Serie 5
Optimal transport, Fall semester
EPFL, Mathematics section, Dr. Xavier Fernández- Real

Exercise 5.1. Let $x_1, x_2, y_1, y_2 \in \mathbb{R}^d$, $x_1 \neq x_2$ and let

$$\mu = \frac{1}{2}\delta_{x_1} + \frac{1}{2}\delta_{x_2} \quad \text{and} \quad \nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}.$$

(i) Describe all maps transporting μ to ν; that is, such that $T_{\#}\mu = \nu$.

(ii) Describe all couplings of μ and ν; that is $\gamma \in \mathcal{P}(X \times Y)$ such that $(\pi_X)_{\#}\gamma = \mu$ and $(\pi_Y)_{\#}\gamma = \nu$.

(iii) Prove that, for any choice of continuous cost $c\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, there exists an optimal transport map (i.e., the optimal coupling has a map structure).

(iv) Assuming that x_1, Tx_1, x_2 and Tx_2 are not colinear, observe that for the linear cost $c(x, y) = |x - y|$, the corresponding optimal transport map does not cross trajectories.

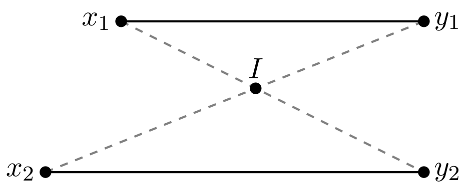

Figure 1: Crossing trajectories

Solution:

(i) We have

$$\frac{1}{2}\varphi(y_1) + \frac{1}{2}\varphi(y_2) = \int \nu\varphi = \int T_{\#}\mu\varphi = \frac{1}{2}\varphi(Tx_1) + \frac{1}{2}\varphi(Tx_2)$$

for all $\varphi \in C_c^\infty(\mathbb{R}^d)$. Suppose now $Tx_1 \notin \{y_1, y_2\}$; by taking φ with $\varphi(Tx_1) > 0$, $\varphi(y_1) = \varphi(y_2) = 0$ and $\varphi(Tx_2) = 0$ if $Tx_2 \neq Tx_1$, we reach a contradiction. Therefore, $Tx_1 \in \{y_1, y_2\}$. Then, we have $\varphi(\bar{y}) = \varphi(Tx_2)$ for some $\bar{y} \in \{y_1, y_2\}$ and for all φ, so $\bar{y} = Tx_2$. Thus, we have two possibilities:

$$\begin{cases} T_1x_1 = y_1 \\ T_1x_2 = y_2 \end{cases} \quad \text{or} \quad \begin{cases} T_1x_1 = y_1 \\ T_1x_2 = y_2. \end{cases}$$

(ii) Notice that $\operatorname{supp}\gamma \subseteq \bigcup_{i,j}(x_i, y_j)$. Indeed, let $(x, y) \notin \bigcup_{i,j}(x_i, y_j)$. Assume without loss of generality that x is different from x_1 and x_2. There is a neighborhood N of x containing neither x_1 nor x_2 and hence $\gamma(N \times \mathbb{R}^d) = 0$. This proves that $(x, y) \in (\mathbb{R}^d \times \mathbb{R}^d) \setminus \operatorname{supp} \gamma$. In particular,

$$\gamma = \alpha_{11}\delta_{x_1 y_1} + \alpha_{12}\delta_{x_1 y_2} + \alpha_{21}\delta_{x_2 y_1} + \alpha_{22}\delta_{x_2 y_2},$$

and γ can be represented by the matrix

$$A_\gamma = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}.$$

Notice that we must have $(\pi_X)_\# \gamma = \mu$, so since $(\pi_X)_\# \gamma = (\alpha_{11} + \alpha_{12})\delta_{x_1} + (\alpha_{21} + \alpha_{22})\delta_{x_2} = \mu$ this implies

$$\begin{cases} \alpha_{11} + \alpha_{12} = 1/2 \\ \alpha_{21} + \alpha_{22} = 1/2. \end{cases}$$

Similarly, since we have $(\pi_Y)_\# \gamma = \nu$,

$$\begin{cases} \alpha_{11} + \alpha_{21} = 1/2 \\ \alpha_{12} + \alpha_{22} = 1/2. \end{cases}$$

In particular, we can take $\alpha_{11} = \alpha$, and then $\alpha_{12} = \alpha_{21} = 1/2 - \alpha$ and $\alpha_{22} = \alpha$, that is

$$A_\gamma = \begin{pmatrix} \alpha & \frac{1}{2} - \alpha \\ \frac{1}{2} - \alpha & \alpha \end{pmatrix} \qquad \text{for some } \alpha \in [0, 1/2],$$

and all such possibilities of A_γ describe a coupling between μ and ν.

(iii) Let us consider

$$\min_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y)\, d\gamma(x, y).$$

Notice that it is well-defined, because γ is supported on $\bigcup_{i,j}(x_i, y_j)$ and c is bounded from below on this finite set. Notice also that by the previous part, the set of couplings is given by $\Gamma(\mu, \nu) = \bigcup_{\alpha \in [0,1/2]} \gamma_\alpha$, with γ_α represented by the matrix

$$A_{\gamma_\alpha} = \begin{pmatrix} \alpha & \frac{1}{2} - \alpha \\ \frac{1}{2} - \alpha & \alpha \end{pmatrix}, \quad \gamma_\alpha = \alpha\delta_{x_1 y_1} + \left(\frac{1}{2} - \alpha\right)\delta_{x_1 y_2} + \left(\frac{1}{2} - \alpha\right)\delta_{x_2 y_1} + \alpha\delta_{x_2 y_2}.$$

Thus, we want

$$\min_{\alpha \in [0,1/2]} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y)\, d\gamma_\alpha(x, y).$$

Let us split

$$A_\alpha := A_{\gamma_\alpha} = 2\alpha A_{1/2} + (1 - 2\alpha)A_0, \qquad \text{so} \qquad \gamma_\alpha = 2\alpha\gamma_{1/2} + (1 - 2\alpha)\gamma_0,$$

and

$$\operatorname{cost}(\gamma_\alpha) = \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y)\, d\gamma_\alpha(x, y) = 2\alpha \operatorname{cost}(\gamma_{1/2}) + (1 - 2\alpha) \operatorname{cost}(\gamma_0).$$

In particular, since $2\alpha \in [0, 1]$, $\operatorname{cost}(\gamma_\alpha) \geq \min\{\operatorname{cost}(\gamma_0), \operatorname{cost}(\gamma_{1/2})\}$ for all $\alpha \in [0, 1/2]$ and

γ_0 or $\gamma_{1/2}$ is an optimal coupling, which has a map structure (T_1 or T_2 from (i)).

(iv) We need to show that the segements $\overline{x_1Tx_1}$ and $\overline{x_2Tx_2}$ do not cross. Since $\{Tx_1, Tx_2\} = \{y_1, y_2\}$ if they were to cross, x_1, x_2, y_1, y_2 would be coplanar, thus it is enough to study the case $d = 2$. Notice that the cost to bring x_1 to Tx_1 is simply $|x_1 - Tx_1|$, thus the result follows by the triangular inequality. Indeed, if the trajectories were to cross at a point I (see figure 1 below), then have

$$|x_1 - y_1| + |x_2 - y_2| \leq |x_1 - I| + |I - y_1| + |x_2 - I| + |I - y_2| = |x_1 - y_2| + |x_2 - y_1|$$

where equality holds only if x_1, x_2, y_1 and y_2 are all colinear, a contradiction. This implies that the trajectories do not cross.

Exercise 5.2. Say if the following sentences are true or false. If they are true, prove it, if they are false, provide a counterexample. The statements below all refer to the quadratic cost.

(i) Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a convex function, then φ is differentiable $\mathcal{L}^d$-a.e. in $\mathbb{R}^d$ and we call $N \subset \mathbb{R}^d$ the Lebesgue measure zero set where φ is not differentiable. For any $x \in N$ take an element $y_x \in \partial\varphi(x)$, and define the map $T : \mathbb{R}^d \to \mathbb{R}^d$ as follows:

$$T(x) = \begin{cases} \nabla\varphi(x) & \text{if } x \in \mathbb{R}^d \setminus N, \\ y_x & \text{if } x \in N. \end{cases}$$

Then, given $\mu \ll \mathcal{L}^d$, the map T is optimal from μ to $T_{\#}\mu$.

(ii) If $T : \mathbb{R} \to \mathbb{R}$ is an optimal map between μ_1 and μ_2 (i.e. $T_{\#}\mu_1 = \mu_2$) and $S : \mathbb{R} \to \mathbb{R}$ is optimal between μ_2 and μ_3, then $S \circ T$ is optimal between μ_1 and μ_3.

(iii) The same as before but in general dimension $d \geq 2$, namely: if $T : \mathbb{R}^d \to \mathbb{R}^d$ is an optimal map between μ_1 and μ_2 (i.e. $T_{\#}\mu_1 = \mu_2$) and $S : \mathbb{R}^d \to \mathbb{R}^d$ is optimal between μ_2 and μ_3, then $S \circ T$ is optimal between μ_1 and μ_3.

Solution:

(i) True. The $\mathcal{L}^d$-a.e. diffentiability follows from the Rademacher Theorem. Then if we define $\gamma := (Id, T)_{\#}\mu$, then by construction we have that supp$\gamma \subset \partial\varphi$, from which thanks to Theorem 2.4.3 (see also Remark 2.4.4) we conclude that γ is optimal because it is contained in a c-cyclically monotone set.

Notice that if we defined for instance, for all $x \in N$, $T(x) = y$ for $y \notin \partial\varphi$ we should have proved also that $\mathrm{supp}(\gamma) = \mathrm{supp}(\gamma) \cap (\mathbb{R}^d \setminus N) \times \mathbb{R}^d \subset \partial\varphi$ where the first equality is not trivial.

(ii) It is true provided that μ_3 is non-atomic.

Suppose that μ_3 is non-atomic. Then, since there exist a transport map from μ_2 to μ_3, μ_2 is non-atomic too. With the same argument we get that μ_1 is non-atomic. As T and S are optimal maps, they have to coincide with the monotone rearrangements from μ_1 to μ_2 and from μ_2 to μ_3 respectively. In particular, $S \circ T$ is monotone, and of course is a transport map from μ_1 to μ_3. This means that $S \circ T$ is optimal from μ_1 to μ_3.

Now we provide a counterexample (suggested by Berk Ceylan) in the case in which μ_3 has atoms. Take

$$\begin{aligned}
\mu_1 &= \frac{1}{6}\delta_1 + \frac{1}{6}\delta_2 + \frac{1}{6}\delta_3 + \frac{1}{6}\delta_4 + \frac{1}{6}\delta_5 + \frac{1}{6}\delta_6,\\
\mu_2 &= \frac{1}{3}\delta_1 + \frac{1}{3}\delta_2 + \frac{1}{6}\delta_5 + \frac{1}{6}\delta_6,\\
\mu_3 &= \frac{1}{2}\delta_1 + \frac{1}{2}\delta_6.
\end{aligned}$$

It can be verified that the following maps are optimal for the Monge problem from μ_1 to μ_2 and from μ_2 to μ_3 respectively:

$$T(x) = \begin{cases} 1 & \text{if } x = 1, 2, \\ 2 & \text{if } x = 3, 4, \\ 5 & \text{if } x = 5, \\ 6 & \text{if } x = 6. \end{cases}, \qquad S(x) = \begin{cases} 1 & \text{if } x = 1, 5, \\ 6 & \text{if } x = 2, 6. \end{cases}$$

However, the map

$$S \circ T(x) = \begin{cases} 1 & \text{if } x = 1, 2, 5, \\ 6 & \text{if } x = 3, 4, 6 \end{cases}$$

is clearly non optimal from μ_1 to μ_3. The problem here is that when there are atoms, in general Monge's optimal maps are not minimizers for the Kantorovich problem. In particular, optimal maps can be non monotone.

(iii) False. We provide a counterexample for Dirac deltas, the exercise can be generalized to absolutely continuous measures.

Let $\mu_1 = \frac{\delta_{(1,0)}+\delta_{(-1,0)}}{2}$, $\mu_2 = \frac{\delta_{(-1,0)}+\delta_{(1,4)}}{2}$, $\mu_3 = \frac{\delta_{(-1,4)}+\delta_{(1,0)}}{2}$. Using point (i) of Exercise 5.1 we can directly prove that T_1 defined as

$$T_1(-1,0) = (-1,0) \qquad T_1(1,0) = (1,4)$$

and extended in whatever way is an optimal map from μ_1 to μ_2, as well as

$$T_2(-1,0) = (1,0) \qquad T_2(1,4) = (-1,4)$$

and extended in whatever way is an optimal map from μ_2 and μ_3, but $T_2 \circ T_1$ is not an optimal map from μ_1 to μ_3 because

$$T_2 \circ T_1(-1,0) = (1,0) \qquad T_2 \circ T_1(1,0) = (-1,4),$$

indeed the optimal map is T_3 from μ_1 to μ_3 is defined as

$$T_3(1,0) = (1,0) \quad T_3(-1,0) = (-1,4)$$

and extended in whatever way.

Exercise 5.3 (Birkhoff - Von Neumann Theorem)**.** A $(n \times n)$-matrix $A \in \mathscr{M}(n, \mathbb{R})$ with nonnegative entries is said to be:

- *a doubly-stochastic matrix* if $\sum_{i=1}^{n} A_{ij} = 1$ for any $j = 1, \ldots, n$, and $\sum_{j=1}^{n} A_{ij} = 1$ for any $i = 1, \ldots, n$.
- *a permutation matrix* if there is a permutation $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that $A_{i\sigma(i)} = 1$ and $A_{ij} = 0$ if $j \neq \sigma(i)$.

Prove that any doubly-stochastic matrix can be written as a finite convex combination of permutation matrices.

Hints: Here is a guideline through a possible proof of the result:

- Use Hall's marriage Theorem[1] to prove that given a doubly-stochastic matrix A, there exists a permutation $\sigma \in S_n$ such that $A_{i\sigma(i)} > 0$ for any $i = 1, \ldots, n$. Deduce that there exists a permutation matrix P and $\lambda > 0$ such that $A_{ij} \geq \lambda P_{ij}$, $\forall i, j \in \{1, \ldots, n\}$.
- Let us now prove the result by induction on the number of non-zero entries k of A. Start by proving that $k \geq n$ and that the result holds for $k = n$.
- Let now $k > n$. Consider the permutation P and λ given in the first bullet above, and define

$$A' = \frac{1}{1-\lambda}(A - \lambda P).$$

  Show that A' is doubly-stochastic with at most $k - 1$ non-zero entries.
- Deduce, by induction, that A is a convex combination of permutation matrices.

Solution: The solution is decomposed into four steps based on the hints.

Step 1: (Application of Hall's marriage theorem) Let us begin with the following lemma.

Lemma 1. Given a doubly-stochastic matrix A, there is a permutation $\sigma \in S_n$ such that $A_{i\sigma(i)} > 0$ for any $i = 1, \ldots, n$.

Proof. Let us construct a bipartite graph as follows: the graph consists of $2n$ vertices labeled by $\{1_r, \ldots, n_r\}$ and $\{1_c, \ldots, n_c\}$ (the indexes r, c stand for row and column). Then, we say that there is an edge between i_r and j_c if and only if $A_{ij} > 0$. We denote the presence of an edge with $i_r \sim j_c$. The first step of the proof consists in showing that such a bipartite graph admits a

[1] See `https://en.wikipedia.org/wiki/Hall%27s_marriage_theorem#Graph_theoretic_formulation`.

perfect matching (i.e., there is a permutation $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ such that $i_r \sim \sigma(i)_c$ for any $i = 1, \dots, n$). In order to do so, we want to apply Hall's marriage theorem. Given a subset $S \subset \{1, \dots, n\}$, let T be the subset defined as

$$T = \{t \in \{1, \dots, n\} : \ s_r \sim t_c \text{ for at least one } s \in S\}$$

Exploiting the fact that the matrix A is doubly-stochastic and the definition of T, we obtain

$$\#S = \sum_{s\in S} \sum_{j=1}^{n} A_{sj} = \sum_{s\in S} \sum_{t\in T} A_{st} \le \sum_{i=1}^{n} \sum_{t\in T} A_{it} = \sum_{t\in T} \sum_{i=1}^{n} A_{it} = \#T$$

Since we can choose S arbitrarily, the inequality $\#S \le \#T$ is exactly the hypothesis necessary to apply Hall's marriage theorem and deduce the existence of a perfect matching. Hence, by definition of perfect matching, there is a permutation σ such that $i_r \sim \sigma(i)_c$ for any $i = 1, \dots, n$. This last fact is equivalent to the desired statement. □

Step 2: We can now prove the statement of the theorem by induction on the number of nonzero entries of the matrix A.

Since A is doubly-stochastic, it is easy to see that it must have at least n nonzero entries. Moreover, if it has exactly n nonzero entries, then it must be already a permutation matrix.

Step 3: Let us assume that the number of nonzero entries of A is $k > n$. Let P^σ be the permutation matrix induced by the permutation σ (that is, $P^\sigma_{i\sigma(i)} = 1$ for all i, and $P^\sigma_{ij} = 0$ if $j \ne \sigma(i)$) whose existence is provided by the lemma. Let $\lambda > 0$ be the maximum value such that $\lambda P^\sigma \le A$ (the inequality must be understood entry-wise, namely $\lambda P^\sigma_{ij} \le A_{ij}$ for all i, j). Notice that, since A is doubly-stochastic, each entry of A is bounded by 1 and therefore $\lambda \le 1$. Also, it must be $\lambda < 1$, as otherwise A would have exactly n nonzero entries.

Let $A' := \frac{1}{1-\lambda}(A - \lambda P^\sigma)$. Since $\lambda P^\sigma \le A$, all entries of A' are nonnegative. Moreover, thanks to the choice of λ, the matrix A' has at most $k-1$ nonzero entries. Finally, one can easily check that A' is doubly-stochastic.

Step 4: By the inductive hypothesis, we are able to write A' as a convex combination of permutation matrices

$$A' = \sum_{i\in I} \lambda_i P^{\sigma_i}, \qquad \lambda_i \ge 0, \qquad \sum_{i\in I} \lambda_i = 1,$$

where I is a finite set of indices and P^{σ_i} are permutation matrices (induced by the permutations σ_i). From the definition of A', it follows that

$$A = \lambda P^\sigma + \sum_{i\in I} \lambda_i (1-\lambda) P^{\sigma_i},$$

thus A is a convex combination of permutation matrices.

Exercise 5.4 (Discrete optimal transport)**.** Given two families $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$ of points in $\mathbb{R}^d$, let $\mu := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu := \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. Prove that, for *any* choice of a continuous cost $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, there exists an optimal transport map from μ to ν.

Hint: Use Exercise 5.3 or Kantorovich duality.

Solution: We present two different solutions of this exercise, the first one uses Birkohoff-Von Neumann's theorem, whereas the second one borrows some ideas from the duality theory.

Note that c is bounded below on the finite set $\{(x_i, y_j)\}_{1\leq i,j\leq n}$, so it follows by Theorem 2.3.2 and Remark 2.3.3 that there exists an optimal coupling $\gamma \in \Gamma(\mu, \nu)$.

Let A_γ be the $n \times n$ matrix defined as $A_{ij} := \gamma(\{x_i, y_j\})$. From the marginal constraints on γ, it follows that nA is a doubly-stochastic matrix. Hence, applying Exercise 5.3, we can express nA as a convex combination of permutation matrices

$$nA = \sum_{k\in I} \lambda_k P^{\sigma_k} ,$$

where I is a finite set of indices, $\sum_{k\in I} \lambda_k = 1$, and P^{σ_k} are permutation matrices (induced by the permutations σ_k).

Let us define the cost of an $n \times n$ matrix B as

$$\mathcal{C}(B) := \sum_{i,j=1}^{n} B_{ij} c(x_i, y_j).$$

By definition, the cost $\mathcal{C}$ is linear and it holds

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} c(x, y)\, d\gamma(x, y) = \mathcal{C}(A) = \frac{1}{n} \sum_{k\in I} \lambda_k \mathcal{C}(P^{\sigma_k}) \geq \frac{1}{n} \min_{k\in I} \mathcal{C}(P^{\sigma_k}).$$

Hence, there is a permutation σ_k such that

$$\frac{1}{n} \sum_{i=1}^{n} c(x_i, y_{\sigma_k(i)}) = \frac{1}{n} \mathcal{C}(P^{\sigma_k}) \leq \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x, y)\, d\gamma(x, y),$$

and therefore the map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T(x_i) = y_{\sigma_k(i)}$ is optimal.

Solution: Without loss of generality (up to relabeling the indices of the points y_i), we can assume that the trivial permutation has the minimum cost among all permutations, that is

$$\sum_{i=1}^{n} c(x_i, y_i) \leq \sum_{i=1}^{n} c(x_i, y_\sigma(i)) \tag{1}$$

for any permutation $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$. Under this assumption, we want to prove that the map $T(x_i) := y_i$ is optimal, in the sense that the coupling induced by it is optimal (in the Kantorovich sense).

We now want to use Theorem 2.6.6. In fact, we only need to use the inequality

$$\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{X\times Y} c\, d\gamma \geq \sup_{\varphi(x)+\psi(y)+c(x,y)\geq 0} \int -\varphi\, d\mu + \int -\psi\, d\nu,$$

which follows immediately from the marginal condition (see the proof of (2.12)). In any case, it suffices to construct two functions $\varphi : \{x_1, \dots, x_n\} \to \mathbb{R}$ and $\psi : \{y_1, \dots, y_n\} \to \mathbb{R}$ such that

$$\varphi(x_i) + \psi(y_j) + c(x_i, y_j) \geq 0 \quad \text{for any } 1 \leq i, j \leq n, \tag{2}$$

$$\sum_{i=1}^{n} c(x_i, y_i) = \sum_{i=1}^{n} -\varphi(x_i) - \psi(y_i)\,. \tag{3}$$

Indeed, from these equations we get

$$\int_{\mathbb{R}^d} c(x, T(x))\, d\mu = \frac{1}{n}\sum_{i=1}^{n} c(x_i, y_i) \overset{(3)}{=} \int_{\mathbb{R}^d} -\varphi\, d\mu + \int_{\mathbb{R}^d} -\psi\, d\nu \overset{(2)}{\leq} \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c\, d\gamma,$$

so the optimality of T follows.

To prove (2)-(3), we claim that it suffices to construct a function φ such that

$$\varphi(x_j) - \varphi(x_i) \leq c(x_i, y_j) - c(x_j, y_j) =: b_{ij} \quad \text{for any } 1 \leq i, j \leq n. \tag{4}$$

Indeed, if the above bound holds, then (2)-(3) hold with the function ψ defined as $\psi(y_i) := -c(x_i, y_i) - \varphi(x_i)$ for any $1 \leq i \leq n$.

So, it remains only to construct a function φ such that (4) holds. To do this, let us consider the weighted oriented complete graph with vertices $\{1, \dots, n\}$ such that the weight of the edge $i \to j$ is b_{ij}, and denote with $d(i, j)$ the distance (i.e. the infimum of the sum of the weights of a path from the first vertex to the second one) between vertex i and vertex j (notice the similarity between this approach and the proof of Rockafellar's theorem, Theorem 2.5.2).

Let us check that, for any $1 \leq i, j \leq n$, it holds $d(i, j) > -\infty$. Since the graph consists of finitely many points, one can note that the distance between two vertices can be $-\infty$ if and only if there is a simple loop (that is, a closed path that visits each vertex at most once) $i_1, i_2, \dots, i_k$ with negative length, that is

$$b_{i_1 i_2} + b_{i_2 i_3} + \cdots + b_{i_k i_1} < 0\,. \tag{5}$$

To rule out this possibility, we have to use the optimality condition (1) (that we have never used until now). Let $\bar\sigma$ be the permutation such that $\bar\sigma(i_1) = i_2, \bar\sigma(i_2) = i_3, \dots, \sigma(i_k) = i_1$, and $\sigma(i) = i$ for all other values of i. Applying 1 with $\sigma = \bar\sigma$, we get

$$0 \leq \sum_{i=1}^{n} c(x_i, y_{\bar\sigma(i)}) - c(x_{\bar\sigma(i)}, y_{\bar\sigma(i)}) = \sum_{i=1}^{n} b_{i\bar\sigma(i)} = b_{i_1 i_2} + b_{i_2 i_3} + \cdots + b_{i_k i_1}\,,$$

which shows that (5) cannot hold.

Hence, we have proven that the distance $d(i, j)$ is finite for every $1 \leq i, j \leq n$. We now observe that, even if this notion of distance on a graph might be negative and not symmetric, it still satisfies the triangle inequality. Therefore we have

$$d(1, j) \leq d(1, i) + d(i, j) \leq d(1, i) + b_{ij} \qquad \forall\, i, j.$$

Hence, if we set $\varphi(x_i) := d(1, i)$ then the desired inequality (4) holds, concluding the proof.