Optimal transport with some directed distances

Wolfgang Stummer

Department of Mathematics, University of Erlangen–Nürnberg, Cauerstrasse 11, 91058 Erlangen, Germany, stummer@math.fau.de

Abstract. We present a toolkit of directed distances between quantile functions. By employing this, we solve some new optimal transport (OT) problems which e.g. considerably flexibilize some prominent OTs expressed through Wasserstein distances.

Keywords: Scaled Bregman divergences · φ–divergences · mass transport.

1 Introduction

Widely used tools for various different tasks in statistics and probability (and thus, in the strongly connected fields of information theory, artificial intelligence and machine learning) are density-based “directed” (i.e. generally asymmetric) distances – called divergences – which measure the dissimilarity/proximity between two probability distributions; some comprehensive overviews can be found in the books of e.g. Liese & Vajda [19], Read & Cressie [30], Vajda [40], Csiszár & Shields [10], Stummer [35], Pardo [24], Liese & Miescke [18], Basu et al. [4]. Amongst others, some important density-based directed-distance classes are:

(1) the Csiszar-Ali-Silvey-Morimoto φ–divergences (CASM divergences) [8],[1],[22]: this includes e.g. the total variation distance, exponentials of Renyi cross-entropies, and the power divergences; the latter cover e.g. the Kullback-Leibler information divergence (relative entropy), the (squared) Hellinger distance, the Pearson chi-square divergence;

(2) the “classical” Bregman distances (CB distances) (see e.g. Csiszar [9], Pardo & Vajda [25,26]); this includes e.g. the density-power divergences cf. [3]) with the squared $L_2$-norm as special case.

More generally, Stummer [37] and Stummer & Vajda [38] introduced the concept of scaled Bregman distances, which enlarges and flexibilizes both the above-mentioned CASM and CB divergence/distance classes; their use for robustness of minimum-distance parameter estimation, testing as well as applications can be found e.g. in Kißlinger & Stummer [15,16,14,13], Roensch & Stummer [31,32,33], Krömer & Stummer [17]. An even much wider framework of directed distances/divergences (BS distances) was introduced in the recent comprehensive paper of Broniatowski & Stummer [6].

Another important omnipresent scientific concept is optimal transport; comprehensive general insights into this (and directly/closely related topics such as e.g. mass transportation problems, Wasserstein distances, optimal coupling problems, extremal copula problems, optimal assignment problems) can be found in the books of e.g. Rachev & Rüschendorf [29], Villani [41,42], Santambrogio [35], Peyre & Cuturi [27], and the survey paper of e.g. Ambrosio & Gigli [2].
In the light of the above explanations, the main goals of this paper are:
(i) To apply (for the sake of brevity, only some subset of) the BS distances to the context of quantile functions (see Section 2).
(ii) To establish a link between (i) and a new class of optimal transport problems where the cost functions are pointwise BS distances (see Section 3).

2 A toolkit of divergences between quantile functions

By adapting and widening the concept of scaled Bregman distances of Stummer [37] and Stummer & Vajda [38], the recent paper of Broniatowski & Stummer [6] introduced a fairly universal, flexible, multi-component system of “directed distances” (which we abbreviate as BS distances) between two arbitrary functions; in the following, we apply (for the sake of brevity, only parts of) this to the important widely used context of quantile functions. We employ the following ingredients:

2.1 Quantile functions

(i1) Let \( \mathcal{X} := [0, 1] \) (open unit interval). For two probability distributions \( \mathbb{P} \) and \( \mathbb{Q} \) on the one-dimensional Euclidean space \( \mathbb{R} \), we denote their cumulative distribution functions (cdf) as \( F^\mathbb{P} \) and \( F^\mathbb{Q} \), and write their quantile functions as

\[
F^{\mathbb{P}}_x := \{ F^{\mathbb{P}}((x)) \}_{x \in \mathcal{X}} := \{ \inf \{ z \in \mathbb{R} : F^\mathbb{P}(z) \geq x \} \}_{x \in \mathcal{X}}
\]

\[
F^{\mathbb{Q}}_x := \{ F^{\mathbb{Q}}((x)) \}_{x \in \mathcal{X}} := \{ \inf \{ z \in \mathbb{R} : F^\mathbb{Q}(z) \geq x \} \}_{x \in \mathcal{X}} ;
\]

if (say) \( \mathbb{P} \) is concentrated on \([0, \infty] \) (i.e., the support of \( \mathbb{P} \) is a subset of \([0, \infty])\), then we alternatively take (for our purposes, without loss of generality)

\[
F^{\mathbb{P}}_x := \{ F^{\mathbb{P}}((x)) \}_{x \in \mathcal{X}} := \{ \inf \{ z \in [0, \infty] : F^\mathbb{P}(z) \geq x \} \}_{x \in \mathcal{X}}
\]

which leads to the nonnegativity \( F^{\mathbb{Q}}_x(x) \geq 0 \) for all \( x \in \mathcal{X} \).

Of course, if the underlying cdf \( z \to F^\mathbb{P}(z) \) is strictly increasing, then \( x \to F^{\mathbb{P}}_x(x) \) is nothing but its “classical” inverse function. Let us also mention that in quantitative finance and insurance, the quantile \( F^{\mathbb{P}}_x(x) \) (e.g. quoted in US dollars units) is called the value-at-risk for confidence level \( x \cdot 100\% \). A detailed discussion on properties and pitfalls of univariate quantile functions can be found e.g. in Embrechts & Hofert [11].

2.2 Directed distances — basic concept

We quantify the dissimilarity between the two quantile functions \( F^{\mathbb{P}}_x,F^{\mathbb{Q}}_x \) in terms of BS distances \( D_\beta(F^{\mathbb{P}}_x,F^{\mathbb{Q}}_x) \) with \( \beta = (\phi, M_1, M_2, M_3) \), defined by

\[
0 \leq D_\beta(F^{\mathbb{P}}_x,F^{\mathbb{Q}}_x) := \int_{\mathcal{X}} \left[ \phi \left( \frac{F^{\mathbb{P}}_x(x)}{M_1(x)} \right) - \phi \left( \frac{F^{\mathbb{Q}}_x(x)}{M_2(x)} \right) \right] \cdot \frac{F^{\mathbb{P}}_x(x) - F^{\mathbb{Q}}_x(x)}{M_1(x) - M_2(x)} \cdot M_3(x) \, d\lambda(x); \quad (1)
\]

the meaning of the integral symbol \( \int \) will become clear in \( \mathcal{X} \) below. Here, in accordance with Broniatowski & Stummer [6] we use:

(i2) the Lebesgue measure \( \lambda \) (it is well known that in general an integral \( \int \ldots \, d\lambda(x) \) turns — except for rare cases — into a classical Riemann integral \( \int \ldots dx \)).

(i3) (measurable) scaling functions \( M_1 : \mathcal{X} \to [-\infty, \infty] \) and \( M_2 : \mathcal{X} \to [-\infty, \infty] \) as well as a nonnegative (measurable) aggregating function \( M_3 : \mathcal{X} \to [0, \infty] \) such that \( M_1(x) \in [-\infty, \infty], M_2(x) \in [-\infty, \infty], M_3(x) \in [0, \infty] \) for \( \lambda \)-a.a. \( x \in \mathcal{X} \). In analogy with the above notation, we use the symbols \( M_i := \{ M_i(x) \}_{x \in \mathcal{X}} \) to refer to the whole functions. In the following, \( \mathcal{A}(G) \) denotes the range (image) of a function \( G := \{ G(x) \}_{x \in \mathcal{X}} \).
the so-called “divergence-generator” \( \phi \) which is a continuous, convex (finite) function \( \phi : E \to \mathbb{R} \) on some appropriately chosen open interval \( E = [a, b] \) such that \( [a, b] \) covers (at least) the union \( \mathcal{A}(F_{\mathcal{F}} M_1) \cup \mathcal{A}(F_{\mathcal{F}} M_2) \) of both ranges \( \mathcal{A}(F_{\mathcal{F}} x) \) of \( x \in \mathcal{F} \) and \( \mathcal{A}(F_{\mathcal{F}} x) \) of \( x \in \mathcal{F} \); for instance, \( E = [0, 1] \), \( E = [0, \infty] \) or \( E = [1, \infty] \); the class of all such functions will be denoted by \( \Phi([a, b]) \). Furthermore, we assume that \( \phi \) 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 \( \overline{\phi}(a) := \lim_{t \to a^+} \phi(t), \overline{\phi}(b) := \lim_{t \to b^-} \phi(t) \) 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 \( \phi \) at the point \( t \in [a, b] \) and \( \phi'_-(t) \) the (always finite) left-hand derivative of \( \phi \) at \( t \in [a, b] \). If \( \phi \in \Phi([a, b]) \) is also continuously differentiable – which we denote by \( \phi \in \Phi_C([a, b]) \) – 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 \( \overline{\phi'}_{+c}(t) := \phi'_{+c}(t) \) (\( t \in [a, b] \)), \( \overline{\phi'}_{+c}(a) := \lim_{t \to a^+} \phi'_{+c}(t), \overline{\phi'}_{+c}(b) := \lim_{t \to b^-} \phi'_{+c}(t) \). 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 \( \psi_{\phi,c}(s, t) := \phi(s) - \phi(t) - \phi'_{+c}(t) \cdot (s - t) \geq 0 \) (\( s, t \in [a, b] \)). To extend this to a lower semi-continuous function \( \psi_{\phi,c} : [a, b] \times [a, b] \to [0, \infty] \) we proceed as follows: firstly, we set \( \psi_{\phi,c}(s, t) := \psi_{\phi,c}(s, t) \) for all \( s, t \in [a, b] \). Moreover, since for fixed \( t \in [a, b] \), the function \( s \mapsto \psi_{\phi,c}(s, t) \) 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)) always interpreted/abbreviated as \( \psi(a) := \phi(a) - \phi(t) - \phi'_{+c}(t) \cdot (a - t) \).

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 short-hand notation \( \psi(b) := \phi(b) - \phi(t) - \phi'_{+c}(t) \cdot (b - t) \). Furthermore, for fixed \( s \in [a, b] \) we interpret \( \phi(s) - \phi(a) - \phi'_{+c}(a) \cdot (s - a) \) as
\[
\psi_{\phi,c}(s, a) := \left\{ \psi(s) - \overline{\psi}_{\phi,c}(a) \cdot (s - a) + \lim_{t \to a^\leftarrow} (t \cdot \overline{\psi}_{\phi,c}(a) - \phi(t)) \right\} \cdot 1_{([-\infty, \psi_{\phi,c}(a)])} + \infty \cdot 1_{(-\infty, 0]}(\overline{\phi'}_{+c}(a)) \text{,}
\]
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
\[
\overline{\psi}_{\phi,c}(s, b) := \left\{ \psi(s) - \overline{\psi}_{\phi,c}(b) \cdot (s - b) + \lim_{t \to b^-} (t \cdot \overline{\psi}_{\phi,c}(b) - \phi(t)) \right\} \cdot 1_{[-\infty, \infty)}(\overline{\phi'}_{+c}(b)) + \infty \cdot 1_{(+\infty)}(\overline{\phi'}_{+c}(b)) \text{,}
\]
where again the involved limit always exists but may be infinite. Finally, we always set \( \psi_{\phi,c}(a, a) := 0, \overline{\psi}_{\phi,c}(b, b) := 0, \psi_{\phi,c}(a, b) := \lim_{s \to a^-} \psi_{\phi,c}(s, b) \), \( \psi_{\phi,c}(b, a) := \lim_{s \to b^+} \psi_{\phi,c}(s, a) \). Notice that \( \psi_{\phi,c}(\cdot, \cdot) \) is lower semi-continuous but not necessarily continuous. Since ratios are ultimately continuous, we also consistently take \( \overline{\psi}_{\phi,c}(0, 0) := 0 \).

With (11) to (14), we define the BS distance (BS divergence) of (11) precisely as
\[
0 \leq D_{\phi,M_1,M_2}(F_{\mathcal{F}} P, F_{\mathcal{F}} Q) = \int_{\mathcal{F}} \psi_{\phi,c} \left( \frac{F_{\mathcal{F}} P}{M_1(x)}, \frac{F_{\mathcal{F}} Q}{M_2(x)} \right) \cdot M_3(x) \, d\lambda(x)
\]
(2)
and
\[
0 \leq D_{\phi,M_1,M_2}(F_{\mathcal{F}} P, F_{\mathcal{F}} Q) = \int_{\mathcal{F}} \psi_{\phi,c} \left( \frac{F_{\mathcal{F}} P}{M_1(x)}, \frac{F_{\mathcal{F}} Q}{M_2(x)} \right) \cdot M_3(x) \, d\lambda(x),
\]
(3)
but mostly use the less clumsy notation with $\int$ given in (11), (2) henceforth, as a shortcut for the implicitly involved boundary behaviour.

Notice that generally (with some exceptions) one has the asymmetry

$$D^c_{\phi,M_1,M_2,M_3}(F^-_P, F^-_Q) \neq D^c_{\phi,M_1,M_2,M_3}(F^-_Q, F^-_P)$$

leading — together with the nonnegativity — to the (already used above) interpretation of $D^c_{\phi,M_1,M_2,M_3}(F^-_P, F^-_Q)$ as a candidate of a “directed” distance/divergence. To make this proper, one needs to verify

(NNg) $D^c_{\phi,M_1,M_2,M_3}(F^-_P, F^-_Q) \geq 0$.

(REg) $D^c_{\phi,M_1,M_2,M_3}(F^-_P, F^-_Q) = 0$ if and only if $F^-_P(x) = F^-_Q(x)$ for $\lambda$-a.a. $x \in \mathcal{X}$.

As already indicated above, the nonnegativity (NNg) holds per construction. For the reflexivity (Reg) one needs further assumptions. Indeed, in a more general context beyond quantile functions and the Lebesgue measure, Broniatowski & Stummer [6] gave conditions such that objects as in (2) and (3) general context beyond quantile functions and the Lebesgue measure, Broniatowski & Stummer [6] gave conditions and (6) satisfy (Reg). We shall adapt this to the current special context, where for the sake of brevity, for the rest of this paper we shall always concentrate on the important adaptive subcase $M_1(x) := W(F^-_P(x), F^-_Q(x))$, $M_2(x) := M_1(x)$, $M_3(x) := W_3(F^-_P(x), F^-_Q(x))$, for some (measurable) functions $W : \mathcal{R}(F^-_P) \times \mathcal{R}(F^-_Q) \to [-\infty, \infty]$ and $W_3 : \mathcal{R}(F^-_P) \times \mathcal{R}(F^-_Q) \to [0, \infty]$.

Accordingly, (1), (2) and (3) simplify to

$$0 \leq D^c_{\phi,W,W_3}(F^-_P, F^-_Q) := D^c_{\phi,W}(F^-_P, F^-_Q) + W(F^-_P, F^-_Q) W_3(F^-_P, F^-_Q) (F^-_P, F^-_Q)$$

$$= \int_{\mathcal{X}} \left[ \phi\left(\frac{F^-_P(x)}{W(F^-_P(x), F^-_Q(x))}\right) - \phi\left(\frac{F^-_Q(x)}{W(F^-_P(x), F^-_Q(x))}\right) \right]$$

$$\cdot W_3(F^-_P(x), F^-_Q(x)) \cdot d\lambda(x)$$

(4)

where we employ $\mathcal{R}_{\phi,W,W_3} : \mathcal{R}(F^-_P) \times \mathcal{R}(F^-_Q) \to [0, \infty]$ defined by

$$\mathcal{R}_{\phi,W,W_3}(u,v) := \int_{\mathcal{X}} \frac{u_\phi}{W(u,v)} \cdot \frac{v}{W(u,v)} \geq 0$$

with

$$\psi_{\phi,c}\left(\frac{u}{W(u,v)} - \frac{v}{W(u,v)}\right) := \left[ a\left(\frac{u}{W(u,v)}\right) - \phi\left(\frac{v}{W(u,v)}\right) - \phi_{\phi,c}\left(\frac{v}{W(u,v)}\right) \right] \cdot \left(\frac{u}{W(u,v)} - \frac{v}{W(u,v)}\right)$$

(7)

We give conditions for the validity of the crucial reflexivity in the following subsection; this may be skipped by the non-specialist (and the divergence expert).

### 2.3 Justification of distance properties

By construction, one gets for all $\phi \in \Phi([a,b])$ and all $c \in [0,1]$ the important assertion $D^c_{\phi,W,W_3}(F^-_P, F^-_Q) \geq 0$ with equality if $F^-_P(x) = F^-_Q(x)$ for $\lambda$-almost all $x \in \mathcal{X}$. As investigated in Broniatowski & Stummer [6], in order to get “sharp identifiability” (i.e. reflexivity) one needs further assumptions on $\phi \in \Phi([a,b])$, $c \in [0,1]$; for instance, if $\phi \in \Phi([a,b])$ is affine linear on the whole interval $[a,b]$ and $W_3$ is constant (say, 1), then $\mathcal{R}_{\phi,W,W_3}$ takes the constant value 0, and hence $D^c_{\phi,W,W_3}(F^-_P, F^-_Q) = 0$ even in cases where $F^-_P(x) \neq F^-_Q(x)$ for $\lambda$-a.a. $x \in \mathcal{X}$.

In order to avoid such and similar phenomena, we use the following
Assumption 1 Let $c \in [0, 1]$, $\phi \in \Phi([a, b])$ and

$$\mathcal{R}(\frac{F^+}{W(F^+(x), F^-(x))}) \cup \mathcal{R}(\frac{F^-}{W(F^-(x), F^+(x))}) \subset [a, b].$$

Moreover, for all $s \in \mathcal{R}(\frac{F^+}{W(F^+(x), F^-(x))})$, all $t \in \mathcal{R}(\frac{F^-}{W(F^-(x), F^+(x))})$, all $u \in \mathcal{R}(F^+_P)$ and all $v \in \mathcal{R}(F^-_Q)$ let the following conditions hold:

(a) $\phi$ is strictly convex at $t$;
(b) if $\phi$ is differentiable at $t$ and $s \neq t$, then $\phi$ is not affine-linear on the interval $[\min(s, t), \max(s, t)]$ (i.e. between $t$ and $s$);
(c) if $\phi$ is not differentiable at $s$, $s > t$ and $\phi$ is affine linear on $[t, s]$, then we exclude $c = 1$ for the ("globally/universally chosen") subderivative $\phi'_{t,c}(\cdot) = c \cdot \phi'_+(\cdot) + (1 - c) \cdot \phi'_-(\cdot)$;
(d) if $\phi$ is not differentiable at $t$, $s < t$ and $\phi$ is affine linear on $[s, t]$, then we exclude $c = 0$ for $\phi'_{t,c}(\cdot)$;
(e) $W_3(u, v) < \infty$;
(f) $W_4(u, v) > 0$ if $u \neq v$;
(g) by employing (with a slight abuse of notation) the function

$$\mathcal{T}(u, v) := W_3(u, v) \cdot \psi_{\phi,c}(\frac{u}{W(u,v)} \cdot \frac{v}{W(u,v)}),$$

we set by convention

$$\mathcal{T}(u, v) := 0 \text{ if } W(u,v) = W(u,v) = a;$$

(h) by convention, $\mathcal{T}(u, v) := 0 \text{ if } W(u,v) = W(u,v) = b$;

(i) $\mathcal{T}(u, v) > 0 \text{ if } \frac{u}{W(u,v)} = a \text{ and } \frac{v}{W(u,v)} = t \notin \{a, b\}$;

this is understood in the following way: for $\frac{u}{W(u,v)} = t \notin \{a, b\}$, we require

$$\lim_{\frac{u}{W(u,v)} \rightarrow a} \mathcal{T}(u, v) > 0 \text{ if this limit exists, or otherwise we set by convention}$$

$$\mathcal{T}(u, v) := 1 \text{ (or any other strictly positive constant) if } \frac{u}{W(u,v)} = a;$$

the following boundary-behaviour conditions have to be interpreted analogously;

(j) $\mathcal{T}(u, v) > 0 \text{ if } \frac{u}{W(u,v)} = b \text{ and } \frac{v}{W(u,v)} = t \notin \{a, b\}$;

(k) $\mathcal{T}(u, v) > 0 \text{ if } \frac{u}{W(u,v)} = a \text{ and } \frac{v}{W(u,v)} = s \notin \{a, b\}$;

(l) $\mathcal{T}(u, v) > 0 \text{ if } \frac{u}{W(u,v)} = b \text{ and } \frac{v}{W(u,v)} = s \notin \{a, b\}$;

(m) $\mathcal{T}(u, v) > 0 \text{ if } \frac{u}{W(u,v)} = b \text{ and } \frac{v}{W(u,v)} = a$ (as limit from (l) or by convention);

(n) $\mathcal{T}(u, v) > 0 \text{ if } \frac{u}{W(u,v)} = a \text{ and } \frac{v}{W(u,v)} = b$ (as limit from (k) or by convention).

Remark 1. We could even work with a weaker assumption obtained by replacing

$s \text{ with } \frac{F^+}{W(F^+(x), F^-(x))}$,
$t \text{ with } \frac{F^-}{W(F^-(x), F^+(x))}$,
$u \text{ with } F^+_P(x)$,
$v \text{ with } F^-_Q(x),$

and by requiring that then the correspondingly plugged-in conditions (a) to (n) hold for $\lambda$-a.a. $x \in \mathcal{X}$.

The following requirement is stronger than the “model-individual/dependent” Assumption 1 but is more “universally applicable”:

Assumption 2 Let $c \in [0, 1]$, $\phi \in \Phi([a, b])$ on some fixed $[a, b] \in (-\infty, +\infty]$ such that

$$\mathcal{R}(\frac{F^+}{W(F^+(x), F^-(x))}) \cup \mathcal{R}(\frac{F^-}{W(F^-(x), F^+(x))}) \subset [a, b].$$

Moreover, for all $s \in [a, b]$, $c \in [a, b]$, all $u \in \mathcal{R}(F^+_P)$ and all $v \in \mathcal{R}(F^-_Q)$ the conditions (a) to (n) of Assumption 1 hold.
By adapting Theorem 4 and Corollary 1 of Broniatowski & Stummer, we obtain (NN) \( D_{\phi,W,W_3}(F_P^+, F_Q^-) \geq 0 \).

(RE) \( D_{\phi,W,W_3}^c(F_P^+, F_Q^-) = 0 \) if and only if \( F_P^+(x) = F_Q^-(x) \) for \( \lambda \)-a.a. \( x \in \mathcal{X} \).

The non-negativity (NN) and the reflexivity (RE) say that we indeed obtain the following:

\[ \min_{X \sim P, Y \sim Q} \mathbb{E} \left[ \Phi_{\phi,c,W,W_3}(X,Y) \right] \]

\[ = \min_{\Psi \in \mathcal{P}(P,Q)} \int_{\mathbb{R} \times \mathbb{R}} \Phi_{\phi,c,W,W_3}(u,v) \, d\Psi(u,v) \]

\[ = \int_{\mathbb{R} \times \mathbb{R}} \Phi_{\phi,c,W,W_3}(u,v) \, d\Psi^{\text{com}}(u,v) \]

\[ = \int_{[0,1]} \Phi_{\phi,c,W,W_3}(F_P^+(x), F_Q^-(x)) \, d\lambda(x) \]

\[ = \int_{[0,1]} \left[ \phi \left( \frac{F_P^+(x)}{W(F_P^+(x), F_Q^-(x))} \right) - \phi \left( \frac{F_Q^-(x)}{W(F_P^+(x), F_Q^-(x))} \right) \right] \cdot W_3(F_P^+(x), F_Q^-(x)) \, d\lambda(x) \]

\[ = D_{\phi,W,W_3}^c(F_P^+, F_Q^-) \geq 0, \]

where the minimum in \( \mathcal{P} \) is taken over all \( \mathbb{R} \)-valued random variables \( X, Y \) on an arbitrary probability space \((\Omega, \mathcal{F}, \mathcal{G})\) such that \( \mathbb{P}[X \in \cdot] = P[\cdot], \mathbb{P}[Y \in \cdot] = Q[\cdot] \). As usual, \( \mathbb{E} \) denotes the expectation with respect to \( \mathbb{P} \).

3 New optimal transport problems

For our applications to optimal transport, we impose henceforth the additional requirement that the nonnegative (extended) function \( \Phi_{\phi,c,W,W_3} \) is continuous and quasi-antitone in the sense

\[ \Phi_{\phi,c,W,W_3}(u_1, v_1) + \Phi_{\phi,c,W,W_3}(u_2, v_2) \leq \Phi_{\phi,c,W,W_3}(u_1, v_1) + \Phi_{\phi,c,W,W_3}(u_2, v_2) \]

for all \( u_1 \leq u_2, v_1 \leq v_2 \); in other words, \(-\Phi_{\phi,c,W,W_3}(\cdot, \cdot) \) is assumed to be continuous and quasi-monotone.

For such a setup, we consider the novel Kantorovich transportation problem (KTP) with the pointwise-BS-distance-type (pBS-type) cost function \( \Phi_{\phi,c,W,W_3}(u,v) \); indeed, we obtain the following:

**Theorem 3.** Let \( \tilde{\Gamma}(\mathcal{P}, \mathcal{Q}) \) be the family of all probability distributions \( \mathcal{P} \) on \( \mathbb{R} \times \mathbb{R} \) which have marginal distributions \( \mathcal{P}[\cdot \times \mathbb{R}] = \mathcal{P}[] \) and \( \mathcal{Q}[\mathbb{R} \times \cdot] = \mathcal{Q}[] \). Moreover, we denote the corresponding upper Hoeffding-Fréchet bound (cf. e.g. Theorem 3.1.1 of Rachev & Rüschendorf [24]) by \( \mathcal{P}^{\text{com}} \) having “comonotonic” distribution function \( F_3^{\text{com}}(u,v) := \min\{F_P(u), F_Q(v)\} \) \( (u,v) \in \mathbb{R} \). Then

\[ \min_{X \sim P, Y \sim Q} \mathbb{E} \left[ \Phi_{\phi,c,W,W_3}(X,Y) \right] \]

\[ = \min_{\Psi \in \mathcal{P}(P,Q)} \int_{\mathbb{R} \times \mathbb{R}} \Phi_{\phi,c,W,W_3}(u,v) \, d\Psi(u,v) \]

\[ = \int_{\mathbb{R} \times \mathbb{R}} \Phi_{\phi,c,W,W_3}(u,v) \, d\Psi^{\text{com}}(u,v) \]

\[ = \int_{[0,1]} \Phi_{\phi,c,W,W_3}(F_P^+(x), F_Q^-(x)) \, d\lambda(x) \]

\[ = \int_{[0,1]} \left[ \phi \left( \frac{F_P^+(x)}{W(F_P^+(x), F_Q^-(x))} \right) - \phi \left( \frac{F_Q^-(x)}{W(F_P^+(x), F_Q^-(x))} \right) \right] \cdot W_3(F_P^+(x), F_Q^-(x)) \, d\lambda(x) \]

\[ = D_{\phi,W,W_3}^c(F_P^+, F_Q^-) \geq 0, \]

where the minimum in \( \mathcal{P} \) is taken over all \( \mathbb{R} \)-valued random variables \( X, Y \) on an arbitrary probability space \((\Omega, \mathcal{F}, \mathcal{G})\) such that \( \mathbb{P}[X \in \cdot] = P[\cdot], \mathbb{P}[Y \in \cdot] = Q[\cdot] \). As usual, \( \mathbb{E} \) denotes the expectation with respect to \( \mathbb{P} \).

1 Other names are: submodular, lattice-subadditive, 2-antitone, 2-negative, \( \Delta \)-antitone, supernegative, “satisfying the (continuous) Monge property/condition”

2 Other names are: supermodular, lattice-superaadditive, 2-increasing, 2-positive, \( \Delta \)-monotone, 2-monotone, “fulfilling the moderate growth property”, “satisfying the measure property”, “satisfying the twist condition”

3 A comprehensive discussion on general quasi-monotone functions can be found e.g. in Chapter 6.C of Marshall et al. [21]
The assertion \([11]\) follows by applying Corollary 2.2a of Tchen [39] (see also – with different regularity conditions and generalizations – Cambanis et al. [7], Rüschendorf [34], Theorem 3.1.2 of Rachev & Rüschendorf [29], Theorem 3.8.2 of Müller & Stoyan [23], Theorem 2.5 of Puccetti & Scarsini [28], Theorem 2.5 of Ambrosio & Gigli [2]).

Remark 2. (i) Notice that \(\mathcal{P}^{\text{com}}\) is \(\mathcal{T}_{\phi,c,W,W}\)-independent, and may not be the unique minimizer in \((10)\). As a (not necessarily unique) minimizer in \((10)\), one can take \(X := F^+_p(U)\), \(Y := F^+_q(U)\) for some uniform random variable \(U\) on \([0,1]\).

(ii) In Theorem 3 we have shown that \(\mathcal{P}^{\text{com}}\) (cf. \((11)\)) is an optimal transport plan of the KTP \((10)\) with the pointwise-BS-distance-type (pBS-type) cost function \(\mathcal{T}_{\phi,c,W,W}(u,v)\). The outcoming minimal value is equal to \(D_{\phi,c,W,W}(F^+_p,F^+_q)\) which is typically straightforward to compute (resp. approximate).

Remark 2(ii) generally contrasts to those prominently used KTP whose cost function is a power \(d(u,v)^p\) of a metric \(d(u,v)\) (denoted as POM-type cost function) which leads to the well-known Wasserstein distances. (Apart from technicalities)

There are some overlaps, though:

Example 1. (i) Take the non-smooth \(\phi(t) := \phi_{TV}(t) := |t - 1| (t \in \mathbb{R})\), \(c = \frac{1}{2}\), \(W(u,v) := |v|, W_3(u,v) := |u - v| := d(u,v)\).

(ii) Take \(\phi(t) := \phi_{2}(t) := \frac{(t - 1)^2}{2} (t \in \mathbb{R}, \gamma \in \mathbb{R})\), \(W(u,v) := 1\) and \(W_3(u,v) := 1\) to end up with \(\mathcal{T}_{\phi,c,W,W}(u,v) = \frac{1}{2} (u - v)^2 = \frac{d(u,v)^2}{2}\).

(iii) The symmetric distances \(d(u,v)\) and \(\frac{d(u,v)^2}{2}\) are convex functions of \(u - v\) and thus continuous quasi-antitone functions on \(\mathbb{R} \times \mathbb{R}\). The correspondingly outcoming Wasserstein distances are thus considerably flexibilized by our new much more general distance \(D_{\phi,c,W,W}(F^+_p,F^+_q)\) of \((14)\).

Depending on the chosen divergence, one may have to restrict the support of \(\mathcal{P}\) respectively \(\mathcal{Q}\) for instance to \([0,\infty[\). We give some further special cases of pBS-type cost functions, which are continuous and quasi-antitone, but which are generally not symmetric and thus not of POM-type:

Example 2. “smooth” pointwise Csiszar-Ali-Silvey-Morimoto divergences (CASM divergences): take \(\phi : [0,\infty[\times \mathbb{R} \to \mathbb{R}\) to be a strictly convex and twice continuously differentiable on \([0,\infty[\) with continuous extension on \(t = 0\), together with \(W(u,v) := v, W_3(u,v) := v (v \in [0,\infty[)\) and \(c\) is obsolete. Accordingly, \(\mathcal{T}_{\phi,c,W,W}(u,v) := v \cdot \phi\left(\frac{u}{v}\right) - v \cdot \phi(1) - v \cdot \phi'(1) \cdot (u - v)\), and hence the second mixed derivative satisfies \(\frac{\partial^2 \mathcal{T}_{\phi,c,W,W}(u,v)}{\partial u \partial v} = \frac{\partial^2 \phi'}{\partial u^2}\left(\frac{u}{v}\right) < 0 (u,v \in [0,\infty[);\) thus, \(\mathcal{T}_{\phi,c,W,W}\) is quasi-antitone on \([0,\infty[\times [0,\infty[\). Accordingly, (9) to (13) applies to (such kind of) CASM divergences concerning \(\mathcal{P}, \mathcal{Q}\) having support in \([0,\infty[\). As an example, take e.g. the power function \(\phi(t) := \frac{t^\gamma - \gamma t^{\gamma - 1}}{\gamma (\gamma - 1)} (\gamma \in \mathbb{R}\setminus \{0,1\})\). A different connection between optimal transport and other kind of CASM divergences can be found in Bertrand et al. [5] in the current GSI2021 volume.

Example 3. “smooth” pointwise classical (i.e. unscaled) Bregman divergences (CBD): take \(\phi : \mathbb{R} \to \mathbb{R}\) to be a strictly convex and twice continuously differentiable function \(W(u,v) := 1, W_3(u,v) := 1, c\) is obsolete. Accordingly,
\[ \Upsilon_{\phi,c,W,W_3}(u,v) := \phi(u) - \phi(v) - \phi'(v) \cdot (u - v) \] and hence \[ \frac{\partial^2 \Upsilon_{\phi,c,W,W_3}(u,v)}{\partial u \partial v} = -\phi''(v) < 0 \quad (u,v \in \mathbb{R}) \] thus, \( \Upsilon_{\phi,c,W,W_3} \) is quasi-antitone on \( \mathbb{R} \times \mathbb{R} \). Accordingly, the representation \((\mathbf{9})\) to \((\mathbf{13})\) applies to (such kind of) CBD. The corresponding special case of \((\mathbf{10})\) is called “a relaxed Wasserstein distance (parameterized by \( \phi \)) between \( \mathbb{P} \) and \( \mathbb{Q} \)” in the recent papers of Lin et al. \([20]\) and Guo et al. \([12]\) for a restrictive setup where \( \mathbb{P} \) and \( \mathbb{Q} \) are supposed to have compact support; the latter two references do not give connections to divergences of quantile functions, but substantially concentrate on applications to topic sparsity for analyzing user-generated web content and social media, respectively, to Generative Adversarial Networks (GANs).

**Example 4.** “Smooth” pointwise Scaled Bregman Distances: for instance, consider \( \mathbb{P} \) and \( \mathbb{Q} \) with support in \([0, \infty[\). Under \( W = W_3 \) one gets that \( \Upsilon_{\phi,c,W,W} \) is quasi-antitone on \([0, \infty[ \times \mathbb{R} \) if the generator function \( \phi \) is strictly convex and thrice continuously differentiable on \([0, \infty[ \) (and hence, \( c \) is obsolete) and the so-called scale connector \( W \) is twice continuously differentiable such that – on \([0, \infty[ \times \mathbb{R} \) – \( \Upsilon_{\phi,c,W,W} \) is twice continuously differentiable and \( \frac{\partial^2 \Upsilon_{\phi,c,W,W}(u,v)}{\partial u \partial v} \leq 0 \) (an explicit formula of the latter is given in the appendix of Kisslinger & Stummer \([14]\)). Illustrative examples of suitable \( \phi \) and \( W \) can be found e.g. in Kisslinger & Stummer \([15]\).

Returning to the general context, it is straightforward to see that if \( \mathbb{P} \) does not give mass to points (i.e. it has continuous distribution function \( F_\mathbb{P} \)) there exists even a deterministic optimal transportation plan: indeed, for the map \( T^{\text{com}} := F_\mathbb{Q}^{-1} \circ F_\mathbb{P} \) one has \( \mathbb{P}^{\text{com}}[\cdot] = \mathbb{P}[(id, T^{\text{com}}) \in \cdot] \) and thus \((\mathbf{11})\) is equal to

\[
\int_{\mathbb{R}} \Upsilon_{\phi,c,W,W_3}(u,T^{\text{com}}(u)) \, d\mathbb{P}(u) = \min_{(T \in \mathcal{F}(\mathbb{P}, \mathbb{Q}))} \int_{\mathbb{R}} \Upsilon_{\phi,c,W,W_3}(u,T(u)) \, d\mathbb{P}(u)
\]

where \((\mathbf{16})\) is called Monge transportation problem (MTP). Here, \( \mathcal{F}(\mathbb{P}, \mathbb{Q}) \) denotes the family of all measurable maps \( T : \mathbb{R} \mapsto \mathbb{R} \) such that \( \mathbb{P}[T \in \cdot] = \mathbb{Q}[\cdot] \).

**Acknowledgement.** I am grateful to the four referees for their comments and suggestions on readability improvements.

**References**

1. Ali, M.S., Silvey, D.: A general class of coefficients of divergence of one distribution from another. J. Roy. Statist. Soc. B-28, 131—140 (1966)
2. Ambrosio, L., Gigli, N.: A user’s guide to optimal transport. In: L. Ambrosio et al.: Modeling and Optimisation of Flows on Networks, LNM 2062, pp. 1—155. Springer, Berlin (2013)
3. 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)
4. Basu, A., Shioya, H., Park, C.: Statistical Inference: The Minimum Distance Approach. CRC Press, Boca Raton (2011)
5. Bertrand, P., Broniatowski, M., Marcotorchino, J.-F.: Divergences minimisation and applications. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information GSI 2021. Lecture Notes in Computer Science, vol. XXXX, pp. XXX—XXX. Springer, International (2021)
6. Broniatowski, M., Stummer, W.: Some universal insights on divergences for statistics, machine learning and artificial intelligence. In: Nielsen, F. (ed.) Geometric Structures of Information, pp. 149—211. Springer Nature, Switzerland (2019)
7. Cambanis, S., Simons, G., Stout, W.: Inequalities for \( \text{Ek}(X,Y) \) when the marginals are fixed. Probab. Theory Rel. Fields 36, 285—294 (1976)
8. 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)
