Electronic Journal of Probability

Electron. J. Probab. **17** (2012), no. 17, 1–20.
ISSN: 1083-6489 DOI: 10.1214/EJP.v17-1797

On optimal stationary couplings between stationary processes*

Ludger Rüschendorf[†] Tomonari Sei[‡]

Abstract

By a classical result of [10] the $\bar{\varrho}$ distance between stationary processes is identified with an optimal stationary coupling problem of the corresponding stationary measures on the infinite product spaces. This is a modification of the optimal coupling problem from Monge–Kantorovich theory. In this paper we derive some general classes of examples of optimal stationary couplings which allow to calculate the $\bar{\varrho}$ distance in these cases in explicit form. We also extend the $\bar{\varrho}$ distance to random fields and to general nonmetric distance functions and give a construction method for optimal stationary $\bar{c}$-couplings. Our assumptions need in this case a geometric positive curvature condition.

Keywords: Optimal stationary couplings; $\bar{\varrho}$-distance; stationary processes; Monge–Kantorovich theory.
AMS MSC 2010: 60E15; 60G10.
Submitted to EJP on May 30, 2011, final version accepted on February 2, 2012.

1 Introduction

[10] introduced the $\bar{\varrho}$ distance between two stationary probability measures μ, ν on $E^{\mathbb{Z}}$, where (E, ϱ) is a separable, complete metric space (Polish space).

The $\bar{\varrho}$ distance extends Ornstein's $\bar{d}$ distance ([14]) and is applied to the information theoretic problem of source coding with a fidelity criterion, when the source statistics are incompletely known. The distance $\bar{\varrho}$ is defined via the following steps. Let $\varrho_n : E^n \times E^n \to \mathbb{R}$ denote the average distance per component on E^n

$$\varrho_n(x, y) := \frac{1}{n} \sum_{i=0}^{n-1} \varrho(x_i, y_i), \quad x = (x_0, \dots, x_{n-1}), y = (y_0, \dots, y_{n-1}). \tag{1.1}$$

*Supported by the leading-researcher program of Graduate School of Information Science and Technology, University of Tokyo, and by KAKENHI 19700258.

†Mathematische Stochastik, University of Freiburg, Eckerstr. 1, 79104 Freiburg, Germany.
E-mail: ruschen@stochastik.uni-freiburg.de

‡Department of Mathematics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522, Japan.
E-mail: sei@math.keio.ac.jp

Let $\bar{\varrho}_n$ denote the corresponding minimal ℓ_1-metric also called Wasserstein distance or Kantorovich distance of the restrictions of μ, ν on E^n, i.e.

$$\bar{\varrho}_n(\mu,\nu) = \inf \left\{ \int \varrho_n(x,y) d\beta(x,y) \mid \beta \in M(\mu^n, \nu^n) \right\}, \tag{1.2}$$

where μ^n, ν^n are the restrictions of μ, ν on E^n, i.e. on the coordinates $(x_0, \dots, x_{n-1})$ and $M(\mu^n, \nu^n)$ is the Fréchet class of all measures on $E^n \times E^n$ with marginals μ^n, ν^n. Then the $\bar{\varrho}$ distance between μ, ν is defined as

$$\bar{\varrho}(\mu,\nu) = \sup_{n \in \mathbb{N}} \bar{\varrho}_n(\mu,\nu). \tag{1.3}$$

In the original Ornstein version ϱ was taken as discrete metric on a finite alphabet. It is known that $\bar{\varrho}(\mu,\nu) = \lim_{n\to\infty} \bar{\varrho}_n(\mu,\nu)$ by Fekete's lemma on superadditive sequences.

The $\bar{\varrho}$-distance has a natural interpretation as average distance per coordinate between two stationary sources in an optimal coupling. This interpretation is further justified by the basic representation result (cp. [10, Theorem 1])

$$\begin{aligned} \bar{\varrho}(\mu,\nu) \;&=\; \bar{\varrho}_s(\mu,\nu) := \inf_{\Gamma \in M_{\mathrm{s}}(\mu,\nu)} \int \varrho(x_0, y_0) d\Gamma(x,y) &(1.4)\\ &=\; \inf\{ \mathrm{E}[\varrho(X_0, Y_0)] \mid (X,Y) \sim \Gamma \in M_{\mathrm{s}}(\mu,\nu)\}. &(1.5) \end{aligned}$$

Here $M_{\mathrm{s}}(\mu,\nu)$ is the set of all jointly stationary (i.e. jointly shift invariant) measures on $E^{\mathbb{Z}} \times E^{\mathbb{Z}}$ with marginals μ, ν and $(X,Y) \sim \Gamma$ means that Γ is the distribution of (X,Y). Thus $\bar{\varrho}(\mu,\nu)$ can be seen as a Monge–Kantorovich problem on $E^{\mathbb{Z}}$ with however a modified Fréchet class $M_{\mathrm{s}}(\mu,\nu) \subset M(\mu,\nu)$. (1.5) states this as an optimal coupling problem between jointly stationary processes X, Y with marginals μ, ν. A pair of jointly stationary processes (X,Y) with distribution $\Gamma \in M_{\mathrm{s}}(\mu,\nu)$ is called *optimal stationary coupling of* μ,ν if it solves problem (1.5), i.e. it minimizes the stationary coupling distance $\bar{\varrho}_s$.

By definition it is obvious (see [10]) that

$$\bar{\varrho}_1(\mu,\nu) \le \bar{\varrho}(\mu,\nu) \le \int \varrho(x_0,y_0) d\mu^0(x_0) d\nu^0(y_0), \tag{1.6}$$

the left hand side being the usual minimal ℓ_1-distance (Kantorovich distance) between the single components μ^0, ν^0.

As remarked in [10, Example 2] the main representation result in (1.4), (1.5) does not use the metric structure of ϱ and ϱ can be replaced by a general cost function c on $E \times E$ implying then the generalized optimal stationary coupling problem

$$\bar{c}_s(\mu,\nu) = \inf\{ \mathrm{E}[c(X_0,Y_0)] \mid (X,Y) \sim \Gamma \in M_{\mathrm{s}}(\mu,\nu)\}. \tag{1.7}$$

Only in few cases information on this optimal coupling problem for $\bar{\varrho}$ resp. $\bar{c}$ is given in the literature. [10] determine $\bar{\varrho}$ for two i.i.d. binary sequences with success probabilities p_1, p_2. They also derive for quadratic cost $c(x_0,y_0) = (x_0 - y_0)^2$ upper and lower bounds for two stationary Gaussian time series in terms of their spectral densities. We do not know of further explicit examples in the literature for the $\bar{\varrho}$ distance. The aim of our paper is to derive optimal couplings and solutions for the $\bar{\varrho}$ metric resp. the generalized $\bar{c}$ distance.

The $\bar{\varrho}$ resp. $\bar{c}$ distance is particularly adapted to stationary processes. One should note that from the general Monge–Kantorovich theory characterizations of optimal couplings for some classes of distances c are available and have been determined for time series and stochastic processes in some cases. For processes with values in a Hilbert space (like the weighted ℓ_2 or the weighted L^2 space) and for general cost functions

c, general criteria for optimal couplings have been given in [18] and [16]. For some examples and extensions to Banach spaces see also [2] and [17]. Some of these criteria have been further extended to measures μ, ν in the Wiener space (W, H, μ) w.r.t. the squared distance $c(x, y) = |x-y|_H^2$ by Feyel and Üstünel (2002, 2004) and [23]. All these results are also applicable to stationary measures and characterize optimal couplings between them. But they do not respect the special stationary structure as described in the representation result in (1.5), (1.7). In the following sections we want to determine optimal stationary couplings between stationary processes.

In Section 2 we consider the optimal stationary coupling of stationary processes on $\mathbb{R}$ and on $\mathbb{R}^m$ with respect to squared distance. In Section 3 we give an extension to the case of random fields. Finally we consider in Section 4 an extension to general cost functions. We interpret an optimal coupling condition by a geometric curvature condition.

2 Optimal couplings of stationary processes w.r.t. squared distance

In this section we consider the optimal stationary coupling of stationary processes on the Euclidean space with respect to squared distance.

We first recall the classical result for optimal couplings on $\mathbb{R}^n$. For two probability distributions μ and ν on $\mathbb{R}^n$ let $M_n(\mu, \nu)$ be the set of joint distributions Γ of random variables $X \sim \mu$ and $Y \sim \nu$. Denote the Euclidean norm on $\mathbb{R}^n$ by $\|\cdot\|_2$. We call a joint distribution Γ in $M_n(\mu, \nu)$ an optimal coupling if Γ attains the minimum of $\int \|x - y\|_2^2 \Gamma(dx, dy)$ over $M_n(\mu, \nu)$.

Theorem 2.1 ([18] and [1]). *For given measures μ and ν on $\mathbb{R}^n$ with existing second moments, there is an optimal coupling $\Gamma \in M_n(\mu, \nu)$ and it is characterized by*

$$Y \in \partial h(X) \ \ \Gamma\text{-a.s.} \tag{2.1}$$

for some convex function h, where the subgradient $\partial h(x)$ at x is defined by

$$\partial h(x) = \{y \in \mathbb{R}^n \mid h(z) - h(x) \geq y \cdot (z - x), \quad \forall z \in \mathbb{R}^n\}. \tag{2.2}$$

Furthermore, if μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$, then the gradient of h is essentially unique.

In the above theorem let μ be absolutely continuous and assume that μ and ν are invariant under the map $x = (x_1, \ldots, x_n) \mapsto L_n x = (x_n, x_1, \ldots, x_{n-1})$. Then, by the uniqueness result, the convex function h in (2.1) must be invariant under L_n. In addition, if h is differentiable, then the gradient ∇h satisfies $(\nabla h) \circ L_n = L_n \circ (\nabla h)$. This identity motivates the following construction of optimal stationary coupling (see (2.3)).

Now we consider stationary processes. For simplicity, we first consider the one-dimensional case $E = \mathbb{R}$. The multi-dimensional case $E = \mathbb{R}^m$ is discussed later. Let $\Omega = E^{\mathbb{Z}} = \mathbb{R}^{\mathbb{Z}}$ and $c(x_0, y_0) = (x_0 - y_0)^2$. Let $L : \Omega \to \Omega$ denote the left shift, $(Lx)_t = x_{t-1}$. Then a pair of processes (X, Y) with values in $\Omega \times \Omega$ is *jointly stationary* when $(X, Y) \stackrel{d}{=} (LX, LY)$ ($\stackrel{d}{=}$ denotes equality in distribution). A Borel measurable map $S : \Omega \to \Omega$ is called *equivariant* if

$$L \circ S = S \circ L. \tag{2.3}$$

This notion is borrowed from the corresponding notion in statistics, where it is used in connection with statistical group models. The following lemma concerns some elementary properties.

Lemma 2.2. *a) A map $S : \Omega \to \Omega$ is equivariant if and only if $S_t(x) = S_0(L^{-t}x)$ for any t, x.*

b) If X is a stationary process and S is equivariant then $(X, S(X))$ is jointly stationary.

Proof. a) If $L \circ S = S \circ L$ then by induction $S = L^t \circ S \circ L^{-t}$ for all $t \in \mathbb{Z}$, and thus $S_t(x) = S_0(L^{-t}x)$. Conversely, if $S_t(x) = S_0(L^{-t}x)$, then $S_{t-1}(x) = S_0(L^{-t+1}x) = S_t(Lx)$. This implies $L(S(x)) = S(Lx)$.

b) Since LX has the same law as X, it follows that $(LX, L(S(X))) = (LX, S(LX)) = (I, S)(LX) \stackrel{d}{=} (I, S)(X) = (X, S(X))$, I denoting the identity. □

For $X \sim \mu$ and $S : \Omega \to \Omega$ the pair $(X, S(X))$ is called *optimal stationary coupling* if it is an optimal stationary coupling w.r.t. μ and $\nu := \mu^S = S_{\#}\mu$, i.e., when ν is the corresponding image (push-forward) measure.

To construct a class of optimal stationary couplings we define for a convex function $f : \mathbb{R}^n \to \mathbb{R}$ an equivariant map $S : \Omega \to \Omega$. For $x \in \Omega$ let

$$\partial f(x) = \{y \in \mathbb{R}^n \mid f(z) - f(x) \geq y \cdot (z - x), \quad \forall z \in \mathbb{R}\} \tag{2.4}$$

denote the subgradient of f at x, where $a \cdot b$ denotes the standard inner product of vectors a and b. By convexity $\partial f(x) \neq \emptyset$. Let $F(x) = (F_k(x))_{0 \leq k \leq n-1}$ be measurable and $F(x) \in \partial f(x)$, $x \in \mathbb{R}^n$. The equivariant map S is defined via Lemma 2.2 by

$$S_0(x) = \sum_{k=0}^{n-1} F_k(x_{-k}, \ldots, x_{-k+n-1}), \qquad S_t(x) = S_0(L^{-t}x), \quad x \in \Omega. \tag{2.5}$$

For terminological reasons we write any map of the form (2.5) as

$$S_0(x) = \sum_{k=0}^{n-1} \partial_k f(x_{-k}, \ldots, x_{-k+n-1}), \qquad S_t(x) = S_0(L^{-t}x), \quad x \in \Omega. \tag{2.6}$$

In particular for differentiable convex f the subgradient set coincides with the derivative of f, $\partial f(x) = \{\nabla f(x)\}$ and $\partial_t f(x) = \frac{\partial}{\partial x_t} f(x)$.

Remark 2.3. *a) In information theory a map of the form $S_t(x) = F(x_{t-n+1}, \ldots, x_{t+n-1})$ is called a sliding block code (see [10]). Thus our class of maps S defined in* (2.6) *are particular sliding block codes.*

b) [19, 20, 21] introduced so-called structural gradient models (SGM) *for stationary time series, which are defined as $\{(S_\vartheta)^{\#} Q \mid \vartheta \in \Theta\}$, where Q is the infinite product of the uniform distribution on $[0, 1]$, on $[0, 1]^{\mathbb{Z}}$, $\{S_\vartheta \mid \vartheta \in \Theta\}$ is a parametric family of transformations of the form given in* (2.6) *and $S_\vartheta^{\#} Q$ denotes the pullback measure of Q by S_ϑ. It turns out that these models have nice statistical properties, e.g. they allow for simple likelihoods and allow the construction of flexible dependencies. The restriction to functions of the form* (2.6) *is well founded by an extended Poincaré lemma (see [21, Lemma 3]) saying in the case of differentiable f that these functions are the only ones with (the usual) symmetry and with an additional stationarity property $S_{t-1}(x) = S_t(Lx)$ for $x \in \mathbb{R}^{\mathbb{Z}}$, which is related to our notion of equivariant mappings.*

c) Even if a map S has a representation of the form (2.6), *the inverse map S^{-1} does not have the same form in general. We give an example. Let $X = (X_t)_{t \in \mathbb{Z}}$ be a* *real-valued stationary process with a spectral representation* $X_t = \int_0^1 \mathrm{e}^{2\pi \mathrm{i}\lambda t} M(\mathrm{d}\lambda)$, *where* $M(\mathrm{d}\lambda)$ *is an* L^2*-random measure. Define a process* $Y = (Y_t)$ *by*

$$Y_t = S_t(X) := X_t + \epsilon(X_{t-1} + X_{t+1}), \quad \epsilon \neq 0.$$

This is of the form (2.6) *with a function* $f(x_0, x_1) = x_0^2/4 + \epsilon x_0 x_1 + x_1^2/4$ *which is convex if* $|\epsilon| < 1/2$*. Under this condition, the map* $X \mapsto Y$ *is shown to be invertible as follows. The spectral representation of* Y *is* $N(\mathrm{d}\lambda) := (1 + \epsilon(\mathrm{e}^{2\pi \mathrm{i}\lambda} + \mathrm{e}^{-2\pi \mathrm{i}\lambda}))M(\mathrm{d}\lambda)$*. Then we have the following inverse representation*

$$X_t = \int_0^1 \frac{\mathrm{e}^{2\pi \mathrm{i}\lambda t}}{1 + \epsilon(\mathrm{e}^{2\pi \mathrm{i}\lambda} + \mathrm{e}^{-2\pi \mathrm{i}\lambda})} N(\mathrm{d}\lambda) = \sum_{s \in \mathbb{Z}} b_s Y_{t-s},$$

where $(b_s)_{s \in \mathbb{Z}}$ *is defined by* $\{1 + \epsilon(\mathrm{e}^{2\pi \mathrm{i}\lambda} + \mathrm{e}^{-2\pi \mathrm{i}\lambda})\}^{-1} = \sum_{s \in \mathbb{Z}} b_s \mathrm{e}^{-2\pi \mathrm{i}\lambda s}$*. By standard complex analysis, the coefficients* (b_s) *are explicitly obtained:*

$$b_s = \frac{z_+^{|s|}}{\epsilon(z_+ - z_-)}, \quad z_\pm := \frac{-1 \pm \sqrt{1 - 4\epsilon^2}}{2\epsilon}.$$

Note that $|z_+| < 1$ *and* $|z_-| > 1$ *since* $|2\epsilon| < 1$*. Hence* $b_s \neq 0$ *for all* $s \in \mathbb{Z}$ *and the inverse map* $S^{-1}(Y) = \sum_s b_s Y_s$ *does not have a representation as in* (2.6)*.*

The following theorem implies that the class of equivariant maps defined in (2.6) gives a class of examples of optimal stationary couplings between stationary processes.

Theorem 2.4 (Optimal stationary couplings of stationary processes on $\mathbb{R}$)**.** *Let* f *be a convex function on* $\mathbb{R}^n$*, let* S *be the equivariant map defined in* (2.6) *and let* X *be a stationary process with law* μ*. Assume that* X_0 *and* $\partial_k f(X^n)$ *(*$k = 0, \ldots, n-1$*) are in* $L^2(\mu)$*. Then* $(X, S(X))$ *is an optimal stationary coupling w.r.t. squared distance between* μ *and* μ^S*, i.e.*

$$\mathrm{E}[(X_0 - S_0(X))^2] = \min_{(X,Y) \sim \Gamma \in M_\mathrm{s}(\mu, \mu^S)} \mathrm{E}[(X_0 - Y_0)^2] = \bar{c}_\mathrm{s}(\mu, \mu^S),$$

Proof. Fix any $\Gamma \in M_\mathrm{s}(\mu, \mu^S)$. By the gluing lemma (see Appendix A), we can construct a jointly stationary process $(X, Y, \tilde{X})$ on a common probability space such that $X \sim \mu$, $Y = S(X)$ and $(\tilde{X}, Y) \sim \Gamma$. From the definition of $Y_0 = S_0(X)$, we have $Y_0 \in L^2(\mu)$. Then by the assumption of identical marginals

$$\begin{aligned} A &:= \frac{1}{2}\mathrm{E}[(X_0 - Y_0)^2 - (\tilde{X}_0 - Y_0)^2] \\ &= \mathrm{E}[-X_0 Y_0 + \tilde{X}_0 Y_0] \\ &= \mathrm{E}[(\tilde{X}_0 - X_0) S_0(X)] \\ &= \mathrm{E}\left[(\tilde{X}_0 - X_0) \sum_{k=0}^{n-1} (\partial_k f)(X_{-k}, \ldots, X_{-k+n-1})\right]. \end{aligned}$$

Using the joint stationarity of $(X, \widetilde{X})$ we get with $X^n = (X_0, \ldots X_{n-1})$, $\widetilde{X}^n = (\widetilde{X}_0, \ldots, \widetilde{X}_{n-1})$ that

$$\begin{aligned} A &= \mathrm{E}\left[\sum_{k=0}^{n-1} (\tilde{X}_k - X_k)(\partial_k f)(X_0, \ldots, X_{n-1})\right] \\ &\leq \mathrm{E}[f(\tilde{X}^n) - f(X^n)] \\ &= 0, \end{aligned}$$

the inequality is a consequence of convexity of f. This implies optimality of (X, Y). We note that the last equality uses integrability of $f(X^n)$, which comes from convexity of f and the L^2-assumptions. This completes the proof. □

Theorem 2.4 allows to determine explicit optimal stationary couplings for a large class of examples. Note that – at least in principle – the $\bar{c}$ distance can be calculated in explicit form for this class of examples.

The construction of Theorem 2.4 can be extended to multivariate stationary sequences in the following way. Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary process, $X_t \in \mathbb{R}^m$ and let $f : (\mathbb{R}^m)^n \to \mathbb{R}$ be a convex function on $(\mathbb{R}^m)^n$. Define an equivariant map $S : (\mathbb{R}^m)^{\mathbb{Z}} \to (\mathbb{R}^m)^{\mathbb{Z}}$ by

$$\begin{aligned} S_0(x) &= \sum_{k=0}^{n-1} \partial_k f(x_{-k}, \dots, x_{-k+n-1}) \\ S_t(x) &= S_0(L^{-t}x), \quad x \in \Omega = (\mathbb{R}^m)^{\mathbb{Z}} \end{aligned} \tag{2.7}$$

where L^{-t} operates on each component of x and $\partial_\ell f$ is (a representative of) the subgradient of f w.r.t. the ℓ-th component. Thus for differentiable f we obtain

$$S_0(x) = \sum_{k=0}^{n-1} \nabla_k f(x_{-k}, \dots, x_{-k+n-1}) \tag{2.8}$$

where $\nabla_\ell f$ is the gradient of f w.r.t. the ℓ-th component.

Then the following theorem is proved similarly to Theorem 2.4.

Theorem 2.5 (Optimal stationary couplings of stationary processes on $\mathbb{R}^m$). *Let f be a convex function on $(\mathbb{R}^m)^n$ and let S be the equivariant map on $\Omega = (\mathbb{R}^m)^{\mathbb{Z}}$ defined in (2.7). Let X be a stationary process on $\mathbb{R}^m$ with distribution μ and assume that X_0 and $\partial_k f(X^n)$, $0 \le k \le n-1$, are square integrable. Then $(X, S(X))$ is an optimal stationary coupling between μ and $\mu^S = S_{\#}\mu$ w.r.t. squared distance, i.e.*

$$\mathrm{E}[\|X_0 - S_0(X)\|_2^2] = \inf\{\mathrm{E}[\|X_0 - Y_0\|_2^2] \mid (X,Y) \sim \Gamma \in M_{\mathrm{s}}(\mu, \mu^S)\} = \bar{c}_{\mathrm{s}}(\mu, \mu^S). \tag{2.9}$$

Remark 2.6. *Multivariate optimal coupling results as in Theorem 2.5 for the squared distance or later in Theorem 4.1 for general distance allow to compare higher dimensional marginals of two real stationary processes. For this purpose we consider a lifting of one-dimensional processes to multi-dimensional processes as follows. For fixed m we define an injective map q from $\mathbb{R}^{\mathbb{Z}}$ to $(\mathbb{R}^m)^{\mathbb{Z}}$ by $q(x) = (q_k(x))_{k\in\mathbb{Z}} = ((x_k, \dots, x_{k+m-1}))_{k\in\mathbb{Z}}$. Note that q satisfies the equivariant condition (2.3). For one-dimensional processes $X = (X_k) \sim \mu$ and $Y = (Y_k) \sim \nu$ define m-dimensional processes $\tilde{X} = q(X)$ and $\tilde{Y} = q(Y)$ and denote their distributions by $\tilde{\mu}$ and $\tilde{\nu}$, respectively. Let $c^{(m)}$ be a cost function on $\mathbb{R}^m \times \mathbb{R}^m$. Then we have the optimal coupling problems between $\tilde{\mu}$ and $\tilde{\nu}$ as*

$$\begin{aligned} \bar{c}_{\mathrm{s}}^{(m)}(\tilde{\mu}, \tilde{\nu}) &= \inf\{\mathrm{E}[c^{(m)}(\tilde{X}_0, \tilde{Y}_0)] \mid (\tilde{X}, \tilde{Y}) \sim \tilde{\Gamma} \in M_{\mathrm{s}}(\tilde{\mu}, \tilde{\nu})\} \\ &= \inf\{\mathrm{E}[c^{(m)}(q_0(X), q_0(Y))] \mid (X,Y) \sim \Gamma \in M_{\mathrm{s}}(\mu, \nu)\}, \end{aligned}$$

where the second equality follows from the fact that any $(\tilde{X}, \tilde{Y}) \sim \tilde{\Gamma} \in M_{\mathrm{s}}(\tilde{\mu}, \tilde{\nu})$ is supported on $q(\mathbb{R}^{\mathbb{Z}}) \times q(\mathbb{R}^{\mathbb{Z}})$, and $q^{-1}(\tilde{X}) \sim \mu, q^{-1}(\tilde{Y}) \sim \nu$. If the cost function $c^{(m)}$ is the squared distance as in Theorem 2.5, then we can solve the lifted problem immediately when we solve the case $m = 1$ since

$$\begin{aligned} \bar{c}_{\mathrm{s}}^{(m)}(\tilde{\mu}, \tilde{\nu}) &= \inf\{\mathrm{E}\|\tilde{X}_0 - \tilde{Y}_0\|_2^2 \mid (\tilde{X}, \tilde{Y}) \sim \tilde{\Gamma} \in M_{\mathrm{s}}(\tilde{\mu}, \tilde{\nu})\} \\ &= \inf\{\mathrm{E}[m(X_0 - Y_0)^2] \mid (X,Y) \sim \Gamma \in M_{\mathrm{s}}(\mu, \nu)\} = m\bar{c}_{\mathrm{s}}^{(1)}(\mu, \nu). \end{aligned}$$

For general c not written as sum of one-dimensional cost functions the quantity $\bar{c}_{\mathrm{s}}^{(m)}(\tilde{\mu}, \tilde{\nu})$ has a meaning different from one-dimensional ones.

3 Optimal stationary couplings of random fields

In the first part of this section we introduce the $\bar{\varrho}$ distance defined on a product space in the case of countable groups and establish an extension of the [10] representation result to random fields. In a second step we extend this result to amenable groups on a Polish function space. This motivates the consideration of the optimal stationary coupling result as in Section 2.

We consider stationary real random fields on an abstract group G. Section 2 was concerned with the case of stationary discrete time processes, where $G = \mathbb{Z}$. Interesting extensions concern the case of stationary random fields on lattices $G = \mathbb{Z}^d$ or the case of stationary continuous time stochastic processes with $G = \mathbb{R}$ or $G = \mathbb{R}^d$.

To state the most general version of the representation result, we prepare some notations and definitions. Let $(G, \mathcal{G})$ be a topological group with the neutral element e. Let B be a Polish space equipped with a continuous and non-negative cost function $c(x, y)$, $x, y \in B$. We assume that the group G continuously acts on B on the left: $(gh).x = g.(h.x)$, $e.x = x$ and the map $x \mapsto g.x$ is continuous. A Borel probability measure μ on B is called stationary if $\mu^g = \mu$ for every $g \in G$, where μ^g is the push-forward measure of μ by g.

Example 3.1. *If G is countable, an example of B is the product space $\Omega = E^G$ of a Polish space E (e.g. $E = \mathbb{R}$) equipped with the product topology. The left group action of G on Ω is defined by $(g.x)_h = x_{g^{-1}h}$. Indeed,*

$$((gh).x)_k = x_{(gh)^{-1}k} = x_{h^{-1}g^{-1}k} = (h.x)_{g^{-1}k} = (g.(h.x))_k.$$

It is easy to see that $e.x = x$ and the function $x \mapsto g.x$ is continuous.

If G is not countable, then $\Omega = E^G$ is not Polish. One can consider a Polish space $B \subset \Omega$ such that the projection $B \to E$, $x \mapsto x_e$, is measurable and $g.B = B$. For example, let $G = \mathbb{R}$, $E = \mathbb{R}$ and B be the set of all continuous functions on $G = \mathbb{R}$ with the compact-open topology, that is, define $f_n \to f$ in B if $\sup_{x \in K} |f_n(x) - f(x)| \to 0$ for each compact K. Then all the requirements are satisfied.

We assume that G is an amenable group, i.e. there exists a sequence λ_n of asymptotically right invariant probability measures on G such that

$$\sup_{A \in \mathcal{G}} |\lambda_n(Ag) - \lambda_n(A)| \to 0 \quad \text{when } n \to \infty. \tag{3.1}$$

The hypothesis of amenability is central for example in the theory of invariant tests. Many of the standard transformation groups are amenable. A typical exception is the free group of two generators. The Ornstein distance can be extended to this class of stationary random fields as follows. Define the average distance w.r.t. λ_n by

$$c_n(x, y) := \int c(g^{-1}.x, g^{-1}.y)\lambda_n(dg). \tag{3.2}$$

For example, if $B = E^G$ and $c(x, y)$ depends only on (x_e, y_e), say $c(x_e, y_e)$, then c_n is given by

$$c_n(x, y) = \int c(x_g, y_g)\lambda_n(dg). \tag{3.3}$$

Let μ and ν be stationary probability measures on B. The induced minimal probability metric is given by

$$\bar{c}_n(\mu, \nu) = \inf\{\mathrm{E}[c_n(X, Y)] \mid (X, Y) \sim \Gamma \in M(\mu, \nu)\}. \tag{3.4}$$

Finally, the natural extension of the $\bar{c}$ metric of [10] is defined as

$$\bar{c}(\mu,\nu) = \sup_n \bar{c}_n(\mu,\nu). \tag{3.5}$$

The optimal stationary coupling problem is introduced similarly as in Section 2 by

$$\bar{c}_{\mathrm{s}}(\mu,\nu) = \inf\{\mathrm{E}[c(X,Y)] \mid (X,Y) \sim \Gamma \in M_s(\mu,\nu)\} \tag{3.6}$$

where $M_{\mathrm{s}}(\mu,\nu) = \{\Gamma \in M(\mu,\nu) \mid \Gamma^{(g,g)} = \Gamma,\ \forall g \in G\}$ is the class of jointly stationary measures with marginals μ and ν. We use the notation $\Gamma(c) = \mathrm{E}[c(X,Y)]$ and $\Gamma(c_n) = \mathrm{E}[c_n(X,Y)]$ for $\Gamma \in M(\mu,\nu)$.

We now can state an extension of the Gray–Neuhoff–Shields representation result for the $\bar{c}$ distance of stationary random fields to amenable groups.

Theorem 3.2 (General representation result for $\bar{c}$ distance). *Let G be an amenable group acting on a Polish space B and c be a non-negative continuous cost function on $B \times B$. Let μ, ν be stationary probability measures on B. Assume that for $X \sim \mu$ (resp. ν), $Ec(X,y) < \infty$ for $y \in B$. Then the extended Ornstein distance $\bar{c}$ defined in* (3.5) *coincides with the optimal stationary coupling distance $\bar{c}_{\mathrm{s}}$,*

$$\bar{c}(\mu,\nu) = \bar{c}_{\mathrm{s}}(\mu,\nu).$$

In particular, $\bar{c}$ does not depend on choice of λ_n.

Proof. To prove that $\bar{c}(\mu,\nu) \le \bar{c}_{\mathrm{s}}(\mu,\nu)$ let for $\varepsilon > 0$ given $\Gamma \in M_{\mathrm{s}}(\mu,\nu)$ be such that $\Gamma(c) \le \bar{c}_{\mathrm{s}}(\mu,\nu) + \varepsilon$. Then using the integrability assumption and stationary of Γ we obtain for all $n \in \mathbb{N}$

$$\begin{aligned}\bar{c}_n(\mu,\nu) &\le \Gamma(c_n) = \mathrm{E}[\int c(g^{-1}.X, g^{-1}.Y)\lambda_n(dg)] \\ &= \int \mathrm{E}[c(g^{-1}.X, g^{-1}.Y)]\lambda_n(dg) = \Gamma(c) \le \bar{c}_{\mathrm{s}}(\mu,\nu) + \varepsilon.\end{aligned}$$

This implies that $\bar{c}(\mu,\nu) \le \bar{c}_{\mathrm{s}}(\mu,\nu)$.

For the converse direction we choose for fixed $\varepsilon > 0$ and $n \ge 0$ an element $\Gamma_n \in M(\mu,\nu)$ such that $\Gamma_n(c_n) \le \bar{c}_n(\mu,\nu) + \varepsilon$. We define probability measures $\{\bar{\Gamma}_n\}$ by

$$\bar{\Gamma}_n(A) := \int_G \Gamma_n(g.A)\lambda_n(dg). \tag{3.7}$$

Note that $\bar{\Gamma}_n(c) = \Gamma_n(c_n)$. Indeed,

$$\begin{aligned}\bar{\Gamma}_n(c) &= \int c(x,y)\bar{\Gamma}_n(dx,dy) = \iint c(x,y)\Gamma_n(g.dx, g.dy)\lambda_n(dg) \\ &= \iint c(g^{-1}.x, g^{-1}.y)\Gamma_n(dx,dy)\lambda_n(dg) = \int c_n(x,y)\Gamma_n(dx,dy) = \Gamma_n(c_n).\end{aligned}$$

Using Fubini's theorem we obtain that

$$\begin{aligned}\bar{\Gamma}_n(h.A) - \bar{\Gamma}_n(A) &= \int_G (\Gamma_n(gh.A) - \Gamma_n(g.A))\lambda_n(dg) \\ &= \int_{B\times B} (\lambda_n(C_{x,y}h^{-1}) - \lambda_n(C_{x,y}))\Gamma_n(dxdy),\end{aligned} \tag{3.8}$$

where $C_{x,y} = \{g \in G | (x,y) \in g.A\}$. By amenability (3.1) of G we have

$$|\bar{\Gamma}_n(h.A) - \bar{\Gamma}_n(A)| \le \int_{B\times B} |\lambda_n(C_{x,y}h^{-1}) - \lambda_n(C_{x,y})|\Gamma_n(dxdy) \to 0 \tag{3.9}$$

as $n \to \infty$, i.e. $\bar{\Gamma}_n$ is asymptotically left invariant on $B \times B$.

We have $\bar{\Gamma}_n \in M(\mu, \nu)$ since projections on finite components of $\bar{\Gamma}_n$ are

$$\begin{aligned}\bar{\Gamma}_n(A_1 \times \Omega) &= \int_G \Gamma_n(g.A_1 \times \Omega)\lambda_n(dg) \\ &= \int_G \mu(g.A_1)\lambda_n(dg) = \mu(A_1)\end{aligned}$$

since μ is stationary. Using tightness of $\{\bar{\Gamma}_n\}$ we get a weakly converging subsequence of $\{\bar{\Gamma}_n\}$. Without loss of generality we assume that $\{\bar{\Gamma}_n\}$ converges weakly to some probability measure $\bar{\Gamma}$ on $B \times B$. In consequence by (3.9) we get $\bar{\Gamma} \in M_{\mathrm{s}}(\mu, \nu)$. Finally,

$$\begin{aligned}\bar{c}_{\mathrm{s}}(\mu, \nu) &\le \bar{\Gamma}(c) \le \limsup \bar{\Gamma}_n(c) = \limsup \Gamma_n(c_n) \\ &\le \limsup \bar{c}_n(\mu, \nu) + \varepsilon \le \bar{c}(\mu, \nu) + \varepsilon\end{aligned}$$

for all $\varepsilon > 0$ which concludes the proof. □

Example 3.3. *Let G be countable and $\lambda_n = \frac{1}{|F_n|}\sum_{g \in F_n} \varepsilon_g$ for some increasing class of finite sets $F_n \subset G$ with $G = \cup_n F_n$, where ε_g denotes the point-mass measure at g. Amenability of G corresponds to the condition that F_n is asymptotically right invariant in the sense that*

$$|F_n \cap (F_n h)|/|F_n| \to 1, \quad \forall h \in G. \tag{3.10}$$

For example, the group $G = \mathbb{Z}$ is amenable because $F_n = \{\lfloor -n/2 \rfloor, \dots, \lfloor n/2 \rfloor - 1\}$ satisfies the above conditions. In the optimal coupling problem, we can take $F_n' = \{0, \dots, n-1\}$ instead of F_n because μ and ν are stationary, although F_n' does not cover $\mathbb{Z}$.

Now take the product space $B = E^G$ and assume that $c(x, y)$ depends only on (x_e, y_e) and is denoted as $c(x, y) = c(x_e, y_e)$. Then we obtain $c_n(x, y) = \frac{1}{|F_n|}\sum_{g \in F_n} c(x_g, y_g) =: c_n(x_{F_n}, y_{F_n})$, where $x_{F_n} = (x_g)_{g \in F_n}$. We now show that $\bar{c}_n(\mu, \nu)$ in (3.4) is equal to

$$\inf\{\mathrm{E}[c_n(X_{F_n}, Y_{F_n})] \mid (X_{F_n}, Y_{F_n}) \sim \Gamma_{F_n} \in M(\mu_{F_n}, \nu_{F_n})\} \tag{3.11}$$

with $X_{F_n} = \pi_{F_n}(X)$, $Y_{F_n} = \pi_{F_n}(Y)$, $\mu_{F_n} = \mu^{\pi_{F_n}}$ and $\nu_{F_n} = \nu^{\pi_{F_n}}$, where π_{F_n} is defined by $\pi_{F_n}(x) = x_{F_n}$. The equation (3.11) follows from the general extension property of probability measures with given marginals, that is, we can construct

$$\Gamma(dx, dy) = \Gamma_{F_n}(dx_{F_n}, dy_{F_n})\mu_{G \setminus F_n}(dx_{G \setminus F_n}|x_{F_n})\nu_{G \setminus F_n}(dy_{G \setminus F_n}|y_{F_n})$$

from any $\Gamma_{F_n} \in M(\mu_{F_n}, \nu_{F_n})$ (see also Appendix A). Finally, the original representation result (1.4) follows from Theorem 3.2 with $G = \mathbb{Z}$ since (3.11) is consistent with (1.2).

Motivated by the representation results in Theorem 3.2 we now consider the optimal stationary coupling problem for general groups G acting on $\Omega = \mathbb{R}^G$ and the squared distance $c(x, y) = (x_0 - y_0)^2$. Let F be a finite subset of G and let $f : \mathbb{R}^F \to \mathbb{R}$ be a convex function. The function f is naturally identified with a function on Ω by $f(x) = f((x_g)_{g \in F})$. As in Section 2 any choice of the subgradient of f is denoted by $((\partial_g f)(x))_{g \in F}$. Define an equivariant Borel measurable function $S : \Omega \to \Omega$ by the shifted sum of gradients

$$S_e(x) = \sum_{g \in F} (\partial_g f)(gx) \quad \text{and} \quad S_h(x) = S_e(h^{-1}x), h \in G. \tag{3.12}$$

Note that $S_e(x)$ depends only on $(x_g)_{g \in G(F)}$, where $G(F)$ is the subgroup generated by F in G. We have $S \circ g = g \circ S$ for any $g \in G$ because

$$S_h(gx) = S_e(h^{-1}gx) = S_{g^{-1}h}(x) = (gS(x))_h.$$

Hence if X is a stationary random field, then $(X, S(X))$ is a jointly stationary random field.

We obtain the following theorem.

Theorem 3.4. *Let μ be a stationary probability measure on $\Omega = \mathbb{R}^G$ with respect to a general group of measurable transformations G. Let S be an equivariant map as defined in* (3.12) *with a convex function f. Let X be a real stationary random field with law μ and assume that X_e and $(\partial_g f(X))_{g\in F}$ are in $L^2(\mu)$. Then $(X, S(X))$ is an optimal stationary coupling w.r.t. squared distance between μ and μ^S, i.e.*

$$\mathrm{E}[(X_e - S_e(X))^2] = \min_{(X,Y)\sim\Gamma\in M_{\mathrm{s}}(\mu,\mu^S)} \mathrm{E}[(X_e - Y_e)^2] = \bar{c}_{\mathrm{s}}(\mu, \mu^S).$$

Proof. The construction of the equivariant mapping in (3.12) and the following remark allow us to transfer the proof of Theorem 2.5 to the class of random field models. Fix $\Gamma \in M_{\mathrm{s}}(\mu, \mu^S)$. Let $G(F)$ be the subgroup generated by F in G. Then $G(F)$ is countable (or finite). We denote the restricted measure of μ on $\mathbb{R}^{G(F)}$ by $\mu|_{G(F)}$. By the gluing lemma, we can consider a jointly stationary random field $(X_g, Y_g, \tilde{X}_g)_{g\in G(F)}$ on a common probability space such that $(X_g)_{g\in G(F)} \sim \mu|_{G(F)}$, $Y_g = S_g(X)$ and $(\tilde{X}_g, Y_g)_{g\in G(F)} \sim \Gamma|_{G(F)}$. Then we have

$$\begin{aligned}
\frac{1}{2}\mathrm{E}[(X_e - S_e(X))^2 - (\tilde{X}_e - S_e(X))^2] &= \mathrm{E}[S_e(X)(\tilde{X}_e - X_e)] \\
&= \sum_{g\in F} \mathrm{E}\big[\big((\partial_g f)(gX)\big)(\tilde{X}_e - X_e)\big] \\
&= \sum_{g\in F} \mathrm{E}\big[\big((\partial_g f)(X)\big)(\tilde{X}_g - X_g)\big] \\
&\le \mathrm{E}[f(\tilde{X}) - f(X)] \\
&= 0.
\end{aligned}$$

This implies that $(X, S(X))$ is an optimal stationary coupling w.r.t. squared distance between the random fields μ and $\mu^S = S_{\#}\mu$. □

The generalization to the multi-dimensional case $E = \mathbb{R}^m$ is now obvious and omitted.

4 Optimal stationary couplings for general cost functions

We consider general cost functions c on general spaces other than the squared distance on $\mathbb{R}^m$. The Monge–Kantorovich problem and the related characterization of optimal couplings have been generalized to general cost functions $c(x, y)$ in [16, 17], while [13] extended the squared loss case to manifolds; see also the surveys in [15] and [24, 25]. Based on these developments we will extend the optimal stationary coupling results in Sections 2, 3 to more general classes of distance functions. Some of the relevant notions from transportation theory are collected in the Appendix B. We will restrict to the case of time parameter $\mathbb{Z}$. As in Section 3 an extension to random fields with general *time* parameter is straightforward.

Let E_1, E_2 be Polish spaces. and let $c : E_1 \times E_2 \to \mathbb{R}$ be a measurable cost function. For $f : E_1 \to \mathbb{R}$ and $x_0 \in E_1$ let

$$\partial^c f(x_0) = \big\{y_0 \in E_2 \mid c(x_0, y_0) - f(x_0) = \inf_{z_0\in E_1} \{c(z_0, y_0) - f(z_0)\}\big\} \tag{4.1}$$

denote the set of c-*supergradients* of f in x_0.

A function $\varphi : E_1 \to \mathbb{R} \cup \{-\infty\}$ is called *c-concave* if there exists a function $\psi : E_2 \to \mathbb{R} \cup \{-\infty\}$ such that

$$\varphi(x) = \inf_{y \in E_2} (c(x,y) - \psi(y)), \quad \forall x \in E_1. \tag{4.2}$$

If $\varphi(x) = c(x, y_0) - \psi(y_0)$, then y_0 is a c-supergradient of φ at x. For squared distance $c(x,y) = \|x-y\|_2^2$ in $\mathbb{R}^m = E_1 = E_2$ c-concavity of φ is equivalent to the concavity of $x \mapsto \varphi(x) - \|x\|_2^2/2$.

Consider $E_1 = E_2 = \mathbb{R}^m$. The characterization of optimal couplings $T(x) \in \partial^c \varphi(x)$ for some c-concave function φ leads for regular φ to a differential characterization of c-optimal coupling functions T

$$(\nabla_x c)(x, T(x)) = \nabla \varphi(x). \tag{4.3}$$

In case (4.3) has a unique solution in $T(x)$ this equation describes optimal c-coupling functions T in terms of differentials of c-concave functions φ and the set of c-supergradients $\partial^c \varphi(x)$ reduces to just one element

$$\partial^c \varphi(x) = \{\nabla_x c^*(x, \varphi(x))\}. \tag{4.4}$$

Here c^* is the Legendre transform of $c(x,\cdot)$ and $\nabla_x c(x,\cdot)$ is invertible and $(\nabla_x c)^{-1}(x, \varphi(x)) = \nabla_x c^*(x, \varphi(x))$ (see [16, 15] and [24, 25]). For functions φ which are not c-concave, the supergradient $\partial^c \varphi(x)$ is empty at some point x. Even if φ is c-concave, the supergradient may be empty. If $c(x,y) = h(x-y)$ with a superlinear strictly convex function h on $\mathbb{R}^m$, the existence of supergradients and regularity of c-concave functions are proved in the appendix of [9].

The construction of optimal stationary c-couplings of stationary processes can be pursued in the following way. Define the average distance per component $c_n : E_1^n \times E_2^n \to \mathbb{R}$ by

$$c_n(x,y) = \frac{1}{n} \sum_{t=0}^{n-1} c(x_t, y_t) \tag{4.5}$$

and assume that for some function $f : E_1^n \to \mathbb{R}$, there exists a function $F^n : E_1^n \to E_2^n$ such that

$$F^n(x) = (F_k(x))_{0 \le k \le n-1} \in \partial^{c_n} f(x), \quad x \in E_1^n. \tag{4.6}$$

Note that (4.6) needs to be satisfied only on the support of (the projection of) the stationary measure μ. In general we can expect $\partial^{c_n} f(x) \neq \emptyset$, $\forall x \in E_1^n$ only if f is c_n-concave. For fixed $y_0, \ldots, y_{n-1} \in E_2$ we introduce the function $h_c(x_0) = \frac{1}{n} \sum_{k=0}^{n-1} c(x_0, y_k)$, $x_0 \in E_1$. $h_c(x)$ describes the average distance of x_0 to the n points $y_0, \ldots, y_{n-1}$ in E_2. We define an equivariant map $S : E_1^{\mathbb{Z}} \to E_2^{\mathbb{Z}}$ by

$$\begin{aligned} &S_0(x) \in \partial^c(h_c(x_0)) \mid_{y_k = F_k(x_{-k}, \ldots, x_{-k+n-1}), 0 \le k \le n-1} \\ &S_t(x) = S_0(L^{-t}x), \quad S(x) = (S_t(x))_{t \in \mathbb{Z}}. \end{aligned} \tag{4.7}$$

Here the c-supergradient is taken for the function $h_c(x_0)$ and the formula is evaluated at $y_k = F_k(x_{-k}, \ldots, x_{-k+n-1})$, $0 \le k \le n-1$. After these preparations we can state the following theorem.

Theorem 4.1 (Optimal stationary c-couplings of stationary processes)**.** *Let $X = (X_t)_{t \in \mathbb{Z}}$ be a stationary process with values in E_1 and with distribution μ, let $c : E_1 \times E_2 \to \mathbb{R}$ be a measurable distance function on $E_1 \times E_2$ and let $f : E_1^n \to \mathbb{R}$ be measurable c_n-concave. If S is the equivariant map induced by f in* (4.7) *and if $c(X_0, S_0(X))$, $\{c(X_k, F_k(X^n))\}_{k=0}^{n-1}$ and $f(X^n)$ are integrable, then $(X, S(X))$ is an optimal stationary c-coupling of the stationary measures μ, μ^S i.e.*

$$\mathrm{E}[c(X_0, S_0(X))] = \inf\{\mathrm{E}[c(Y_0, Z_0)] \mid (Y,Z) \sim \Gamma \in M_{\mathrm{s}}(\mu, \mu^S)\} = \bar{c}_{\mathrm{s}}(\mu, \mu^S). \tag{4.8}$$

Proof. The construction of the equivariant function in (4.7) allows us to extend the basic idea of the proof of Theorem 2.4 to the case of general cost function. Fix any $\Gamma \in M_s(\mu, \mu^S)$. By the gluing lemma, we can consider a jointly stationary process $(X, Y, \tilde{X})$ on a common probability space with properties $X \sim \mu$, $Y = S(X)$ and $(\tilde{X}, Y) \sim \Gamma$. Then we have by construction in (4.7) and using joint stationarity of $(X, \tilde{X})$

$$\begin{aligned}
&\mathrm{E}[c(X_0, S_0(X)) - c(\tilde{X}_0, S_0(X))] \\
&\quad \le \mathrm{E}\left[n^{-1} \sum_{k=0}^{n-1} \{c(X_0, y_k) - c(\tilde{X}_0, y_k)\}\Big|_{y_k = F_k(X_{-k}, \dots, X_{-k+n-1})} \right] \\
&\quad = \mathrm{E}\left[n^{-1} \sum_{k=0}^{n-1} \{c(X_k, y_k) - c(\tilde{X}_k, y_k)\}\Big|_{y_k = F_k(X_0, \dots, X_{n-1})} \right] \\
&\quad = \mathrm{E}\left[c_n(X^n, F^n(X^n)) - c_n(\tilde{X}^n, F^n(X^n)) \right] \\
&\quad \le \mathrm{E}[f(X^n) - f(\tilde{X}^n)] \\
&\quad = 0.
\end{aligned}$$

The first inequality is a consequence of $S_0 \in \partial^c(h_c)(x_0)$. The last inequality follows from c_n-concavity of f while the last equality is a consequence of the assumption that $X \stackrel{d}{=} \tilde{X}$. As consequence we obtain that $(X, S(X))$ is an optimal stationary c-coupling. □

The conditions in the construction (4.7) of optimal stationary couplings in Theorem 4.1 (conditions (4.6), (4.7)) simplify essentially in the case $n = 1$. In this case we get as corollary of Theorem 4.1

Corollary 4.2. *Let $X = (X_t)_{t \in U}$ be a stationary process with values in E_1 and distribution μ and let $c : E_1 \times E_2 \to \mathbb{R}$ be a cost function as in Theorem 4.1. Let $f : E_1 \to \mathbb{R}$ be measurable c-concave and define*

$$S_0(x) \in \partial^c f(x_0), \quad S_t(x) = S_0(L^{-t}x) \in \partial^c f(x_t), \quad S(x) = (S_t(x))_{t \in \mathbb{Z}}. \tag{4.9}$$

Then $(X, S(X))$ is an optimal stationary c-coupling of the stationary measures μ, μ^S.

Thus the equivariant componentwise transformation of a stationary process by supergradients of a c-concave function is an optimal stationary coupling. In particular in the case that $E_1 = \mathbb{R}^k$ several examples of c-optimal transformations are given in [17] resp. [15] which can be used to apply Corollary 4.2.

In case $n \ge 1$ conditions (4.6), (4.7) are in general not obvious. In some cases c_n-convexity of a function $f : E_1^n \to \mathbb{R}$ is however easy to see.

Lemma 4.3. *Let $f(x) = \sum_{k=0}^{n-1} f_k(x_k)$, $f_k : E_1 \to \mathbb{R}$, $0 \le k \le n-1$. If the f_k's are c-concave, $0 \le k \le n-1$, then f is c_n-concave and*

$$\partial^{c_n} f(x) = \sum_{k=0}^{n-1} \partial^c f(x_k). \tag{4.10}$$

Proof. Let $y_k \in \partial^c f_k(x_k)$, $0 \le k \le n-1$, then with $y = (y_k)_{0 \le k \le n-1}$ by definition of c-supergradients

$$c_n(x, y) - f(x) = \frac{1}{n} \sum_k (c(x_k, y_k) - f_k(x_k)) = \inf\{c_n(z, y) - f(z); z \in E_1^n\}$$

and thus $y \in \partial^{c_n} f(x)$. The converse inclusion is obvious. □

Lemma 4.3 allows to construct some examples of functions F^n satisfying condition (4.5). For $n > 1$ non-emptiness of the c-supergradient of $h_c(x_0) = \frac{1}{n}\sum_{k=0}^{n-1} c(x_0, y_k)$ has to be established. The condition $u_0 \in \partial^c h_c(x_0)$ is equivalent to

$$c(x_0, u_0) - h_c(x_0) = \inf_z (c(z, u_0) - h_c(z)). \tag{4.11}$$

In the differentiable case (4.11) implies the necessary condition

$$\nabla_x c(x_0, u_0) = \nabla_x h_c(x_0) = \frac{1}{n}\sum_{k=0}^{n-1} \nabla_x c(x_0, y_k). \tag{4.12}$$

If the map $u \to \nabla_x c(x_0, u)$ is invertible then equation (4.12) implies

$$u_0 = (\nabla_x c)^{-1}(x_0, \cdot)\left(\frac{1}{n}\sum_{k=0}^{n-1} \nabla_x c(x_0, y_k)\right) \tag{4.13}$$

(see (4.4)). Thus in case that (4.11) has a solution, it is given by (4.13).

Lemma 4.4. *Suppose that for some $x_0 \in E_1$, $\partial^c h_c(x_0) \neq \emptyset$, and that the map $E_2 \to E_2 : u \mapsto \nabla_x c(x_0, u)$ is one to one, then $\partial^c h_c(x_0)$ is reduced to the single point u_0 defined by*

$$u_0 = (\nabla_x c)^{-1}(x_0, \cdot)\left(\frac{1}{n}\sum_{k=0}^{n-1} \nabla_x c(x_0, y_k)\right). \tag{4.14}$$

Example 4.5. *If $c(x, y) = H(x - y)$ for a superlinear strictly convex function H, then $\nabla_x c(x, \cdot)$ is invertible and we can construct the necessary c-supergradients of h_c. The c-concavity of h_c is not discussed here. If for example $c(x, y) = \|x - y\|_2^2$, where $\|\cdot\|_2$ is the Euclidean norm, then we get for any $x_0 \in \mathbb{R}^m$,*

$$u_0 = u_0(x_0) = \frac{1}{n}\sum_{k=0}^{n-1} y_k = \overline{y} \tag{4.15}$$

is independent of x_0 and

$$\overline{y} \in \partial^c h_c(x_0), \quad \forall x_0 \in \mathbb{R}^m. \tag{4.16}$$

If $c(x, y) = \|x - y\|_2^p$, $p > 1$, then we get for $x_0 \in \mathbb{R}^m$

$$u_0 = u_0(x_0) = x_0 + \|a(x_0)\|_2^{\frac{1}{p-1}} \frac{a(x_0)}{\|a(x_0)\|_2}, \tag{4.17}$$

where $a(x_0) = \frac{1}{n}\sum_{k=0}^{n-1} \|x_0 - y_k\|_2^{p-1} \frac{x_0 - y_k}{\|x_0 - y_k\|_2}$. For this and related further examples see [17] and [9].

The c-concavity of h_c has a geometrical interpretation. $u_0 \in \partial^c h_c(x_0)$ if the difference of the distance of z_0 in E_1 to u_0 in E_2 and the average distance of z_0 to the given points $y_0, \dots, y_{n-1}$ in E_2 is minimized in x_0. The c-concavity of h_c can be interpreted as a positive curvature condition for the distance c. To handle this condition we introduce the notion of convex stability.

Definition 4.6. *The cost function c is called* convex stable *of index $n \geq 1$ if $\partial^c h_c(x_0) \neq \emptyset$ for any $x_0 \in E_1$ and $y \in E_2^n$, where*

$$h_c(x_0) = \frac{1}{n}\sum_{k=0}^{n-1} c(x_0, y_k), \quad x_0 \in E_1. \tag{4.18}$$

The cost c is called convex stable if it is convex stable of index n for all $n \geq 1$.

Example 4.7. *Let $E_1 = E_2 = H$ be a Hilbert space, as for example $H = \mathbb{R}^m$, let $c(x,y) = \|x-y\|_2^2/2$ and fix $y \in H^n$, then*

$$\begin{aligned} h_c(x_0) &= \frac{1}{n}\sum_{k=0}^{n-1} c(x_0, y_k) \\ &= c(x_0, \bar{y}) + \frac{1}{n}\sum_{k=0}^{n-1} c(\bar{y}, y_k), \end{aligned} \tag{4.19}$$

where $\bar{y} = \frac{1}{n}\sum_{k=0}^{n-1} y_k$ Thus by definition (4.2) h_c is c-concave and a c-supergradient of h_c is given by $\bar{y}$ independent of x_0, i.e.

$$\bar{y} \in \partial^c h_c(x_0), \quad \forall x_0 \in H. \tag{4.20}$$

Thus the squared distance c is convex stable.

The property of a cost function to be convex stable is closely connected with the geometric property of non-negative cross curvature. Let E_1 and E_2 be open connected subsets in $\mathbb{R}^m$ ($m \geq 1$) with coordinates $x = (x^i)_{i=1}^m$ and $y = (y^j)_{j=1}^m$. Let $c : E_1 \times E_2 \to \mathbb{R}$ be $C^{2,2}$, i.e. c is two times differentiable in each variable. Denote the cross derivatives by $c_{ij,k} = \partial^3 c/\partial x^i \partial x^j \partial y^k$ and so on. Define $c_x(x,y) = (\partial c/\partial x^i)_{i=1}^m$, $c_y(x,y) = (\partial c/\partial y^j)_{j=1}^m$, $U_x = \{c_x(x,y) \mid y \in E_2\} \subset \mathbb{R}^m$, $V_y = \{c_y(x,y) \mid x \in E_1\} \subset \mathbb{R}^m$. Assume the following two conditions.

[B1] The maps $c_x(x,\cdot) : E_2 \to U_x$ and $c_y(\cdot, y) : E_1 \to V_y$ are diffeomorphic, i.e., they are injective and the matrix $(c_{i,j}(x,y))$ is invertible everywhere.

[B2] The sets U and V are convex.

The conditions [B1] and [B2] are called bi-twist and bi-convex conditions, respectively. Now we define the *cross curvature* $\sigma(x,y;u,v)$ in $x \in E_1$, $y \in E_2$, $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^m$ by

$$\sigma(x,y;u,v) := \sum_{i,j,k,l}\left(-c_{ij,kl} + \sum_{p,q} c_{ij,q}c^{p,q}c_{p,kl}\right) u^i u^j v^k v^l \tag{4.21}$$

where $(c^{i,j})$ denotes the inverse matrix of $(c_{i,j})$.

The following result is given by [11]. Note that these authors use the terminology *time-convex sliding-mountain* instead of the notion convex-stability as used in this paper.

Proposition 4.8. *Assume the conditions [B1] and [B2]. Then c is convex stable if and only if the cross curvature is non-negative, i.e.,*

$$\sigma(x,y;u,v) \geq 0, \quad \forall x,y,u,v. \tag{4.22}$$

The cross-curvature is related to the Ma-Trudinger-Wang tensor ([12]), which is the restriction of $\sigma(x,y;u,v)$ to $\sum_{i,j} u^i v^j c_{i,j} = 0$. Known examples that have non-negative cross-curvature are the n-sphere with the squared Riemannian distance ([11], [7]), its perturbation ([3], [8]), their tensorial product and their Riemannian submersion.

If $E_1, E_2 \subset \mathbb{R}$, then the conditions [B1] and [B2] are implied from a single condition in case $c_{x,y} = \partial^2 c(x,y)/\partial x \partial y \neq 0$. Hence we have the following result as a corollary. A selfcontained simplified proof of this result is given in Appendix C.

Proposition 4.9. *Let E_1, E_2 be open intervals in $\mathbb{R}$ and let $c \in C^{2,2}$, $c : E_1 \times E_2 \to \mathbb{R}$. Assume that $c_{x,y} \neq 0$ for all x,y. Then c is convex stable if and only if $\sigma(x,y) := -c_{xx,yy} + c_{xx,y}c_{x,yy}/c_{x,y} \geq 0$.*

Example 4.10. *Let $E_1, E_2 \subset \mathbb{R}$ be open intervals and let $E_1 \cap E_2 = \emptyset$. Consider $c(x,y) = \frac{1}{p}|x-y|^p$ with $p \geq 2$ or $p < 1$. Then c is convex stable. In fact $c_{x,y} = -(p-1)|x-y|^{p-2} \neq 0$ for all x, y and $\sigma(x,y) = (p-1)(p-2)|x-y|^{p-4} \geq 0$ for all x, y. As $p \to 0$, we also have a convex stable cost $c(x,y) = \log|x-y|$.*

If the cost function c is a metric then the optimal coupling in the case $E_1 = E_2 = \mathbb{R}$ can be reduced to the case of $E_1 \cap E_2 = \emptyset$ as in the classical Kantorovich–Rubinstein theorem. This is done by subtracting (and renormalizing) from the marginals μ_0, ν_0 the lattice infimum, i.e. defining

$$\mu_0' := \frac{1}{a}(\mu_0 - \mu_0 \wedge \nu_0), \quad \nu_0' := \frac{1}{a}(\nu_0 - \mu_0 \wedge \nu_0). \tag{4.23}$$

The new probability measures live on disjoint subsets to which the previous proposition can be applied.

Some classes of optimal c-couplings for various distance functions c have been discussed in [17], see also [15]. The examples discussed in these papers can be used to establish c_n-concavity of f in some cases. This is an assumption used in Theorem 4.1 for the construction of the optimal stationary couplings. Note that c_n is convex-stable if c is convex-stable. Therefore the following proposition due to [6] (partially [22]) is also useful to construct a c_n-concave function f.

Proposition 4.11. *Assume [B1] and [B2]. Then c satisfies the non-negative cross curvature condition if and only if the space of c-concave functions is convex, that is, $(1-\lambda)f + \lambda g$ is c-concave as long as f and g are c-concave and $\lambda \in [0,1]$.*

Example 4.12. *Consider Example 4.10 again. Let $E_1 = (0,1)$, $E_2 = (-\infty, 0)$, $c(x_1, y_1) = p^{-1}(x_1 - y_1)^p$ ($p \geq 2$) and $c_n(x,y) = (np)^{-1}\sum_{k=0}^{n-1}(x_k - y_k)^p$. An example of c_n-concave functions of the form $f(x) = \sum_{k=0}^{n-1} f_k(x_k)$ with suitable real functions f_k is given in [17] Example 1 (b). We add a further example here. Put $\bar{x} = n^{-1}\sum_{k=0}^{n-1} x_k$ and let $f(x) = A(\bar{x})$ with a real function A. We prove $f(x)$ is c_n-concave if $A' \geq 1$ and $A'' \leq 0$. For example, $A(\xi) = \xi + \sqrt{\xi}$ satisfies this condition. Equation (4.3) becomes*

$$n^{-1}(x_i - y_i)^{p-1} = n^{-1}A'(\bar{x}) \tag{4.24}$$

which uniquely determines $y_i \in E_2$ since $A' \geq 1$ and $x_i \in E_1$. To prove c_n-concavity of f, it is sufficient to show convexity of $x \mapsto c_n(x,y) - f(x)$ for each y. Indeed, the Hessian is

$$\delta_{ij} n^{-1}(p-1)(x_i - y_i)^{p-2} - n^{-2}A''(\bar{x}) \succeq -n^{-2}A''(\bar{x}) \succeq 0$$

in matrix sense. Note that the set of functions A satisfying $A' \geq 1$ and $A'' \leq 0$ is convex, which is consistent with Proposition 4.11. Therefore, any convex combination of $A(\bar{x})$ and the c_n-concave function $\sum_k f_k(x_k)$ discussed above is also c_n-concave by Proposition 4.11.

Appendix

A Gluing lemma for stationary measures

The gluing lemma is a well known construction of joint distributions. We repeat this construction in order to derive an extension to the gluing of jointly stationary processes. For given probability measures P and Q on some measurable spaces E_1 and E_2, we denote the set of joint probability measures on $E_1 \times E_2$ with marginals P and Q by $M(P,Q)$.

Lemma A.1 (Gluing lemma)**.** *Let P_1, P_2, P_3 be Borel probability measures on Polish spaces E_1, E_2, E_3, respectively. Let $P_{12} \in M(P_1, P_2)$ and $P_{23} \in M(P_2, P_3)$. Then there exists a probability measure P_{123} on $E_1 \times E_2 \times E_3$ with marginals P_{12} on $E_1 \times E_2$ and P_{23} on $E_2 \times E_3$.*

Proof. Let $P_{1|2}(\cdot|\cdot)$ be the regular conditional probability measure such that

$$P_{12}(A_1 \times A_2) = \int_{A_2} P_{1|2}(A_1|x) P_2(\mathrm{d}x)$$

and $P_{3|2}(\cdot|\cdot)$ be the regular conditional probability measure such that

$$P_{23}(A_2 \times A_3) = \int_{A_2} P_{3|2}(A_3|x) P_2(\mathrm{d}x).$$

Then a measure P_{123} uniquely defined by

$$P_{123}(A_1 \times A_2 \times A_3) := \int_{A_2} P_{1|2}(A_1|x) P_{3|2}(A_3|x) P_2(\mathrm{d}x) \tag{A.1}$$

satisfies the required condition. □

Next we consider an extension of the gluing lemma to stationary processes. We note that even if a measure P_{123} on $E_1^{\mathbb{Z}} \times E_2^{\mathbb{Z}} \times E_3^{\mathbb{Z}}$ has stationary marginals P_{12} on $E_1^{\mathbb{Z}} \times E_2^{\mathbb{Z}}$ and P_{23} on $E_2^{\mathbb{Z}} \times E_3^{\mathbb{Z}}$, it is not necessarily true that P is stationary. For example, consider the $\{-1, 1\}$-valued fair coin processes $X = (X_t)_{t \in \mathbb{Z}}$ and $Y = (Y_t)_{t \in \mathbb{Z}}$ independently, and let $Z_t = (-1)^t X_t Y_t$. Then (X, Y) and (Y, Z) have stationary marginal distributions respectively, but (X, Y, Z) is not jointly stationary because $X_t Y_t Z_t = (-1)^t$.

For given stationary measures P and Q on some product spaces, let $M_{\mathrm{s}}(P, Q)$ be the jointly stationary measures with marginal distributions P and Q on the corresponding product spaces.

Lemma A.2. *Let E_1, E_2, E_3 be Polish spaces. Let P_1, P_2, P_3 be stationary measures on $E_1^{\mathbb{Z}}, E_2^{\mathbb{Z}}, E_3^{\mathbb{Z}}$, respectively. Let $P_{12} \in M_{\mathrm{s}}(P_1, P_2)$ and $P_{23} \in M_{\mathrm{s}}(P_2, P_3)$. Then there exists a jointly stationary measure P_{123} on $E_1^{\mathbb{Z}} \times E_2^{\mathbb{Z}} \times E_3^{\mathbb{Z}}$ with marginals P_{12} and P_{23}.*

Proof. We define P_{123} by (A.1) and check joint stationarity of P_{123}. First, since P_{12} is stationary, the conditional probability $P_{1|2}$ is stationary in the sense that $P_{1|2}(LA_1|Lx) = P_{1|2}(A_1|x)$ for any A_1 and x (P_2-a.s.). Indeed, for any A_1 and A_2,

$$\begin{aligned}
\int_{A_2} P_{1|2}(A_1|x) P_2(\mathrm{d}x) &= P_{12}(A_1 \times A_2) \\
&= P_{12}(LA_1 \times LA_2) \\
&= \int_{LA_2} P_{1|2}(LA_1|x) P_2(\mathrm{d}x) \\
&= \int_{A_2} P_{1|2}(LA_1|Lx) P_2(\mathrm{d}x),
\end{aligned}$$

where the second and last equality is due to stationarity of P_{12} and P_2, respectively. Now joint stationarity of P_{123} follows from (A.1) and stationarity of $P_{1|2}$, $P_{3|2}$ and P_2. □

B c-concave function

We review some basic results on c-concavity. See [16, 17, 15, 24, 25] for details.

Let E_1 and E_2 be two Polish spaces and $c : E_1 \times E_2 \to \mathbb{R}$ be a measurable function.

Definition B.1. *We define the c-transforms of functions f on E_1 and g on E_2 by*

$$f^c(y) := \inf_{x\in E_1}\{c(x,y)-f(x)\} \quad \textit{and} \quad g^c(x) := \inf_{y\in E_2}\{c(x,y)-g(y)\}.$$

A function f on E_1 is called c-concave if there exists some function g on E_2 such that $f(x) = g^c(x)$.

In general, $f^{cc} \geq f$ holds. Indeed, for any x and y, we have $c(x,y) - f^c(y) \geq f(x)$. Then $f^{cc}(x) = \inf_y\{c(x,y) - f^c(y)\} \geq f(x)$.

Lemma B.2. *Let f be a function of E_1. Then f is c-concave if and only if $f^{cc} = f$.*

Proof. The "if" part is obvious. We prove the "only if" part. Assume $f = g^c$. Then $f^c = g^{cc} \geq g$, and therefore

$$f^{cc}(x) = \inf_y\{c(x,y) - f^c(y)\} \leq \inf_y\{c(x,y) - g(y)\} = g^c(x) = f(x).$$

Since $f^{cc} \geq f$ always holds, we have $f^{cc} = f$. □

Define the c-supergradient of any function $f : E_1 \to \mathbb{R}$ by

$$\partial^c f(x) = \{y \in E_2 \mid c(x,y) - f(x) = f^c(y)\}.$$

Lemma B.3. *Assume that $\partial^c f(x) \neq \emptyset$ for any $x \in E_1$. Then f is c-concave.*

Proof. Fix $x \in E_1$ and let $y \in \partial^c f(x)$. Then we have

$$f(x) = c(x,y) - f^c(y) \geq f^{cc}(x) \geq f(x).$$

Hence $f^{cc} = f$ and thus f is c-concave. □

The converse of Lemma B.3 does not hold in general. For example, consider $E_1 = [0,\infty)$, $E_2 = \mathbb{R}$ and $c(x,y) = -xy$. Then c-concavity is equivalent to usual concavity. The function $f(x) = \sqrt{x}$ is concave but the supergradient at $x = 0$ is empty.

C Proof of Proposition 4.9

Consider the cost function $c(x,y)$ on $E_1 \times E_2$ with the assumptions in Proposition 4.9. Since $c_{x,y} \neq 0$, the map $y \mapsto c_x(x,y)$ is injective. Denote its image and inverse function by $U_x = \{c_x(x,y) \mid y \in E_2\}$ and $\eta_x = (c_x(x,\cdot))^{-1} : U_x \mapsto E_2$, respectively. Hence $c_x(x,\eta_x(u)) = u$ for all $u \in U_x$ and $\eta_x(c_x(x,y)) = y$ for all $y \in E_2$. Note that U_x is an interval and therefore convex. Also note that the subscript x of η_x does not mean the derivative. By symmetry, we can define $V_y = \{c_y(x,y) \mid x \in E_1\}$ and $\xi_y = (c_y(\cdot,y))^{-1} : V_y \mapsto E_1$.

We first characterize the c-gradient of a differentiable c-concave function f. Let $x \in E_1$ and $y \in \partial^c f(x)$. Then $c(x,y) - f(x) \leq c(z,y) - f(z)$ for any $z \in E_1$. By the tangent condition at $z = x$, we have $c_x(x,y) - f'(x) = 0$, or equivalently, $y = \eta_x(f'(x))$. Hence we have $\partial^c f(x) = \{\eta_x(f'(x))\}$. We denote the unique element also by $\partial^c f(x) = \eta_x(f'(x))$.

To prove Proposition 4.9, it is sufficient to show that the following conditions are equivalent:

(i) c is convex stable for any index n

(ii) The map $u \mapsto c(x,\eta_x(u)) - c(z,\eta_x(u))$ is convex for all $x, z \in E_1$.

(iii) $-c_{xx,yy} + c_{xx,y}c_{x,yy}/c_{x,y} \geq 0$.

We first prove (i) $\Leftrightarrow$ (ii). Assume (i). Let $\mathbb{Q}$ be the set of rational numbers. By the definition of convex stability, for any $u_0, u_1 \in U_x$ and $\lambda \in [0,1] \cap \mathbb{Q}$, the function

$$\phi(z) := (1-\lambda)c(z, \eta_x(u_0)) + \lambda c(z, \eta_x(u_1))$$

is c-concave. The c-gradient of ϕ is given by

$$\partial^c \phi(x) = \eta_x((1-\lambda)c_x(x, \eta_x(u_0)) + \lambda c_x(x, \eta_x(u_1))) = \eta_x((1-\lambda)u_0 + \lambda u_1).$$

Then c-concavity, $c(x, \partial^c\phi(x)) - \phi(x) \leq c(z, \partial^c\phi(x)) - \phi(z)$ for any z, is equivalent to

$$\begin{aligned} &c(x, \eta_x((1-\lambda)u_0 + \lambda u_1)) - c(z, \eta_x((1-\lambda)u_0 + \lambda u_1)) \\ &\quad \leq (1-\lambda)\{c(x, \eta_x(u_0)) - c(z, \eta_x(u_0))\} + \lambda\{c(x, \eta_x(u_1)) - c(z, \eta_x(u_1))\}. \end{aligned}$$

Since both hand side is continuous with respect to λ, (ii) is obtained. The converse is similar.

Next we prove (ii) $\Leftrightarrow$ (iii). Assume (ii). Fix $x, z \in E_1$ and $u_0 \in U_x$. Let $y_0 = \eta_x(u_0)$ and therefore $u_0 = c_x(x, y_0)$. Since $u \mapsto c(x, \eta_x(u)) - c(z, \eta_x(u))$ is convex for any z, its second derivative at $u = u_0$ is non-negative:

$$\begin{aligned} &\partial_u^2 \{c(x, \eta_x(u)) - c(z, \eta_x(u))\}\big|_{u=u_0} \\ &= \{c_{yy}(x, y_0) - c_{yy}(z, y_0)\}(\eta_x^{(1)}(u_0))^2 + \{c_y(x, y_0) - c_y(z, y_0)\}\eta_x^{(2)}(u_0) \\ &\geq 0. \end{aligned} \tag{C.1}$$

On the other hand, by differentiating the identity $c_x(x, \eta_x(u)) = u$ twice at $u = u_0$, we have

$$c_{x,yy}(x, y_0)(\eta_x^{(1)}(u_0))^2 + c_{x,y}(x, y_0)\eta_x^{(2)}(u_0) = 0.$$

Combining the two relations, we have

$$\left[\{c_{yy}(x, y_0) - c_{yy}(z, y_0)\} - \frac{c_{x,yy}(x, y_0)}{c_{x,y}(x, y_0)}\{c_y(x, y_0) - c_y(z, y_0)\}\right](\eta_x^{(1)}(u_0))^2 \geq 0.$$

Since $\eta_x^{(1)}(u_0) = 1/c_{x,y}(x, y_0) \neq 0$, we obtain

$$\{c_{yy}(x, y_0) - c_{yy}(z, y_0)\} - \frac{c_{x,yy}(x, y_0)}{c_{x,y}(x, y_0)}\{c_y(x, y_0) - c_y(z, y_0)\} \geq 0.$$

Now let $v_0 = c_y(x, y_0)$ and $v = c_y(z, y_0)$. Then $x = \xi_{y_0}(v_0)$ and $z = \xi_{y_0}(v)$ from the definition of ξ_y. We have

$$\{c_{yy}(\xi_{y_0}(v_0), y_0) - c_{yy}(\xi_{y_0}(v), y_0)\} - \frac{c_{x,yy}(\xi_{y_0}(v_0), y_0)}{c_{x,y}(\xi_{y_0}(v_0), y_0)}(v_0 - v) \geq 0. \tag{C.2}$$

This means convexity of the map $v \mapsto -c_{yy}(\xi_{y_0}(v), y_0)$. Hence its second derivative is non-negative. Therefore

$$-c_{xx,yy}(z, y_0)(\xi_{y_0}^{(1)}(v))^2 - c_{x,yy}(z, y_0)\xi_{y_0}^{(2)}(v) \geq 0.$$

On the other hand, by differentiating the identity $c_y(\xi_{y_0}(v), y_0) = v$ twice, we have

$$c_{xx,y}(z, y_0)(\xi_{y_0}^{(1)}(v))^2 + c_{x,y}(z, y_0)\xi_{y_0}^{(2)}(v) = 0.$$

Combining the two relations, we have

$$\left[-c_{xx,yy}(z, y_0) + \frac{c_{xx,y}(z, y_0)}{c_{x,y}(z, y_0)}c_{x,yy}(z, y_0)\right](\xi_{y_0}^{(1)}(v))^2 \geq 0. \tag{C.3}$$

Since $\xi_{y_0}^{(1)}(v) = 1/c_{x,y}(z, y_0) \neq 0$, we conclude

$$-c_{xx,yy}(z, y_0) + c_{x,yy}(z, y_0)\frac{c_{xx,y}(z, y_0)}{c_{x,y}(z, y_0)} \geq 0.$$

Since z and $y_0(= \eta_x(u_0))$ are arbitrary, we obtain (iii).

The proof of (iii) ⇒ (ii) is just the converse. First, (C.3) follows from (iii). Since (C.3) is the second derivative of the left hand side of (C.2), the convexity condition (C.2) follows. The condition (C.2) is equivalent to (C.1), and (C.1) is also equivalent to (ii). This completes the proof.

References

[1] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Commun. Pure Appl. Math.*, 44(4):375–417, 1991. MR-1100809

[2] J. A. Cuesta-Albertos, L. Rüschendorf, and A. Tuero-Diaz. Optimal coupling of multivariate distributions and stochastic processes. *Journal of Multivariate Analysis*, 46(2):335–361, 1993. MR-1240428

[3] Ph. Delanoë and Y. Ge. Locally nearly spherical surfaces are almost-positively c-curved. Preprint: arXiv:1009.3586, 2010. MR-2719556

[4] D. Feyel and A. S. Üstünel. Measure transport on Wiener space and the Girsanov theorem. *C. R., Math., Acad. Sci. Paris*, 334(11):1025–1028, 2002. MR-1913729

[5] D. Feyel and A. S. Üstünel. Monge Kantorovich measure transport and Monge–Ampere equation on Wiener space. *Probability Theory and Related Fields*, 128:347–385, 2004. MR-2036490

[6] A. Figalli, Y.-H. Kim, and R. J. McCann. When is multidimensional screening a convex program? To appear in: *J. Econom. Theory*, Preprint: arXiv: 0912.3033, 2010a.

[7] A. Figalli and L. Rifford. Continuity of optimal transport maps and convexity of injectivity domains on small deformations of 2-sphere. *Comm. Pure Appl. Math.*, 62(12):1670–1706, 2009. MR-2569074

[8] A. Figalli, L. Rifford, and C. Villani. Nearly round spheres look convex. To appear in: *Amer. J. Math.*, 2010b. MR-2729302

[9] W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta. Math.*, 177: 113–161, 1996. MR-1440931

[10] R. M. Gray, D. L. Neuhoff, and P. C. Shields. A generalization of Ornstein's $\bar{d}$ distance with applications to information theory. *Annals of Probability*, 3(2):315–318, 1975. MR-0368127

[11] Y. H. Kim and R. J. McCann. Towards the smoothness of optimal maps on Riemmanian submersions and Riemannian products (of round spheres in particular). To appear in: *J. Reine Angew. Math.*, Preprint; arXiv:0806.0351v1, 2008.

[12] X.-N. Ma, N. S. Trudinger, and X.-J. Wang. Regularity of potential functions of the optimal transportation problem. *Arch. Rational Mech. Anal.*, 177:151–183, 2005. MR-2188047

[13] R. J. McCann. Polar factorization of maps on riemannian manifolds. *Geom. Funct. Anal.*, 11 (3):589–608, 2001. MR-1844080

[14] D. S. Ornstein. An application of ergodic theory to probability theory. *Annals of Probability*, 1(1):43–58, 1973. MR-0348831

[15] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems. Vol. 1: Theory. Vol. 2: Applications.* Springer, 1998.

[16] L. Rüschendorf. Fréchet-bounds and their applications. In *Advances in Probability Distributions with given Marginals. Beyond the Copulas*, volume 67, pages 151–188. Math. Appl., 1991. MR-1215951

[17] L. Rüschendorf. Optimal solutions of multivariate coupling problems. *Applicationes Mathematicae*, 23(3):325–338, 1995. MR-1360058

[18] L. Rüschendorf and S. T. Rachev. A characterization of random variables with minimum L^2-distance. *Journal of Multivariate Analysis*, 32:48–54, 1990. MR-1035606

[19] T. Sei. Parametric modeling based on the gradient maps of convex functions. To appear in: *Annals of the Institute of Statistical Mathematics* with changed title: Gradient modeling for multivariate quantitative data; Preprint: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/data/2006/METR06-51.pdf, 2006.

[20] T. Sei. A structural model on a hypercube represented by optimal transport. To appear in: *Statistica Sinica*, Preprint: arXiv: 0901.4715, 2010a. MR-2827524

[21] T. Sei. Structural gradient model for time series. *Proceedings of the International Symposium on Statistical Analysis of Spatio-Temporal Data*, November 4–6, 2010, Kamakura, Japan, 2010b.

[22] T. Sei. A Jacobian inequality for gradient maps on the sphere and its application to directional statistics. To appear in: *Communications in Statistics – Theory and Methods*, Preprint: arXiv: 0906.0874, 2010c.

[23] A. S. Üstünel. Estimation for the additive Gaussian channel and Monge–Kantorovich measure transportation. *Stochastic Processes Appl.*, 117:1316–1329, 2007. MR-2343942

[24] C. Villani. *Topics in Optimal Transportation*. AMS, 2003. MR-1964483

[25] C. Villani. *Optimal Transport. Old and New.* Springer, 2009. MR-2459454

Acknowledgments. The authors are grateful to the anonymous referees for very careful reading and many detailed comments which led to an improvement of the paper.