*Submitted to Bernoulli*

# General construction and classes of explicit $L^1$-optimal couplings

GIOVANNI PUCCETTI[1] and LUDGER RÜSCHENDORF[2]

[1]*University of Milano, Via Conservatorio 7, 20122 Milano, Italy.*
*E-mail:* giovanni.puccetti@unimi.it
[2]*University of Freiburg, Ernst-Zermelo-Straße 1, 79104 Freiburg, Germany.*
*E-mail:* ruschen@stochastik.uni-freiburg.de

The main scope of this paper is to give some explicit classes of examples of $L^1$-optimal couplings. Optimal transportation w.r.t. the Kantorovich metric $\ell_1$ (resp. the Wasserstein metric $W_1$) between two absolutely continuous measures is known since the basic papers of Kantorovich and Rubinstein (1957) and Sudakov (1979) to occur on rays induced by a decomposition of the basic space (and more generally to higher dimensional decompositions in the case of general measures) induced by the corresponding dual potentials. Several papers have given this kind of structural result and established existence and uniqueness of solutions in varying generality. Since the dual problems pose typically too strong challenges to be solved in explicit form, these structural results have so far been applied for the solution of few particular instances.

First, we give a self-contained review of some basic optimal coupling results and we propose and investigate in particular some basic principles for the construction of $L^1$-optimal couplings given by a reduction principle and some usable forms of the decomposition method. This reduction principle, together with symmetry properties of the reduced measures, gives a hint to the decomposition of the space into sectors and via the non crossing property of optimal transport leads to the choice of transportation rays. The optimality of the induced transports is then a consequence of the characterization results of optimal couplings.

Then, we apply these principles to determine in explicit form $L^1$-optimal couplings for several classes of examples of elliptical distributions. In particular, we give for the first time a general construction of $L^1$-optimal couplings between two bivariate Gaussian distributions. We also discuss optimality of special constructions like shifts and scalings, and provide an extended class of dual functionals allowing for the closed-form computation of the $\ell_1$-metric or of accurate lower bounds of it in a variety of examples.

*Keywords:* Kantorovich $\ell_1$-metric; $L^1$-Wasserstein distance; optimal mass transportation; optimal couplings; Gaussian distributions; Monge-Kantorovich problem; Kantorovich-Rubinstein Theorem

## 1. Introduction

Let $P, Q \in M^1(\mathbb{R}^d)$ be two probability measures on $\mathbb{R}^d$ with finite first moments. In this paper we study the mass transportation problem given by

$$\ell_1(P,Q) = \inf\left\{ \int \|x-y\|\, d\mu(x,y); \mu \in \mathcal{M}(P,Q) \right\}, \tag{1.1}$$

where $\mathcal{M}(P,Q)$ is the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $P$ and $Q$, and $\|\cdot\|$ denotes the Euclidean norm. $\ell_1$ is the minimal version of the $L^1$-metric $L^1(X,Y) = \mathbb{E}\|X-Y\|$ for random vectors $X, Y$ in $\mathbb{R}^d$.

More generally, for $p \in [1,\infty)$, the *Kantorovich $\ell_p$-metric* $\ell_p(P,Q)$ is defined as

$$\ell_p(P,Q)^p = \inf\left\{ \int \|x-y\|^p \, d\mu(x,y); \mu \in \mathcal{M}(P,Q) \right\}. \tag{1.2}$$

The mass transportation problem (1.2) was introduced in Kantorovich (1942, 1948) in the general case where the cost function is a metric and Kantorovich derived a dual representation for it. This led in 1957 to the fundamental representation of the minimal $L^1$-metric $\ell_1$ by the dual Lipschitz metric, the Kantorovich–Rubinstein Theorem. This duality result arose a lot of interest and led to extensions over the years to come in various areas from probability, statistics over analysis and geometry to various application areas like clustering or image analysis. From some point on, in much of the literature the term $L^p$-Wasserstein distance was used for $\ell_p(P,Q)$ instead of the historically correct notion of Kantorovich $\ell_p$-metric; see the detailed historical remark given at the end of Chapter 6 in Villani (2009) and Rüschendorf (1998). To keep the paper in connection with the recent literature, we remark that $\ell_p(P,Q)$ is also known as the *$L^p$-Wasserstein distance* $W_p(P,Q)$.

*Notation*

A measure $\mu \in \mathcal{M}(P,Q)$ is called an *$L^p$-optimal coupling* of $P$ and $Q$ or an optimal coupling with respect to the $l_p$-metric if

$$\int \|x-y\|^p \, d\mu(x,y) = \ell_p(P,Q)^p.$$

Similarly, a pair of $\mathbb{R}^d$-valued random vectors $(X,Y)$ on a probability space, $X \sim P$, $Y \sim Q$, is called an *$L^p$-optimal coupling* of $P$ and $Q$ if $E\|X-Y\|^p = \ell_p(P,Q)^p$.

In general, any measure (coupling) $\mu \in \mathcal{M}(P,Q)$ describes a transportation plan for the mass distribution from $P$ to $Q$ or, equivalently, the joint distribution of a pair of random vectors $(X,Y)$ on $\mathbb{R}^d$ with $X \sim P$ and $Y \sim Q$. Using conditional distributions, one obtains

$$\int \|x-y\|^p \, d\mu(x,y) = \int \left( \int \|x-y\|^p \mu(dy|x) \right) P(dx).$$

Any mass at point $x$ is transported to (possibly many locations) $y$ according to the conditional distribution $\tau(x,\cdot) = \mu(\cdot|x)$, which is called a *transport kernel* of $P$ and $Q$. Notice that $P \times \mu(\cdot|x)$ is a coupling of $P$ and $Q$ and let $\mathcal{K}(P,Q)$ be set of all transport kernels $\tau(x,\cdot)$ with $P \times \tau(x,\cdot) \in \mathcal{M}(P,Q)$.

A transport kernel $\tau(x,\cdot)$ is called an *$L^p$-optimal transport kernel* if $P \times \tau(x,\cdot)$ is an $L^p$-optimal coupling. If the optimal transport kernel is deterministic, i.e. $\tau(x,\cdot) = \varepsilon_{T(x)}$, where $T$ is a measurable function which transports $P$ to $Q$, $P^T = Q$, then $T$ is said to be an *$L^p$-optimal transport map* from $P$ to $Q$. In this case, the mass in $x$ is not split and $(X,T(X))$ is an $L^p$-optimal coupling of $P,Q$. Throughout the text, $\varepsilon_x$ indicates the Dirac measure at $x$.

*The cases $p = 1,2$*

Problem (1.2) has a closed analytical solution on the line for any $p \geq 1$, but is in general difficult to determine for higher dimensions $d > 1$. Most of the literature on explicit solutions has focused on the case $p = 2$, i.e. on the minimization of the $L^2$-distance, which allows for an easier mathematical treatment. For quadratic cost functions, a basic and general characterization of optimal couplings, resp. transports, was given in Rüschendorf and Rachev (1990) and Brenier (1991).

A pair of random vectors $X \sim P$, $Y \sim Q$ on $\mathbb{R}^d$ is an $L^2$-optimal coupling, i.e. $(E\|X-Y\|^2)^{\frac{1}{2}} = \ell_2(P,Q)$, if and only if for some lower semi-continuous convex function $f$ one has

$$Y \in \partial f(X) \text{ a.s.,}$$

where $\partial f$ denotes the subdifferential of $f$; equivalently, if and only if the joint distribution of $(X,Y)$ has a cyclically monotone support. This characterization was given first under the assumption of finite second moments but generalized later on to measures $P,Q$, with $\ell_2(P,Q) < \infty$.

This criterion allows for a fairly general construction of optimal couplings for $p=2$. In particular it implies the construction of optimal couplings of two multivariate normals, a case which had been solved first by analytical tools in Olkin and Pukelsheim (1982) and Dowson and Landau (1982). The optimal transport between the centered Gaussian distributions $N(0,\Sigma_1)$ and $N(0,\Sigma_2)$ is a linear map; see for instance Corollary 3.2.13 in Rachev and Rüschendorf (1998).

The geometry of optimal transport on $\mathbb{R}^d$ varies a lot with the choice of $p$ and the solution in the Gaussian case for $p=2$ does not extend to $p \neq 2$; see Figure 1.1.

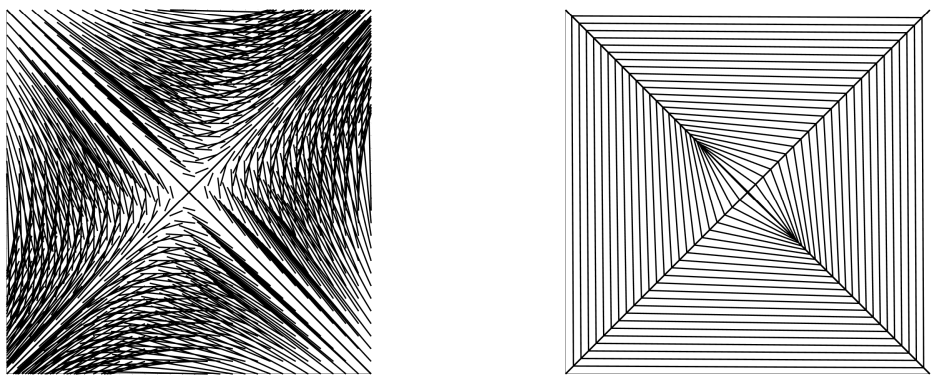

**Figure 1.1** Optimal transportation curves with respect to the squared Euclidean distance ($p=2$, left) and the Euclidean distance ($p=1$, right), between the same Gaussian distributions (corresponding to case B in Table 6.1).

In this paper we focus on the case $p=1$ which has received less attention in the literature, probably because of its more cumbersome nature. In computer vision, the $\ell_1$-metric (1.1) is also known as *Earth's mover distance* (EMD), and is widely used in content-based image retrieval to compute distances between the color histograms of two digital images; see for instance Rubner et al. (2000). Relevant implementations of the EMD also exist in other fields like biology; see Orlova et al. (2016). However, for such applications, the EMD is typically computed between point clouds via numerical techniques. Particularly relevant in these problems is the case of Gaussian distributions (Ruttenberg and Singh, 2011), for which we provide a general method to compute the EMD on the plane.

For the $\ell_1$-metric the first approach to show the existence and to characterize optimal transports was proposed in Sudakov (1979). Its main idea was to reduce via the Kantorovich-Rubinstein theorem the optimal coupling problem for the measures $P,Q$, by the introduction of a suitable decomposition $(R_t)$ of the basic space – in our case $\mathbb{R}^d$ – by means of the dual potentials. These are one-dimensional rays in the case of strict norms and absolutely continuous distributions. Then by disintegration one gets a family of simpler transportation problems for $P_t, Q_t$, which however in general are not Lebesgue continuous (even when $P,Q$ are) contrary to the absolute continuity statement in Sudakov (1979). If they are then one gets optimal maps $T_t$ on $R_t$ which can be pasted together to an optimal map on the basic space; see Caravenna (2011). With additional regularity properties on the densities of $P, Q$, this approach was successfully followed in Trudinger and Wang (2001), Caffarelli et al. (2002), Ambrosio and Pratelli (2003), and Bianchini and Daneri (2018). Evans and Gangbo (1997) develop a PDE approach to derive a potential $u$ for the associated duality problem. As a result they obtain the existence of a potential $u$ describing the direction of the optimal transport and derive an ODE for the "transport density", i.e., the length of transport in the corresponding directions. Even if counterexamples were given to Proposition 78 in Sudakov (1979), it is also true that the strategy outlined by Sudakov to build optimal transportation maps can indeed be justified for the Euclidean norm, and for general norms in dimension $d = 2$; see for instance Ambrosio (2003) and Ambrosio et al. (2004).

If measured by the Euclidean distance $\|x-y\|$, the shortest path between two points $x, y \in \mathbb{R}^d$ is given by the segment $S_{xy}$. Moreover, an $L^1$-optimal transport must satisfy a no-crossing condition; see Villani (2009, Ch. 8). This suggests that optimal transportations for $p = 1$ should be given on a family of non-intersecting rays.

The Sudakov approach to $\ell_1$-transportation problems as described above is based on a solution of the dual problem as given in the Kantorovich-Rubinstein Theorem. However, a dual solution is not available in explicit form in typical examples.

In our paper we introduce a modified approach to this problem. In a first step, a reduction theorem allows us to restrict to the case of so-called reduced measures with disjoint supports. The symmetry properties of the reduced measures give a hint to the decomposition of the space into sectors. In a second step, we make use of the non-crossing property of optimal transports for the choice of transportation directions within the sectors. Finally, in a third step, we verify the optimality of the induced construction by means of the characterization results of optimal transportation.

We apply this three-step approach to derive explicit optimal coupling results for a series of examples of distributions as classes of normals or elliptical distributions as a counterpart to the $p = 2$ case.

*Summary*

In Section 2 of this paper we remind and review some of the basic notions of optimal $c$-couplings as $c$-cyclical monotonicity, $c$-convexity, $c$-transform, and the specialization of these notions to the metric case, i.e., to $\ell_1$-convexity, $\ell_1$-subgradients, and $\ell_1$-subdifferentials which allows a manageable description as in the $\ell_2$-case. Our aim is to establish a variety of simple tools and descriptions of suitable admissible dual functions which allow to construct optimal couplings between two given probability measures $P, Q$.

In Section 3 we state, based on the Kantorovich–Rubinstein Theorem, a reduction principle allowing to reduce optimal coupling problems to the case of measures with disjoint supports. In fact versions of this result are given already in early work of Rüschendorf (1991) and in Gangbo and McCann (1995), but here we give a self-contained proof. This reduction principle is of particular importance for the construction of $L^1$-optimal couplings and is complemented with the description of the $\ell_1$-subdifferential of (admissible) dual Lipschitz functions $f$.

As applications in Section 4 we describe the quantile optimal coupling technique on the line and, based on the three-step approach as introduced above, give concrete optimal coupling results for a variety of examples including shifts, radial transformations, and some classes of elliptical distributions. We also provide a general dual bound on $\ell_1(P, Q)$ which is shown to be sharp in some specific examples.

In Section 5, we give a general formulation of our three-step decomposition approach leading to the transport on rays in a suitable decomposition of the space into sectors. The length of the transportation on the rays is given by the one-dimensional quantile coupling of the conditional distributions on the rays in the sector. Note that the optimal transport is in general not given by a transport map $T$ but by a transport plan $\mu$.

As a main class of examples for this construction method, in Section 6 we give the $L^1$-optimal coupling between general bivariate Gaussian distributions.

In Section 7 we discuss optimality of simple transportations as the scaling of one or more components. The scaling of one component can be shown to be optimal. We strengthen a result in Alfonsi and Jourdain (2014) showing that even the scaling of just two components is not optimal in general. We then discuss some related constructions also in higher dimensions.

*Related results*

While for the $\ell_2$(resp. $W_2$)-metric several applicable results are available, the main literature concerning the $\ell_1$-metric is mostly concerned with general existence, uniqueness, and structural characterizations; see Rüschendorf (1991, 1995), Gangbo and McCann (1996), and Ambrosio and Pratelli (2003). Numerical methods for the solution are given in Peyré and Cuturi (2018) and Eckstein and Kupper (2021).

Few exceptions containing concrete examples are available in the literature; see Cuesta-Albertos et al. (1993), Rüschendorf (1995), Uckelmann (1998), and the present paper. In Cuesta-Albertos et al. (1993) the authors show $\ell_1$-optimality of radial transformations. Several criteria for determining $L^1$-optimal couplings were given in Rüschendorf (1995) but so far were not elaborated to obtain solutions for classes of concrete probability measures. Uckelmann (1998) in his thesis derived some usable practical criteria and investigated some examples.

Our paper introduces a usable method for the solution of $L^1$-optimal couplings. It includes or generalizes the concrete examples of $L^1$-optimal couplings given in the literature up to now, and introduces some novel ones. In particular we derive $L^1$-optimal couplings for some classes of elliptical distributions and we give for the first time a general construction of $L^1$-optimal couplings between two bivariate Gaussian distributions.

## 2. Optimal $c$-couplings and $c$-convexity

For the general transportation problem

$$\sup\{\mathbb{E}[c(X,Y)] : X \sim P, Y \sim Q\} \tag{2.1}$$

of maximizing the expectation of a measurable cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, optimal couplings have been first characterized in Rüschendorf (1991, Theorem 18). Under weak integrability conditions on $c$, it holds that a pair $X \sim P$, $Y \sim Q$ is an optimal $c$-coupling (i.e., a solution of (2.1)) if and only if

$$Y \in \partial_c f(X) \text{ a.s.,}$$

for some $c$-convex function $f$, where $\partial_c f(x)$ denotes the $c$-subdifferential of $f$, or equivalently if and only if

$$f(X) + f^c(Y) = c(X,Y) \text{ a.s.,}$$

where $f^c(y) = \sup_x\{c(x,y) - f(x)\}$ is the $c$-*conjugate* of $f$. Optimality is further equivalent to the joint distribution of $(X,Y)$ having a $c$-cyclically monotone support. For these notions and more details we refer to Rüschendorf (1991, 1996).

These characterizations hold similarly for the corresponding $\inf$ problem by switching the sign of $c$. The notion of $c$-convexity of $f$ is then transformed to $c$-concavity and hence in the $c$-cyclicality conditions the sign changes; see Rüschendorf (1995) or Gangbo and McCann (1996). In the case of the minimal $\ell_1$-metric, one takes in (2.1) the function $c_1(x,y) = -\|x-y\|$. Let $\mathrm{Lip}_1$ denote the set of Lipschitz functions $f$ with constant 1, i.e., such that $|f(x)-f(y)| \leq \|x-y\|$. Recall that a differentiable function $f$ is in $\mathrm{Lip}_1$ if and only if $\|\nabla f\| \leq 1$.

The $\ell_1$-subdifferential of $f \in \mathrm{Lip}_1$ is the set

$$\begin{aligned} \partial_1 f(x) &= \{y;\ \ f(z) - f(y) \geq \|x-z\| - \|z-y\|,\ \ \forall z\} \\ &= \{y;\ \ |f(y) - f(x)| = \|x-y\|\}. \end{aligned} \tag{2.2}$$

The elements of $\partial_1 f(x)$ are called $\ell_1$-subgradients.

The $c_1$-conjugate of a measurable function $f$ is in $\mathrm{Lip}_1$, and the $c_1$-conjugate of a $\mathrm{Lip}_1$ function $f$ is given by $f^c = -f$. From these results, and the fact that a function is c-convex if and only if $f = (f^c)^c$ (Dietrich, 1988) the following equivalent optimality conditions immediately follow.

**Theorem 2.1.** *The following conditions are equivalent.*

a) *$f$ is $\ell_1$-convex (i.e., $c_1$-convex, with $c_1(x,y) = -\|x-y\|$);*
b) *$f \in \mathrm{Lip}_1$;*
c) *$f^c = -f \in \mathrm{Lip}_1$.*

From the above statements and the characterizations in Rüschendorf (1991), the following theorem directly follows.

**Theorem 2.2.** *The following conditions are equivalent.*

a) *The pair $X \sim P, Y \sim Q$ is an $L^1$-optimal coupling;*
b) *$Y \in \partial_1 f(X)$ a.s. for some $f \in \mathrm{Lip}_1$;*
c) *$X \in \partial_1(-f)(Y)$ a.s. for some $f \in \mathrm{Lip}_1$;*
d) *$f(Y) - f(X) = \|X-Y\|$ a.s. for some $f \in \mathrm{Lip}_1$.*

The $\ell_1$-convex or, equivalently, the $\mathrm{Lip}_1$ functions $f$ allow a representation of the form

$$f(x) = \sup_{(y,a)\in A} (-\|x-y\| + a), \tag{2.3}$$

for some $A \subset \mathbb{R}^{d+1}$, and $y \in \partial_1 f(x)$ iff the sup in (2.3) is attained in $y$, i.e.

$$f(x) = -\|x-y\| + a, \tag{2.4}$$

and in this case $a = f(y)$. In particular, (2.3) and (2.4) imply a method of determining $\ell_1$-subdifferentials by considering inequalities as induced by (2.3), (2.4).

Finally, there exists a quite applicable condition to verify the optimality of a transport plan: the support of a $L^1$-optimal coupling has to be *c-cyclically monotone*; see for instance Theorem 2.3 and Corollary 2.4 in Gangbo and McCann (1996) or Theorem 2.2 in Rüschendorf (1996).

If $\ell_1(P,Q) < \infty$, a measure $\mu \in \mathcal{M}(P,Q)$ is an $L^1$-optimal coupling of $P$ and $Q$ if and only if $\mu$ has a *c-cyclically monotone* support, that is for any finite number of points $(x_i,y_i), i = 1,\ldots,m$, in the support of $\mu$, and any possible permutation $\sigma$ of $\{1,\ldots,m\}$, we have

$$\sum_{i=1}^{m} \|x_i - y_i\| \leq \sum_{i=1}^{m} \|x_i - y_{\sigma(i)}\|. \tag{2.5}$$

Having a *c-cyclically monotone* support also implies that optimal transportation lines cannot intersect.

# 3. Criteria for $L^1$-optimal couplings

The basic result for $L^1$-optimal couplings is the Kantorovich–Rubinstein Theorem, going back in its basic form to the classical work of Kantorovich and Rubinstein (1957). For its history see Rüschendorf (1991, around Theorem 11) or Rachev and Rüschendorf (1998). In our framework of $\mathbb{R}^d$ it reads as follows.

**Theorem 3.1** (Kantorovich–Rubinstein Theorem). *Let $P, Q \in M^1(\mathbb{R}^d)$. Then, the $\ell_1$-metric $\ell_1(P,Q)$ is identical to the dual Lipschitz metric, i.e.*

$$\ell_1(P,Q) = \sup\left\{\int f d(P-Q);\ \ f \in \mathrm{Lip}_1\right\}. \tag{3.1}$$

Notice that, by the symmetry of $\ell_1$, i.e. $\ell_1(P,Q) = \ell_1(Q,P)$, it equivalently holds that

$$\ell_1(P,Q) = \sup\left\{\int f d(Q-P);\ \ f \in \mathrm{Lip}_1\right\}. \tag{3.2}$$

Theorems 2.1, 2.2, and Theorem 3.1 give useful tools to determine $\ell_1$ and the corresponding optimal couplings/transports. Any $\mathrm{Lip}_1$ *dual* function $f$ in (3.1) provides a general lower bound on $\ell_1$, that is

$$\ell_1(P,Q) \geq \int f d(P-Q).$$

This bound is sharp (the above $\geq$ holds with $=$) if and only if, for $X \sim P, Y \sim Q$, one finds that

$$\mathbb{E}\|X - Y\| = \int f d(P-Q).$$

The dual solutions $f$ in (3.1) are also called Kantorovich potentials in the analysis literature; see for instance Ambrosio and Pratelli (2003).

**Remark 3.2.** Two bounds are readily available for the $\ell_1$-distance between general distributions. For general random vectors $X \sim P, Y \sim Q$, we have

$$\Big|\mathbb{E}\|Y\| - \mathbb{E}\|X\|\Big| \leq \ell_1(P,Q) \leq \ell_2(P,Q). \tag{3.3}$$

The left inequality is implied by (3.1) and (3.2) via the basic $\mathrm{Lip}_1$ function $f(x) = \|x\|$, the right one is derived from Jensen inequality.

The proof in Rüschendorf (1991, Theorem 11) derives Theorem 3.1 as consequence of the general Monge–Kantorovich duality theorem and is based on an interesting reduction principle derived within the proof, which is valuable for the application to concrete examples. A proof of this principle based on properties of $c$-cyclical monotonicity of optimal couplings is also given in Gangbo and McCann (1995, Proposition 2.9). We give in the following a simple proof of this reduction principle by means of the Kantorovich–Rubinstein Theorem.

For $P, Q \in M^1(\mathbb{R}^d)$ let $P \wedge Q$ denote the inf of both measures, i.e., for $P = g_1\nu, Q = g_2\nu$, one has $P \wedge Q = (g_1 \wedge g_2)\nu$, where $g_1 \wedge g_2(x) = \min\{g_1(x), g_2(x)\}$.

**Theorem 3.3** (Reduction principle). *Let $0 < \ell_1(P,Q) < \infty$ and define $k = \int(g_1 - g_1 \wedge g_2)d\nu$, that is $k = \|P - P\wedge Q\|_\infty \le 1$, where $\|\cdot\|_\infty$ denotes the* sup*-distance. Define the* reduced *measures*

$$P_1 = \frac{1}{k}(P - P\wedge Q) \text{ and } Q_1 = \frac{1}{k}(Q - Q\wedge P). \tag{3.4}$$

*Then:*

a)
$$\ell_1(P,Q) = k\,\ell_1(P_1,Q_1);$$

b) *If $\tau(x,\cdot) \in \mathcal{K}(P_1,Q_1)$ is an $L^1$-optimal transport kernel of the reduced measures $P_1$ and $Q_1$, and*

$$A = \{g_1 > g_2\},\ B = \{g_2 > g_1\}, \tag{3.5}$$

*then*

$$\mu(\cdot|x) = (g_1 \wedge g_2(x))\,\varepsilon_x + 1_A(x)(g_1(x) - g_1\wedge g_2(x))\,\tau(x,\cdot) \in \mathcal{K}(P,Q) \tag{3.6}$$

*is an $L^1$-optimal transport kernel of $P$ and $Q$.*

*In particular, if $T$ is an $L^1$-optimal transport map from $P_1$ to $Q_1$, then (3.6) with $\tau(x,\cdot) = \varepsilon_{T(x)}$ is an $L^1$-optimal transport kernel of $P$ and $Q$.*

Theorem 3.3 allows to reduce the $L^1$-optimal coupling of $P,Q$, to that of the reduced measures $P_1,Q_1$, which have disjoint supports $A$ and $B$ in the sense that $P(A^c) = 0$ and $Q(B^c) = 0$ (not necessarily closed sets); see Figure 3.1.

Equation (3.6) simply means that the mass which is common to $P$ and $Q$ is not moved. Starting from $P$, the excess mass $(g_1 - g_2)$ in $A$ is then moved to $B$ according to the transport kernel $\tau(x,\cdot)$ or (if available) the transport map $T$.

***Proof of Theorem 3.3.*** When $k = 1$, we have $P = P_1$ and $Q = Q_1$. With $R = \frac{1}{1-k}(P\wedge Q)$ we have the decomposition $P = (P - P\wedge Q) + P\wedge Q = k\,P_1 + (1-k)\,R$ and, similarly, $Q = k\,Q_1 + (1-k)\,R$. This implies, using the Kantorovich-Rubinstein theorem, that

$$\begin{aligned}\ell_1(P,Q) &= \sup\left\{\int f d(P-Q); f \in \mathrm{Lip}_1\right\}\\ &= \sup\left\{\int f d\big[(k\,P_1 + (1-k)\,R) - (k\,Q_1 + (1-k)\,R)\big];\ \ f \in \mathrm{Lip}_1\right\}\\ &= \sup\left\{k\int f d(P_1 - Q_1);\ \ f \in \mathrm{Lip}_1\right\} = k\,\ell_1(P_1,Q_1). \quad \square\end{aligned}$$

The following theorem gives a useful tool for the determination of $L^1$-optimal couplings by stating that optimal transports are concentrated on rays in the direction of the gradient of a dual solution $f$. Even if it is known in the literature that $L^1$-optimal transports have this property (Caffarelli et al., 2002; Ambrosio and Pratelli, 2003; Caravenna, 2011), we find it useful to formalize this result on $\mathbb{R}^d$ by the following theorem, which is a modification and extension of Proposition 1.14 in Uckelmann (1998).

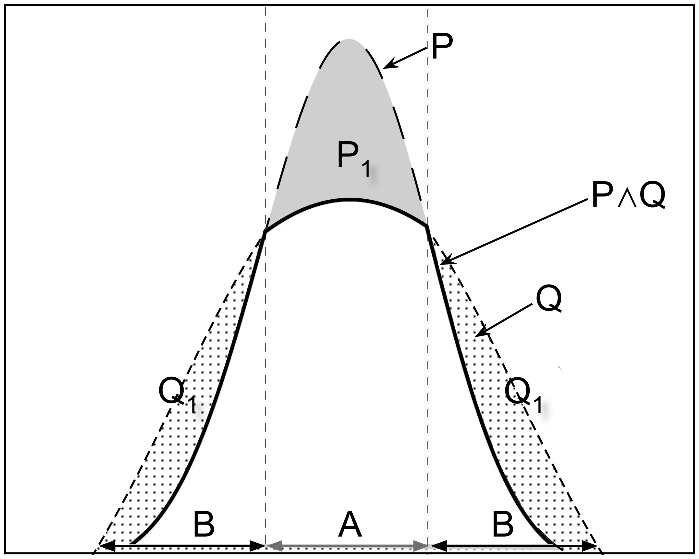


**Figure 3.1** Reduction of measures $P, Q$ to $P_1, Q_1$ with disjoint supports. To find an optimal transport w.r.t to the $l_1$-metric, one does not need to move the white probability mass which is common to $P$ and $Q$.

**Theorem 3.4.** *Let $f \in \mathrm{Lip}_1$. Then*

a) $x \in \partial_1 f(x)$;
b) *If $y \in \partial_1 f(x)$, then*

$$S_{x,y} = \{z; z = x + \alpha(y - x), 0 \leq \alpha \leq 1\} \subset \partial_1 f(x),$$

*and for $z \in S_{x,y}$ it holds that:*

$$z \in \partial_1 f(x), y \in \partial_1 f(z), \text{ and } z \in \partial_1 f(y).$$

c) *If $f$ is differentiable in $x$ and $\|\nabla f(x)\| < 1$, then $\partial_1 f(x) = \{x\}$.*
*If $\|\nabla f(x)\| = 1$, then for some $T_1(x) \leq 0 \leq T_2(x) \leq \infty$, we have that*

$$\partial_1 f(x) = \{y; \;\; y = x + s\nabla f(x), \; T_1(x) \leq s \leq T_2(x)\}.$$

***Proof.*** a) is immediate from the description of the $\ell_1$-subdifferential in (2.2).
b) If $y \in \partial_1 f(x)$, $f \in \mathrm{Lip}_1$, then by (2.2) one has $|f(y) - f(x)| = \|y - x\|$. For $z \in S_{x,y}$ this implies

$$\begin{aligned}\|y - x\| &= |f(y) - f(x)| \\ &= |f(y) - f(z) + (f(z) - f(x))| \\ &\leq |f(y) - f(z)| + |f(z) - f(x)| \leq \|y - z\| + \|z - x\|.\end{aligned}$$

On the other hand, since $z \in S_{x,y}$ it follows that

$$\|y - x\| = \|y - z\| + \|z - x\|,$$

which, being $f \in \mathrm{Lip}_1$, implies that

$$|f(y) - f(z)| = \|y - z\| \text{ and } |f(z) - f(x)| = \|x - z\|.$$

c) Assume there exists $y \neq x$ with $y \in \partial_1 f(x)$. Then, by point b), $S_{x,y} \subset \partial_1 f(x)$ and thus for $v = \frac{y-x}{\|y-x\|}$ and $0 \leq t \leq \|y - x\|$, one has that $x + tv \in \partial_1 f(x)$ and by (2.2) also that

$$|f(x+tv) - f(x)| = \|x + tv - x\| = \|tv\| = t\|v\| = t.$$

Since $f$ is differentiable in $x$, we can write

$$D_v f(x) = \lim_{t=0} \frac{f(x+tv) - f(x)}{t} = \lim_{t=0^+} \frac{|f(x+tv) - f(x)|}{t} = 1.$$

On the other hand, we have that

$$1 = |D_v f(x)| = |\langle v, \nabla f(x) \rangle| \leq \|v\| \, \|\nabla f(x)\| = \|\nabla f(x)\|. \tag{3.7}$$

If $\|\nabla f(x)\| < 1$, then (3.7) leads to a contradiction and, therefore, $\partial_1 f(x) = \{x\}$.

If $\|\nabla f(x)\| = 1$, then again by (3.7) one has that $\nabla f(x) = v = \frac{y-x}{\|y-x\|}$. As a consequence $\partial_1 f(x)$ lies in the ray in $x$ generated by $\nabla f(x)$, i.e. $\partial_1 f(x) \subset x + \mathbb{R}^1 \nabla f(x)$ and $\partial_1 f(x)$ is by b) connected. This implies c). □

**Remark 3.5.** We have the following remarks on Theorem 3.4.

1) The statement in a) is equivalent to $\ell_1(P, P) = 0$ and is a direct consequence of $\| \cdot \|$ being a norm. In other words, the identity is an optimal map.
2) Point b) is an analytical statement of the fact that the shortest path between two points under the Euclidean distance is a straight line. Notice that this is not true for any $p > 1$. For instance, there is no shortest path between two points under the squared Euclidean distance ($p = 2$).
3) Point c) implies that optimal transports are concentrated in each point on lines in the direction of the gradient of a dual solution $f$ and the gradient of $f$ on this line is constant. In the following sections we show how this construction principle can be applied.

# 4. Examples of optimal transports

The following applications of the construction principles given in Section 3 concern some examples like the one-dimensional case, the optimality of shifts and radial transformations, and optimal coupling results for some classes of elliptical distributions. Based on them we give in Section 5 some more general construction results for $L^1$-optimal couplings.

## 4.1. One-dimensional case

If $P, Q \in M^1(\mathbb{R}^1)$ with distribution functions $F, G$, then it is well-known since Dall'Aglio (1956) that an optimal coupling w.r.t. $\ell_1$ (and in fact w.r.t. any $\ell_p$, $1 \leq p \leq \infty$) is given by the *quantile coupling*

$$(F^{-1}(U), G^{-1}(U)), \quad U \sim \mathcal{U}([0,1]),$$

which gives

$$\ell_1(P,Q) = \mathbb{E}\Big|F^{-1}(U) - G^{-1}(U)\Big|.$$

To transport $P$ to $Q$ it is natural to transport mass in $F^{-1}(t)$ to the right, i.e.

$$f'(t) = +1 \quad \text{if} \quad F^{-1}(t) < G^{-1}(t), \tag{4.1}$$

and to the left, i.e.

$$f'(t) = -1 \quad \text{if} \quad F^{-1}(t) > G^{-1}(t), \tag{4.2}$$

while choosing

$$f'(t) = 0 \quad \text{if} \quad F^{-1}(t) = G^{-1}(t), \tag{4.3}$$

i.e. no transport takes place here. By (4.1)–(4.3) the function $f(t) = \int_0^t f'(u)du$ is defined up to possible countable many sign changes of $(F^{-1} - G^{-1})$.

Such $f$ is in $\mathrm{Lip}_1$ and for $t \in [0,1]$ with $F^{-1}(t) > G^{-1}(t)$, the set

$$\Big\{u \geq t;\ \ F^{-1}|_{[t,u]} \geq G^{-1}|_{[t,u]}\Big\}$$

lies in the $\ell_1$-subdifferential of $f$ at $F^{-1}(t)$ and similarly for the other domain. This implies that $G^{-1}(U) \in \partial_1 f(F^{-1}(U))$ a.s. and since $(F^{-1}(U), G^{-1}(U))$ is a coupling of $P,Q$, it is an $L^1$-optimal coupling. In the case of finitely many sign changes of $(F^{-1} - G^{-1})$ in the points $a_1, \ldots, a_n$, one can represent $f$ in the form

$$f(x) = \sup_{a \in \{a_1, \ldots, a_n\}} \{-|x-a| + c_a\},$$

for constants $c_a \in \mathbb{R}$, which is in $\mathrm{Lip}_1$ and hence $\ell_1$-convex as the supremum of $\mathrm{Lip}_1$ functions; see Figure 4.1.

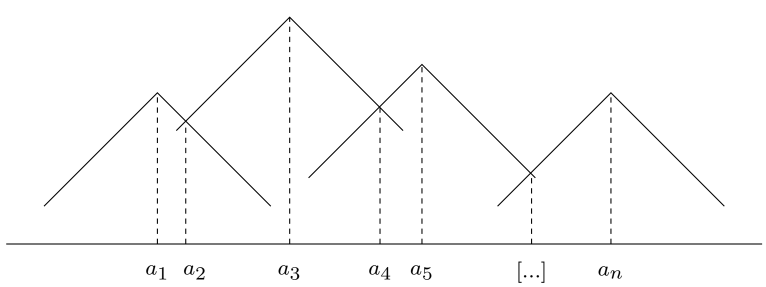


**Figure 4.1** An optimal choice of the dual $\ell_1$-convex function $f$ for the $\ell_1$-optimal coupling of two distributions intersecting at a finite number of points $a_1, \ldots, a_n$. The gradient of $f$ shows the direction into which mass should be moved.

In Figure 4.2 we consider the case $P = N(0,\sigma^2)$, $Q = U([-a,a])$, for $a > 0$. The corresponding distributions have three intersection points at $a_1 = -R, a_2 = 0, a_3 = R$, where $R$ is the positive root of

$$\Phi\left(\frac{R}{\sigma}\right) = \frac{R+a}{2a}.$$

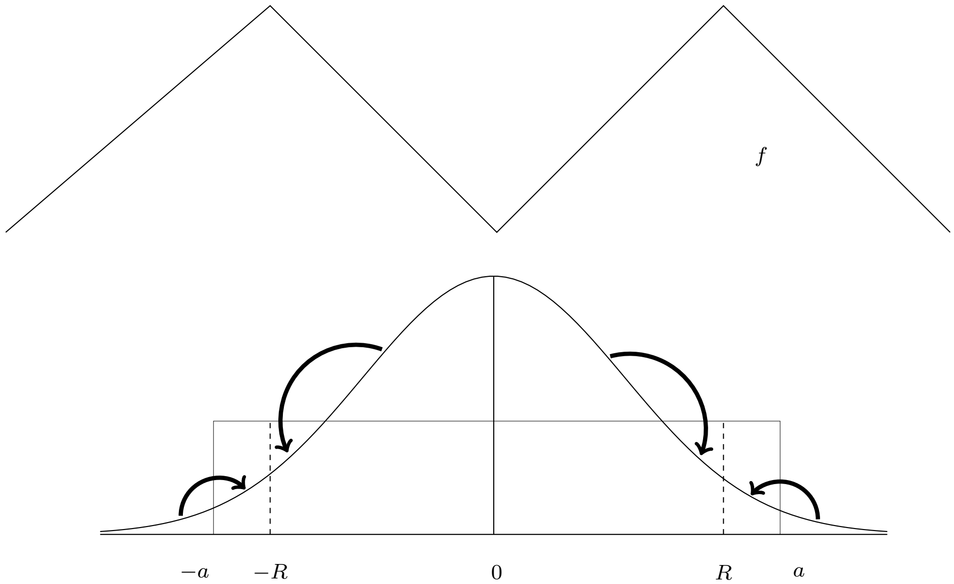


**Figure 4.2** Optimal dual function $f$ and optimal transport between a normal $N(0,\sigma^2)$ and a uniform $\mathcal{U}([-a,a])$.

The one-dimensional quantile coupling on the line will play a fundamental role also in the construction of optimal couplings for multivariate measures as described in Sections 5–6.

## 4.2. Optimality of shifts

The following result on the optimality of shifts allows that in general one can restrict to the couplings of distributions $P, Q$ with zero mean.

**Proposition 4.1** (Optimality of shifts). *Let $(X,Y)$, with $X \sim P, Y \sim Q$, be an $L^1$-optimal coupling of $P$ and $Q$. Then, for $a,b \in \mathbb{R}^d$, $(X+a, Y+b)$ with $X+a \sim P_a, Y+b \sim Q_b$, is an $L^1$-optimal coupling of $P_a$ and $Q_b$.*

***Proof.*** Since $\ell_1(P_a,Q_b) = \ell_1(P_{a-b},Q)$, we can assume w.l.o.g. that $b=0$. Assume that $(X,Y)$ is an $L^1$-optimal coupling of $P$ and $Q$. By Theorem 2.2, b), there exists $f \in \mathrm{Lip}_1$ such that $Y \in \partial_1 f(X)$ a.s.. Define the function $g(x) = f(x-a) \in \mathrm{Lip}_1$. For $Z = X+a$ it holds a.s. that

$$Y \in \partial_1 f(X) = \partial_1 f(Z-a) = \partial_1 g(Z).$$

Hence $(Z = X+a, Y)$ is a $L^1$-optimal coupling of $P_a$ and $Q$. □

**Remark 4.2.** The optimality of shifts for the coupling of measures $P_a, P$ is given already in Cuesta-Albertos et al. (1993). Their simple direct proof however does not extend to the optimal coupling result for shifts as in Proposition 4.1, which is needed in the following for the reduction to distributions with zero mean.

## 4.3. Radial transformations

Let $R(x) = \alpha(\|x\|)\frac{x}{\|x\|}$, $x \in \mathbb{R}^d$, with $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ increasing, be a radial transformation. Consider the $\mathrm{Lip}_1$ function $f(x) = \|x\|$, $x \in \mathbb{R}^d$. Then for $x \neq 0$, $\nabla f(x) = \frac{x}{\|x\|}$ and the $l_1$-subdifferential of $f$ in $x$ is the ray $R_t = \mathbb{R}^1_+ t$, generated by $t = \frac{x}{\|x\|}$. Note that

$$\|x - R(x)\| = \big\|(\|x\| - \alpha(\|x\|))t\big\| = \big|\|x\| - \alpha(\|x\|)\big| = |f(x) - f(R(x))|.$$

By (2.2) this implies that

$$R(x) \in \partial_1 f(x). \tag{4.4}$$

As consequence we obtain that radial transportations are optimal. A general version of this result is given in Cuesta-Albertos et al. (1993, Theorem 3.2).

**Theorem 4.3.** *Let $P, Q \in M^1(\mathbb{R}^d)$; if for $X \sim P, Y \sim Q$ it holds $\lambda^d$-a.s. that*

$$\frac{X}{\|X\|} = \frac{Y}{\|Y\|}, \tag{4.5}$$

*and $\|X\|, \|Y\|$ are optimally coupled, then $(X, Y)$ is an $L^1$-optimal coupling.*

Under assumption (4.5) one can construct an optimal pair $(X, Y)$ in the following way. Let $X \sim P, Y \sim Q$, and define $Z = \frac{Y}{\|Y\|}$ and let $V \sim \mathcal{U}([0,1])$ be independent of $Z$. Then define $X = LZ$, with

$$L = F^{-1}_{\|X\|} \circ \tau \quad \text{where} \quad \tau = F_{\|Y\|}(\|Y\|, V)$$

is the distributional transform of $F_{\|Y\|}$. Then $(X, Y)$ is an $L^1$-optimal coupling of $P$ and $Q$. Cuesta-Albertos et al. (1993) prove this result using some simple inequalities, but note that as in (4.4) we find that $Y \in \partial_1 f(x)$ for $f$ as defined above. Therefore $(X, Y)$ is an $L^1-$optimal coupling and Theorem (4.3) thus follows also from the mass transportation approach as used above.

The general insight of this example is that an optimal transport can be constructed pathwise on a infinite family of rays $R_t = \mathbb{R}^1_+ t$, $t \in B^d$, the unit ball in $\mathbb{R}^d$. In this case the transformation on the rays are easy to determine and are not parallel.

A particular case of an optimal radial transformation is of interest for the remainder of the paper. Consider the two Gaussian measures $P = N(0, a\, I_d)$ and $Q = N(0, b\, I_d)$, with $0 < a \leq b$, where by $I_d$ we denote the $d-$dimensional identity matrix. Let also $X \sim P$ and $Y \sim Q$. For $Z \sim N(0, I_d)$, recall that

$$\mathbb{E}\|Z\| = \sqrt{2}\,\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}.$$

Using the lower bound in (3.3), one obtains

$$\ell_1(P, Q) \geq \mathbb{E}\|Y\| - \mathbb{E}\|X\| = \sqrt{b}\,\mathbb{E}\|Z\| - \sqrt{a}\,\mathbb{E}\|Z\| = \left(\sqrt{b} - \sqrt{a}\right)\mathbb{E}\|Z\|. \tag{4.6}$$

On the other hand, by choosing the coupling $(X^*, Y^*)$ where $Y_i^* = \sqrt{b/a}\, X_i^*, i = 1, 2$, one finds

$$\begin{aligned}\ell_1(P,Q) &\le \mathbb{E}\|Y^* - X^*\| \\ &= \left(\sqrt{b/a} - 1\right) \mathbb{E}\|X\| = \left(\sqrt{b/a} - 1\right) \sqrt{a}\, \mathbb{E}\|Z\| = \left(\sqrt{b} - \sqrt{a}\right) \mathbb{E}\|Z\|. \end{aligned} \tag{4.7}$$

## 4.4. Optimal transports between two elliptical distributions

The class of elliptical distributions is an important class of models in the statistical analysis including many relevant models like multivariate normal, t- and Cauchy distributions. Let $\mathcal{E}(\mu, \Sigma, \phi)$ denote the elliptical distribution with location $\mu$, scaling matrix $\Sigma$ and generator (of its radial part) $\phi$. Any $X \sim \mathcal{E}(\mu, \Sigma, \phi)$ has a representation of the form

$$X \sim \mu + RAU,$$

where $U$ is uniformly distributed on the unit sphere $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d : x'x = 1\}$, $R \ge 0$ is radial random variable independent of $U$, and $A$ is a deterministic $d \times d$ matrix with $AA' = \Sigma$. The generator $\phi$ of an elliptical distribution determines the distribution of $R$. In this section we determine $L^1$-optimal couplings for two subclasses of elliptical distributions.

*Same scaling matrix, different generators*

Let $X \sim P = \mathcal{E}(0, \Sigma, \phi_1)$ and $Y \sim Q = \mathcal{E}(0, \Sigma, \phi_2)$ be elliptically distributed having representations $X = R_1 A_1 U_1, Y = R_2 A_2 U_2$, with the same scaling matrix

$$\Sigma = A_1 A_1' = A_2 A_2'.$$

In this case an $L^1$-optimal coupling of $P$ and $Q$ is obtained from Theorem 4.3 on radial transformations.

**Proposition 4.4.** *Let $X \sim P = \mathcal{E}(0, \Sigma, \phi_1)$ and $Y \sim Q = \mathcal{E}(0, \Sigma, \phi_2)$ with $R_i \sim F_{R_i}$ being the radial parts corresponding to $\phi_i, i = 1, 2$. Then for $U \sim \mathcal{U}(S^{d-1})$ independent of $V \sim \mathcal{U}([0, 1])$, a $L^1$-optimal coupling of $P$ and $Q$ is given by*

$$(X, Y) = \left( F^{-1}_{R_1 L}(V) \frac{AU}{\|AU\|}, F^{-1}_{R_2 L}(V) \frac{AU}{\|AU\|} \right), \tag{4.8}$$

*where $L = \|AU\|$ and $F_{R_i L}(x) = \int F_{R_i}(x/l) dF_L(l)$, and $A$ is any matrix such that $AA' = \Sigma$. The $\ell_1$-distance between $P$ and $Q$ is then given by*

$$\ell_1(P, Q) = \mathbb{E}\left| F^{-1}_{R_1 L}(V) - F^{-1}_{R_2 L}(V) \right|. \tag{4.9}$$

***Proof.*** For the proof we note that $\left(F^{-1}_{R_1 L}(V), F^{-1}_{R_2 L}(V)\right)$ is the optimal coupling of $\|X\|, \|Y\|$, so that (4.8) is a consequence of Theorem 4.3. □

**Remark 4.5.** We have the following remarks on Proposition 4.4.

1) Note that the optimal coupling in (4.8) is different from the "natural" coupling

$$\left(F^{-1}_{R_1}(V) AU, F^{-1}_{R_2}(V) AU\right),$$

where the radial parts are optimally coupled. This can be seen by calculating some simple examples.

2) For the calculation of the optimal coupling in (4.8) and distance in (4.9), one needs the distributions of $L = \|AU\|$ and of $R_i L, i = 1, 2$. In some examples, these can be given in explicit form.
In the case that $X = RAU \sim N(0, I_d)$, it holds that $\|X\| = R\|U\| = R \sim \sqrt{\chi_d^2}$, where $\chi_d^2$ is a chi-squared distribution with $d$ degrees of freedom. As a consequence $L = 1$ a.s. and $F_{RL} = F_R = \sqrt{\chi_d^2}$. $F_R$ has density $f_R(r) = \frac{r^{d-1} e^{-r^2/2}}{2^{d/2-1}\Gamma(d/2)}$.
In the case that $X = RAU \sim N(0, \Sigma)$, with $A = \Sigma^{1/2}$, it holds using an eigenvalue decomposition that $\|X\| \sim \sqrt{\chi_{d,\alpha}^2}$, where $\chi_{d,\alpha}^2$ is a weighted chi-squared, i.e. the distribution of $\sum_{i=1}^d \alpha_i \overline{X}_i^2$, with $\alpha_i$ being the eigenvalues of $\Sigma$ and $\overline{X}_i^2$ i.i.d. $\chi_1^2$-distributed random variables. As a consequence one obtains that

$$L = \|AU\| \sim \frac{\sqrt{\chi_{d,\alpha}^2}}{\sqrt{\chi_d^2}} = \sqrt{F_{\alpha,d,d}}$$

is a quotient of two (weighted) independent $\sqrt{\chi^2}$ distributions and thus is a non-central $\sqrt{F_{\alpha,d,d}}$ distribution. As a result, for two elliptical distributions $P$ and $Q$ as in Proposition 4.4, explicit formulas are given by (4.9) for $\ell_1(P, Q)$ involving the difference of two inverse couplings $F_{R_i L}^{-1}(V)$ of a multiplicative mixture of a $\sqrt{F_{\alpha,d,d}}$ distribution with the distribution of the radial part. If $P = t_d(0, \Sigma, v)$ is a t-distribution then the radial part $R$ has the density

$$f_R(r) = \frac{2r}{d} f_F(r^2/d), F \sim F_{d,v},$$

(see Kotz and Nadarajah, 2004) allowing for a similar explicit result as in the normal case. For related formulas for several further elliptical distributions we refer to Fang (2017).

*Symmetric bivariate elliptical distributions*

In this subsection we determine the optimal transport on the plane between two bivariate elliptical measures $P = \mathcal{E}(0, \Sigma_1, \phi)$ and $Q = \mathcal{E}(0, \Sigma_2, \phi)$, with the same generator $\phi$ and symmetric scaling matrices

$$\Sigma_1 = \begin{pmatrix} 1 & -\varrho \\ -\varrho & 1 \end{pmatrix}, \; \Sigma_2 = \begin{pmatrix} 1 & +\varrho \\ +\varrho & 1 \end{pmatrix},$$

for $0 \le \varrho \le 1$. By the reduction principle, we can reduce the transportation problem for $P$ and $Q$ to the transportation problem between the reduced measures $P_1$ and $Q_1$ as in (3.4). Let $g_1$ and $g_2$ denote the densities of $P$ and $Q$ and

$$A = \{g_1 > g_2\}, \; B = \{g_2 > g_1\}$$

be the domains of $P_1$ and $Q_1$.

Since $\Sigma_2$ is obtained from $\Sigma_1$ by applying a rotation of $\pi/2$, we have that the sets $A$ and $B$ can be characterized as

$$A = \{x = (x_1, x_2) \in \mathbb{R}^2 : x_1 x_2 < 0\},$$
$$B = \{x = (x_1, x_2) \in \mathbb{R}^2 : x_1 x_2 > 0\};$$

see Figure 4.3. Now, define the four sectors

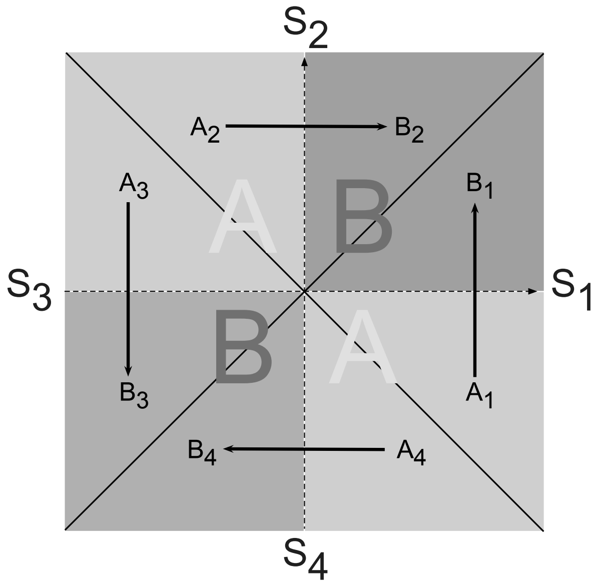


**Figure 4.3** Optimal coupling of two symmetric elliptical distributions with the same generator. The probability mass of the reduced measure $P_1$ in $A_i$ is reflected onto $B_i$, for $1 \le i \le 4$.

$$\begin{aligned} S_1 &= \{x = (x_1, x_2) \in \mathbb{R}^2; x_1 > 0, -x_1 \le x_2 \le x_1\}, \\ S_2 &= \{x = (x_1, x_2) \in \mathbb{R}^2; x_2 > 0, -x_2 \le x_1 \le x_2\}, \\ S_3 &= \{x = (x_1, x_2) \in \mathbb{R}^2; x_1 < 0, x_1 \le x_2 \le -x_1\}, \\ S_4 &= \{x = (x_1, x_2) \in \mathbb{R}^2; x_2 < 0, x_2 \le x_1 \le -x_2\}. \end{aligned} \tag{4.10}$$

We will see that the $L^1$-optimal transport in this symmetric case consists in moving the probability mass from $A$ to $B$ on directions parallel to the axis. For $A_i = A \cap S_i$ and $B_i = B \cap S_i$, for $1 \le i \le 4$, define the map $T$ as

$$T(x) = \begin{cases} (x_1, -x_2), & \text{for } x \in A_1 \cup A_3; \\ (-x_1, x_2), & \text{for } x \in A_2 \cup A_4, \end{cases} \tag{4.11}$$

and the $\mathrm{Lip}_1$ function $f : \mathbb{R}^2 \to \mathbb{R}$ as

$$f(x) = \frac{|x_1 + x_2| - |x_1 - x_2|}{2}. \tag{4.12}$$

We now show that $T$ and $f$ are optimal in this symmetric case.

**Theorem 4.6.** *An $L^1$-optimal transport kernel of $P = \mathcal{E}(0, \Sigma_1, \phi)$ and $Q = \mathcal{E}(0, \Sigma_2, \phi)$, with*

$$\Sigma_1 = \begin{pmatrix} 1 & -\varrho \\ -\varrho & 1 \end{pmatrix}, \; \Sigma_2 = \begin{pmatrix} 1 & +\varrho \\ +\varrho & 1 \end{pmatrix}, \; 0 \le \varrho \le 1, \tag{4.13}$$

*is given by*

$$\tau(x,\cdot) = (g_1 \wedge g_2(x))\varepsilon_x + 1_A(x)(g_1(x) - g_1 \wedge g_2(x))\varepsilon_{T(x)},$$

*where $T$ is the map defined in (4.11). Moreover, an optimal dual function is given by $f$ in (4.12), hence*

$$\ell_1(P,Q) = 1/2\,(\mathbb{E}|Y_1+Y_2| - \mathbb{E}|Y_1-Y_2| - \mathbb{E}|X_1+X_2| + \mathbb{E}|X_1-X_2|)\,. \tag{4.14}$$

***Proof.*** As a result of the symmetry of $P_1$ and $Q_1$, and since both elliptical distributions have the same generator, we obtain by construction of $T$ that $P_1^T = Q_1$. On sector $S_1$, probability mass is to be transported from $A_1 = A \cap S_1$ to $B_1 = B \cap S_1$. For $x = (x_1, x_2) \in A_1$, we have

$$T\binom{x_1}{x_2} = \binom{x_1}{-x_2} = x - 2x_2 \cdot k_1,$$

with $k_1 = \binom{0}{1}$, i.e. $T/A_1$ is just a shift of length $r(x) = -2x_2$ in the $y$-direction. Notice that, for $x = (x_1, x_2) \in A_1$, we have that $\nabla f(x) = k_1$. By Theorem 3.4, c), this implies that

$$T(x) \in \partial_1 f(x) \text{ for } r(x) \ \text{ such that } \ x + r(x)k_1 = \binom{x_1}{x_2 + r(x)} \in S_1.$$

Then $T_1(x) \in B_1 = B \cap S_1$ and $T_1$ maps any $x_1$-line $R^1_{x_1} = \{(x_1,x_2); -x_1 \le x_2 \le 0\} \cap A_1$ into $R^2_{x_1} = \{(x_1,x_2); 0 \le x_2 \le x_1\} \cap B_1$. Similarly, one checks that

$$T(x) = \begin{cases} x - 2x_1 \cdot k_2 = (-x_1, x_2), & \text{for } x \in A_2; \\ x + 2x_2 \cdot k_3 = (x_1, -x_2), & \text{for } x \in A_3; \\ x + 2x_1 \cdot k_4 = (-x_1, x_2), & \text{for } x \in A_4, \end{cases}$$

with directions $k_2 = (1,0)$, $k_3 = (0,-1)$, $k_4 = (-1,0)$ satisfies $T(x) \in \partial_1 f(x), x \in A$. As a consequence, we obtain that $T \in \partial_1 f(x)$ is an optimal transport of $P_1, Q_1$ and by Theorem 3.3 we get the optimal transport kernel of $P$ and $Q$. At this point one easily checks that

$$\begin{aligned}\ell_1(P,Q) &= \int f d(Q-P) = 1/2(\mathbb{E}|Y_1+Y_2| - \mathbb{E}|Y_1-Y_2| - \mathbb{E}|X_1+X_2| + \mathbb{E}|X_1-X_2|) \\ &= \mathbb{E}\,(\|X - T(X)\| 1_A(X)) = k\,\ell_1(P_1,Q_1). \quad \square\end{aligned}$$

**Remark 4.7.** We have the following remarks on Theorem 4.6.

1) For Gaussian distributions having correlation matrices as in (4.13) the value in (4.14) takes the simple expression

$$\ell_1(P,Q) = \frac{2}{\sqrt{\pi}}\left(\sqrt{1+\varrho} - \sqrt{1-\varrho}\right). \tag{4.15}$$

   A similar formula holds for Student's t distributions. For more general elliptical distributions with identical radial part corresponding formulas for the $\ell_1$-distance can be given directly based on the formulas in Remark 4.5.
2) Up to the rotation described in Proposition 6.3, the construction given in the proof of Theorem 4.6 is optimal in a more general framework. In fact, one only needs that the matrices $\Sigma_1$ and $\Sigma_2$ are

obtained as a negative and, resp., positive rotation of a common matrix $\Sigma$. Formally, for an arbitrary dispersion matrix $\Sigma$, one can assume that $\Sigma_1 = \Sigma_{-\varphi}$ and $\Sigma_2 = \Sigma_\varphi$, where

$$\Sigma_{-\varphi} = D_{-\varphi}\Sigma D'_{-\varphi}, \quad \Sigma_\varphi = D_\varphi \Sigma D'_\varphi,$$

where, for $\varphi \in [0, \frac{\pi}{2}]$, $D_\varphi = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$ denotes the rotation matrix by $\varphi$. At this point one applies the rotation described in Proposition 6.3 to obtain new matrices $\Sigma'_{-\varphi}, \Sigma'_\varphi$ for which the $L^1$-distance is the same and the construction in Theorem 4.6 is optimal.

By the Kantorovich-Rubinstein theorem, any admissible dual function $f$ yields a lower bound on $\ell_1(P,Q)$, i.e., for *general* distributions $P$ and $Q$ on $\mathbb{R}^2$ and $X = (X_1, X_2) \sim P, Y = (Y_1, Y_2) \sim Q$, one has

$$\ell_1(P,Q) \geq 1/2 \Big| \mathbb{E}|Y_1 + Y_2| - \mathbb{E}|Y_1 - Y_2| - \mathbb{E}|X_1 + X_2| + \mathbb{E}|X_1 - X_2| \Big|.$$

By considering an extended family of dual functions, one finds the following general dual bound.

**Theorem 4.8** (Dual bound). *For general measures $P,Q$ on $\mathbb{R}^2$ and $X = (X_1, X_2)' \sim P$ and $Y = (Y_1, Y_2)' \sim Q$, we have that*

$$\ell_1(P,Q) \geq \sup_{a^2+b^2 \leq 1/2} \Big| a\left(\mathbb{E}|Y_1 + Y_2| - \mathbb{E}|X_1 + X_2|\right) - b\left(\mathbb{E}|Y_1 - Y_2| - \mathbb{E}|X_1 - X_2|\right) \Big|. \tag{4.16}$$

***Proof.*** Define the function $f_{a,b} : \mathbb{R}^2 \to \mathbb{R}$ as

$$f_{a,b}(x_1, x_2) = a|x_1 + x_2| - b|x_1 - x_2|.$$

The bound immediately follows from Theorem 3.1 by checking that $\|\nabla f_{a,b}(x)\| = 2(a^2 + b^2)$ and, consequently, $f_{a,b} \in \mathrm{Lip}_1$ for $a^2 + b^2 \leq 1/2$. □

Notice that for the case of two symmetric elliptical distributions we obtain that the bound in (4.16) coincides (for $a = b = 1/2$) with the sharp bound in (4.14). Sharpness of (4.16) can indeed only occur when transportation rays are parallel in each sector $S_i$; see also the case C illustrated in Figure 6.5 and Table 6.1.

*Uniform distributions on ellipsoids in the plane*

A particular case of elliptical distributions are uniform distributions on ellipsoids. Consider the case $d = 2$ and, for $a, b > 0$, let

$$E_{ab} = \{x \in \mathbb{R}^2;\ \langle x, Dx \rangle \leq 1\}, \text{ with } D = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix},$$

be an ellipsoid with center in the origin. For an angle $\varphi \in [0, \frac{\pi}{2}]$, define the rotated ellipsoids

$$C_1 = \{y = D_{-\varphi}x, x \in E_{ab}\} \text{ and } C_2 = \{y = D_\varphi x, x \in E_{ab}\}.$$

Let $P = \mathcal{U}(C_1)$, $Q = \mathcal{U}(C_2)$ be the uniform distributions on $C_1$ and $C_2$, and define

$$A = C_1 \setminus C_2 \quad \text{and} \quad B = C_2 \setminus C_1.$$

Then the optimal transport kernel $\tau$ in Theorem 4.6 simplifies and is identical to the optimal map

$$\overline{T}(x) = x 1_C(x) + T(x) 1_A(x),$$

where $C = C_1 \cap C_2$ and $T$ is given in (4.11). On the intersection $C$ of the ellipsoids the mass is not moved while on $A$ it is reflected about the corresponding axis; see Figure 4.4.

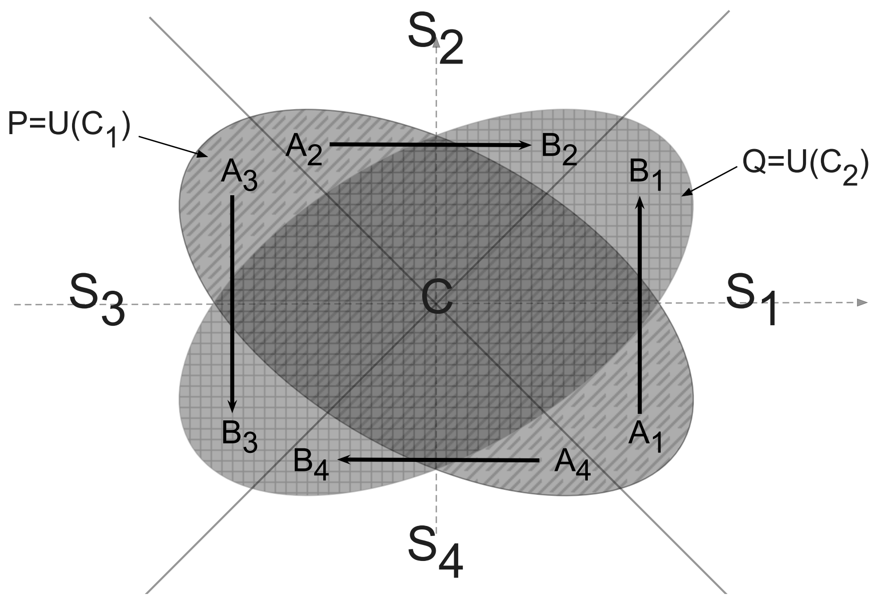


**Figure 4.4** Optimal coupling of two uniform distributions on two rotated ellipsoids. The darker probability mass is not moved.

**Remark 4.9.** In Uckelmann (1998) this case is also dealt with in a different way based on the Lipschitz function

$$f(x) = \sup_{t \in \mathbb{R}} f_t(x), \quad f_t(x) = -\|x - t \cdot \mathbf{1}\| + t,$$

where $\mathbf{1} = (1,1)^t$. By the reduction method we get an explanation on how to choose the directions for the optimal transport.

## 5. General construction method of optimal couplings

In this section, we generalize the reduction and decomposition technique used to construct optimal transports in the examples given in Section 4.

A situation arising in these examples is the following. By the reduction principle, one considers the optimal coupling problem for the reduced measures $P_1$ on $A$ and $Q_1$ on $B$ as defined in Theorem 3.3. Furthermore, for some (possibly infinite) decomposition of $\mathbb{R}^d$ into sectors (cones) $S_i$, by Theorem 3.4 the transport is done on rays within these sectors. In Section 7.5 below, we give some insight on how to find suitable decompositions and transportation directions.

Now, let $R_{t,i}, t \in \Gamma_i$ be a (continuous) family of disjoint rays partitioning the sector $S_i$, $\cup_{t\in\Gamma_i} R_{t,i} = S_i$, and such that

$$R_{t,i} = \{x = x_t + sk_{t,i},\ s \in \mathbb{R}^1_+\} \cap S_i.$$

Each ray $R_{t,i}$ in sector $S_i$ is generated by the point $x_t \in \mathbb{R}^d$ and by the direction $k_{t,i}$, with $\|k_{t,i}\| = 1, t \in \Gamma_i$; see Figure 5.1. In Section 6 we discuss how to obtain a conjecture for the sector directions $k_{t,i}$.

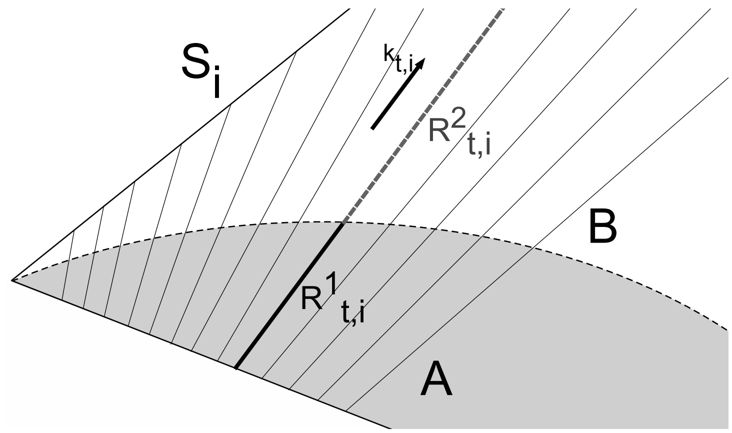


**Figure 5.1** Family of rays $R_{t,i}$ with direction $k_{t,i}$ partitioning the sector $S_i$.

Assume that

$$\{s \in \mathbb{R}^1_+; x_t + sk_{t,i} \in S_i\} = [s_0, s_1],\ \text{with } s_0 = s_0(t,i), s_1 = s_1(t,i) \in \mathbb{R}^*_+, t \in \Gamma_i.$$

Let

$$R^1_{t,i} = R_{t,i} \cap A \text{ and } R^2_{t,i} = R_{t,i} \cap B$$

denote the intersections of $R_{t,i}$ with $A$ and $B$, i.e. $R_{t,i} = R^1_{t,i} + R^2_{t,i}$, w.r.t $\lambda^d$. For each sector $S_i$, define Lipschitz functions $f_i$ on $R_{t,i}$ by

$$f_i(x_t + sk_{t,i}) = s. \tag{5.1}$$

and merge all $f_i$'s together such that for two neighboring sectors $S_i, S_j$ and $x \in S_i \cap S_j$, $f_i(x) = f_j(x)$ are continuous at the boundaries, and satisfy

$$|f_i(x) - f_j(y)| = O(\|x - y\|).$$

Notice that, by (3.7), the gradient of $f_i$ is the assigned direction $k_{t,i}$. In order to have that $f_i$ are $\mathrm{Lip}_1$ it is sufficient that for any $t_1, t_2 \in \Gamma_i$ the vector $x_{t_1} - x_{t_2}$ is orthogonal to the direction $k_{t,i}$ in $S_i$. For instance, in the example of the radial transformation $x_t = t$, for $t \in B^d$, $S_t = t\mathbb{R}^1_+$ and $k_{t,i} = t$. Also for possibly infinitely many sectors, one has to construct $f_\varepsilon$ continuously fitted together. Defining

$$f(x) = f_i(x),\ \text{for } x \in S_i,$$

then $f \in \mathrm{Lip}_1$.

This way we partition the sector $S_i$ with given directions $k_{t,i}$ into disjoint lines $R_{t,i}$. To construct an optimal coupling between the reduced measures on the rays $R_{t,i} = R^1_{t,i} + R^2_{t,i}$ we define for fixed $i$ the reduced conditional distributions $F_{t,i} \sim P_1(\cdot \mid R^1_{t,i})$ and $G_{t,i} \sim Q_1(\cdot \mid R^2_{t,i})$. In order to construct an optimal coupling, it is necessary that the choice of the directions in the corresponding sector $S_i$ is such that

$$P_1(R^1_{t,i} \mid R_{t,i}) = Q_1(R^2_{t,i} \mid R_{t,i}), \tag{5.2}$$

$t \in \Gamma_i$. Then, the optimal coupling on $R_{t,i}$ between the conditional distributions $F_{t,i}, G_{t,i}$ is given by the quantile coupling (see Section 4.1)

$$(\overline{F}^{-1}_{t,i}(U), \overline{G}^{-1}_{t,i}(U)), \tag{5.3}$$

where $\overline{F}_{t,i} = (F_{t,i}(\cdot|R^i_t))^{H_{t,i}}$, with $H_{t,i}(x_t + s\,k_{t,i}) = s$, and similarly $\overline{G}_{t,i}$. Denote by $\gamma_{t,i}$ the distribution of the coupling in (5.3). Note that in general this coupling is not given by a map, but by a plan. As consequence we obtain that the measure induced by fitting these couplings together is an optimal coupling of $P_1, Q_1$.

Defining a random variable $Z$ independent of $U \sim \mathcal{U}([0,1])$ with $P(Z = t) = 2P_1(R^1_{t,i} \mid R_{t,i})$ for $x_t \in S_i$ we obtain that an optimal coupling of $P_1, Q_1$ is given by

$$\big(\overline{F}^{-1}_Z(U), \overline{G}^{-1}_Z(U)\big), \tag{5.4}$$

where $\overline{F}_t = \overline{F}_{t,i}, \overline{G}_t = \overline{G}_{t,i}$ on $S_i$. The induced optimal coupling for $P, Q$ then is given by the corresponding optimal transport kernel as in (3.6).

**Theorem 5.1.** *If the coupling problem $\ell_1(P,Q)$ can be split into rays $R_{t,i}$ with directions $k_{t,i}$ in sectors $S_i$ such that the function $f(x) = f_i(x)$, $x \in S_i$, defined as in (5.1), is in $\mathrm{Lip}_1$, then an optimal coupling of the reduced measures $P_1, Q_1$ is given by the quantile coupling $(\overline{F}^{-1}_Z(U), \overline{G}^{-1}_Z(U))$ as in (5.4). An optimal transport kernel of $P$ and $Q$, for $x \in R_{t,i}$, is given by*

$$\tau(x,\cdot) = (g_1 \wedge g_2(x))\,\varepsilon_x + 1_A(x)(g_1(x) - g_1 \wedge g_2(x))\,\gamma_{t,i},$$

*where $\gamma_{t,i}$ is the distribution of the coupling in (5.3).*

Theorem 5.1 states that an optimal coupling is obtained by letting as much mass stay as it can and transporting the remaining mass by quantile transportation on the rays within the sectors. To apply this technique one needs to know the decomposition into sectors $S_i$ and, most importantly, the optimal directions within each sector. These are non-trivial tasks in general. We give solutions in some particular instances in Section 4 (see Figures 4.3-4.4). In the next section we provide solutions for general Gaussian distributions on the plane.

## 6. Coupling of bivariate Gaussian distributions

In this section we determine an $\ell_1$-optimal coupling between two bivariate Gaussian distributions $P = N(0,\Sigma_1)$ and $Q = N(0,\Sigma_2)$. The solution for the corresponding problem for the $\ell_2$-metric is known since long time; see Olkin and Pukelsheim (1982) and Dowson and Landau (1982). By our method we also provide an accurate numerical estimate of the value of $\ell_1(P,Q)$.

We distinguish two main classes of construction of optimal couplings for Gaussian distributions depending on the domains $A$ and $B$ of the reduced measures as defined in (3.5). Figure 6.5 gives some illustrations of the cases a) and b) in the following Proposition (which holds in generality for Gaussian distributions on $\mathbb{R}^d$).

**Proposition 6.1.** *Let $P = N(0,\Sigma_1)$ and $Q = N(0,\Sigma_2)$ on $\mathbb{R}^d$ have invertible covariance matrices $\Sigma_1,\Sigma_2$ and densities $g_1, g_2$ w.r.t $\lambda^2$. Then:*

*a) Both sets $A = \{g_1 > g_2\}$ and $B = \{g_2 > g_1\}$ are unbounded if and only if*

$$\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right) \text{ is neither positive nor negative definite.} \tag{6.1}$$

*b) One of the sets $A$ or $B$ is bounded if and only if*

$$\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right) \text{ is either positive or negative definite.} \tag{6.2}$$

***Proof.*** Note that $x = (x_1, x_2) \in A$ if and only if

$$\frac{\exp\left(-1/2\, x'\Sigma_1^{-1}x\right)}{S_1} > \frac{\exp\left(-1/2\, x'\Sigma_2^{-1}x\right)}{S_2},$$

where $S_1 = |\Sigma_1|^{1/2}, S_2 = |\Sigma_2|^{1/2} > 0$. Taking the logarithms in the above inequality and simplifying, we have that

$$A = \{g_1 > g_2\} = \left\{x \in \mathbb{R}^2;\, x'\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right)x > 2\log\frac{S_1}{S_2}\right\}.$$

Noting that $C = (\Sigma_2^{-1} - \Sigma_1^{-1})$ is symmetric, the set $A$ represents a strict superlevel set of the quadratic form $Q(x) = x'Cx$. If $C$ is positive definite, then $Q$ is strictly convex and hence $A$ is unbounded. Similarly, if $C$ is negative definite, then $Q$ is strictly concave and the set $B$ is unbounded. If $C$ is neither positive or negative definite, then both sets are unbounded. □

## 6.1. Choice of sectors

In order to construct optimal couplings for bivariate Gaussian measures we divide the plane into the four sectors $S_1,\ldots,S_4$, already defined in (4.10). In fact, for Gaussian measures it is always possible to reduce to the case in which the domains $A$ and $B$ of the reduced measures are symmetric with respect to these sectors.

**Proposition 6.2.** *Let $P = N(0,\Sigma_1)$ and $Q = N(0,\Sigma_2)$ on $\mathbb{R}^2$ have invertible covariance matrices*

$$\Sigma_1 = \begin{pmatrix} \sigma_{11}^P & \sigma_{12}^P \\ \sigma_{12}^P & \sigma_{22}^P \end{pmatrix} \text{ and } \Sigma_2 = \begin{pmatrix} \sigma_{11}^Q & \sigma_{12}^Q \\ \sigma_{12}^Q & \sigma_{22}^Q \end{pmatrix},$$

*and densities $g_1, g_2$ w.r.t $\lambda^2$. Then both sets $A = \{g_1 > g_2\}$ and $B = \{g_2 > g_1\}$ are symmetric with respect to the lines $\{x_2 = x_1\}$ and $\{x_2 = -x_1\}$ if and only if*

$$V(\Sigma_1,\Sigma_2) = \frac{(\sigma_{22}^P - \sigma_{11}^P)}{|\Sigma_1|} - \frac{(\sigma_{22}^Q - \sigma_{11}^Q)}{|\Sigma_2|} = 0. \tag{6.3}$$

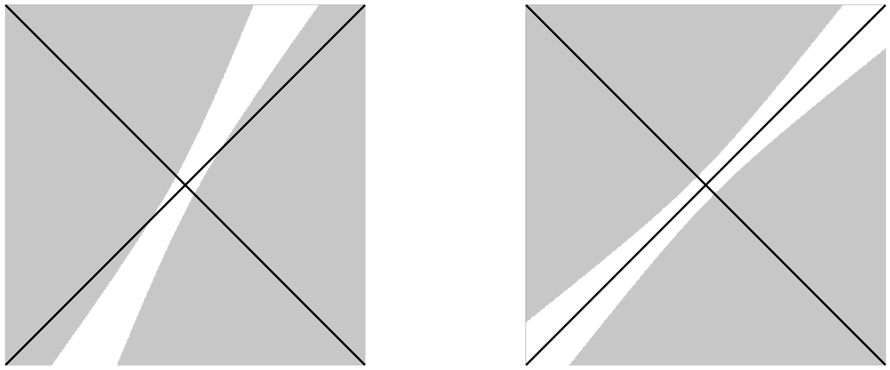

**Figure 6.1** The domains of two reduced Gaussian densities can be rotated (without changing their $L^1$-distance) in order to fulfill the symmetry condition (6.3).

***Proof.*** Let $Q(x) = x'Cx$ be the quadratic form defined by the symmetric matrix $C = (\Sigma_2^{-1} - \Sigma_1^{-1})$. $A$ and $B$ have the required symmetry iff $Q$ is symmetric. This is achieved iff $c_{11} = c_{22}$. Elementary calculations give condition (6.3). □

When (6.3) is not satisfied, one can rotate both measures $P$ and $Q$ so that the symmetry condition (6.3) is satisfied and the $\ell_1$-distance between the two distributions is left unchanged; see Figure 6.1.

Recall that $D_\varphi = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$ denotes the rotation matrix by the angle $\varphi \in [0, 2\pi]$. Cumbersome but elementary calculations prove the following proposition.

**Proposition 6.3.** *Let $P = N(0, \Sigma_1)$ and $Q = N(0, \Sigma_2)$ on $\mathbb{R}^2$ have invertible covariance matrices $\Sigma_1$ and $\Sigma_2$. Let*

$$\Sigma_1' = D_\varphi \Sigma_1 D_\varphi' \ \text{ and } \ \Sigma_2' = D_\varphi \Sigma_2 D_\varphi'$$

*be the new covariance matrices obtained after a rotation of angle*

$$\varphi = \frac{1}{2} \arctan \frac{V(\Sigma_1, \Sigma_2)}{2\left(\frac{\sigma_{12}^Q}{|\Sigma_2|} - \frac{\sigma_{12}^P}{|\Sigma_1|}\right)}. \tag{6.4}$$

*Then, we have that:*

*a)* $\ell_1(N(0, \Sigma_1), N(0, \Sigma_2)) = \ell_1(N(0, \Sigma_1'), N(0, \Sigma_2'))$;
*b)* $V(\Sigma_1', \Sigma_2') = 0$.

As a consequence of Proposition 6.3, in the remainder of this section we assume that condition (6.3) on the covariance matrices of the two Gaussian distributions $P, Q$ is satisfied, so that the sets $A$ and $B$, i.e. the domains of the reduced measures $P_1, Q_1$, are symmetric with respect to the four sectors $S_i, 1 \le i \le 4$, defined in (4.10).

## 6.2. Coupling of Gaussian measures, unbounded case

We now construct an optimal coupling between two reduced Gaussian measures with unbounded domains, that is for covariance matrices satisfying condition (6.1); see Figure 6.2 for an example. The other case with one bounded domain is analogous and treated below. Let $g_1$ and $g_2$ denote the densities w.r.t. $\lambda^2$ of $P = N(0, \Sigma_1)$ and $Q = N(0, \Sigma_2)$.

The idea following the general framework given in Section 5 is to move probability mass between the reduced measures $P_1$ and $Q_1$ within each sector $S_i$ along a given family of directions $k_{t,i}, t \in \Gamma_i$, and according to the quantile coupling for the conditional distributions as in Theorem 5.1. The difficulty, as compared for instance to the symmetric case illustrated in Section 4.4, is that, in the general Gaussian case, the directions are not parallel to the axis nor constant within each sector. We illustrate a methodology to find optimal directions for sector $S_1$, the procedure being symmetric for the other sectors. To ease the notation, in what follows we drop the indication of the sector for the rays and the directions, i.e. we write $R_t = R_{t,i}$ and $k(t) = k_{t,i}$.

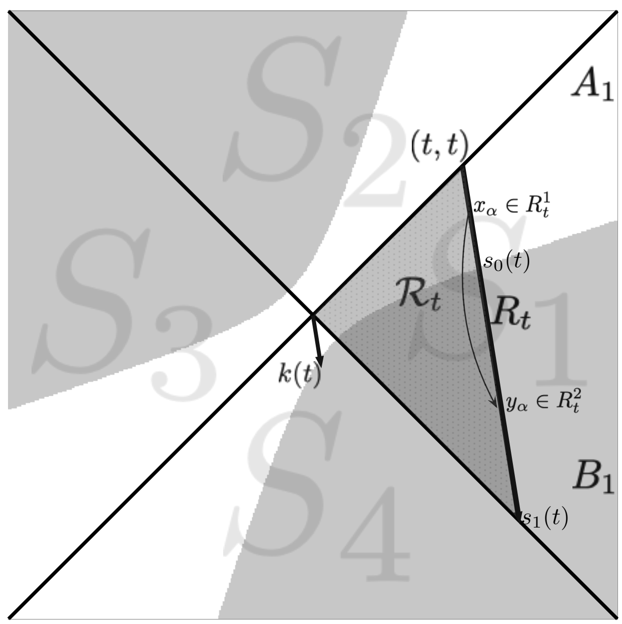


**Figure 6.2** Transport of probability mass along ray $R_t$ between the reduced measures in sector $S_1$ when the domains $A$ (in white) and $B$ (in grey) of the reduced measures $P_1$ and $Q_1$ are both unbounded.

In order to find a candidate for a family of optimal directions, for each point $(t,t)$ in $\Gamma = \{x = (x_1, x_2) \in \mathbb{R}^2 : x_2 = x_1, x_1 > 0\}$, we choose a ray $R_t$ with corresponding direction $k(t) = (k_1(t), k_2(t))$ so that $R_t$ splits the sector $S_1$ into two regions each one having the same measure with respect to both $P_1$ and $Q_1$ (or, equivalently, to both $P$ and $Q$). This also follows the fact that another transport ray, going from the side having more initial mass toward the part side having more final mass, would cross $R_t$ hence violating the $c$-cyclically monotone condition (2.5) of the transport plan.

Formally, we define

$$R_t = \{(x_1, x_2) = (t,t) + s\,(k_1(t), k_2(t)), s \in \mathbb{R}^1_+\} \cap S_1.$$

For a fixed $t \geq 0$, the direction $k(t) = (k_1(t), k_2(t))$ satisfies

$$\int_{\mathcal{R}_t} g_1(x)dx = \int_{\mathcal{R}_t} g_2(x)dx, \tag{6.5}$$

where we denote by $\mathcal{R}_t$ the region of $S_1$ to the left of the ray $R_t$; see Figure 6.2.

In order to construct an optimal coupling via a quantile coupling on each ray, it is necessary that the above defined directions $k(t)$ satisfy condition (5.2). For $x \in S_1$, by a change of coordinates

$$x = (t,t) + s\,(k_1(t), k_2(t)),$$

equation (5.2) reads as

$$\overline{g}_1(t) - \overline{g}_2(t) = 0,\; t \geq 0, \tag{6.6}$$

where we define the functionals

$$\begin{aligned} \overline{g}_1(t) &= \int_0^{s_0(t)} (g_1 - g_2)_+(t + k_1(t)\,v, t + k_2(t)\,v) \cdot \left|\det \begin{pmatrix} 1 + vk_1'(t) & k_1(t) \\ 1 + vk_2'(t) & k_2(t) \end{pmatrix}\right| dv, \\ \overline{g}_2(t) &= \int_{s_0(t)}^{s_1(t)} (g_2 - g_1)_+(t + k_1(t)\,v, t + k_2(t)\,v) \cdot \left|\det \begin{pmatrix} 1 + vk_1'(t) & k_1(t) \\ 1 + vk_2'(t) & k_2(t) \end{pmatrix}\right| dv, \end{aligned} \tag{6.7}$$

with $s_1(t) = -\frac{2t}{k_1(t)+k_2(t)}$. Recall that $R_t = R_t^1 + R_t^2$ denotes the intersection of $R_t$ with $A$ and, resp., $B$, with

$$\begin{aligned} &\{s \in \mathbb{R}^1_+ : (t,t) + s\,(k_1(t), k_2(t)) \in A_1\} = [0, s_0(t)), \\ &\{s \in \mathbb{R}^1_+ : (t,t) + s\,(k_1(t), k_2(t)) \in B_1\} = (s_0(t), s_1(t)]. \end{aligned}$$

Moreover, on the set $A_1$ we define the quantile map $T$ so that, on ray $R_t$, it moves point

$$x_\alpha = (t + k(t)\underline{s}_\alpha) \in R_t^1 \quad \text{to} \quad y_\alpha = (t + k(t)\overline{s}_\alpha) \in R_t^2,$$

i.e.

$$T(x_\alpha) = y_\alpha, \tag{6.8}$$

where $\underline{s}_\alpha$ and $\overline{s}_\alpha$ satisfy the quantile coupling equation

$$\begin{aligned} &\frac{\int_0^{\underline{s}_\alpha} (g_1 - g_2)_+(t + k_1(t)\,v, t + k_2(t)\,v) \cdot \left|\det \begin{pmatrix} 1 + vk_1'(t) & k_1(t) \\ 1 + vk_2'(t) & k_2(t) \end{pmatrix}\right| dv}{\overline{g}_1(t)} \\ &\qquad = \frac{\int_{s_0(t)}^{\overline{s}_\alpha} (g_2 - g_1)_+(t + k_1(t)\,v, t + k_2(t)\,v) \cdot \left|\det \begin{pmatrix} 1 + vk_1'(t) & k_1(t) \\ 1 + vk_2'(t) & k_2(t) \end{pmatrix}\right| dv}{\overline{g}_2(t)} = \alpha, \end{aligned} \tag{6.9}$$

for all $\alpha \in [0,1]$. Note that by (6.6) one has $\overline{g}_1(t) = \overline{g}_2(t)$ and, for fixed $\alpha \in [0,1]$, the cost of transporting $x_\alpha$ to $y_\alpha$ is $(\overline{s}_\alpha - \underline{s}_\alpha)$.

Having defined the map $T(x)$ for $x = (x_1, x_2) \in A_1$, and building on the symmetry condition (6.3), in all the other sectors $S_i$, the map $T$ can be defined on $\mathbb{R}^2$ as

$$T(x) = \begin{cases} (T_2(x_2,x_1), T_1(x_2,x_1)), & \text{for } (x_1,x_2) \in A_2, \\ (-T_1(-x_1,-x_2), -T_2(-x_1,-x_2)), & \text{for } (x_1,x_2) \in A_3, \\ (-T_2(-x_2,-x_1), -T_1(-x_2,-x_1)), & \text{for } (x_1,x_2) \in A_4. \end{cases} \tag{6.10}$$

If (6.6) holds, by symmetry with respect to the sectors we have that $P_1^T = Q_1$ and, by Theorem 3.4, also that

$$T(x) \in \partial_1 f(x), x \in A,$$

where, for $(x_1,x_2) = (t,t) + s(k_1(t), k_2(t))$, the function $f : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$f(x_1,x_2) = \left(\frac{k_1(t)+k_2(t)}{2}\right)|x_1+x_2| - \left(\frac{k_2(t)-k_1(t)}{2}\right)|x_1-x_2|.$$

In fact, it is easy to check that $\|\nabla f(x)\| = \|k(t)\| = 1$, and therefore mass transportation according to $T$ occurs on rays in the direction of the gradient of $f \in \mathrm{Lip}_1$. As a consequence, $T$ is an $L^1$-optimal transport between $P_1$ and $Q_1$. By Theorem 3.3, the total cost of the (symmetric) transportation in all the sectors $S_i$ is given by

$$\ell_1(P,Q) = 4\int_0^\infty \int_0^{s_0(t)} (\overline{s}_\alpha - \underline{s}_\alpha)\cdot(g_1-g_2)_+(t+k_1(t)\underline{s}_\alpha, u+k_2(t)\underline{s}_\alpha) \cdot \left|\det\begin{pmatrix} 1+\underline{s}_\alpha k_1'(t) & k_1(t) \\ 1+\underline{s}_\alpha k_2'(t) & k_2(t)\end{pmatrix}\right| d\underline{s}_\alpha\, dt, \tag{6.11}$$

where $\overline{s}_\alpha$ is determined via (6.9).

*Numerical estimate of $\ell_1$*

In practice, it seems cumbersome to work with equation (6.6) as it includes via (6.7) the derivatives $k_i'(t)$ of the directions, which vary with $t$ and are chosen based on equation (6.5). To circumvent this problem, we discretize the problem and construct an approximate transport on a finite family of rays splitting the sector $S_1$ in a number of slices each one having the same probability with respect to $P_1$ and $Q_1$; see Figure 6.3.

Thus, we choose a positive number $M > 0$ and we discretize the interval $[0,M]$ into $N+1$ points $0 \le t_1, \ldots, t_{N+1} \le M \in \Gamma$ with $\Delta_N^t = \sup_{1\le n\le N}\{t_{n+1} - t_n\}$. For each $n \ge 1$, we compute the direction $k(t_n)$ by solving numerically (6.5) for $t = t_n$. At this point we substitute the two derivatives $k_i'(t)$ in (6.7) and in (6.9) by their discrete variations $\Delta_i = \frac{k_i(t_{n+1}) - k_i(t_n)}{t_{n+1}-t_n}$, $1 \le n \le N$, and define the quantile coupling transportation on the rays $R_{t_n}, 0 \le n \le N$, by means of (6.9).

To compute an approximate estimate of $\ell_1(P,Q)$, for a fixed $n \ge 1$ we also discretize the interval $[0, s_0(t_n)]$ into $N$ points $0 \le v_1, \ldots, v_N \le s_0(t_n)$ with $\Delta_N^v = \sup_{1\le n\le N}\{v_n - v_{n-1}\}$. For each point

$$x_{i,j} = t_i + k(t_i)v_j, \quad 1 \le i,j \le N,$$

we then compute the cost of moving the mass according to equation (6.9).

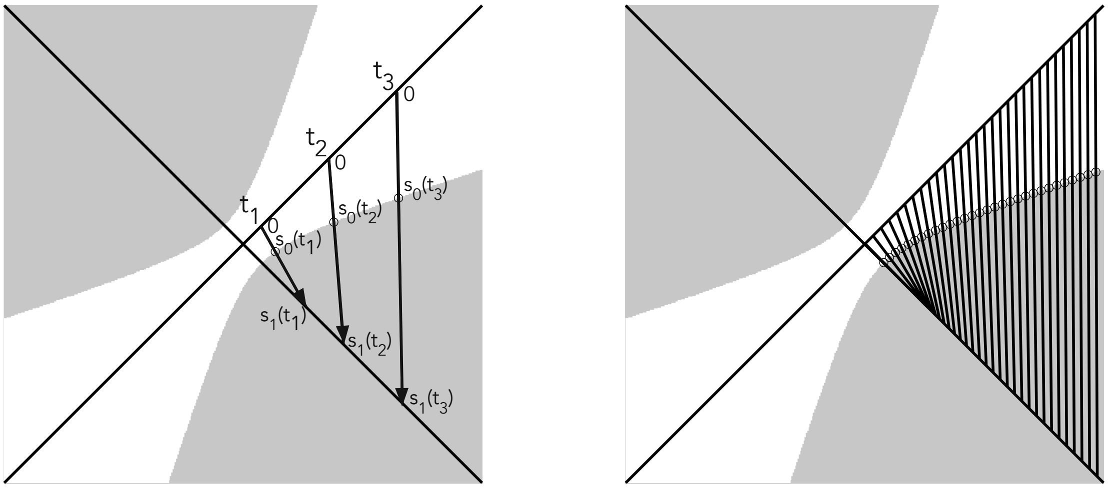


**Figure 6.3** Construction of an approximate optimal coupling on a finite family of rays in the unbounded case. Each slice has the same probability with respect to $P_1$ and $Q_1$

At this point, we perform a trapezoidal integration via MatLab to estimate numerically (6.11). We expect that, for $M$ sufficiently large and $\Delta_N^{t,v} \to 0$, the difference in (6.6) goes to zero, and thus the discretization procedure will result into an estimate of the transportation cost $\ell_1(P,Q)$ in the continuous case.

Numerical estimates obtained in MatLab are given in Table 6.1 and directions of optimal transport are shown in Figure 6.5 for different choices of covariance matrices. In Table 6.1 we also report the numerical estimates obtained via the swapping algorithm in Puccetti (2017) and via the lower dual bounds in (3.3) and (4.16). The swapping algorithm is a relatively simple numerical procedure which typically produces an accurate upper bound on the $L^p$-Wasserstein distance between two measures. The so obtained numerical estimates are consistent with optimality of our construction of transports between the Gaussian distributions.

## 6.3. Coupling of Gaussian measures, bounded case

When the covariances of the two Gaussian measures satisfy (6.2) and (6.3), the domain of one of the two reduced measures is an ellipse symmetric with respect to the sectors $S_i$; see Figure 6.4 for an example.

The construction of an optimal coupling is similar to the one illustrated above with one fundamental difference in the geometry of transportation with respect to the unbounded case: the transportation rays start from the subset $\Gamma_M \subset \Gamma$ where

$$\Gamma_M = \{(x_1, x_2) \in \mathbb{R}^2; x_2 = x_1, 0 < x_1 < M\}.$$

The value $M$ is the sup of all positive values $t$ for which it is possible to define a ray $R_t$ and corresponding direction $k(t)$ so that (6.5) is satisfied. Thus the value of $M$ here is not chosen arbitrarily but is univocally determined based on $P_1, Q_1$. We have that

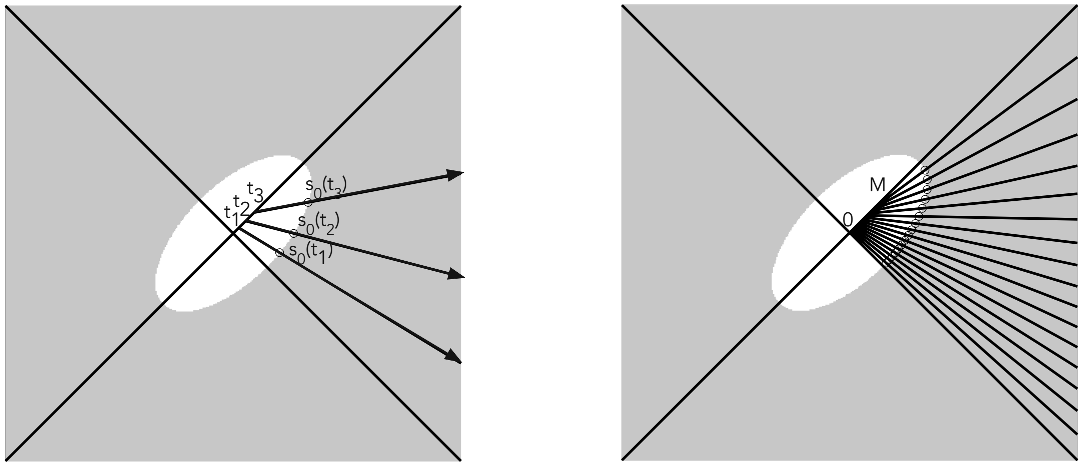


**Figure 6.4** Construction of an approximate optimal coupling on a finite family of rays in the bounded case. Each slice has the same probability with respect to $P_1$ and $Q_1$.

$$\ell_1(P,Q) = 4\int_0^M \int_0^{s_0(t)} (\overline{s}_\alpha - \underline{s}_\alpha)\cdot(g_1-g_2)_+(t+k_1(t)\underline{s}_\alpha, u+k_2(t)\underline{s}_\alpha) \cdot \left|\det\begin{pmatrix} 1+\underline{s}_\alpha k_1'(t) & k_1(t) \\ 1+\underline{s}_\alpha k_2'(t) & k_2(t)\end{pmatrix}\right| d\underline{s}_\alpha\, dt, \quad (6.12)$$

Estimates of $\ell_1(P_1,Q_1)$ in this case are obtained similarly as in the unbounded case by splitting the sector in a finite number of slices each one having the same probability with respect to the reduced measures (see Figure 6.4). They are reported in Table 6.1.

## 6.4. Coupling of Gaussian measures, summary of results

The following theorem summarizes our main result about the $L^1$-optimal coupling of two bivariate Gaussian measures.

**Theorem 6.4.** *Assume that $P = N(0,\Sigma_1)$ and $Q = N(0,\Sigma_2)$ are two bivariate Gaussian measures with invertible covariance matrices satisfying condition* (6.3) *(one can reduce to this case by a rotation as described in Proposition 6.3). Let $P_1$ and $Q_1$ be the corresponding reduced measures as in* (3.5)*. Then:*

*(a) An $L^1$-optimal transport map from $P_1$ to $Q_1$ is given by the function $T$ defined in* (6.8) *and* (6.10).

*(b) An $L^1$-optimal transport kernel of $P$ and $Q$ is given by*

$$\tau(x,\cdot) = (g_1\wedge g_2(x))\,\varepsilon_x + 1_A(x)(g_1(x) - g_1\wedge g_2(x))\,\varepsilon_{T(x)};$$

*(c) If $(\Sigma_2^{-1} - \Sigma_1^{-1})$ is neither positive nor negative definite (unbounded reduced domains), an analytical formula for $\ell_1(P,Q)$ is given in* (6.11)*;*

*(d) If $(\Sigma_2^{-1} - \Sigma_1^{-1})$ is either positive or negative definite (one bounded reduced domain), an analytical formula for $\ell_1(P,Q)$ is given in (6.12).*

The following example shows the typical change of optimal transport maps depending on the geometry of the problem induced by the covariance matrices.

**Example 6.5.** Consider the class of cases for the covariance matrices as given in Table 6.1. The corresponding optimal transport directions are given in Figure 6.5, and illustrate the difference in the geometry of the unbounded and bounded cases.

In Figure 6.5 one can see the transition between the symmetric case A treated in Section 4.4, the general unbounded case (corresponding to $M = \infty$ and $\Gamma_\infty = \Gamma$), the bounded case ($0 < M < \infty$) and the radial case F treated in Section 4.3 ($M = 0$). The case C corresponds to the situation when one of the eigenvalues of $\Sigma_2^{-1} - \Sigma_1^{-1}$ is null.

We notice that in the cases A and C transportation lines are parallel and hence the dual bound given in (4.16) is sharp; in case F the lower dual bound in (3.3) is sharp. In all these cases, the $L^1$-distance between Gaussian distributions can be computed analytically.

| case | $\Sigma_1$ | $\Sigma_2$ | dual bound | optimal coupling | swapping alg. |
|---|---|---|---|---|---|
| A | $\begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$ | $\begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}$ | 0.4611 (4.15) | 0.4611 (4.15) | 0.4625 |
| B | $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$ | $\begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}$ | 0.7413 (4.16) | 0.7482 (6.11) | 0.7498 |
| C | $\begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$ | $\begin{pmatrix} 1.8 & -0.4 \\ -0.4 & 1.8 \end{pmatrix}$ | 0.5654 (4.16) | 0.5654 (6.11) | 0.6131 |
| D | $\begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$ | $\begin{pmatrix} 2 & -0.4 \\ -0.4 & 2 \end{pmatrix}$ | 0.6215 (4.16) | 0.6226 (6.12) | 0.6715 |
| E | $\begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$ | $\begin{pmatrix} 1.3 & 0.4 \\ 0.4 & 1.3 \end{pmatrix}$ | 0.1799 (4.6) | 0.1844 (6.12) | 0.1907 |
| F | $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ | $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ | 0.5191 (4.6) | 0.5191 (4.7) | 0.5192 |

**Table 6.1.** Figures for $\ell_1(N(0,\Sigma_1), N(0,\Sigma_2))$ computed via the dual bounds and the optimal coupling constructions given in this paper, and via the swapping algorithm. Estimates for the swapping algorithm are averages evaluated over 50 identical runs with a discretization of $10^5$ points and fixed accuracy $\xi = 10^{-5}$. Estimates for the optimal couplings are analytical for cases A and F, are computed numerically using $N = 500$ (cases B, C, D, E) and $M = 4$ (for cases B, C). For each dual bound and optimal coupling estimate we indicate the corresponding reference formula.

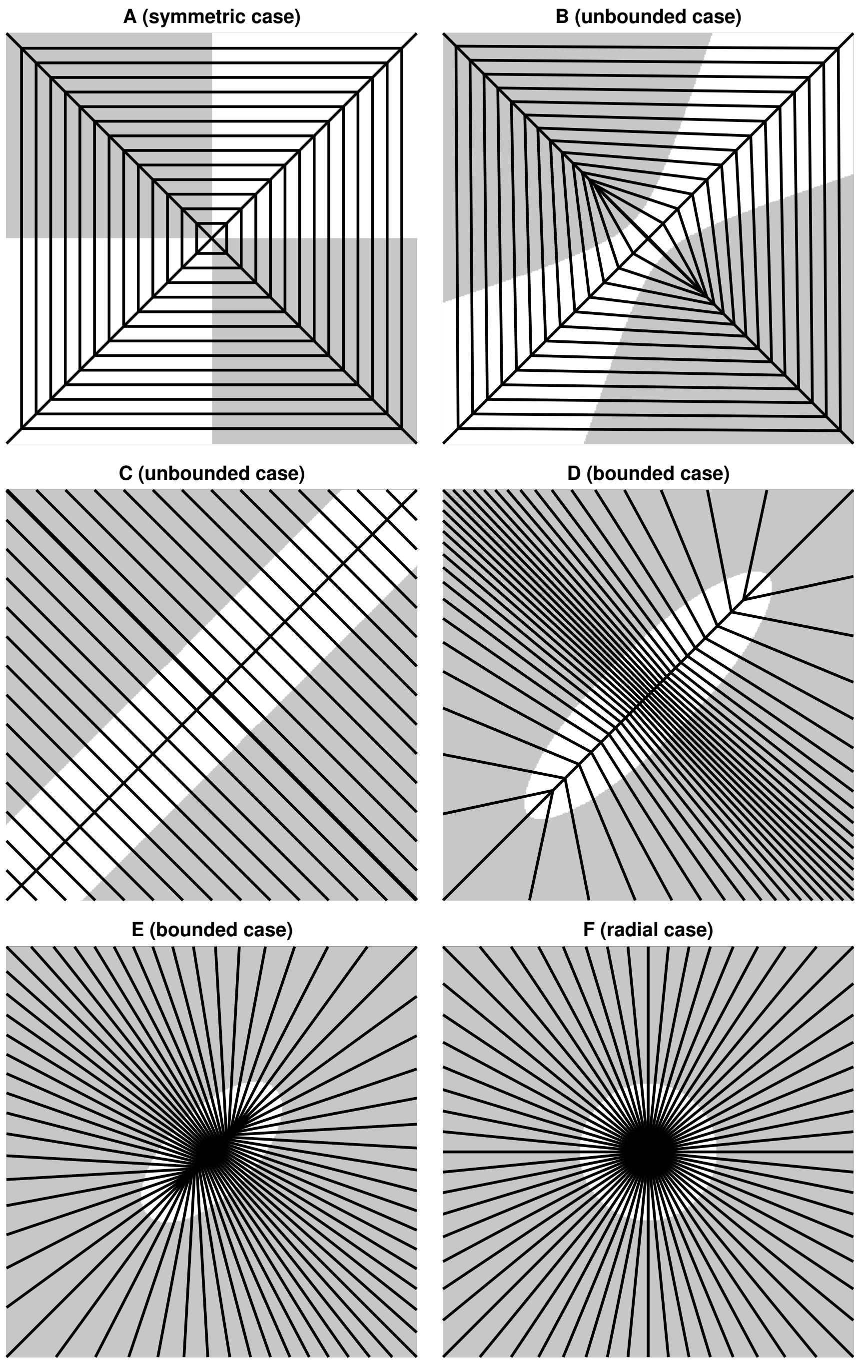


**Figure 6.5** Directions of optimal transport of probability mass between couples of Gaussian measures. Case A reduces to the symmetric transport as in Section 4.4. Case E reduces to the radial transport as in Section 4.3. The various cases correspond to the couples of covariance matrices given in Table (6.1).

# 7. Coupling by scalings

The geometry of $L^1$-optimal couplings is non-trivial. In Alfonsi and Jourdain (2014) it is shown that coupling by increasing transformations of the components is in general not $\ell_1$-optimal, i.e., if $X \sim P$ and $Y = (Y_i)$, $Y_i = h_i(X_i)$, $h_i$ increasing, then $(X,Y)$ is not an $L^1$-optimal coupling in general.

In this section, we strengthen the result of Alfonsi and Jourdain by showing that even the special case of increasing linear scalings does not produce optimal couplings. In particular, we show that, in general, scaling in one component is optimal whereas scaling in two components is not.

## 7.1. Scaling of one component

We show that scaling in one component is optimal.

**Theorem 7.1.** *Let $X \sim P$ and $c > 0$.*

*a) For $X_c = \begin{pmatrix} cX_1 \\ \vdots \\ X_n \end{pmatrix} \sim Q$, the pair $(X, X_c)$ is an $L^1$-optimal coupling of $P$ and $Q$.*

*b) For $\tilde{X}_c = c \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim \tilde{Q}$, the pair $(X, \tilde{X}_c)$ is an $L^1$-optimal coupling of $P$ and $\tilde{Q}$.*

***Proof.*** We give two proofs of a). For all $X \sim P$, $Y \sim Q$, it holds that

$$E\|X-Y\| \geq E|X_1 - Y_1| \geq E|X_1 - cX_1| = E\|X - X_c\|,$$

where the second inequality is a consequence of one-dimensional optimal coupling. The second proof is based on $\ell_1$-convexity. Let $f(x) = |x_1|$, then $f \in \mathrm{Lip}_1$ and

$$\nabla f(x) = \begin{cases} (1,0), & \text{if } x_1 > 0; \\ (-1,0), & \text{if } x_1 > 0. \end{cases}$$

It follows that

$$|f(x) - f(y)| = ||x_1| - |y_1|| \leq |x_1 - y_1| \leq \|x - y\|$$

with equality if and only if $x_2 = y_2$ and $x_1, y_1$ have the same sign. Thus $y \in \partial_1 f(x) \Leftrightarrow y = x + s\nabla f(x)$ for some $s \geq 0$. This implies by Theorem 3.4, c), that $T_c(x) = \begin{pmatrix} cx_1 \\ x_2 \end{pmatrix}$ is an optimal transport. b) follows from the optimality of radial transformations in Theorem 4.3. □

**Remark 7.2.** The second proof of a) can be generalized to show the optimality of $(X, X_c^m)$ where

$$X_c^m = (cX_1, \ldots, cX_m, X_{m+1}, \ldots, X_n),$$

for $1 < m < n$. Also notice that the linear scaling of components $h_1(x) = cx$ can be replaced by any increasing function $h_1(x)$. Roughly speaking, applying the same change of scale to any number of components of a random vector is an $L^1$-optimal transport. Applying just two different scalings is no more optimal in general, as we show in the remainder.

## 7.2. Scaling of two components

The scaling of two components is not optimal in general, i.e., the Alfonsi and Jourdain result even holds for simple scalings. To show this, one can consider the following (counter)example.

Let $P = \mathcal{U}([0,1]^2)$ and $T_1(x) = (2x_1, \frac{1}{2}x_2)$ the scaling of the unit square $[0,1]^2$ to the rectangle $[0,2] \times [0,\frac{1}{2}]$. Let $Q = P^T$, then $Q \sim \mathcal{U}([0,2] \times [0,\frac{1}{2}])$ is the uniform distribution on the rectangle. The transportation cost of the scaling map $T_1$ is given by

$$C_1 = \int_0^1 \int_0^1 \|x - T_1(x)\| dx = \int_0^1 \int_0^1 \left\| \binom{x_1}{x_2} - \binom{2x_1}{\frac{1}{2}x_2} \right\| dx_1 dx_2$$
$$= \int_0^1 \int_0^1 \sqrt{x_1^2 + \frac{x_2^2}{4}}\, dx_1\, dx_2 \simeq 0.5932.$$

We notice that this scaling transport map cannot be optimal since it is not $c$-cyclically monotone, e.g. because $(1/2,1),(1,1/2)$ and $(1,1/2),(2,1/4)$ are in the support of the proposed plan but it is strictly more convenient moving $(1/2,1)$ to $(2,1/4)$ and leaving $(1,1/2)$ fixed.

One then considers the alternative transportation given by $T_2(x) = (x_1+1, x_2-\frac{1}{2})'$ defined on $[0,1] \times [\frac{1}{2},1] \longrightarrow [1,2] \times [0,\frac{1}{2}]$. By the reduction principle we should optimally transport $P_1 = \mathcal{U}([0,1] \times [\frac{1}{2},1])$ to $Q_1 = \mathcal{U}([1,2],[0,\frac{1}{2}])$, leaving the mass in $[0,1] \times [0,\frac{1}{2}]$ unaltered. As transportation cost for $T_2$ we get

$$C_2 = \int_0^1 \int_{1/2}^1 \|x - T_2(x)\| dx = \int_0^1 \int_{1/2}^1 \left\| \binom{x_1}{x_2} - \binom{x_1+1}{x_2-\frac{1}{2}} \right\| dx_1 dx_2$$
$$= \int_0^1 \int_{1/2}^1 \sqrt{\frac{5}{4}}\, dx_1 dx_2 = \frac{\sqrt{5}}{4} \simeq 0.5590 < C_1.$$

We show that $T_2$ is an optimal transport from $P_1$ to $Q_1$. Note that

$$T_2(x) = x + s\nabla f(x) \in \partial_1 f(x),$$

with $s = \frac{\sqrt{5}}{2}$ and the $\mathrm{Lip}_1$ function $f$ defined by

$$f(x) = \frac{\sqrt{5}}{5}(2x_1 - x_2).$$

Moreover, for $X \sim P, Y \sim Q$, one computes

$$\int f d(Q-P) = \frac{\sqrt{5}}{5}(2\mathbb{E}[Y_1] - \mathbb{E}[Y_2] - 2\mathbb{E}[X_1] + \mathbb{E}[X_2]) = \frac{\sqrt{5}}{4} = C_2,$$

implying that

$$T(x) = x 1_{[0,1]\times[0,1/2]}(x) + T_2(x) 1_{[0,1]\times[1/2,1]}(x)$$

is an optimal transport from $P$ to $Q$.

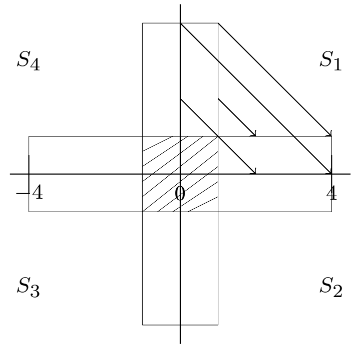


**Figure 7.1** Optimal transport between rotated rectangles

## 7.3. General scalings

The scaling example can be extended to different kind of scalings. For instance, let $P \sim \mathcal{U}([-1,1],[-4,4])$ and $Q \sim \mathcal{U}([-4,4]\times[-1,1])$; see Figure 7.1. By Theorem 3.3 we have to transport $P_1 = \mathcal{U}(A)$, with $A = [-1,1]\times\{[-4,-1]\cup[1,4]\}$ to $Q_1 = \mathcal{U}(B)$, with $B = \{[-4,-1]\cup[1,4]\}\times[-1,-1]$. Again, we divide the plane into the sectors $S_1,\ldots,S_4$, as in (4.10). On sector $S_1$, we define

$$T_1 : A\cap S_1 \to B\cap S_1,\ T(x_1,x_2) = (x_2,x_1).$$

$T_1$ is the reflection about the diagonal of $S_1$.

Similarly, we define $T_i, 2\le i\le 4$, on the other sectors as the corresponding reflections about the diagonals. Then by Theorem 3.3, resp. Theorem 5.1, $T = \sum_{i=1}^4 1_{A\cap S_i} T_i$ is an optimal transportation of $P$ to $Q$ since $T$ is of the form

$$T(x) = \sum_{i=1}^{4} 1_{A\cap S_i}(x + s(x)k_i), \tag{7.1}$$

with directions $k_1 = k\binom{1}{-1}$, $k_2 = k\binom{1}{1}$, $k_3 = k\binom{-1}{1}$, $k_4 = k\binom{-1}{-1}$, with $k = \sqrt{2}/2$.

Within sector $S_i$, $T(x) = T_i(x) \in \partial_1 f_i(x)$ for $\ell_1$-convex functions $f_i, i = 1,\ldots,4$. Defining $f(x) = \sum_{i=1}^4 1_{S_i} f_i(x)$, we find that $f \in \mathrm{Lip}_1$ and $T(x) \in \partial_1 f(x)$.

This principle can be extended to general rectangles $[-a,a]\times[-b,b]$ and $[-c,c]\times[-d,d]$, and also leads to the idea of how to couple two normal vectors in the 3- or general $d$-dimensional case.

## 7.4. Couplings in higher dimensions

The general form of optimal couplings for scalings as in Section 7.3 also allows the optimal coupling of various classes of multivariate normal distributions.

Let $P = N(0,\Sigma_1)$, $Q = N(0,\Sigma_2)$ be $d$-dimensional normal distributions with $\Sigma_1,\Sigma_2$ being simultaneously diagonalizable. A particular case is when $\Sigma_1,\Sigma_2$ are commutable, i.e. $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$. Let $Q$ be the orthogonal matrix of eigenvectors such that $Q\Sigma_1 Q' = D$, $Q\Sigma_2 Q' = S$ with $D = \mathrm{diag}(d_i)$, $S = \mathrm{diag}(s_i)$ being the $d$-dimensional diagonal matrices of corresponding eigenvalues. For $N \sim N(0,I)$, define $X = Q'D^{1/2}N \sim P$ and $Y = Q'S^{1/2}N \sim Q$. The pair $(X,Y)$ is a natural scaling type coupling of $P, Q$, and

$$\mathbb{E}\|X - Y\| = \mathbb{E}\|Q'(D^{1/2} - S^{1/2})N\| = \mathbb{E}\sqrt{\chi^2_{d,\alpha}}, \tag{7.2}$$

with $\alpha_i = \sqrt{d_i} - \sqrt{s_i}, 1 \leq i \leq d$. The coupling in (7.2) will however not be an optimal coupling as one can see from the scaling example in Section 7.2.

Now, in the new coordinate system induced by the the pairwise orthogonal normed eigenvectors $r_1, \dots, r_d$, we get that the coupling problem of $P$ and $Q$ reduces to a scaling problem similar to the one treated in Section 7.3 and can be treated analogously with parallel transportation directions in each one of the sectors. We give a rough sketch of the argument. In the sectors $S_i$ induced by the new coordinate system directions $k_i$ allow to transform the reduced distributions $P_1 = P/_A$ to $Q_1 = Q/_B$ by transformations $T_i$ on $S_i \cap A$ of the form $T_i(x) = x + s(x)k_i$. $T_i$ can be identified by the symmetry of the normal distribution as reflection on the diagonal in the component where between two neighboring sectors a sign change takes place. Identifying the length of the shift $s(x), x \in A_i$ by the quantile equation leads by Theorem 5.1 to an optimal coupling in this case.

This idea suggests to consider $2^3 = 8$ sectors resp. $2^d$ in the $d$-dimensional case corresponding to the various sign combinations. As in the case $d = 2$ one obtains optimal solutions as in (7.1) with direction vectors $k_i$ in the various sectors.

## 7.5. Finding sectors and directions of optimal transport

The construction of optimal couplings following our three-step approach as described in Section 5 is applied in a variety of examples in this paper. This approach requires to have a suitable division of $\mathbb{R}^d$ into sectors and, for each sector, the directions along which the probability mass is to be moved.

An useful hint to the choice of sectors is induced in the first step of our approach by the geometry and symmetry of the reduced transportation problems, as indicated in the various examples in the paper and in particular in the Gaussian application. The non crossing property of optimal transports leads in the second step to a hint for the choice of transportation rays within the sectors. In the third and final step, this hint is verified by the characterization result for optimal couplings.

A further useful tool to confirm or get a hint for the choice of sectors and transport directions is to look at the numerical solution of a discretized version of the problem. In particular in more complicated cases, this helps to support the indications from symmetry and non-crossing property. In some cases it also may lead to a conjecture for the optimal dual solution, hence rejoining the Sudakov framework.

For discrete measures $P = P_N$ and $Q = Q_N$ supported on $N$ points, the computation of (1.1) is obtained through the solution of a finite linear program (LP) that can in principle be solved exactly by a linear solver; see for instance Carlier et al. (2015, Sec. 2.3). Such LP quickly becomes intractable for linear programming software with increasing values of $N$, but a standard laptop can quickly solve it for relatively small $N < 1000$ and provide the solution, as well as the (discretized) dual function solving the corresponding dual (3.1).

In Figure 7.2 we plot optimal transportation lines (left) and the optimal dual function solving problem (1.1) for a discrete version of two Gaussian measures with correlation matrices

$$\Sigma_1 = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}.$$

First, by looking at the transportation lines on the left, one can conjecture the subdivision of the plane into four sectors. Second, by looking at the discretized dual functional, one can conjecture that

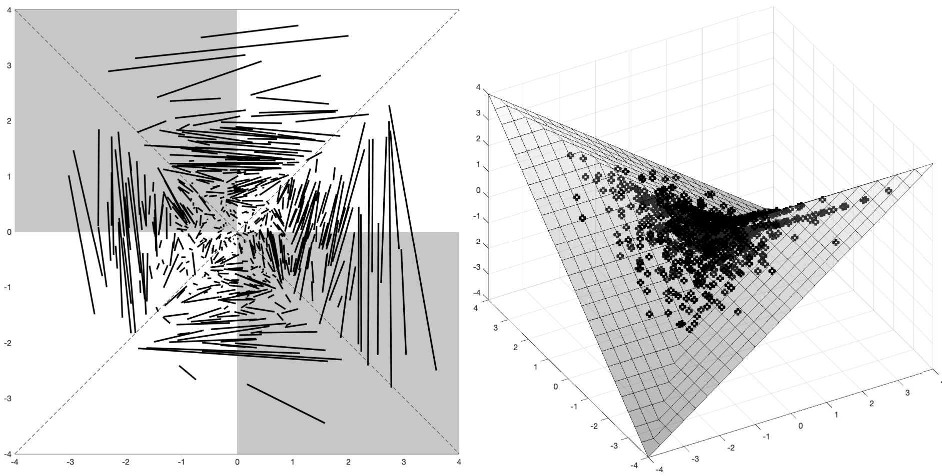

**Figure 7.2** Directions of optimal transport (left) and optimal dual function (right) between discrete versions of the two Gaussian marginals as in case A in Table 6.1. Discrete Gaussian distributions are obtained via $N = 500$ simulations. On the left picture we plot the domains of the reduced densities, whereas on the right the exact dual function (7.3) is superimposed on the discretized dual one.

an optimal dual function $f$ in this case is

$$f(x) = \frac{|x_1 + x_2| - |x_1 - x_2|}{2}. \tag{7.3}$$

The function $f$ also implies (conversely) the subdivision into sectors $S_i$, where it has constant partial derivatives.

Even if finding the solution in the general case of bivariate Gaussian marginals is much more involved, this tool can in principle be applied to study general problems of optimal transport, also with a different metric or general cost function.

*Acknowledgments*

The authors are grateful to two anonymous reviewers for their accurate reading of a previous version of this paper and many insightful suggestions for improvement. The code to reproduce the figures of this paper can be downloaded at `https://github.com/giovannipuccetti/L1optimalcoupling`.

# References

Alfonsi, A. and B. Jourdain (2014). A remark on the optimal transport between two probability measures sharing the same copula. *Stat. Probab. Lett. 84*, 131–134.

Ambrosio, L. (2003). Lecture notes on optimal transport problems. In *Mathematical aspects of evolving interfaces (Funchal, 2000)*, Volume 1812 of *Lecture Notes in Math.*, pp. 1–52. Springer, Berlin.

Ambrosio, L., B. Kirchheim, and A. Pratelli (2004). Existence of optimal transport maps for crystalline norms. *Duke Math. J. 125*(2), 207–241.

Ambrosio, L. and A. Pratelli (2003). Existence and stability results in the $L^1$ theory of optimal transportation. In *Optimal Transportation and Applications*, pp. 123–160. Springer, Berlin.

Bianchini, S. and S. Daneri (2018). *On Sudakov's Type Decomposition of Transference Plans with Norm Costs*. American Mathematical Society, Providence RI.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math. 44*(4), 375–417.

Caffarelli, L. A., M. Feldman, and R. J. McCann (2002). Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs. *J. Am. Math. Soc. 15*(1), 1–26.

Caravenna, L. (2011). A proof of Sudakov theorem with strictly convex norms. *Math. Z. 268*, 371–407.

Carlier, G., A. Oberman, and E. Oudet (2015). Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM Math. Model. Numer. Anal. 49*(6), 1621–1642.

Cuesta-Albertos, J. A., L. Rüschendorf, and A. Tuero-Díaz (1993). Optimal coupling of multivariate distributions and stochastic processes. *J. Multivariate Anal. 46*(2), 335–361.

Dall'Aglio, G. (1956). Sugli estremi dei momenti delle funzioni di ripartizione doppie. *Ann. Scuola Normale Superiore di Pisa, Cl. Sci. 3*(1), 33–74.

Dietrich, H. (1988). Zur c-konvexität und c-subdifferenzierbarkelt von funktionalen. *Optimization 19*(3), 355–371.

Dowson, D. C. and B. V. Landau (1982). The Fréchet distance between multivariate normal distributions. *J. Multivariate Anal. 12*(3), 450–455.

Eckstein, S. and M. Kupper (2021). Computation of optimal transport and related hedging problems via penalization and neural networks. *Appl. Math. Optim. 83*, 639–887.

Evans, L. C. and W. Gangbo (1997). *Differential Equations Methods for the Monge–Kantorovich Mass Transfer problem*. American Mathematical Society, Providence RI.

Fang, K. W. (2017). *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC.

Gangbo, W. and R. J. McCann (1995). Optimal maps in Monge's mass transport problem. *C. R. Acad. Sci. Paris Sér. I Math. 321*(12), 1653–1658.

Gangbo, W. and R. J. McCann (1996). The geometry of optimal transportation. *Acta Math. 177*(2), 113–161.

Kantorovich, L. V. (1942). On the translocation of masses. *C. R. (Dokl.) Acad. Sci. URSS, n. Ser. 37*, 199–201.

Kantorovich, L. V. (1948). On a problem of Monge. *Uspekhi Mat. Nauk 3*, 225–226. In Russian.

Kantorovich, L. V. and G. S. Rubinstein (1957). On a function space in certain extremal problems. *Dokl. Akad. Nauk USSR 115*(6), 1058–1061.

Kotz, S. and S. Nadarajah (2004). *Multivariate T-Distributions and Their Applications*. Cambridge University Press.

Olkin, I. and F. Pukelsheim (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl. 48*, 257–263.

Orlova, D. Y., N. Zimmerman, S. Meehan, C. Meehan, J. Waters, and et al. (2016). Earth mover's distance (EMD): A true metric for comparing biomarker expression levels in cell populations. *PLoS ONE 11*(3), e0151859.

Peyré, G. and M. Cuturi (2018). Computational optimal transport. With applications to data sciences. *Found. Trends Mach. Learn. 11*(5-6), 1–262.

Puccetti, G. (2017). An algorithm to approximate the optimal expected inner product of two vectors with given marginals. *J. Math. Anal. Appl. 451*, 132–145.

Rachev, S. T. and L. Rüschendorf (1998). *Mass Transportation Problems. Vol. I-II*. Springer-Verlag, New York.

Rubner, Y., C. Tomasi, and L. J. Guibas (2000). The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision 40*(2), 99–121.

Rüschendorf, L. (1991). Fréchet-bounds and their applications. In *Advances in Probability Distributions with Given Marginals*, pp. 151–187. Dordrecht: Kluwer Acad. Publ.

Rüschendorf, L. (1995). Optimal solutions of multivariate coupling problems. *Appl. Math. 23*(3), 325–338.

Rüschendorf, L. (1996). Developments on Fréchet bounds. In *Distribution with Fixed Marginals and Related Topics*, pp. 273–296. Institute of Mathematical Statistics, Hayward CA.

Rüschendorf, L. (1998). Wasserstein metric. In M. Hazewinkel (Ed.), *Encyclopaedia of Mathematics. Supplement, I, II, III.* Kluwer Academic Publishers (1997–2001).

Rüschendorf, L. and S. T. Rachev (1990). A characterization of random variables with minimum $L^2$-distance. *J. Multivariate Anal. 32*(1), 48–54.

Ruttenberg, B. E. and A. K. Singh (2011). Indexing the earth mover's distance using normal distributions. *Proc. VLDB Endow. 5*(3), 205–216.

Sudakov, V. N. (1979). Geometric problems in the theory of infinite-dimensional probability distributions. *Proc. Steklov Inst. Math. (translation from Tr. Mat. Inst. Steklov 141 (1976)) 141*, 1–178.

Trudinger, N. S. and X.-J. Wang (2001). On the Monge mass transfer problem. *Calc. Var. Partial Differ. Equ. 13*(1), 19–31.

Uckelmann, L. (1998). *Über das Monge-Kantorovich Transportproblem und dessen Verallgemeinerungen.* Dissertation, University of Freiburg.

Villani, C. (2009). *Optimal Transport. Old and New.* Springer-Verlag, Berlin.