\alpha}\Big(\overrightarrow{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\overrightarrow{\mathbb{Q}}}_\theta\Big) \\ &= \tfrac{1}{\alpha}\int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)^\alpha \,\mathrm{d}\lambda_2(x) - \tfrac{1}{N\cdot(\alpha-1)}\sum_{i=1}^N \Big(h_2(Y_i(\omega))\Big)^{\alpha-1} \\ &+\tfrac{1}{N}\sum_{i=1}^N \frac{\Big(\overrightarrow{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega)) + h_0\Big(Y_i(\omega), \overrightarrow{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))\Big)\Big)^\alpha + (\alpha-1)\cdot\Big(h_1(Y_i(\omega))\Big)^\alpha}{\alpha\cdot(\alpha-1)\cdot\overrightarrow{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))}, \end{aligned} \tag{114}$$

and in particular the special case (112) turns into

$$\begin{aligned} &D_{\lambda,\alpha}\Big(\overrightarrow{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\overrightarrow{\mathbb{Q}}}_\theta\Big) = \tfrac{1}{\alpha}\int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)^\alpha \,\mathrm{d}\lambda_2(x) - \tfrac{1}{N\cdot(\alpha-1)}\sum_{i=1}^N \Big(\vec{\mathbb{q}}_\theta(Y_i(\omega))\Big)^{\alpha-1} \\ &+\tfrac{1}{N\cdot\alpha\cdot(\alpha-1)}\sum_{i=1}^N \Big(\overrightarrow{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))\Big)^{\alpha-1}, \end{aligned} \tag{115}$$

whereas the special case (113) simplifies to

$$\begin{aligned} &0 < D_{\lambda,\alpha}\Big(\overrightarrow{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\overrightarrow{\mathbb{Q}}}_\theta\Big) = \tfrac{1}{\alpha}\int_{\mathscr{X}} \vec{\mathbb{q}}_\theta(x)^\alpha \,\mathrm{d}\lambda_2(x) \\ &+\tfrac{1}{N}\sum_{i=1}^N \left[\frac{\Big(\vec{\mathbb{q}}_\theta(Y_i(\omega))\Big)^\alpha}{\alpha\cdot\overrightarrow{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))} - \frac{\Big(\vec{\mathbb{q}}_\theta(Y_i(\omega))\Big)^{\alpha-1}}{\alpha-1}\right] + \tfrac{1}{N}\sum_{i=1}^N \frac{\Big(\overrightarrow{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))\Big)^{\alpha-1}}{\alpha\cdot(\alpha-1)}. \end{aligned} \tag{116}$$

Notice that (116) coincides with (96), but both were derived within quite different frameworks: to obtain (116) we have used the concept of decomposable pseudo-divergences (which may generally become negative at the boundary) together with $\overrightarrow{\mathbb{Q}} := \widetilde{\overrightarrow{\mathbb{Q}}}_\theta$ which leads to total mass of 1 (cf. (88)); on the other hand, for establishing (96) we have employed the concept of divergences (which are generally always strictly positive at the boundary) together with $\overrightarrow{\mathbb{Q}} := \overrightarrow{\mathbb{Q}}_\theta$ which amounts to total mass greater than 1 (cf. (91)). Moreover, choosing $h_0(x,s) \equiv 0$, $h_1(x) \equiv 0$, $h_2(x) \equiv 0$ in (114) gives exactly the divergence (90) for the current generator $\phi(t) := \phi_\alpha(t)$ with $\alpha > 1$; recall that the latter has been a starting motivation for the search of repairs. For $c_0 = c_1 = \bar{c}_0 = -1$ and the limit case $\alpha \to 1$ one gets $\phi(t) := \phi_1(t) := t\cdot\log t + 1 - t\,(t \in]0,\infty[)$ of (18), for which by (22) and (103) we obtain the decomposition $\psi^0(t) := \psi^0_1(t) := t\cdot\log t - t$, $\psi^1(t) := \psi^1_1(t) := t > 0$, $\rho(t) := \rho_1(t) := -\log t$. Accordingly, (111) simplifies to

$$D_\lambda\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big) := D_{\lambda,1}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big)$$
$$= 1 - \tfrac{1}{N}\sum_{i=1}^N \log\Big(h_2(Y_i(\omega))\Big)$$
$$+\tfrac{1}{N}\sum_{i=1}^N \frac{\psi_1^0\Big(\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega)) + h_0\Big(Y_i(\omega), \vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))\Big)\Big) + h_1(Y_i(\omega))}{\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))}, \tag{117}$$

and in particular the special case (112) turns into

$$D_{\lambda,1}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big) = \frac{1}{N}\sum_{i=1}^N \log\Big(\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))\Big) - \frac{1}{N}\sum_{i=1}^N \log\Big(\vec{\mathbb{q}}_\theta(Y_i(\omega))\Big), \tag{118}$$

whereas the special case (113) becomes

$$0 < D_{\lambda,1}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big) = \tfrac{1}{N}\sum_{i=1}^N \log\Big(\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))\Big) - \tfrac{1}{N}\sum_{i=1}^N \log\Big(\vec{\mathbb{q}}_\theta(Y_i(\omega))\Big)$$
$$+\tfrac{1}{N}\sum_{i=1}^N \frac{\vec{\mathbb{q}}_\theta(Y_i(\omega))}{\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))}. \tag{119}$$

To end up this subsection, let us briefly indicate that choosing in step (Enc2) a decomposable pseudo-divergence of the form (respectively) (111)–(119), and in the course of (Enc3) minimize this over $\theta \in \Theta$, we end up at the corresponding min $-dec D_\lambda$-estimator (109). For the special case (118) (i.e. $\alpha = 1$) this leads to the omnipresent, celebrated *maximum-likelihood-estimator* (MLE) which is known to be efficient but not robust. The particular choice (115) for $\alpha > 1$ gives the density-power divergence estimator DPDE of Basu et al. [10], where $\alpha = 2$ amounts to the (squared) L_2-estimator which is robust but not efficient (see e.g. Hampel et al. [33]); accordingly, taking $\alpha \in]1, 2[$ builds a smooth bridge between the robustness and efficiency. The reversed version of the DPDE can be analogously imbedded in our context, by employing our new approach with $\phi(t) := \widetilde{\widetilde{\phi}}_\alpha(t)$ (cf. (79)).

4.6 *Minimum Divergences - Generalized Subdivergence Method*

One can flexibilize some of the methods of the previous Sect. 4.5, by employing an additional (a.s.) strictly positive density function $\mathbb{M}$ to define a pseudo-divergence $D_{\mathbb{M},\lambda} : \widetilde{\mathscr{P}}^\lambda \otimes \mathscr{P}^\lambda \to \mathbb{R}$ of the form $D_{\mathbb{M},\lambda}(\mathbb{P}, \mathbb{Q}) =$
$\int_{\mathscr{X}} \psi^{dec}\Big(\frac{\mathbb{p}(x)}{\mathbb{m}(x)}, \frac{\mathbb{q}(x)}{\mathbb{m}(x)}\Big) \cdot \mathbb{m}(x)\, \mathrm{d}\lambda(x)$ for some (measurable) mapping
$\psi^{dec} : [0, \infty[\times [0, \infty[\mapsto \mathbb{R}$ with representation

$$\begin{aligned}\psi^{dec}(s,t) &:= \psi^0\Big(s + h_0(x,s)\cdot \mathbf{1}_{\{0\}}(t)\Big)\cdot \mathbf{1}_{]\overline{c}_0,\infty[}(s)\cdot \mathbf{1}_{]c_0,\infty[}(t)\\ &\quad +\psi^1\Big(t + h_1(x)\cdot \mathbf{1}_{\{0\}}(t)\Big)\cdot \mathbf{1}_{]c_1,\infty[}(t)\\ &\quad +\rho\Big(t + h_2(x)\cdot \mathbf{1}_{\{0\}}(t)\Big)\cdot s \quad \text{for all } (s,t)\in [0,\infty[\times[0,\infty[\setminus\{(0,0)\}\,, \quad (cf.(102))\\ \psi^{dec}(0,0) &:= 0.\end{aligned}$$

It is straightforward to see that $D_{\mathbb{M},\lambda}(\cdot,\cdot)$ is a pointwise decomposable pseudo-divergence in the sense of Definition 3(b), and one gets for fixed $m > 0$

$$\psi_m^{dec}(s,t) := m\cdot\psi^{dec}\Big(\tfrac{s}{m},\tfrac{t}{m}\Big) = m\cdot\psi^0\Big(\tfrac{s}{m}\Big) + m\cdot\psi^1\Big(\tfrac{t}{m}\Big) + \rho\Big(\tfrac{t}{m}\Big)\cdot s \geqslant 0 \quad \text{for all } (s,t)\in\,]0,\infty[\times]0,\infty[\,, \tag{120}$$

$$\begin{aligned}&\psi_m^{dec}(s,t) = 0 \quad \text{if and only if} \quad s = t\,,\\ &\tfrac{s}{m} + h_0\Big(x,\tfrac{s}{m}\Big) \geqslant 0 \qquad \text{for all } s\in[0,\infty[\text{ and } \lambda\text{-almost all } x\in\mathscr{X}\,,\\ &\mathbb{R}\ni\psi_m^{dec}(s,0) = m\cdot\psi^0\Big(\tfrac{s}{m} + h_0\Big(x,\tfrac{s}{m}\Big)\Big)\cdot\breve{c}_0 + m\cdot\psi^1(h_1(x))\cdot\breve{c}_1 + \rho(h_2(x))\cdot s \\ &\qquad\qquad \text{for all } s > 0\,, \end{aligned} \tag{121}$$

$$\mathbb{R}\ni\psi_m^{dec}(0,t) = m\cdot\psi^0(0)\cdot\breve{\overline{c}}_0 + m\cdot\psi^1\Big(\tfrac{t}{m}\Big) \quad \text{for all } t > 0\,. \tag{122}$$

For each class-family member $\mathbb{M} := \vec{\mathbb{Q}}_\tau$ with arbitrarily fixed $\tau\in\Theta$, we can apply Definition 4 to $D_\lambda(\cdot,\cdot) := D_{\underset{\mathbb{Q}_\tau,\lambda}{\rightharpoonup}}(\cdot,\cdot)$, and arrive at the corresponding $\min-dec D_{\underset{\mathbb{Q}_\tau,\lambda}{\rightharpoonup}}$-estimators

$$\widehat{\theta}_{N,dec D_{\underset{\mathbb{Q}_\tau,\lambda}{\rightharpoonup}}}(\omega) := T_{D_{\underset{\mathbb{Q}_\tau,\lambda}{\rightharpoonup}}}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}\Big) \quad \text{for } \vec{\mathbb{P}}_N^{\overline{emp}(\omega)}\in\mathscr{P}_{emp}^{\lambda\perp} \tag{123}$$

of the true unknown parameter θ_0. Hence, analogously to the derivation of (111), we obtain from (102), (121), (122) for each $\tau\in\Theta$

$$\begin{aligned}&D_{\underset{\mathbb{Q}_\tau,\lambda}{\rightharpoonup}}\Big(\mathbb{P}_N^{\overline{emp}(\omega)},\widetilde{\mathbb{Q}}_\theta\Big) = \int_{\mathscr{X}}\ \psi^{dec}\Big(\tfrac{\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x)}{\vec{\mathbb{Q}}_\tau(x)},\tfrac{\vec{\widetilde{\mathbb{Q}}}_\theta(x)}{\vec{\mathbb{Q}}_\tau(x)}\Big)\cdot\mathfrak{Q}_\tau(x)\,\mathrm{d}\lambda(x)\\
&= \int_{\mathscr{X}}\ \psi^1\Big(\tfrac{\vec{\mathbb{Q}}_\theta(x)}{\vec{\mathbb{Q}}_\tau(x)}\Big)\cdot\mathfrak{Q}_\tau(x)\,\mathrm{d}\lambda_2(x) + \sum_{x\in\mathscr{X}}\Big[\mathfrak{Q}_\tau(x)\cdot\psi^0\Big(\tfrac{\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x)}{\vec{\mathbb{Q}}_\tau(x)} + h_0\Big(x,\tfrac{\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x)}{\vec{\mathbb{Q}}_\tau(x)}\Big)\Big)\cdot\breve{c}_0\\
&+\mathfrak{Q}_\tau(x)\cdot\psi^1(h_1(x))\cdot\breve{c}_1 + \rho(h_2(x))\cdot\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}(x)\Big]\cdot\mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x) + \psi^0(0)\cdot\breve{\overline{c}}_0\\
&= \int_{\mathscr{X}}\ \psi^1\Big(\tfrac{\vec{\mathbb{Q}}_\theta(x)}{\vec{\mathbb{Q}}_\tau(x)}\Big)\cdot\mathfrak{Q}_\tau(x)\,\mathrm{d}\lambda_2(x) + \tfrac{1}{N}\sum_{i=1}^N\rho(h_2(Y_i(\omega))) + \psi^0(0)\cdot\breve{\overline{c}}_0\\
&+\tfrac{1}{N}\sum_{i=1}^N\frac{\psi^0\Big(\frac{\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))}{\vec{\mathbb{Q}}_\tau(Y_i(\omega))} + h_0\Big(Y_i(\omega),\frac{\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))}{\vec{\mathbb{Q}}_\tau(Y_i(\omega))}\Big)\Big)\cdot\breve{c}_0 + \psi^1(h_1(Y_i(\omega)))\cdot\breve{c}_1}{\vec{\mathbb{P}}_N^{emp(\omega)}(Y_i(\omega))}\cdot\mathfrak{Q}_\tau(Y_i(\omega))\,.\end{aligned} \tag{124}$$

Just as in the derivation of (112) respectively (113), reasonable choices for the "boundary-functions" in (124) are $h_0(x,s) \equiv 0$, $h_1(x) \equiv 0$, $h_2(x) = \frac{\vec{q}_\theta(x)}{\vec{q}_\tau(x)}$, respectively $h_0(x,s) \equiv 0$, $h_1(x) \equiv \frac{\vec{q}_\theta(x)}{\vec{q}_\tau(x)}$, $h_2(x) = \frac{\vec{q}_\theta(x)}{\vec{q}_\tau(x)}$. As for example, consider for all $\theta_0, \theta, \tau \in \Theta$ the scaled Bregman divergences in the sense of Stummer [81], Stummer and Vajda [84] (cf. Remark (2)(b)), for which we get from (36) with $r(x) \equiv 1$

$$
\begin{aligned}
&0 \leqslant D^c_{\phi, \vec{\mathbb{Q}}_\tau, \vec{\mathbb{Q}}_\tau, \mathbb{1}\cdot\vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_\theta) \\
&:= \int_{\mathscr{X}} \left[\phi\Big(\frac{\vec{q}_{\theta_0}(x)}{\vec{q}_\tau(x)}\Big) - \phi\Big(\frac{\vec{q}_\theta(x)}{\vec{q}_\tau(x)}\Big) - \phi'_{+,c}\Big(\frac{\vec{q}_\theta(x)}{\vec{q}_\tau(x)}\Big) \cdot \Big(\frac{\vec{q}_{\theta_0}(x)}{\vec{q}_\tau(x)} - \frac{\vec{q}_\theta(x)}{\vec{q}_\tau(x)}\Big) \right] \cdot \vec{q}_\tau(x)\, \mathrm{d}\lambda_2(x)\,, \\
&=: D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_\theta)\,, \qquad (125)
\end{aligned}
$$

from which – together with (120) – one can identify immediately the pointwise decomposability with $\psi^0(s) := \psi^0_\phi(s) := \phi(s), \psi^1(t) := \psi^1_\phi(t) := t \cdot \phi'_{+,c}(t) - \phi(t), \rho(t) := \rho_\phi(t) := -\phi'_{+,c}(t)$; by plugging this into (124), one obtains the objective $D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big)$, which in the course of (Enc3) should be – for fixed $\tau \in \Theta$ – minimized over $\theta \in \Theta$ in order to obtain the corresponding τ-individual" $\min - dec D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda}$-estimator $\widehat{\theta}_{N,\tau}(\omega) := \operatorname{arginf}_{\theta\in\Theta} D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big)$. Recall that this choice can be motivated by $0 = \min_{\theta\in\Theta} D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_\theta)$ and $\theta_0 = \operatorname{argmin}_{\theta\in\Theta} D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_\theta)$. Furthermore, one gets even $0 = \min_{\theta\in\Theta} \min_{\tau\in\Theta} D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_\theta)$, $\theta_0 = \operatorname{argmin}_{\theta\in\Theta} \min_{\tau\in\Theta} D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_\theta)$, and in case of $\max_{\tau\in\Theta} D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_\theta) < \infty$ also $0 = \min_{\theta\in\Theta} \max_{\tau\in\Theta} D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_\theta)$, $\theta_0 = \operatorname{argmin}_{\theta\in\Theta} \max_{\tau\in\Theta} D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_\theta)$. This suggests the alternative, "τ-uniform" estimators $\widehat{\theta}_N(\omega) := \operatorname{argmin}_{\theta\in\Theta} \min_{\tau\in\Theta} D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big)$, respectively $\widehat{\theta}_N(\omega) := \operatorname{argmin}_{\theta\in\Theta} \max_{\tau\in\Theta} D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big)$. As a side remark, let us mention that in general, (say) $\min_{\tau\in\Theta} D_{\phi, \vec{\mathbb{Q}}_\tau, \lambda}\Big(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)}, \widetilde{\vec{\mathbb{Q}}}_\theta\Big)$ is not necessarily decomposable anymore, and therefore the standard theory of M-estimators is not applicable to this class of estimators.

With our approach, we can generate numerous further estimators of the true unknown parameter θ_0, by permuting the positions – but not the roles (!) – of the parameters (θ_0, θ, τ) in the (pseudo-)divergences of the above investigations. For the sake of brevity, we only sketch two further cases; the full variety will appear elsewhere. To start with, consider the adaptively scaled and aggregated divergence

$$0 \leqslant D^{rev}_{\phi,\vec{\mathbb{Q}}_\tau,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0},\vec{\mathbb{Q}}_\theta) := D^{c}_{\phi,\vec{\mathbb{Q}}_{\theta_0}{}^2/\vec{\mathbb{Q}}_\tau,\vec{\mathbb{Q}}_\theta{}^2/\vec{\mathbb{Q}}_\tau,\mathbb{1}\cdot\vec{\mathbb{Q}}_{\theta_0},\lambda_2}(\vec{\mathbb{Q}}_{\theta_0},\vec{\mathbb{Q}}_\theta)$$

$$:= \int_{\mathscr{X}} \left[\phi\left(\frac{\vec{\mathbb{q}}_{\theta_0}(x)}{\frac{\vec{\mathbb{q}}_{\theta_0}(x)^2}{\vec{\mathbb{q}}_\tau(x)}}\right) - \phi\left(\frac{\vec{\mathbb{q}}_\theta(x)}{\frac{\vec{\mathbb{q}}_\theta(x)^2}{\vec{\mathbb{q}}_\tau(x)}}\right) - \phi'_{+,c}\left(\frac{\vec{\mathbb{q}}_\theta(x)}{\frac{\vec{\mathbb{q}}_\theta(x)^2}{\vec{\mathbb{q}}_\tau(x)}}\right)\cdot\left(\left(\frac{\vec{\mathbb{q}}_{\theta_0}(x)}{\frac{\vec{\mathbb{q}}_{\theta_0}(x)^2}{\vec{\mathbb{q}}_\tau(x)}}\right) - \left(\frac{\vec{\mathbb{q}}_\theta(x)}{\frac{\vec{\mathbb{q}}_\theta(x)^2}{\vec{\mathbb{q}}_\tau(x)}}\right)\right)\right]$$

$$\cdot\vec{\mathbb{q}}_\tau(x)\,\mathrm{d}\lambda_2(x)$$

$$= \int_{\mathscr{X}} \left[\phi\left(\frac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_{\theta_0}(x)}\right) - \phi\left(\frac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\right) - \phi'_{+,c}\left(\frac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\right)\cdot\left(\frac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_{\theta_0}(x)} - \frac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\right)\right]\cdot\vec{\mathbb{q}}_{\theta_0}(x)\,\mathrm{d}\lambda_2(x)$$

$$=: \int_{\mathscr{X}} \left[\psi^{0,rev}_{\vec{\mathbb{q}}_\tau(x)}(\vec{\mathbb{q}}_{\theta_0}(x)) + \psi^{1,rev}_{\vec{\mathbb{q}}_\tau(x)}(\vec{\mathbb{q}}_\theta(x)) + \rho^{rev}_{\vec{\mathbb{q}}_\tau(x)}(\vec{\mathbb{q}}_\theta(x))\cdot\vec{\mathbb{q}}_{\theta_0}(x)\right]\mathrm{d}\lambda_2(x)$$

(indeed, by Theorem 4 and (80) this is zero if and only if $\theta = \theta_0$). By means of the involved mappings $\psi^0(s) := \psi^{0,rev}_m(s) := s\cdot\phi(\frac{m}{s})$, $\psi^1(t) := \psi^{1,rev}_m(t) := -m\cdot\phi'_{+,c}(\frac{m}{t})$, $\rho(t) := \rho^{rev}_m(t) := \frac{m}{t}\cdot\phi'_{+,c}(\frac{m}{t}) - \phi(\frac{m}{t}) =: \phi^{\circledcirc}(\frac{m}{t})$ $(s,t,m>0)$, the properties (103), (104) are applicable and thus $D^{rev}_{\phi,\vec{\mathbb{Q}}_\tau,\lambda_2}(\cdot,\cdot)$ can be extended to a pointwise decomposable pseudo-divergence on $\widetilde{\mathscr{P}}^\lambda\otimes\mathscr{P}^\lambda$ by using (102) with appropriate functions h_0,h_1,h_2 and constants c_0,c_1,$\overline{c}_0$. Furthermore, by minimizing over $\theta\in\Theta$ the objective (111) with these choices $\psi^{0,rev}_m(\cdot)$, $\psi^{1,rev}_m(\cdot)$, $\rho^{rev}_m(\cdot)$, in the course of (Enc3) we end up at the corresponding $\min -dec D^{rev}_{\phi,\vec{\mathbb{Q}}_\tau,\lambda}$-estimator. In particular, the corresponding special case $h_0(x,s)\equiv 0$, $h_1(x)\equiv 1$, $h_2(x) = \vec{\mathbb{q}}_\theta(x)$ $(\neq\widetilde{\vec{\mathbb{q}}}_\theta(x))$ leads to the objective (cf. (112) but with $\psi^1(1)$ instead of $\psi^1(0)$)

$$D^{rev}_{\phi,\vec{\mathbb{Q}}_\tau,\lambda_2}\left(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)},\widetilde{\vec{\mathbb{Q}}}_\theta\right) = \phi^*(0)\cdot\breve{c}_0 - \int_{\mathscr{X}}\vec{\mathbb{q}}_\tau(x)\cdot\phi'_{+,c}\left(\frac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\right)\mathrm{d}\lambda_2(x)$$
$$+\tfrac{1}{N}\textstyle\sum_{i=1}^N \phi^{\circledcirc}\left(\frac{\vec{\mathbb{q}}_\tau(Y_i(\omega))}{\vec{\mathbb{q}}_\theta(Y_i(\omega))}\right)$$
$$+\tfrac{1}{N}\textstyle\sum_{i=1}^N\left[\phi\left(\frac{\vec{\mathbb{q}}_\tau(Y_i(\omega))}{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))}\right)\cdot\breve{c}_0 - \frac{\vec{\mathbb{q}}_\tau(Y_i(\omega))\cdot\phi'_{+,c}(\vec{\mathbb{q}}_\tau(Y_i(\omega)))}{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))}\cdot\breve{c}_1\right]$$

to be minimized over θ. As a second possibility to permutate the positions of the parameters (θ_0,θ,τ), let us consider

$$0 \leqslant D^{c}_{\phi,\vec{\mathbb{Q}}_\theta,\vec{\mathbb{Q}}_\theta,\mathbb{1}\cdot\vec{\mathbb{Q}}_\theta,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0},\vec{\mathbb{Q}}_\tau)$$
$$:= \int_{\mathscr{X}}\left[\phi\left(\frac{\vec{\mathbb{q}}_{\theta_0}(x)}{\vec{\mathbb{q}}_\theta(x)}\right) - \phi\left(\frac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\right) - \phi'_{+,c}\left(\frac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\right)\cdot\left(\frac{\vec{\mathbb{q}}_{\theta_0}(x)}{\vec{\mathbb{q}}_\theta(x)} - \frac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\right)\right]\cdot\vec{\mathbb{q}}_\theta(x)\,\mathrm{d}\lambda_2(x)\,; \quad (126)$$

this is a pointwise decomposable divergence between $\vec{\mathbb{Q}}_{\theta_0}$ and $\vec{\mathbb{Q}}_\tau$, but it is *not* a divergence – yet still a nonnegative and obviously *not* pointwise decomposable functional – between $\vec{\mathbb{Q}}_{\theta_0}$ and $\vec{\mathbb{Q}}_\theta$. Indeed, for $\theta=\theta_0\neq\tau$ one obtains $D^{c}_{\phi,\vec{\mathbb{Q}}_{\theta_0},\vec{\mathbb{Q}}_{\theta_0},\mathbb{1}\cdot\vec{\mathbb{Q}}_{\theta_0},\lambda_2}(\vec{\mathbb{Q}}_{\theta_0},\vec{\mathbb{Q}}_\tau) > 0$. Notice that from (126) one gets

$$\int_{\mathscr{X}} \phi\Big(\frac{\vec{\mathfrak{q}}_{\theta_0}(x)}{\vec{\mathfrak{q}}_{\theta}(x)}\Big) \cdot \vec{\mathfrak{q}}_{\theta}(x)\, \mathrm{d}\lambda_2(x) \geqslant \int_{\mathscr{X}} \Big\{\Big[\phi\Big(\frac{\vec{\mathfrak{q}}_{\tau}(x)}{\vec{\mathfrak{q}}_{\theta}(x)}\Big) - \phi'_{+,c}\Big(\frac{\vec{\mathfrak{q}}_{\tau}(x)}{\vec{\mathfrak{q}}_{\theta}(x)}\Big) \cdot \frac{\vec{\mathfrak{q}}_{\tau}(x)}{\vec{\mathfrak{q}}_{\theta}(x)}\Big] \cdot \vec{\mathfrak{q}}_{\theta}(x)$$
$$+\phi'_{+,c}\Big(\frac{\vec{\mathfrak{q}}_{\tau}(x)}{\vec{\mathfrak{q}}_{\theta}(x)}\Big) \cdot \vec{\mathfrak{q}}_{\theta_0}(x)\Big\}\, \mathrm{d}\lambda_2(x) =: \mathscr{D}^c_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta})\,, \tag{127}$$

provided that the integral on the right-hand side exists and is finite. If moreover $\phi(1) = 0$, then by (54) the inequality (127) rewrites as

$$D^c_{\phi,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) := D^c_{\phi, \vec{\mathbb{Q}}_\theta, \vec{\mathbb{Q}}_\theta, \mathbb{1}\cdot\vec{\mathbb{Q}}_\theta, \lambda}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) \geqslant D^c_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) \tag{128}$$

with (for fixed θ) equality if and only if $\theta_0 = \tau$; this implies that

$$D^c_{\phi,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) = \max_{\tau\in\Theta} \mathscr{D}^c_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) \tag{129}$$
$$= \max_{\tau\in\Theta} \int_{\mathscr{X}} \Big[\psi^{1,sub}_{\vec{\mathfrak{q}}_\tau(x)}(\vec{\mathfrak{q}}_\theta(x)) + \rho^{sub}_{\vec{\mathfrak{q}}_\tau(x)}(\vec{\mathfrak{q}}_\theta(x)) \cdot \vec{\mathfrak{q}}_{\theta_0}(x)\Big]\, \mathrm{d}\lambda_2(x) \tag{130}$$

with $\psi^0(s) := \psi^{0,sub}_m(s) \equiv 0, \psi^1(t) := \psi^{1,sub}_m(t) := t \cdot \phi(\frac{m}{t}) - m \cdot \phi'_{+,c}(\frac{m}{t}), \rho(t) := \rho^{sub}_m(t) := \phi'_{+,c}(\frac{m}{t})$ $(s, t, m > 0)$. In other words, this means that the Csiszar-Ali-Silvey divergence CASD $D^c_{\phi,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta})$ can be represented as the τ-maximum over – not necessarily nonnegative – pointwise decomposable (in the sense of (103), (104)) functionals $\mathscr{D}^c_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta})$ between $\vec{\mathbb{Q}}_{\theta_0}$ and $\vec{\mathbb{Q}}_{\theta}$. Furthermore, from Theorem 5 and (130) we arrive at

$$0 = \min_{\theta\in\Theta} D^c_{\phi,\lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}) = \min_{\theta\in\Theta} \max_{\tau\in\Theta} \mathscr{D}^c_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta})$$
$$= \min_{\theta\in\Theta} \max_{\tau\in\Theta} \int_{\mathscr{X}} \Big[\psi^{1,sub}_{\vec{\mathfrak{q}}_\tau(x)}(\vec{\mathfrak{q}}_\theta(x)) + \rho^{sub}_{\vec{\mathfrak{q}}_\tau(x)}(\vec{\mathfrak{q}}_\theta(x)) \cdot \vec{\mathfrak{q}}_{\theta_0}(x)\Big]\, \mathrm{d}\lambda_2(x)\,,$$
$$\theta_0 = \operatorname{argmin}_{\theta\in\Theta} \max_{\tau\in\Theta} \mathscr{D}^c_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2}(\vec{\mathbb{Q}}_{\theta_0}, \vec{\mathbb{Q}}_{\theta}). \tag{131}$$

Accordingly, in analogy to the spirit of (81), (82), (86), respectively Definition 4 and (110), in order to achieve an estimator of the true unknown parameter θ_0 we first extend the "pure parametric case" $\mathscr{D}^c_{\phi, \vec{\mathbb{Q}}_\tau, \lambda_2} : \mathscr{P}^\lambda_\Theta \otimes \mathscr{P}^\lambda_\Theta \mapsto \mathbb{R}$ to a singularity-covering functional $\mathscr{D}^c_{\phi, \vec{\mathbb{Q}}_\tau, \lambda} : \mathscr{P}^\lambda_{\Theta,emp} \otimes \mathscr{P}^\lambda_\Theta \mapsto \mathbb{R}$, although it is not a pseudo-divergence anymore; indeed, by employing the reduced form of (102) we take

$$\mathscr{D}^c_{\phi, \vec{\mathbb{Q}}_\tau, \lambda}(\vec{\mathbb{P}}, \vec{\mathbb{Q}}) := \int_{\mathscr{X}} \Big[\psi^{1,sub}_{\vec{\mathfrak{q}}_\tau(x)}\Big(\vec{\mathfrak{q}}(x) + h_1(x) \cdot \mathbf{1}_{\{0\}}(\vec{\mathfrak{q}}(x))\Big) \cdot \mathbf{1}_{]c_1,\infty[}(\vec{\mathfrak{q}}(x))$$
$$+\rho^{sub}_{\vec{\mathfrak{q}}_\tau(x)}\Big(\vec{\mathfrak{q}}(x) + h_2(x) \cdot \mathbf{1}_{\{0\}}(\vec{\mathfrak{q}}(x))\Big) \cdot \vec{\mathfrak{p}}(x)\Big]\, \mathrm{d}\lambda(x) \quad \text{for all } \vec{\mathbb{P}} \in \mathscr{P}^\lambda_{\Theta,emp}, \vec{\mathbb{Q}} \in \mathscr{P}^\lambda_\Theta\,. \tag{132}$$

Hence, analogously to the derivation of (111), we obtain from (132)

$$\sup_{\tau\in\Theta}\ \mathscr{D}^{c}_{\phi,\vec{\mathbb{Q}}_\tau,\lambda}\left(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)},\widetilde{\vec{\mathbb{Q}}}_\theta\right) = \sup_{\tau\in\Theta}\int_{\mathscr{X}}\psi^{1,sub}_{\vec{\mathbb{q}}_\tau(x)}\left(\vec{\mathbb{q}}_\theta(x)\right)\mathrm{d}\lambda_2(x)$$

$$+\sum_{x\in\mathscr{X}}\left[\psi^{1,sub}_{\vec{\mathbb{q}}_\tau(x)}\left(h_1(x)\right)\cdot\breve{c}_1+\rho^{sub}_{\vec{\mathbb{q}}_\tau(x)}(h_2(x))\cdot\vec{\mathbb{p}}_N^{\overline{emp}(\omega)}(x)\right]\cdot\mathbf{1}_{\mathscr{R}(Y_1(\omega),\ldots,Y_N(\omega))}(x)$$

$$=\sup_{\tau\in\Theta}\int_{\mathscr{X}}\left[\vec{\mathbb{q}}_\theta(x)\cdot\phi\left(\frac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\right)-\vec{\mathbb{q}}_\tau(x)\cdot\phi'_{+,c}\left(\frac{\vec{\mathbb{q}}_\tau(x)}{\vec{\mathbb{q}}_\theta(x)}\right)\right]\mathrm{d}\lambda_2(x) \tag{133}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\phi'_{+,c}\left(\frac{\vec{\mathbb{q}}_\tau(Y_i(\omega))}{h_2(Y_i(\omega))}\right)$$

$$+\frac{1}{N}\sum_{i=1}^{N}\frac{h_1(Y_i(\omega))\cdot\phi\left(\frac{\vec{\mathbb{q}}_\tau(Y_i(\omega))}{h_1(Y_i(\omega))}\right)-\vec{\mathbb{q}}_\tau(Y_i(\omega))\cdot\phi'_{+,c}\left(\frac{\vec{\mathbb{q}}_\tau(Y_i(\omega))}{h_1(Y_i(\omega))}\right)}{\vec{\mathbb{p}}_N^{emp(\omega)}(Y_i(\omega))}\cdot\breve{c}_1 \tag{134}$$

to be minimized over $\theta\in\Theta$. In the view of (131), we can estimate (respectively learn) the true unknown parameter θ_0 by the estimator

$$\widehat{\theta}_{N,sup\mathscr{D}_{\phi,\lambda}}(\omega):=\operatorname{arginf}_{\theta\in\Theta}\sup_{\tau\in\Theta}\mathscr{D}^{c}_{\phi,\vec{\mathbb{Q}}_\tau,\lambda}\left(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)},\widetilde{\vec{\mathbb{Q}}}_\theta\right)\ \text{for}\ \vec{\mathbb{P}}_N^{\overline{emp}(\omega)}\in\mathscr{P}^{\lambda\perp}_{emp}, \tag{135}$$

which under appropriate technical assumptions (integrability, etc.) exists, is finite, unique, and Fisher consistent; moreover, this method can be straightforwardly extended to non-parametric setups. Similarly to the derivation of (112) respectively (113), reasonable choices for the "boundary-functions" in (134) are $h_2(x):=\vec{\mathbb{q}}_\theta(x)$ together with $h_1(x)\equiv 1$ respectively $h_1(x):=\vec{\mathbb{q}}_\theta(x)$ (where the nominator in the last sum becomes $-\vec{\mathbb{q}}_\tau(Y_i(\omega))\cdot\phi'_{+,c}(1)$)). In the special case with $c_1=0=\breve{c}_1$ – where the choice of $h_1(\cdot)$ is irrelevant – and $h_2(x):=\vec{\mathbb{q}}_\theta(x)$, the estimator $\widehat{\theta}_{N,sup\mathscr{D}_{\phi,\lambda}}(\omega)$ was first proposed independently by Liese and Vajda [42] under the name *modified ϕ-divergence estimator* and Broniatowski and Keziou [16, 17] under the name *minimum dual ϕ-divergence estimator*; furthermore, within this special-case setup, Broniatowski and Keziou [17] also introduced for each fixed $\theta\in\Theta$ the related, so-called *dual ϕ-divergence estimator* $\widehat{\theta}_{N,\theta,\mathscr{D}_{\phi,\lambda}}(\omega):=\operatorname{argsup}_{\tau\in\Theta}\mathscr{D}^{c}_{\phi,\vec{\mathbb{Q}}_\tau,\lambda}\left(\vec{\mathbb{P}}_N^{\overline{emp}(\omega)},\widetilde{\vec{\mathbb{Q}}}_\theta\right)$. The latter four references also work within a nonparametric framework. Let us also mention that by (128) and (129), $\widehat{\theta}_{N,\mathscr{D}_{\phi,\lambda}}(\omega)$ can be interpreted as *maximum sub-ϕ-divergence estimator*, whereas $\widehat{\theta}_{N,sup\mathscr{D}_{\phi,\lambda}}(\omega)$ can be viewed as *minimum super-ϕ-divergence estimator* (cf. Vajda [90], Broniatowski and Vajda [18] for the probability-measure-theoretic context of footnote 15).

Remark 6 Making use of the escort parameter τ proves to be useful in statistical inference under the model; its use under misspecification has been considered in Toma and Broniatowski [86], Al Mohamad [5], for Csiszar-Ali-Silvey divergences.

As a final example, consider $c_1=0$, $h_2(x):=\vec{\mathbb{q}}_\theta(x)$, and $\phi(t):=t\log t+1-t$, for which we can deduce

$$\widehat{\theta}_{N,sup\mathscr{D}_{\phi,\lambda}}(\omega)=\widehat{\theta}_{N,\theta,\mathscr{D}_{\phi,\lambda}}(\omega)=\operatorname{argsup}_{\xi\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\log\left(\vec{\mathbb{q}}_\xi(Y_i(\omega))\right)$$

for all $\theta \in \Theta$, i.e. in this case all maximum sub-ϕ-divergence estimators and the minimum super-ϕ-divergence estimator exceptionally coincide, and give the celebrated maximum-likelihood estimator.

5 Conclusions

Motivated by fields of applications from statistics, machine learning, artificial intelligence and information geometry, we presented for a wide audience a new unifying framework of divergences between functions. Within this, we illuminated several important subcases – such as scaled Bregman divergences and Csiszar-Ali-Silvey ϕ-divergences – as well as involved subtleties and pitfalls. For the often desired task of finding the "continuous" model with best divergence-proximity to the observed "discrete" data, we summarized existing and also derived new approaches. As far as potential future studies is concerned, the kind of universal nature of our introduced toolkit suggests quite a lot of possibilities for further adjacent developments and concrete applications.

Acknowledgements We are grateful to three anonymous referees for their very useful suggestions and comments. W. Stummer wants to thank very much the Sorbonne Universite Pierre et Marie Curie Paris for its partial financial support.

Appendix: Proofs

Proof of Theorem 4. Assertion (1) and the "if-part" of (2) follow immediately from Theorem 1 which uses less restrictive assumptions. In order to show the "only-if" part of (2) (and the "if-part" of (2) in an alternative way), one can use the straightforwardly provable fact that the Assumption 2 implies

$$\overline{\mathbb{w}_3 \cdot \psi_{\phi,c}}(x, s, t) \;=\; 0 \qquad \text{if and only if} \qquad s \;=\; t \tag{136}$$

for all $s \in \mathscr{R}\big(\frac{P}{M_1}\big)$, all $t \in \mathscr{R}\big(\frac{Q}{M_2}\big)$ and λ-a.a. $x \in \mathscr{X}$. To proceed, assume that $D^c_{\phi,M_1,M_2,\mathbb{M}_3,\lambda}(P, Q) = 0$, which by the non-negativity of $\overline{\mathbb{w}_3 \cdot \psi_{\phi,c}}(\cdot,\cdot)$ implies that $\overline{\mathbb{w}_3 \cdot \psi_{\phi,c}}\big(\frac{p(x)}{m_1(x)}, \frac{q(x)}{m_2(x)}\big) = 0$ for λ-a.a. $x \in \mathscr{X}$. From this and the "only-if" part of (136), we obtain the identity $\frac{p(x)}{m_1(x)} = \frac{q(x)}{m_2(x)}$ for λ-a.a. $x \in \mathscr{X}$. □

Proof of Theorem 5. Consistently with Theorem 1 (and our adaptions) the "if-part" follows from (51). By our above investigations on the adaptions of the Assumptions 2 to the current context, it remains to investigate the "only-if" part (2) for the following four cases (recall that ϕ is strictly convex at $t = 1$):
(ia) ϕ is differentiable at $t = 1$ (hence, c is obsolete and $\phi'_{+,c}(1)$ collapses to $\phi'(1)$) and the function ϕ is affine linear on $[1, s]$ for some $s \in \mathscr{R}\big(\frac{P}{Q}\big)\backslash[a, 1]$;

(ib) ϕ is differentiable at $t = 1$, and the function ϕ is affine linear on $[s, 1]$ for some $s \in \mathscr{R}\big(\frac{P}{Q}\big)\backslash[1, b]$;
(ii) ϕ is not differentiable at $t = 1$, $c = 1$, and the function ϕ is affine linear on $[1, s]$ for some $s \in \mathscr{R}\big(\frac{P}{Q}\big)\backslash[a, 1]$;
(iii) ϕ is not differentiable at $t = 1$, $c = 0$, and the function ϕ is affine linear on $[s, 1]$ for some $s \in \mathscr{R}\big(\frac{P}{Q}\big)\backslash[1, b]$.
It is easy to see from the strict convexity at 1 that for (ii) one has $\phi(0) + \phi'_{+,1}(1) - \phi(1) > 0$, whereas for (iii) one gets $\phi^*(0) - \phi'_{+,0}(1) > 0$; furthermore, for (ia) there holds $\phi(0) + \phi'(1) - \phi(1) > 0$ and for (ib) $\phi^*(0) - \phi'(1) > 0$. Let us first examine the situations (ia) respectively (ii) under the assumptive constraint $D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P}, \mathbb{Q}) = 0$ with $c = 1$ respectively (in case of differentiability) obsolete c, for which we can deduce from (51)

$$
\begin{aligned}
0 &= D^c_{\phi,\mathbb{Q},\mathbb{Q},\mathbb{R}\cdot\mathbb{Q},\lambda}(\mathbb{P}, \mathbb{Q}) \\
&\geqslant \int_{\mathscr{X}} \mathbb{r}(x) \cdot \Big[\mathbb{q}(x) \cdot \phi\big(\tfrac{\mathbb{p}(x)}{\mathbb{q}(x)}\big) - \mathbb{q}(x) \cdot \phi\big(1\big) - \phi'_{+,c}\big(1\big) \cdot \big(\mathbb{p}(x) - \mathbb{q}(x)\big)\Big] \\
&\qquad\qquad \cdot \mathbf{1}_{]0,\infty[}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{]\mathbb{p}(x),\infty[}\big(\mathbb{q}(x)\big)\, \mathrm{d}\lambda(x) \\
&+ \big[\phi(0) + \phi'_{+,c}(1) - \phi(1)\big] \cdot \int_{\mathscr{X}} \mathbb{r}(x) \cdot \mathbb{q}(x) \cdot \mathbf{1}_{\{0\}}\big(\mathbb{p}(x)\big) \cdot \mathbf{1}_{]\mathbb{p}(x),\infty[}\big(\mathbb{q}(x)\big)\, \mathrm{d}\lambda(x) \;\geqslant 0,
\end{aligned}
$$

and hence $\int_{\mathscr{X}} \mathbf{1}_{]\mathbb{p}(x),\infty[}\big(\mathbb{q}(x)\big) \cdot \mathbb{r}(x)\, \mathrm{d}\lambda(x) = 0$. From this and (55) we obtain

$$
0 = \int_{\mathscr{X}} \big(\mathbb{p}(x) - \mathbb{q}(x)\big) \cdot \mathbb{r}(x)\, \mathrm{d}\lambda(x) = \int_{\mathscr{X}} \big(\mathbb{p}(x) - \mathbb{q}(x)\big) \cdot \mathbf{1}_{]\mathbb{q}(x),\infty[}\big(\mathbb{p}(x)\big) \cdot \mathbb{r}(x)\, \mathrm{d}\lambda(x)
$$

and therefore $\int_{\mathscr{X}} \mathbf{1}_{]\mathbb{q}(x),\infty[}\big(\mathbb{p}(x)\big) \cdot \mathbb{r}(x)\, \mathrm{d}\lambda(x) = 0$. Since for λ-a.a. $x \in \mathscr{X}$ we have $\mathbb{r}(x) > 0$, we arrive at $\mathbb{p}(x) = \mathbb{q}(x)$ for λ-a.a. $x \in \mathscr{X}$. The remaining cases (ib) respectively (iii) can be treated analogously. □

References

1. Amari, S.-I.: Information Geometry and Its Applications. Springer, Japan (2016)
2. Amari, S.-I., Karakida, R., Oizumi, M.: Information geometry connecting Wasserstein distance and Kullback-Leibler divergence via the entropy-relaxed transportation problem. Info. Geo. (2018). https://doi.org/10.1007/s41884-018-0002-8
3. Amari, S.-I., Nagaoka, H.: Methods of Information Geometry. Oxford University Press, Oxford (2000)
4. Ali, M.S., Silvey, D.: A general class of coefficients of divergence of one distribution from another. J. R. Stat. Soc. **B–28**, 131–140 (1966)
5. Al Mohamad, D.: Towards a better understanding of the dual representation of phi divergences. Stat. Papers (2016). https://doi.org/10.1007/s00362-016-0812-5
6. Avlogiaris, G., Micheas, A., Zografos, K.: On local divergences between two probability measures. Metrika **79**, 303–333 (2016)
7. Avlogiaris, G., Micheas, A., Zografos, K.: On testing local hypotheses via local divergence. Stat. Methodol. **31**, 20–42 (2016)
8. Ay, N., Jost, J., Le, H.V., Schwachhöfer, L.: Information Geometry. Springer, Berlin (2017)
9. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. J. Mach. Learn. Res. **6**, 1705–1749 (2005)

10. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C.: Robust and efficient estimation by minimizing a density power divergence. Biometrika **85**(3), 549–559 (1998)
11. Basu, A., Lindsay, B.G.: Minimum disparity estimation for continuous models: efficiency, distributions and robustness. Ann. Inst. Stat. Math. **46**(4), 683–705 (1994)
12. Basu, A., Mandal, A., Martin, N., Pardo, L.: Robust tests for the equality of two normal means based on the density power divergence. Metrika **78**, 611–634 (2015)
13. Basu, A., Shioya, H., Park, C.: Statistical Inference: The Minimum Distance Approach. CRC Press, Boca Raton (2011)
14. Birkhoff, G.D: A set of postulates for plane geometry, based on scale and protractor. Ann. Math. **33**(2) 329–345 (1932)
15. Boissonnat, J.-D., Nielsen, F., Nock, R.: Bregman Voronoi diagrams. Discret. Comput. Geom. **44**(2), 281–307 (2010)
16. Broniatowski, M., Keziou, A.: Minimization of ϕ-divergences on sets of signed measures. Stud. Sci. Math. Hungar. **43**, 403–442 (2006)
17. Broniatowski, M., Keziou, A.: Parametric estimation and tests through divergences and the duality technique. J. Multiv. Anal. **100**(1), 16–36 (2009)
18. Broniatowski, M., Vajda, I.: Several applications of divergence criteria in continuous families. Kybernetika **48**(4), 600–636 (2012)
19. Broniatowski, M., Toma, A., Vajda, I.: Decomposable pseudodistances in statistical estimation. J. Stat. Plan. Inf. **142**, 2574–2585 (2012)
20. Buckland, M.K.: Information as thing. J. Am. Soc. Inf. Sci. **42**(5), 351–360 (1991)
21. Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning and Games. Cambridge University Press, Cambridge (2006)
22. Chhogyal, K., Nayak, A., Sattar, A.: On the KL divergence of probability mixtures for belief contraction. In: Hölldobler,S., et al. (eds.) KI 2015: Advances in Artificial Intelligence. Lecture Notes in Artificial Intelligence, vol. 9324, pp. 249–255. Springer International Publishing (2015)
23. Cliff, O.M., Prokopenko, M., Fitch, R.: An information criterion for inferring coupling in distributed dynamical systems. Front. Robot. AI **3**(71). https://doi.org/10.3389/frobt.2016.00071 (2016)
24. Cliff, O.M., Prokopenko, M., Fitch, R.: Minimising the Kullback-Leibler divergence for model selection in distributed nonlinear systems. Entropy **20**(51). https://doi.org/10.3390/e20020051 (2018)
25. Collins, M., Schapire, R.E., Singer, Y.: Logistic regression, AdaBoost and Bregman distances. Mach. Learn. **48**, 253–285 (2002)
26. Cooper, V.N., Haddad, H.M., Shahriar, H.: Android malware detection using Kullback-Leibler divergence. Adv. Distrib. Comp. Art. Int. J., Special Issue 3(2) (2014)
27. Csiszar, I.: Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Publ. Math. Inst. Hungar. Acad. Sci. **A-8**, 85–108 (1963)
28. DasGupta, A.: Some results on the curse of dimensionality and sample size recommendations. Calcutta Stat. Assoc. Bull. **50**(3–4), 157–178 (2000)
29. De Groot, M.H.: Uncertainty, information and sequential experiments. Ann. Math. Stat. **33**, 404–419 (1962)
30. Ghosh, A., Basu, A.: Robust Bayes estimation using the density power divergence. Ann. Inst. Stat. Math. **68**, 413–437 (2016)
31. Ghosh, A., Basu, A.: Robust estimation in generalized linear models: the density power divergence approach. TEST **25**, 269–290 (2016)
32. Ghosh, A., Harris, I.R., Maji, A., Basu, A., Pardo, L.: A generalized divergence for statistical inference. Bernoulli **23**(4A), 2746–2783 (2017)
33. Hampel, F.R., Ronchetti, E.M., Rousseuw, P.J., Stahel, W.A.: Robust Statistics: The Approach Based on Influence Functions. Wiley, New York (1986)
34. Karakida, R., Amari, S.-I.: Information geometry of Wasserstein divergence. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 119–126. Springer International (2017)

35. Kißlinger, A.-L., Stummer, W.: Some decision procedures based on scaled Bregman distance surfaces. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2013. Lecture Notes in Computer Science, vol. 8085, pp. 479–486. Springer, Berlin (2013)
36. Kißlinger, A.-L., Stummer, W.: New model search for nonlinear recursive models, regressions and autoregressions. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2015. Lecture Notes in Computer Science, vol. 9389, pp. 693–701. Springer International (2015)
37. Kißlinger, A.-L., Stummer, W.: Robust statistical engineering by means of scaled Bregman distances. In: Agostinelli, C., Basu, A., Filzmoser, P., Mukherjee, D. (eds.) Recent Advances in Robust Statistics - Theory and Applications, pp. 81–113. Springer, India (2016)
38. Kißlinger, A.-L., Stummer, W.: A new toolkit for robust distributional change detection. Appl. Stochastic Models Bus. Ind. **34**, 682–699 (2018)
39. Kuchibhotla, A.K., Basu, A.: A general setup for minimum disparity estimation. Stat. Prob. Lett. **96**, 68–74 (2015)
40. Liese, F., Miescke, K.J.: Statistical Decision Theory: Estimation, Testing, and Selection. Springer, New York (2008)
41. Liese, F., Vajda, I.: Convex Statistical Distances. Teubner, Leipzig (1987)
42. Liese, F., Vajda, I.: On divergences and informations in statistics and information theory. IEEE Trans. Inf. Theory **52**(10), 4394–4412 (2006)
43. Lin, N., He, X.: Robust and efficient estimation under data grouping. Biometrika **93**(1), 99–112 (2006)
44. Liu, M., Vemuri, B.C., Amari, S.-I., Nielsen, F.: Total Bregman divergence and its applications to shape retrieval. In: Proceedings of 23rd IEEE CVPR, pp. 3463–3468 (2010)
45. Liu, M., Vemuri, B.C., Amari, S.-I., Nielsen, F.: Shape retrieval using hierarchical total Bregman soft clustering. IEEE Trans. Pattern Anal. Mach. Intell. **34**(12), 2407–2419 (2012)
46. Lizier, J.T.: JIDT: an information-theoretic toolkit for studying the dynamcis of complex systems. Front. Robot. AI **1**(11). https://doi.org/10.3389/frobt.2014.00011 (2014)
47. Menendez, M., Morales, D., Pardo, L., Vajda, I.: Two approaches to grouping of data and related disparity statistics. Comm. Stat. - Theory Methods **27**(3), 609–633 (1998)
48. Menendez, M., Morales, D., Pardo, L., Vajda, I.: Minimum divergence estimators based on grouped data. Ann. Inst. Stat. Math. **53**(2), 277–288 (2001)
49. Menendez, M., Morales, D., Pardo, L., Vajda, I.: Minimum disparity estimators for discrete and continuous models. Appl. Math. **46**(6), 439–466 (2001)
50. Millmann, R.S., Parker, G.D.: Geometry - A Metric Approach With Models, 2nd edn. Springer, New York (1991)
51. Minka, T.: Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd., Cambridge, UK (2005)
52. Morales, D., Pardo, L., Vajda, I.: Digitalization of observations permits efficient estimation in continuous models. In: Lopez-Diaz, M., et al. (eds.) Soft Methodology and Random Information Systems, pp. 315–322. Springer, Berlin (2004)
53. Morales, D., Pardo, L., Vajda, I.: On efficient estimation in continuous models based on finitely quantized observations. Comm. Stat. - Theory Methods **35**(9), 1629–1653 (2006)
54. Murata, N., Takenouchi, T., Kanamori, T., Eguchi, S.: Information geometry of U-boost and Bregman divergence. Neural Comput. **16**(7), 1437–1481 (2004)
55. Nielsen, F., Barbaresco, F. (eds.): Geometric Science of Information GSI 2013. Lecture Notes in Computer Science, vol. 8085. Springer, Berlin (2013)
56. Nielsen, F., Barbaresco, F. (eds.): Geometric Science of Information GSI 2015. Lecture Notes in Computer Science, vol. 9389. Springer International (2015)
57. Nielsen, F., Barbaresco, F. (eds.): Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589. Springer International (2017)
58. Nielsen, F., Bhatia, R. (eds.): Matrix Information Geometry. Springer, Berlin (2013)
59. Nielsen, F., Nock, R.: Bregman divergences from comparative convexity. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 639–647. Springer International (2017)

60. Nielsen, F., Sun, K., Marchand-Maillet, S.: On Hölder projective divergences. Entropy **19**, 122 (2017)
61. Nielsen, F., Sun, K., Marchand-Maillet,S.: K-means clustering with Hölder divergences. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 856–863. Springer International (2017)
62. Nock, R., Menon, A.K., Ong, C.S.: A scaled Bregman theorem with applications. Advances in Neural Information Processing Systems 29 (NIPS 2016), pp. 19–27 (2016)
63. Nock, R., Nielsen, F.: Bregman divergences and surrogates for learning. IEEE Trans. Pattern Anal. Mach. Intell. **31**(11), 2048–2059 (2009)
64. Nock, R., Nielsen, F., Amari, S.-I.: On conformal divergences and their population minimizers. IEEE Trans. Inf. Theory **62**(1), 527–538 (2016)
65. Österreicher, F., Vajda, I.: Statistical information and discrimination. IEEE Trans. Inf. Theory **39**, 1036–1039 (1993)
66. Pal, S., Wong, T.-K.L.: The geometry of relative arbitrage. Math. Financ. Econ. **10**, 263–293 (2016)
67. Pal, S., Wong, T.-K.L.: Exponentially concave functions and a new information geometry. Ann. Probab. **46**(2), 1070–1113 (2018)
68. Pardo, L.: Statistical Inference Based on Divergence Measures. Chapman & Hall/CRC, Boca Raton (2006)
69. Park, C., Basu, A.: Minimum disparity estimation: asymptotic normality and breakdown point results. Bull. Inf. Kybern. **36**, 19–33 (2004)
70. Patra, S., Maji, A., Basu, A., Pardo, L.: The power divergence and the density power divergence families: the mathematical connection. Sankhya 75-B Part 1, 16–28 (2013)
71. Peyre, G., Cuturi M.: Computational Optimal Transport (2018). arXiv:1803.00567v1
72. Read, T.R.C., Cressie, N.A.C.: Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer, New York (1988)
73. Reid, M.D., Williamson, R.C.: Information, divergence and risk for binary experiments. J. Mach. Learn. Res. **12**, 731–817 (2011)
74. Roensch, B., Stummer, W.: 3D insights to some divergences for robust statistics and machine learning. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 460–469. Springer International (2017)
75. Rüschendorf, L.: On the minimum discrimination information system. Stat. Decis. Suppl. Issue **1**, 263–283 (1984)
76. Scott, D.W.: Multivariate Density Estimation - Theory, Practice and Visualization, 2nd edn. Wiley, Hoboken (2015)
77. Scott, D.W., Wand, M.P.: Feasibility of multivariate density estimates. Biometrika **78**(1), 197–205 (1991)
78. Stummer, W.: On a statistical information measure of diffusion processes. Stat. Decis. **17**, 359–376 (1999)
79. Stummer, W.: On a statistical information measure for a generalized Samuelson-Black-Scholes model. Stat. Decis. **19**, 289–314 (2001)
80. Stummer, W.: Exponentials, Diffusions, Finance. Entropy and Information. Shaker, Aachen (2004)
81. Stummer, W.: Some Bregman distances between financial diffusion processes. Proc. Appl. Math. Mech. **7**(1), 1050503–1050504 (2007)
82. Stummer, W., Kißlinger, A-L.: Some new flexibilizations of Bregman divergences and their asymptotics. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 514–522. Springer International (2017)
83. Stummer, W., Vajda, I.: On divergences of finite measures and their applicability in statistics and information theory. Statistics **44**, 169–187 (2010)
84. Stummer, W., Vajda, I.: On Bregman distances and divergences of probability measures. IEEE Trans. Inf. Theory **58**(3), 1277–1288 (2012)
85. Sugiyama, M., Suzuki, T., Kanamori, T.: Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. Ann. Inst. Stat. Math. **64**, 1009–1044 (2012)

86. Toma, A., Broniatowski, M.: Dual divergence estimators and tests: robustness results. J. Multiv. Anal. **102**, 20–36 (2011)
87. Tsuda, K., Rätsch, G., Warmuth, M.: Matrix exponentiated gradient updates for on-line learning and Bregman projection. J. Mach. Learn. Res. **6**, 995–1018 (2005)
88. van der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes. Springer, Berlin (1996)
89. Vajda, I.: Theory of Statistical Inference and Information. Kluwer, Dordrecht (1989)
90. Vajda, I.: Modifications of divergence criteria for applications in continuous families. Research Report No. 2230, Institute of Information Theory and Automation, Prague (2008)
91. Vemuri, B.C., Liu, M., Amari, S.-I., Nielsen, F.: Total Bregman divergence and its applications to DTI analysis. IEEE Trans. Med. Imag. **30**(2), 475–483 (2011)
92. Victoria-Feser, M.-P., Ronchetti, E.: Robust estimation for grouped data. J. Am. Stat. Assoc. **92**(437), 333–340 (1997)
93. Weller-Fahy, D.J., Borghetti, B.J., Sodemann, A.A.: A survey of distance and similarity measures used within network intrusion anomaly detection. IEEE Commun. Surv. Tutor. **17**(1), 70–91 (2015)
94. Wu, L., Hoi, S.C.H., Jin, R., Zhu, J., Yu, N.: Learning Bregman distance functions for semi-supervised clustering. IEEE Trans. Knowl. Data Engin. **24**(3), 478–491 (2012)
95. Zhang, J., Naudts, J.: Information geometry under monotone embedding, part I: divergence functions. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2017. Lecture Notes in Computer Science, vol. 10589, pp. 205–214. Springer International (2017)
96. Zhang, J., Wang, X., Yao, L., Li, J., Shen, X.: Using Kullback-Leibler divergence to model opponents in poker. Computer Poker and Imperfect Information: Papers from the AAAI-14 Workshop (2014)