PROVABLY CONVERGENT ALGORITHM FOR FREE-SUPPORT WASSERSTEIN BARYCENTER OF CONTINUOUS NON-PARAMETRIC MEASURES

ZEYI CHEN, ARIEL NEUFELD, AND QIKUN XIANG

ABSTRACT. We propose a provably convergent algorithm for approximating the Wasserstein barycenter of continuous non-parametric probability measures, and consider its application in probabilistic forecasts aggregation. Our algorithm is inspired by the fixed-point iterative scheme of Álvarez-Esteban, Del Barrio, Cuesta-Albertos, and Matrán (2016) whose convergence relies on obtaining optimal transport (OT) maps exactly which is computationally intractable for general non-parametric measures. To circumvent this difficulty, we develop tailored approximation techniques including consistent OT map estimators. Replacing the exact OT maps in the fixed-point iterative scheme with our estimated counterparts then gives rise to a computationally tractable stochastic fixed-point algorithm which is provably convergent to the true Wasserstein barycenter. Our algorithm remarkably does not restrict the support of the barycenter to be fixed and can be implemented in a distributed computing environment, which makes it suitable for large-scale information aggregation problems. In our numerical experiments, we apply the algorithm to aggregate the probabilistic forecasts on market sales and evaluate the aggregated forecast. Our results showcase that our algorithm is capable of developing high quality forecasts out of individual forecasts of lower quality, and thus harnesses the "wisdom of crowds".

Keywords: information aggregation, Wasserstein barycenter, optimal transport, shape-constrained regression

1. INTRODUCTION

Aggregating information from multiple heterogeneous data sources is commonly encountered in many application scenarios. Typical instances include coordinating amongst sensor networks in signal processing [36, 46], forming group consensus from expert judgements or forecasts in decision analysis [28, 69, 98], combining subsample posteriors for large datasets in Bayesian inferences and learning [11, 61], averaging hypotheses from base algorithms via ensemble models in supervised machine learning [18, 99, 101], etc. A common technique in information aggregation utilizes a type of barycenter of data sources represented in a prescribed metric space. [To be discussed: add more examples of barycenter?] In this paper, we are interested in the particular case of information aggregation where information from $K > 2$ data sources represented by probability measures $\nu_1, \dots, \nu_K$ on $\mathbb{R}^d$ are aggregated via their 2-Wasserstein barycenter ($\mathcal{W}_2$-Barycenter) [1] defined as follows.

Definition 1.1 ($\mathcal{W}_2$-distance and $\mathcal{W}_2$-barycenter [1])**.** *The 2-Wasserstein distance, or $\mathcal{W}_2$-distance, between two probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with finite second moments is defined via the following optimal transport problem (see, e.g., [78, 95, 96]) with squared-distance cost (here $\Pi(\mu,\nu)$ represents the set of couplings of μ and ν; see Definition 2.2 later):*

$$\mathcal{W}_2(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \, \pi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \right)^{\frac{1}{2}}. \tag{1.1}$$

For $\nu_1, \dots, \nu_K \in \mathcal{P}_2(\mathbb{R}^d)$, weights $w_1, \dots, w_K > 0$ satisfying $\sum_{k=1}^K w_k = 1$, and for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, let $V(\mu)$ denote the convex combination of the squared $\mathcal{W}_2$-distances between μ and $\nu_1, \dots, \nu_K$ given by

$$V(\mu) := \sum_{k=1}^K w_k \mathcal{W}_2(\mu, \nu_k)^2. \tag{1.2}$$

Then, $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ is called a $\mathcal{W}_2$-barycenter of $\nu_1, \dots, \nu_K$ with weights $w_1, \dots, w_K$ if

$$\bar{\mu} \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} V(\mu).$$

Essentially, the $\mathcal{W}_2$-distance between two probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as the minimal transport cost of moving probability mass from μ to ν under the squared-distance cost function, which induces a metric on the space of probability measures with finite second moments that metrizes the weak convergence;

Date: August 4, 2024.

see, e.g., [96, Theorem 6.9]. The seminal work of Agueh and Carlier [1] established the existence and uniqueness results of $\mathcal{W}_2$-barycenter together with its characterization under mild assumptions. Due to its appealing geometric and statistical properties, there soon emerged a series of subsequent studies developing the computational aspects of the $\mathcal{W}_2$-barycenter and its associated variants; see Section 1.1 for a review. Moreover, the $\mathcal{W}_2$-barycenter has been serving as a powerful tool in widespread applications in terms of distribution aggregation and representation tasks, including but not limited to computer graphics [76, 83], statistical machine learning [33, 74, 100], natural language processing [63, 81], theoretical economics [68, 73], Bayesian statistics [10, 85, 86], network analysis [84], financial risk management [8, 70], etc. In the sequel, we consider $\mathcal{W}_2$-barycenters with identical weights $w_1 = \cdots = w_K = \frac{1}{K}$ for simplicity, but we remark that all our results naturally generalize to the case with non-identical weights.

Our work was inspired by the work of Álvarez-Esteban, Del Barrio, Cuesta-Albertos, and Matrán [5], which demonstrated that the $\mathcal{W}_2$-barycenter of absolutely continuous probability measures $\nu_1, \dots, \nu_K \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ can be computed via a fixed-point of the operator $G : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ defined by

$$G(\mu) := \left[\tfrac{1}{K} \textstyle\sum_{k=1}^{K} T^{\mu}_{\nu_k}\right] \sharp \mu \qquad \forall \mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d), \tag{1.3}$$

where $T^{\mu}_{\nu_k}$ corresponds to Monge's optimal transport (OT) map from μ to ν_k (see Brenier's theorem discussed later in Theorem 2.4), i.e.,

$$T^{\mu}_{\nu_k} \in \operatorname*{arg\,min}_{T} \left\{ \int_{\mathbb{R}^d} \big\| \boldsymbol{x} - T(\boldsymbol{x}) \big\|^2 \mu(\mathrm{d}\boldsymbol{x}) : T : \mathbb{R}^d \to \mathbb{R}^d \text{ is Borel measurable and } T\sharp\mu = \nu_k \right\}.$$

We summarize the nice properties of the G-operator developed by Álvarez-Esteban et al. [5] in the following theorem.

Theorem 1.2 (Properties of the G-operator [5, Corollary 3.5 and Theorem 3.6])**.** *Let $\nu_1, \dots, \nu_K \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. The G-operator defined in (1.3) satisfies the following properties.*

(i) $G : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is continuous with respect to the $\mathcal{W}_2$-metric.

(ii) The unique $\mathcal{W}_2$-barycenter $\bar{\mu} \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ (see Theorem 2.3 later) of $\nu_1, \dots, \nu_K$ is a fixed-point of G, i.e., $\bar{\mu} = G(\bar{\mu})$.

(iii) For any $\mu_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, the sequence $(\mu_t)_{t\in\mathbb{N}}$ generated by the iteration

$$\mu_{t+1} := G(\mu_t) \qquad \forall t \in \mathbb{N}_0 \tag{1.4}$$

is tight. Moreover, every accumulation point of the sequence $(\mu_t)_{t\in\mathbb{N}_0}$ with respect to the $\mathcal{W}_2$-metric is a fixed-point of G.

Theorem 1.2(iii) thus leads to a simple iterative scheme for $\mathcal{W}_2$-barycenter, where one begins with an arbitrary $\mu_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ and iterates (1.4) to generate $(\mu_t)_{t\in\mathbb{N}_0}$. When G has a unique fixed-point, Theorem 1.2(ii) and Theorem 1.2(iii) then guarantee that $(\mu_t)_{t\in\mathbb{N}_0}$ converges in $\mathcal{W}_2$ to the $\mathcal{W}_2$-barycenter of $\nu_1, \dots, \nu_K$. However, when $\nu_1, \dots, \nu_K$ are general non-parametric probability measures, the iteration (1.4) is a theoretical but computationally intractable "oracle" since it suffers from the difficulty in computing the OT map $T^{\mu}_{\nu_k}$ exactly. Therefore, numerical implementations of this scheme are either limited to particular parametric measures (see, e.g., [5, Section 4]), or carried out via neural network approximations without convergence guarantees (see, e.g., [52]). This drawback has motivated our development of an estimator-based stochastic extension of this deterministic fixed-point iterative scheme with rigorous convergence guarantee. The idea of our stochastic fixed-point iterative scheme is sketched in Conceptual Algorithm 1, where we approximate each true OT map $T^{\widehat{\mu}_t}_{\nu_k}$ by an OT map estimator $\widehat{T}_{t+1,k}$ (Line 6) and approximate the G-operator when updating from $\widehat{\mu}_t$ to $\widehat{\mu}_{t+1}$ (Line 7). In particular, letting $\widehat{T}_{t+1,k} = T^{\widehat{\mu}_t}_{\nu_k}$ in Line 6 and letting $\widehat{\mu}_{t+1} = \left[\frac{1}{K}\sum_{k=1}^{K} \widehat{T}_{t+1,k}\right]\sharp\widehat{\mu}_t$ in Line 7 recovers the deterministic fixed-point iterative scheme. Our objective is to develop a concrete setting as well as a computationally tractable implementation of Conceptual Algorithm 1 such that the resultant stochastic sequence of probability measures $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$ will converge to the $\mathcal{W}_2$-barycenter of $\nu_1, \dots, \nu_K$ in an appropriate sense.

[To discuss: shall we highlight the main difficulties?]

Specifically, our contributions can be summarized as follows:

(i) We provide a computationally tractable stochastic fixed-point algorithm (i.e., Algorithm 2) for approximately computing the $\mathcal{W}_2$-barycenter of $\nu_1, \dots, \nu_K$. Unlike most existing approximation approaches,

Conceptual Algorithm 1: Stochastic fixed-point iterative scheme

```
Input: $K$ input measures $\nu_1, \dots, \nu_K \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, initial measure $\mu_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$.
Output: $(\widehat{\mu}_t)_{t \in \mathbb{N}_0}$.
Initialize $\widehat{\mu}_0 \leftarrow \mu_0$.
for $t = 0, 1, 2, \dots$ do
    for $k = 1, \dots, K$ do
        Randomly generate $N_{t,k}$ i.i.d. samples $\{\boldsymbol{X}_{t+1,k,i}\}_{i=1:N_{t,k}}$ from $\widehat{\mu}_t$.
        Randomly generate $N_{t,k}$ i.i.d. samples $\{\boldsymbol{Y}_{t+1,k,i}\}_{i=1:N_{t,k}}$ from $\nu_k$.
        Approximate $T_{\nu_k}^{\widehat{\mu}_t}$ with an estimator $\widehat{T}_{t+1,k} \approx T_{\nu_k}^{\widehat{\mu}_t}$ using the samples $\{\boldsymbol{X}_{t+1,k,i}\}_{i=1:N_{t,k}}$ and
          $\{\boldsymbol{Y}_{t+1,k,i}\}_{i=1:N_{t,k}}$.
    Choose $\widehat{\mu}_{t+1} \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ such that $\widehat{\mu}_{t+1} \approx \big[\frac{1}{K}\sum_{k=1}^{K} \widehat{T}_{t+1,k}\big]\sharp\widehat{\mu}_t$.
return $(\widehat{\mu}_t)_{t \in \mathbb{N}_0}$.
```

we neither restrict the support of the approximate $\mathcal{W}_2$-barycenter to be a finite collection of points, nor restrict the input measures $\nu_1, \dots, \nu_K$ to be discrete or to specific parametric families of measures. Instead, our algorithm belongs to the class of "free-support" approaches and adapts to general continuous non-parametric probability measures as inputs.

(ii) We perform rigorous convergence analysis of our algorithm to show that it converges to the true $\mathcal{W}_2$-barycenter of $\nu_1, \dots, \nu_K$ in an almost sure sense (see Setting 3.12 and Theorem 3.13). To the best of our knowledge, our algorithm is the first computationally tractable extension of the fixed-point iterative scheme by Álvarez-Esteban et al. [5] with convergence guarantee.

(iii) [To discuss: are our estimators considered as contributions? One novelty is that we incorporate smoothing techniques into the classical shape-constrained convex least squares estimator to make them infinitely differentiable.]

(iv) We demonstrate via numerical experiment that our algorithm is effective and can be implemented in a distributed and paralleled computing environment, which has the potential for scalable and efficient applications in distributed systems under diverse application scenarios; see, e.g., [13] for a comprehensive reference on the related topics.

1.1. **Literature review.** In this subsection, we review two main streams of literature related to our study, namely the numerical methods for approximating Wasserstein barycenters and the statistical estimation approaches of the optimal transport map.

[To be done: detailed comparison with a few existing approaches]

Numerical methods for approximating $\mathcal{W}_2$-barycenters. Recent years have witnessed an extensive and rapid growth in the literature on numerically computing $\mathcal{W}_2$-barycenters since the seminal work by Agueh and Carlier [1]. Exact or regularized methods applicable to specific parametric families of probability measures have been developed, such as Gaussian [1, 25], the location-scatter family [5], or φ-exponential distributions [53]. For non-parametric measures, many studies have developed efficient algorithms for the case when the input measures are discrete with finite supports on $\mathbb{R}^d$; see, e.g., [7, 12, 31, 38, 42, 44, 47, 56, 76, 93], etc. In particular, given discrete measures, Borgwardt [15, Proposition 1] showed the existence of a sparsely-supported discrete $\mathcal{W}_2$-barycenter, and Altschuler and Boix-Adsera [2] and Altschuler and Boix-Adsera [3] proved that the $\mathcal{W}_2$-barycenter can be computed in polynomial time for any fixed d yet has exponential dependence on d.

Regarding possibly continuous input measures, a popular strategy is to discretize the support of the underlying $\mathcal{W}_2$-barycenter over a fixed number of atoms hence it suffices to optimize the histogram weights over a finite-dimensional simplex, which is computationally favorable; see a variety of relevant algorithms in, e.g.,[27, 35, 87, 89]. Despite their advantages in computational speed, such "fixed-support" algorithms incorporates inductive biases since the true support of the barycenter is unknown a priori. In contrast, "free-support" algorithms impose no restrictions on the support of the underlying Wasserstein barycenter, and approximative approaches include alternating optimization with Newton's method [31], stochastic gradient descent under entropic or quadratic regularization [54], approximation schemes via variational distributions [26], Sinkhorn divergence based Frank–Wolfe algorithm [57], etc. Recently, there exhibits a growing interest in the numerical methods for approximating continuous $\mathcal{W}_2$-barycenters leveraging neural network parametrization or generative neural networks, while the non-convex training objectives pose challenges to further theoretical analyses; see, e.g., [24, 29, 51, 52]. To provide theoretic insights, some studies detour to investigate algorithms in tackling the more general MMOT problem and view numerically solving the $\mathcal{W}_2$-barycenter a specific application of it; see, e.g., [4, 67, 68, 97], etc.

Estimations of the optimal transport map. In this thesis, we are specially interested in estimators of the optimal transport (OT) map as they constitute a crucial ingredient of our proposed stochastic fixed-point algorithm. The problem of finding an optimal measure-preserving map without mass splits dates back to Monge's formulation in [62], which could be infeasible. Studies by Knott and Smith [50] and Brenier [19] showed that, for arbitrary measures $\mu, \nu \in \mathcal{P}_2(\mathcal{X})$ where $\mathcal{X}$ is a Polish space, an OT map always exists and is uniquely determined μ-almost everywhere as the gradient of a convex function referred as the Brenier potential (up to a constant shift), whenever μ is absolutely continuous with respect to the Lebesgue measure (see details in Theorem 2.4). These classical results provide a useful perspective for computing the Wasserstein distance between measures by solving the underlying OT map, which turns out to be the object of interest in many statistical and machine learning applications on its own right; see a comprehensive review by Peyré et al. [74]. However, computing the true OT map has proved to be exceptionally hard and suffer from the curse of dimensionality; see, e.g., [88, 90] for rigorous complexity analyses. As such, many studies focus on designing statistically well-behaved estimators to the OT map out of independent samples from respective probability measures.

In literature, various efficient heuristic and learning-based methods were discussed for the computational sake; see, e.g., [48, 58, 72]. Recently, diverse types of OT map estimators with rigorous statistical guarantees have been proposed. Under pre-specified regularity assumptions, Hütter and Rigollet [45] disclosed a minimax $\mathcal{L}^2(\mu)$-convergence rate (of increasing data) for any measurable function of samples as a lower bound (see Theorem 6 therein), and proposed a near-optimal OT map estimator via the truncated wavelet approximation. On the other hand, Gunsilius [41] obtained from kernel density estimations an upper bound on the $\mathcal{L}^2(\mu)$-risk of plug-in estimators extended over $\mathbb{R}^d$. Subsequently, Pooladian and Niles-Weed [75] and Deb, Ghosal, and Sen [32] derived distinguished minimax optimal map candidates via barycentric projection techniques [6, Definition 5.4.2] in regularized and non-regularized settings. Manole, Balakrishnan, Niles-Weed, and Wasserman [59] sharpened the upper bound risk in [41] and provided in addition so-called "plug-in" estimators, which are built upon the optimal transport plan when the source and target measures are replaced by their empirical counterparts yet extended over $\mathbb{R}^d$. Moreover, by employing kernel sum-of-squares [14, Chapter 3] as building blocks, Vacher, Muzellec, Rudi, Bach, and Vialard [94] and Muzellec, Vacher, Bach, Vialard, and Rudi [64] proposed estimators with comparable $\mathcal{L}^2(\mu)$-convergence rate plus dimension-free computational complexity up to a constant term potentially exponential on d. Regardless of the convergence rate issues, another remarkable stream of works developed a class of plug-in estimators built from smooth and strongly convex regression and interpolation (see recent advances in, e.g., [55, 60, 79, 91, 92], and references therein) to approximate OT maps, in light of the underlying geometric properties of Brenier potential as implied by the prominent Caffarelli's regularity theory [22]. For instance, Paty, d'Aspremont, and Cuturi [71] formulated a quadratically constrained quadratic program (QCQP) for approximating Brenier potential leveraging the convex interpolability framework developed by Taylor [91]; Curmei and Hall [30] developed a hierarchy of semi-definite programs against the shape constraints using sum-of-square polynomials; González-Sanz, De Lara, Béthune, and Loubes [40] deployed state-of-the-art Lipschitz-constrained generative adversarial networks (GAN) for the regression task.

1.2. **Organization.** This paper is organized as follows. Section 2 introduces the notations used in this paper and crucial preliminary results on which our contributions are based. Section 3 presents our stochastic fixed-point algorithm for approximating the $\mathcal{W}_2$-barycenter. We also perform detailed analysis of its convergence. In Section 4, we develop two concrete plug-in OT map estimators that can guarantee the convergence of our stochastic fixed-point algorithm. Finally, we compare our proposed algorithm and estimators with other state-of-the-art algorithms in $\mathcal{W}_2$-barycenter approximations through numerical experiments in Section 5, thus verify their efficacy.

2. Notations and Preliminaries

2.1. **Notations.** In the following, we introduce the terminologies and notations that are used throughout this paper. All vectors are assumed to be column vectors and are denoted by boldface symbols. In particular, for

$k \in \mathbb{N}$, $\mathbf{0}_k$ denotes the vector in $\mathbb{R}^k$ with all entries equal to zero. We also use $\mathbf{0}$ when the dimension can be inferred from the context. We denote by $\langle \cdot, \cdot \rangle$ the Euclidean dot product, i.e., $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^\mathsf{T}\boldsymbol{y}$ and we denote by $\|\cdot\|$ the Euclidean norm, i.e., $\|\boldsymbol{x}\| := (\langle \boldsymbol{x}, \boldsymbol{x}\rangle)^{\frac{1}{2}}$. Open and closed balls centered at $\boldsymbol{x}$ with radius r are denoted by $B(\boldsymbol{x}, r)$ and $\bar{B}(\boldsymbol{x}, r)$, respectively. For any set $\mathcal{X} \subseteq \mathbb{R}^d$, we let $\mathrm{int}(\mathcal{X})$, $\mathrm{cl}(\mathcal{X})$, and $\mathrm{bd}(\mathcal{X})$ denote its interior, closure, and boundary, respectively. Moreover, for two matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \succeq \mathbf{B}$ indicates that $\mathbf{A} - \mathbf{B}$ is positive semi-definite. Furthermore, for $k \in \mathbb{N}$, $\mathbf{I}_k$ denotes the k-by-k identity matrix.

For a closed subset $\mathcal{X}$ of a Euclidean space, let $\mathcal{B}(\mathcal{X})$ denote the Borel subsets of $\mathcal{X}$, and let $\mathcal{P}(\mathcal{X})$ denote the set of Borel probability measures on $\mathcal{X}$, while $\mathcal{P}_2(\mathcal{X}) \subseteq \mathcal{P}(\mathcal{X})$ consists of the ones with finite second moments. The associated set $\mathcal{P}_{2,\mathrm{ac}}(\mathcal{X})$ contains probability measures in $\mathcal{P}_2(\mathcal{X})$ which are absolutely continuous with respect to the Lebesgue measure. For any $\mu \in \mathcal{P}(\mathcal{X})$ and any $\mathcal{Y} \in \mathcal{B}(\mathcal{X})$ with $\mu(\mathcal{Y}) > 0$, let $\mu|_{\mathcal{Y}}$ denote the probability measure formed by truncating μ to $\mathcal{Y}$, i.e., $\mu|_{\mathcal{Y}}(A) := \frac{\mu(\mathcal{Y}\cap A)}{\mu(\mathcal{Y})}$ for all $A \in \mathcal{B}(\mathcal{X})$. Moreover, for two closed subsets $\mathcal{X}, \mathcal{Y}$ of Euclidean spaces and a Borel measurable function $T : \mathcal{X} \to \mathcal{Y}$, the pushforward of a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ by T is denoted by $T\sharp\mu$, i.e., $T\sharp\mu := \mu \circ T^{-1}$.

Let us introduce the notations for the following families of functions. For an open bounded set $\mathcal{X} \subset \mathbb{R}^d$ and for $k \in \mathbb{N}_0$, $\alpha \in (0, 1]$, let $\mathcal{C}^k(\mathrm{cl}(\mathcal{X}))$ denote the set of $\mathbb{R}$-valued continuous functions on $\mathrm{cl}(\mathcal{X})$ that are k-times continuously differentiable on $\mathcal{X}$, and let $\mathcal{C}^{k,\alpha}(\mathrm{cl}(\mathcal{X}))$ denote the set of $\mathbb{R}$-valued continuous functions on $\mathrm{cl}(\mathcal{X})$ that are k-times continuously differentiable on $\mathcal{X}$ whose k-th order partial derivatives are α-Hölder continuous. In particular, $\mathcal{C}^{k,\alpha}(\mathrm{cl}(\mathcal{X}))$ is a Banach space with respect to the following norm (see, e.g., Evans [37, Theorem 5.1.1]):

$$\|\varphi\|_{\mathcal{C}^{k,\alpha}(\mathrm{cl}(\mathcal{X}))} := \max_{|\boldsymbol{\beta}|\le k} \sup_{\boldsymbol{x}\in\mathcal{X}} \big|\partial^{\boldsymbol{\beta}}\varphi(\boldsymbol{x})\big| + \max_{|\boldsymbol{\beta}|=k} \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}} \frac{\big|\partial^{\boldsymbol{\beta}}\varphi(\boldsymbol{x}) - \partial^{\boldsymbol{\beta}}\varphi(\boldsymbol{y})\big|}{\|\boldsymbol{x}-\boldsymbol{y}\|^{\alpha}} \qquad \forall \varphi \in \mathcal{C}^{k,\alpha}(\mathrm{cl}(\mathcal{X})),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{N}_0^d$ is a multi-index, $|\boldsymbol{\beta}| := \beta_1 + \cdots + \beta_d$, and $\partial^{\boldsymbol{\beta}}\varphi := \frac{\partial^{|\boldsymbol{\beta}|}\varphi}{\partial x_1^{\beta_1}\cdots\partial x_d^{\beta_d}}$ denotes the partial derivative of φ with respect to the multi-index $\boldsymbol{\beta}$. We refer to $\mathcal{C}^{k,\alpha}(\mathrm{cl}(\mathcal{X}))$ as the set of (k, α)-Hölder functions on $\mathrm{cl}(\mathcal{X})$. Moreover, let $\mathcal{C}^{k,\alpha}_{\mathrm{loc}}(\mathbb{R}^d)$ denote the set of $\mathbb{R}$-valued functions on $\mathbb{R}^d$ that are (k, α)-Hölder when restricted to $\mathrm{cl}(\mathcal{X})$ for any bounded open set $\mathcal{X} \subset \mathbb{R}^d$. We refer to $\mathcal{C}^{k,\alpha}_{\mathrm{loc}}(\mathbb{R}^d)$ as the set of locally (k, α)-Hölder functions on $\mathbb{R}^d$. In addition, let $\mathcal{C}^\infty(\mathbb{R}^d)$ denote the set of infinitely differentiable $\mathbb{R}$-valued functions on $\mathbb{R}^d$. Lastly, we denote by $\mathcal{C}_{\mathrm{lin}}(\mathbb{R}^d, \mathbb{R}^d)$ the set of continuous functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ that have at most linear growth, i.e., $T \in \mathcal{C}_{\mathrm{lin}}(\mathbb{R}^d, \mathbb{R}^d)$ if and only if $\sup_{\boldsymbol{x}\in\mathbb{R}^d} \frac{\|T(\boldsymbol{x})\|}{1+\|\boldsymbol{x}\|} < \infty$. Note that $\mathcal{C}_{\mathrm{lin}}(\mathbb{R}^d, \mathbb{R}^d)$ is a Banach space with respect to the norm $\|T\|_{\mathcal{C}_{\mathrm{lin}}(\mathbb{R}^d,\mathbb{R}^d)} := \sup_{\boldsymbol{x}\in\mathbb{R}^d} \frac{\|T(\boldsymbol{x})\|}{1+\|\boldsymbol{x}\|}$ $\forall T \in \mathcal{C}_{\mathrm{lin}}(\mathbb{R}^d, \mathbb{R}^d)$.

2.2. **Smooth and strongly convex functions.**

We give an overview of the notion of smooth and strongly convex functions that is frequently used in our discussions.

Definition 2.1 (Smooth and strongly convex functions)**.** *For $0 \le \underline{l} \le \bar{l} \le \infty$, a proper, lower semi-continuous (l.s.c.) and convex function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is called $\bar{l}$-smooth if*

$$\varphi(\boldsymbol{y}) \le \varphi(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{\bar{l}}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,\ \forall \boldsymbol{g} \in \partial\varphi(\boldsymbol{x}),$$

and is called $\underline{l}$-strongly convex if

$$\varphi(\boldsymbol{y}) \ge \varphi(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{\underline{l}}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,\ \forall \boldsymbol{g} \in \partial\varphi(\boldsymbol{x}).$$

We denote by $\mathfrak{C}_{\underline{l},\bar{l}}(\mathbb{R}^d)$ the collection of proper l.s.c. convex functions on $\mathbb{R}^d$ which are $\bar{l}$-smooth and $\underline{l}$-strongly convex. In particular, $\mathfrak{C}_{0,\infty}(\mathbb{R}^d)$ contains all proper l.s.c. convex functions on $\mathbb{R}^d$. Moreover, we denote $\mathfrak{C}^\infty_{\underline{l},\bar{l}}(\mathbb{R}^d) := \mathcal{C}^\infty(\mathbb{R}^d) \cap \mathfrak{C}_{\underline{l},\bar{l}}(\mathbb{R}^d)$, $\mathfrak{C}^k_{\underline{l},\bar{l}}(\mathbb{R}^d) := \mathcal{C}^k(\mathbb{R}^d) \cap \mathfrak{C}_{\underline{l},\bar{l}}(\mathbb{R}^d)$, $\mathfrak{C}^{k,\alpha}_{\underline{l},\bar{l}}(\mathbb{R}^d) := \mathcal{C}^{k,\alpha}_{\mathrm{loc}}(\mathbb{R}^d) \cap \mathfrak{C}_{\underline{l},\bar{l}}(\mathbb{R}^d)$ for $k \in \mathbb{N}_0$, $\alpha \in (0, 1]$. It follows from classical results (see, e.g., [65, Lemma 1.2.3 & Theorem 2.1.5]) that for $\bar{l} < \infty$, every $\varphi \in \mathfrak{C}_{0,\bar{l}}(\mathbb{R}^d)$ is continuously differentiable and $\nabla\varphi$ is $\bar{l}$-Lipschitz continuous.

2.3. **Optimal transport and Wasserstein distance.**

Many of our discussions in this paper are established upon results from optimal transport theory, especially properties around the Wasserstein distance between probability measures; see, e.g., the seminal work of Villani [95, 96] and Santambrogio [78]. We start by introducing the notion of couplings, which is detailed as follows.

Definition 2.2 (Coupling)**.** *Given $m \in \mathbb{N}$ probability measures $\nu_1 \in \mathcal{P}(\mathcal{X}_1), \dots, \nu_m \in \mathcal{P}(\mathcal{X}_m)$ on closed subsets $\mathcal{X}_1, \dots, \mathcal{X}_m$ of $\mathbb{R}^d$, the set of couplings of $\nu_1, \dots, \nu_m$ is denoted by $\Pi(\nu_1, \dots, \nu_m)$, which is defined as*

$$\Pi(\nu_1, \dots, \nu_m) := \big\{\pi \in \mathcal{P}(\mathcal{X}_1 \times \cdots \times \mathcal{X}_m) : \text{ the marginal of } \pi \text{ on } \mathcal{X}_i \text{ is } \nu_i \text{ for } i = 1, \dots, m\big\}.$$

The minimization problem embedded in the formulation (1.1) is known as Kantorovich's optimal transport problem [49] with respect to the squared-distance cost, and the infimum is well known to be attained by an optimal coupling; see, e.g., [96, Theorem 4.1]. In the rest of this paper, the optimality of a coupling is always considered with respect to the squared-distance cost.

The existence of a $\mathcal{W}_2$-barycenter is shown by Agueh and Carlier [1, Proposition 2.3]. In general, there may exist more than one $\mathcal{W}_2$-barycenters of $\nu_1, \nu_2, \dots, \nu_K$. A sufficient condition to guarantee the uniqueness of the $\mathcal{W}_2$-barycenter is given as follows.

Theorem 2.3 ([1, Proposition 3.5 & Theorem 5.1])**.** *Among $\nu_1, \dots, \nu_K \in \mathcal{P}_2(\mathbb{R}^d)$, if there exists at least one index $k \in \{1, \dots, K\}$ such that $\nu_k \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, then the $\mathcal{W}_2$-barycenter $\bar{\mu}$ in Definition 1.1 is unique. Moreover, if there exists at least one index $k \in \{1, \dots, K\}$ such that ν_k has bounded density, then the unique $\bar{\mu} \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$.*

Next, let us present Brenier's theorem which characterizes optimal couplings with gradient of convex functions when the source measure μ is absolutely continuous with respect to the Lebesgue measure.

Theorem 2.4 (Brenier's theorem [19] & [95, Theorem 2.12])**.** *Let $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Then, there is a unique optimal coupling $\pi^\star \in \Pi(\mu, \nu)$ that minimizes (1.1). Moreover, $\pi \in \Pi(\mu, \nu)$ minimizes (1.1) if and only if there exists a lower semi-continuous (l.s.c.) and convex function $\varphi_\nu^\mu : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ such that $\pi = \big[I_d, T_\nu^\mu\big]\sharp\mu$ where $I_d : \mathbb{R}^d \to \mathbb{R}^d$ denotes the identity map on $\mathbb{R}^d$ and $T_\nu^\mu = \nabla\varphi_\nu^\mu$ is the gradient of φ_ν^μ (that is uniquely determined μ-almost everywhere). In this case, T_ν^μ is also the μ-almost everywhere unique optimal solution of Monge's optimal transport problem:*

$$\inf\left\{\int_{\mathbb{R}^d} \left\|\boldsymbol{x} - T(\boldsymbol{x})\right\|^2 \mu(\mathrm{d}\boldsymbol{x}) : T : \mathbb{R}^d \to \mathbb{R}^d \text{ is Borel measurable and } T\sharp\mu = \nu\right\}.$$

We refer to $\varphi_\nu^\mu : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and $T_\nu^\mu : \mathbb{R}^d \to \mathbb{R}^d$ in Theorem 2.4 as the optimal Brenier potential from μ to ν and the optimal transport (OT) map from μ to ν, respectively. Note that the optimal Brenier potential φ_ν^μ is μ-almost everywhere uniquely determined up to the addition of an arbitrary constant.

The convergence of our proposed algorithm requires regularity properties of the OT map T_ν^μ. Regarding this matter, a series of studies by Caffarelli [20, 21, 22, 23] developed foundations of the regularity theory of OT maps under suitable geometric assumptions on supports and densities of the measures. Here, we report a part of these results in [96].

Theorem 2.5 (Caffarelli's global regularity theory; see, e.g., [96, Theorem 12.50])**.** *Let $\mathcal{X}_\mu$ and $\mathcal{X}_\nu$ be two connected bounded open sets in $\mathbb{R}^d$ that both have $\mathcal{C}^2$-boundaries and are both uniformly convex*[1]*. Let $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ be concentrated on $\mathcal{X}_\mu$ and let $\nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ be concentrated on $\mathcal{X}_\nu$, i.e., $\mu(\mathbb{R}^d \backslash \mathcal{X}_\mu) = \nu(\mathbb{R}^d \backslash \mathcal{X}_\nu) = 0$. Suppose that for $k \in \mathbb{N}_0$, $\alpha \in (0, 1]$, $f_\mu \in \mathcal{C}^{k,\alpha}\big(\mathrm{cl}(\mathcal{X}_\mu)\big)$ and $f_\nu \in \mathcal{C}^{k,\alpha}\big(\mathrm{cl}(\mathcal{X}_\nu)\big)$ are the density functions of μ and ν with respect to the Lebesgue measure, respectively. Moreover, suppose that there exists $\gamma > 1$ such that $\gamma^{-1} \le f_\mu(\boldsymbol{x}) \le \gamma$ for all $\boldsymbol{x} \in \mathrm{cl}(\mathcal{X}_\mu)$ and that $\gamma^{-1} \le f_\nu(\boldsymbol{x}) \le \gamma$ for all $\boldsymbol{x} \in \mathrm{cl}(\mathcal{X}_\nu)$. Then, the optimal Brenier potential φ_ν^μ satisfies $\varphi_\nu^\mu \in \mathcal{C}^{k+2,\alpha}\big(\mathrm{cl}(\mathcal{X}_\mu)\big)$.*

3. Stochastic Fixed-Point Algorithm for $\mathcal{W}_2$-Barycenter

In this section, we will present our computationally tractable stochastic fixed-point algorithm for $\mathcal{W}_2$-Barycenter and show its convergence. Section 3.1 introduces the specifications of the approximation steps in Line 6 and Line 7 of Conceptual Algorithm 1 as well as additional assumptions. In Section 3.2, we develop sufficient conditions for the convergence of our stochastic fixed-point algorithm.

[1] A set $\mathcal{X} \subset \mathbb{R}^d$ is said to have $\mathcal{C}^p$ boundary with $p \in [0, +\infty)$ if $\mathrm{bd}(\mathcal{X})$ is locally the graph of a $\mathcal{C}^p$ function, and is said to be uniformly convex if for every $\epsilon > 0$, there exists $\delta > 0$ such that, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ with $\|\boldsymbol{x} - \boldsymbol{y}\| < \epsilon$, the distance from the mid-point $(\boldsymbol{x} + \boldsymbol{y})/2$ to $\mathrm{bd}(\mathcal{X})$ is at least δ.

3.1. **Settings.** Conceptual Algorithm 1 illustrates the conceptual procedure of our stochastic fixed-point iterative scheme. Before presenting its computationally tractable implementation as a concrete algorithm, let us introduce the following additional notions in Definition 3.1, Assumption 3.3, and Assumption 3.4.

Definition 3.1 (Admissible support sets and admissible probability measures)**.** *Let $\mathcal{S}$ denote the collection of subsets of $\mathbb{R}^d$ defined as follows:*

$$\mathcal{S} := \big\{\mathrm{cl}(\mathcal{Y}) : \mathcal{Y} \subset \mathbb{R}^d \textit{ is open, bounded, uniformly convex, and has a } \mathcal{C}^2\textit{-boundary}\big\}.$$

We will refer to $\mathcal{S}$ as the admissible support sets. Let $\mathcal{M}$ denote the collection of probability measures on $\mathbb{R}^d$ defined as follows:

$$\mathcal{M} := \left\{ \mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) : \begin{array}{l} \mathrm{supp}(\mu) \in \mathcal{S},\ \exists \alpha \in (0,1],\ \exists \gamma > 1,\ \exists f_\mu \in \mathcal{C}^{0,\alpha}(\mathrm{supp}(\mu)), \\ \gamma^{-1} \le f_\mu(\boldsymbol{x}) \le \gamma\ \forall \boldsymbol{x} \in \mathrm{supp}(\mu),\ f_\mu \textit{ is the density function of } \mu \end{array} \right\}.$$

We will refer to $\mathcal{M}$ as the admissible compactly supported probability measures. Moreover, let $\mathcal{M}_{\mathrm{full}}$ denote the collection of probability measures on $\mathbb{R}^d$ defined as follows:

$$\mathcal{M}_{\mathrm{full}} := \left\{ \rho \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) : \begin{array}{l} \mathrm{supp}(\rho) = \mathbb{R}^d,\ \exists \alpha \in (0,1],\ \exists f_\rho \in \mathcal{C}^{0,\alpha}_{\mathrm{loc}}(\mathbb{R}^d), \\ f_\rho(\boldsymbol{x}) > 0\ \forall \boldsymbol{x} \in \mathbb{R}^d,\ f_\rho \textit{ is the density function of } \rho \end{array} \right\}.$$

We will refer to $\mathcal{M}_{\mathrm{full}}$ as the admissible fully supported probability measures.

The conditions in the definitions of the admissible support sets and the admissible compactly supported probability measures are motivated by the conditions in Caffarelli's global regularity theory (Theorem 2.5). As a result, we can derive the following curvature properties of the optimal Brenier potential φ^μ_ν from $\mu \in \mathcal{M}$ to $\nu \in \mathcal{M}$, which will serve as a crucial premise when controlling the estimation errors of OT map estimators; see the results in Section 4.

Lemma 3.2 (Curvature properties of φ^μ_ν [59, Lemma 2] & [39, Corollary 3.2])**.** *Let $\mu, \nu \in \mathcal{M}$ be arbitrary. Then, $\varphi^\mu_\nu \in \mathcal{C}^2(\mathrm{supp}(\mu))$ and there exist $0 < \lambda_{\mathrm{LB}} \le \lambda_{\mathrm{UB}} < \infty$ such that $\lambda_{\mathrm{LB}}\mathbf{I}_d \preceq \nabla^2\varphi^\mu_\nu(\boldsymbol{x}) \preceq \lambda_{\mathrm{UB}}\mathbf{I}_d$ for all $\boldsymbol{x} \in \mathrm{supp}(\mu)$, where $\varphi^\mu_\nu : \mathbb{R}^d \to \mathbb{R}$ is the optimal Brenier potential from μ to ν (that is unique μ-almost everywhere up to the addition of an arbitrary constant). Moreover, there exists $\widetilde{\varphi}^\mu_\nu \in \mathfrak{C}_{\lambda_{\mathrm{LB}},\lambda_{\mathrm{UB}}}(\mathbb{R}^d)$ that is equal to φ^μ_ν μ-almost everywhere, and therefore one can let $\varphi^\mu_\nu \in \mathfrak{C}_{\lambda_{\mathrm{LB}},\lambda_{\mathrm{UB}}}(\mathbb{R}^d)$ without loss of generality.*

Proof of Lemma 3.2. Let f_μ and f_ν denote the density functions of μ and ν which satisfy the conditions in Definition 3.1. The properties of μ and ν in Definition 3.1 and Caffarelli's global regularity theory (Theorem 2.5) imply that the optimal Brenier potential φ^μ_ν satisfies $\varphi^\mu_\nu \in \mathcal{C}^{2,\alpha}(\mathrm{supp}(\mu))$ for some $\alpha \in (0,1]$. Thus, there exists $\lambda_{\mathrm{UB}} < \infty$ such that $\nabla^2\varphi^\mu_\nu(\boldsymbol{x}) \preceq \lambda_{\mathrm{UB}}\mathbf{I}_d$ for all $\boldsymbol{x} \in \mathrm{supp}(\mu)$. Moreover, φ^μ_ν needs to satisfy the following Monge–Ampère type equation as implied by the change of variable formula for pushforward (see, e.g., [6, Lemma 5.5.3]):

$$\det\big(\nabla^2\varphi^\mu_\nu(\boldsymbol{x})\big) = \frac{f_\mu(\boldsymbol{x})}{f_\nu\big(\nabla\varphi^\mu_\nu(\boldsymbol{x})\big)} \qquad \forall \boldsymbol{x} \in \mathrm{supp}(\mu).$$

Since both f_μ and f_ν are bounded from above and bounded away from 0, it follows that $\det\big(\nabla^2\varphi^\mu_\nu(\boldsymbol{x})\big)$ is bounded away from 0 on $\mathrm{supp}(\mu)$. Consequently, there exists $\lambda_{\mathrm{LB}} > 0$ such that $\nabla^2\varphi^\mu_\nu(\boldsymbol{x}) \succeq \lambda_{\mathrm{LB}}\mathbf{I}_d$ for all $\boldsymbol{x} \in \mathrm{supp}(\mu)$. Lastly, due to the closedness and convexity of $\mathrm{supp}(\mu)$, φ^μ_ν can be extended to a λ_{UB}-smooth, λ_{LB}-strongly convex function $\widetilde{\varphi}^\mu_\nu \in \mathfrak{C}_{\lambda_{\mathrm{LB}},\lambda_{\mathrm{UB}}}(\mathbb{R}^d)$ such that $\widetilde{\varphi}^\mu_\nu(\boldsymbol{x}) = \varphi^\mu_\nu(\boldsymbol{x})$ and $\nabla\widetilde{\varphi}^\mu_\nu(\boldsymbol{x}) = \nabla\varphi^\mu_\nu(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathrm{supp}(\mu)$; see, e.g., [91, Corollary 2.60]. The proof is now complete. □

Our algorithm uses a family of increasing sets for truncating probability measures in $\mathcal{M}_{\mathrm{full}}$ to probability measures in $\mathcal{M}$, which satisfies the assumption below.

Assumption 3.3 (Family of increasing sets)**.** *$(\mathcal{X}_r)_{r\in\mathbb{N}}$ is an infinite collection of subsets of $\mathbb{R}^d$ that satisfies the following conditions:*

(i) for all $r \in \mathbb{N}$, $\mathcal{X}_r \in \mathcal{S}$ where $\mathcal{S}$ is the collection of admissible support sets in Definition 3.1;
(ii) $\mathbf{0}_d \in \mathcal{X}_1$;
(iii) for all $r \in \mathbb{N}$, $\mathcal{X}_{r+1} \supseteq \mathcal{X}_r$;
(iv) $\bigcup_{r\in\mathbb{N}} \mathcal{X}_r = \mathbb{R}^d$.

Algorithm 2: Computationally tractable stochastic fixed-point iterative scheme

```
Input: $K$ input measures $\nu_1,\dots,\nu_K \in \mathcal{M}$, $\rho_0 \in \mathcal{M}_{\text{full}}$, family $(\mathcal{X}_r)_{r\in\mathbb{N}}$ of increasing sets, plug-in OT
        map estimator $\widehat{T}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta)$.
Output: $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$.
Initialize $\widehat{\rho}_0 \leftarrow \rho_0$.
for $t = 0, 1, 2, \dots$ do
    Choose $R_t \in \mathbb{N}$.
    $\widehat{\mu}_t \leftarrow \widehat{\rho}_t|_{\mathcal{X}_{R_t}}$.
    for $k = 1, \dots, K$ do
        Choose $N_{t,k} \in \mathbb{N}$ and $\Theta_{t,k} \in \mathbb{N}$.
        Randomly generate $N_{t,k}$ i.i.d. samples $\{\boldsymbol{X}_{t+1,k,i}\}_{i=1:N_{t,k}}$ from $\widehat{\mu}_t$.
        Randomly generate $N_{t,k}$ i.i.d. samples $\{\boldsymbol{Y}_{t+1,k,i}\}_{i=1:N_{t,k}}$ from $\nu_k$.
        $\widehat{T}_{t+1,k} \leftarrow \widehat{T}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta)\big|_{\mu=\widehat{\mu}_t,\,\nu=\nu_k,\,m=n=N_{t,k},\,\theta=\Theta_{t,k}}$.
    $\widehat{\rho}_{t+1} \leftarrow \Big[\frac{1}{K}\sum_{k=1}^{K}\widehat{T}_{t+1,k}\Big]\sharp\widehat{\rho}_t$.
return $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$.
```

A concrete example of such a family of increasing sets is $\big(\bar{B}(\mathbf{0}_d, r)\big)_{r\in\mathbb{N}}$. Similarly, a family of nested ellipsoids in $\mathbb{R}^d$ containing the origin also satisfies Assumption 3.3.

With respect to any pair of admissible compactly supported probability measures $\mu, \nu \in \mathcal{M}$, we consider plug-in estimators of the true OT map T^{μ}_{ν} from μ to ν which satisfy the conditions in the assumption below.

Assumption 3.4 (Plug-in OT map estimator). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For $\mu, \nu \in \mathcal{M}$ with $\mathbf{0}_d \in \operatorname{supp}(\mu)$, and for $m, n \in \mathbb{N}$, let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_m, \boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n : \Omega \to \mathbb{R}^d$ be independent random variables such that $\operatorname{law}(\boldsymbol{X}_i) = \mu$ for $i = 1, \dots, m$ and $\operatorname{law}(\boldsymbol{Y}_j) = \nu$ for $j = 1, \dots, n$. For any $\theta \in \mathbb{N}$, let $\widehat{T}^{\mu,m}_{\nu,n}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_m, \boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n; \theta)$ be a $\mathcal{C}_{\text{lin}}(\mathbb{R}^d, \mathbb{R}^d)$-valued random variable that estimates the OT map T^{μ}_{ν} from μ to ν based on m independent samples from μ and n independent samples from ν, where θ denotes a parameter that, for example, represents the extend of smoothing/regularization. We assume that θ takes value in $\mathbb{N}$ for simplicity; see Section 4 for details about this parameter in concrete plug-in OT map estimators. For the sake of notational simplicity, we make the dependence of this estimated OT map on the samples implicit and use $\widehat{T}^{\mu,m}_{\nu,n}(\boldsymbol{x};\theta)$ to denote $\widehat{T}^{\mu,m}_{\nu,n}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_m, \boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n; \theta)$ evaluated at $\boldsymbol{x} \in \mathbb{R}^d$.*

We assume that $\widehat{T}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta)$ satisfies the following conditions.

(i) ***Shape condition:*** *there exist constants $\alpha \equiv \alpha(\mu,\nu) \in (0,1]$, $\underline{\lambda} \equiv \underline{\lambda}(\mu,\nu)$, $\overline{\lambda} \equiv \overline{\lambda}(\mu,\nu)$ with $0 < \underline{\lambda} \le \overline{\lambda} < \infty$, all of which depend on μ and ν, such that for all $m, n \in \mathbb{N}$ and all $\theta \in \mathbb{N}$, it holds $\mathbb{P}$-almost surely that $\widehat{T}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta) = \nabla\widehat{\varphi}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta)$ for $\widehat{\varphi}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta) \in \mathfrak{C}^{2,\alpha}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$.*

(ii) ***Growth condition:*** *there exist constants $u_1(\nu), u_2(\nu) \in \mathbb{R}_+$ that only depend on ν such that, for all $m, n \in \mathbb{N}$ and all $\theta \in \mathbb{N}$, it holds that $\mathbb{E}\Big[\big\|\widehat{T}^{\mu,m}_{\nu,n}(\boldsymbol{x};\theta)\big\|^2\Big] \le u_1(\nu) + u_2(\nu)\|\boldsymbol{x}\|^2$ $\forall \boldsymbol{x} \in \mathbb{R}^d$.*

(iii) ***Consistency condition:*** *for any $\epsilon > 0$, there exist $\overline{n}(\mu,\nu,\epsilon) \in \mathbb{N}$ that depends on μ, ν, ϵ and $\overline{\theta}(\mu,\nu,m,n,\epsilon) \in \mathbb{N}$ that depends on μ, ν, ϵ as well as the sample sizes m, n such that*

$$\mathbb{E}\Big[\big\|\widehat{T}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta) - T^{\mu}_{\nu}\big\|^2_{\mathcal{L}^2(\mu)}\Big] \le \epsilon \qquad \forall m \ge \overline{n}(\mu,\nu,\epsilon),\ \forall n \ge \overline{n}(\mu,\nu,\epsilon),\ \forall \theta \ge \overline{\theta}(\mu,\nu,m,n,\epsilon),$$

where

$$\big\|\widehat{T}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta) - T^{\mu}_{\nu}\big\|_{\mathcal{L}^2(\mu)} := \bigg(\int_{\mathbb{R}^d} \big\|\widehat{T}^{\mu,m}_{\nu,n}(\boldsymbol{x};\theta) - T^{\mu}_{\nu}(\boldsymbol{x})\big\|^2\,\mu(\mathrm{d}\boldsymbol{x})\bigg)^{\frac{1}{2}}.$$

Concrete examples of plug-in OT map estimators which satisfy Assumption 3.4 will be introduced later in Section 4. Note that the consistency condition in Assumption 3.4(iii) is possible due to the curvature properties of $T^{\mu}_{\nu} = \nabla\varphi^{\mu}_{\nu}$ in Lemma 3.2.

With these notions, Algorithm 2 describes a computationally tractable algorithm which completes Conceptual Algorithm 1. The setting for Algorithm 2 is presented below.

Setting 3.5 (Inputs of Algorithm 2)**.** *In Algorithm 2, we assume that the inputs $\nu_1, \dots, \nu_K$ are admissible compactly supported probability measures in $\mathcal{M}$. The input ρ_0 is an arbitrary admissible fully supported probability measure in $\mathcal{M}_{\text{full}}$. Moreover, we assume that $(\mathcal{X}_r)_{r\in\mathbb{N}}$ is a family of increasing sets satisfying the conditions in Assumption 3.3, and $\widehat{T}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta)$ is a plug-in OT map estimator satisfying the conditions in Assumption 3.4.*

Definition 3.1 and the shape condition in Assumption 3.4(i) imply the following property of $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$ which shows that the procedure in Algorithm 2 is well-defined.

Proposition 3.6 (Well-definedness of Algorithm 2)**.** *Let the inputs of Algorithm 2 satisfy Setting 3.5 and let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space on which the random samples on Line 7 and Line 8 of Algorithm 2 are defined. Then, it holds $\mathbb{P}$-almost surely that $\widehat{\rho}_t \in \mathcal{M}_{\text{full}}$ and $\widehat{\mu}_t \in \mathcal{M}$ for all $t \in \mathbb{N}_0$, and thus $\widehat{T}_{t+1,k}$ on Line 9 of Algorithm 2 is $\mathbb{P}$-almost surely well-defined.*

Proof of Proposition 3.6. For $t = 0$, $\widehat{\rho}_0 = \rho_0 \in \mathcal{M}_{\text{full}}$ holds by assumption. Since $\operatorname{supp}(\widehat{\rho}_0) = \mathbb{R}^d$, it holds by Line 4 that $\operatorname{supp}(\widehat{\mu}_0) = \operatorname{supp}\big(\widehat{\rho}_0|_{\mathcal{X}_{R_0}}\big) = \mathcal{X}_{R_0} \in \mathcal{S}$. Moreover, $f_{\widehat{\mu}_0} := \frac{f_{\widehat{\rho}_0}\mathbb{1}_{\mathcal{X}_{R_0}}}{\widehat{\rho}_0(\mathcal{X}_{R_0})} \in \mathcal{C}^{0,\alpha}(\mathcal{X}_{R_0})$ is the density function of $\widehat{\mu}_0$, where $f_{\widehat{\rho}_0} \in \mathcal{C}^{0,\alpha}_{\text{loc}}(\mathbb{R}^d)$ is the density function of $\widehat{\rho}_0$. Since $f_{\widehat{\rho}_0}(\boldsymbol{x}) > 0\ \forall \boldsymbol{x} \in \mathbb{R}^d$, it holds by the compactness of $\mathcal{X}_{R_0}$ that $0 < \inf_{\boldsymbol{x}\in\mathcal{X}_{R_0}} f_{\widehat{\mu}_0}(\boldsymbol{x}) \le \sup_{\boldsymbol{x}\in\mathcal{X}_{R_0}} f_{\widehat{\mu}_0}(\boldsymbol{x}) < \infty$ and thus $\widehat{\mu}_0 \in \mathcal{M}$. Consequently, for $k = 1, \dots, K$, $\widehat{T}_{1,k} = \widehat{T}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta)\big|_{\mu=\widehat{\mu}_0,\,\nu=\nu_k,\,m=n=N_{0,k},\,\theta=\Theta_{0,k}}$ is well-defined.

Next, let us assume that $\widehat{\rho}_t \in \mathcal{M}_{\text{full}}$ and $\widehat{\mu}_t \in \mathcal{M}$ $\mathbb{P}$-almost surely for some $t \in \mathbb{N}_0$. For $k = 1, \dots, K$, since $\widehat{T}_{t+1,k}$ is $\mathbb{P}$-almost surely well-defined and satisfies the shape condition in Assumption 3.4(i), it holds $\mathbb{P}$-almost surely that there exists $\widehat{\varphi}_{t+1,k} \in \mathfrak{C}^{2,\alpha_k}_{\underline{\lambda}_k,\overline{\lambda}_k}(\mathbb{R}^d)$ such that $\widehat{T}_{t+1,k} = \nabla\widehat{\varphi}_{t+1,k}$, $\alpha_k \in (0,1]$, $0 < \underline{\lambda}_k \le \overline{\lambda}_k < \infty$. Subsequently, let us denote $\bar{T}_{t+1} := \frac{1}{K}\sum_{k=1}^K \widehat{T}_{t+1,k}$ and $\bar{\varphi}_{t+1} := \frac{1}{K}\sum_{k=1}^K \widehat{\varphi}_{t+1,k}$. It thus follows that $\bar{T}_{t+1} = \nabla\bar{\varphi}_{t+1}$ and $\bar{\varphi}_{t+1} \in \mathfrak{C}^{2,\alpha}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$ $\mathbb{P}$-almost surely for $\alpha := \min_{1\le k\le K}\{\alpha_k\} \in (0,1]$, $\underline{\lambda} := \min_{1\le k\le K}\{\underline{\lambda}_k\} > 0$, $\overline{\lambda} := \max_{1\le k\le K}\{\overline{\lambda}_k\} < \infty$. By the well-known duality between smooth convex functions and strongly convex functions (see, e.g., [77, Proposition 12.60]), it holds that the convex conjugate $\bar{\varphi}^*_{t+1}$ of $\bar{\varphi}_{t+1}$ is $\underline{\lambda}^{-1}$-smooth, $\overline{\lambda}^{-1}$-strongly convex and $\nabla\bar{\varphi}^*_{t+1} = \bar{T}^{-1}_{t+1}$. This shows that $\bar{T}_{t+1} : \mathbb{R}^d \to \mathbb{R}^d$ is a homeomorphism and $\bar{T}^{-1}_{t+1}$ is $\underline{\lambda}^{-1}$-Lipschitz continuous on $\mathbb{R}^d$. Moreover, we have by the second-order characterization of smooth and strongly convex functions (see, e.g., [65, Theorem 2.1.6]) that $\underline{\lambda}\mathbf{I}_d \preceq \nabla^2\bar{\varphi}_{t+1}(\boldsymbol{x}) \preceq \overline{\lambda}\mathbf{I}_d$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Now, since Line 10 sets $\widehat{\rho}_{t+1} := \bar{T}_{t+1}\sharp\widehat{\rho}_t$, the change of variable formula for pushforward (see, e.g., [6, Lemma 5.5.3]) yields

$$f_{\widehat{\rho}_{t+1}}(\boldsymbol{y}) = \frac{f_{\widehat{\rho}_t}\big(\bar{T}^{-1}_{t+1}(\boldsymbol{y})\big)}{\det\big(\nabla^2\bar{\varphi}_{t+1}(\bar{T}^{-1}_{t+1}(\boldsymbol{y}))\big)} \qquad \forall \boldsymbol{y} \in \mathbb{R}^d, \tag{3.1}$$

where $f_{\widehat{\rho}_{t+1}}$ and $f_{\widehat{\rho}_t}$ denote the density functions of $\widehat{\rho}_{t+1}$ and $\widehat{\rho}_t$, respectively. Since $f_{\widehat{\rho}_t}(\boldsymbol{x}) > 0\ \forall \boldsymbol{x} \in \mathbb{R}^d$ by assumption and since $\det\big(\nabla^2\bar{\varphi}_{t+1}(\boldsymbol{x})\big) \ge \underline{\lambda}^d > 0\ \forall \boldsymbol{x} \in \mathbb{R}^d$, it holds that $f_{\widehat{\rho}_{t+1}}(\boldsymbol{y}) > 0\ \forall \boldsymbol{y} \in \mathbb{R}^d$. Let us now show the local Hölder continuity of $f_{\widehat{\rho}_{t+1}}$. On the one hand, combining the induction hypothesis that $f_{\widehat{\rho}_t} \in \mathcal{C}^{0,\alpha'}_{\text{loc}}(\mathbb{R}^d)$ for some $\alpha' \in (0,1]$ and the Lipschitz continuity of $\bar{T}^{-1}_{t+1}$ on $\mathbb{R}^d$ shows that the numerator of (3.1) satisfies $f_{\widehat{\rho}_t} \circ \bar{T}^{-1}_{t+1} \in \mathcal{C}^{0,\alpha'}_{\text{loc}}(\mathbb{R}^d)$. On the other hand, since $\bar{\varphi}_{t+1} \in \mathcal{C}^{2,\alpha}_{\text{loc}}(\mathbb{R}^d)$ and $\det(\cdot)$ is a polynomial in all entries of the input matrix, the denominator of (3.1) satisfies $\det \circ \nabla^2\bar{\varphi}_{t+1} \circ \bar{T}^{-1}_{t+1} \in \mathcal{C}^{0,\alpha''}_{\text{loc}}(\mathbb{R}^d)$ for some $\alpha'' \in (0,1]$. Consequently, since the denominator is also bounded from below by $\underline{\lambda}^d > 0$, we get $f_{\widehat{\rho}_{t+1}} \in \mathcal{C}^{0,\widetilde{\alpha}}_{\text{loc}}(\mathbb{R}^d)$ for some $\widetilde{\alpha} \in (0,1]$. Furthermore, we have by the $\mathbb{P}$-almost sure $\overline{\lambda}$-Lipschitz continuity of $\bar{T}_{t+1}$ that

$$\int_{\mathbb{R}^d} \|\boldsymbol{y}\|^2\,\widehat{\rho}_{t+1}(\mathrm{d}\boldsymbol{y}) = \int_{\mathbb{R}^d} \big\|\bar{T}_{t+1}(\boldsymbol{x})\big\|^2\,\widehat{\rho}_t(\mathrm{d}\boldsymbol{x}) \le \int_{\mathbb{R}^d} \big(\big\|\bar{T}_{t+1}(\mathbf{0})\big\| + \overline{\lambda}\|\boldsymbol{x}\|\big)^2\,\widehat{\rho}_t(\mathrm{d}\boldsymbol{x}) < \infty \qquad \mathbb{P}\text{-a.s.},$$

and thus $\widehat{\rho}_{t+1} \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$. Therefore, we have shown that $\widehat{\rho}_{t+1} \in \mathcal{M}_{\text{full}}$ $\mathbb{P}$-almost surely. Lastly, it can be shown via the same argument used for the case $t = 0$ that $\widehat{\mu}_{t+1} = \widehat{\rho}_{t+1}|_{\mathcal{X}_{R_{t+1}}} \in \mathcal{M}$ $\mathbb{P}$-almost surely.

We conclude by induction that $\widehat{\rho}_t \in \mathcal{M}_{\text{full}}$ and $\widehat{\mu}_t \in \mathcal{M}$ for all $t \in \mathbb{N}_0$ and that Line 9 is well-defined for all iterations $\mathbb{P}$-almost surely. The proof is now complete. □

Remark 3.7. *In Algorithm 2, rather than directly updating $\widehat{\mu}_{t+1}$ to $\widehat{\mu}_{t+1} \leftarrow \Big[\frac{1}{K}\sum_{k=1}^K \widehat{T}_{t+1,k}\Big]\sharp\widehat{\mu}_t$, we first apply the pushforward of $\widehat{\rho}_t \in \mathcal{M}_{\text{full}}$ by $\Big[\frac{1}{K}\sum_{k=1}^K \widehat{T}_{t+1,k}\Big]$ in Line 10 to obtain $\widehat{\rho}_{t+1} \in \mathcal{M}_{\text{full}}$, and then truncate*

$\widehat{\rho}_{t+1}$ to $\mathcal{X}_{R_{t+1}}$ to get $\widehat{\mu}_{t+1}$. The truncation step guarantees that $\widehat{\mu}_{t+1} \in \mathcal{M}$ so that the consistency condition of the plug-in OT map estimator in Assumption 3.4(iii) can be satisfied (see our results and discussions in Section 4). Note that the support of the pushforward $\Big[\frac{1}{K}\sum_{k=1}^{K}\widehat{T}_{t+1,k}\Big]\sharp\widehat{\mu}_t$ is not necessarily an admissible support set in $\mathcal{S}$; specifically, the uniform convexity condition may fail.

Let us now examine the stochastic processes generated by Algorithm 2. To begin, let us consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which the random samples on Line 7 and Line 8 are defined. Let $\mathcal{F}_0 := \{\emptyset, \Omega\}$. $\widehat{\rho}_0 : \Omega \to \mathcal{M}_{\mathrm{full}}$ initialized in Line 1 takes a pre-specified value ρ_0 and is thus $\mathcal{F}_0$-measurable. For each $t = 0, 1, 2, \ldots$, the index $R_t : \Omega \to \mathbb{N}$ in Line 3 is an $\mathcal{F}_t$-measurable random variable. After R_t has been chosen, $\widehat{\mu}_t : \Omega \to \mathcal{M}$ is set to $\widehat{\rho}_t|_{\mathcal{X}_{R_t}}$ in Line 4, which is also $\mathcal{F}_t$-measurable. Next, for $k = 1, \ldots, K$, the sample size $N_{t,k} : \Omega \to \mathbb{N}$ and the parameter $\Theta_{t,k} : \Omega \to \mathbb{N}$ in Line 6 are both $\mathcal{F}_t$-measurable random variables. Subsequently, after $N_{t,k}$ and $\Theta_{t,k}$ have been chosen, $N_{t,k}$ i.i.d. samples $\boldsymbol{X}_{t+1,k,1}, \ldots, \boldsymbol{X}_{t+1,k,N_{t,k}} : \Omega \to \mathbb{R}^d$ from $\widehat{\mu}_t$ and $N_{t,k}$ i.i.d. samples $\boldsymbol{Y}_{t+1,k,1}, \ldots, \boldsymbol{Y}_{t+1,k,N_{t,k}} : \Omega \to \mathbb{R}^d$ from ν_k are randomly generated in Line 7 and Line 8. We require $\big\{\boldsymbol{X}_{t+1,k,1}, \ldots, \boldsymbol{X}_{t+1,k,N_{t,k}}, \boldsymbol{Y}_{t+1,k,1}, \ldots, \boldsymbol{Y}_{t+1,k,N_{t,k}}\big\}_{k=1:K}$ to be jointly independent conditional on $\mathcal{F}_t$. Let $\mathcal{F}_{t+1}$ be the σ-algebra generated by all the random samples up to the iteration $t+1$, i.e., $\mathcal{F}_{t+1} := \sigma\Big(\bigcup_{0 \le s \le t}\big(\{\boldsymbol{X}_{s+1,k,i}\}_{i=1:N_{s,k},\, k=1:K} \cup \{\boldsymbol{Y}_{s+1,k,i}\}_{i=1:N_{s,k},\, k=1:K}\big)\Big)$. For $k = 1, \ldots, K$, the plug-in OT map estimator $\widehat{T}_{t+1,k} : \Omega \to \mathcal{C}_{\mathrm{lin}}(\mathbb{R}^d, \mathbb{R}^d)$ in Line 9 is thus $\mathcal{F}_{t+1}$-measurable. Since $\widehat{\rho}_{t+1} : \Omega \to \mathcal{M}_{\mathrm{full}}$ in Line 10 depends on $\widehat{\rho}_t$ and $\big(\widehat{T}_{t+1,k}\big)_{k=1:K}$, it is $\mathcal{F}_{t+1}$-measurable. Iteratively repeating the above construction for $t = 0, 1, 2, \ldots$ leads to a filtered probability space with filtration $(\mathcal{F}_t)_{t \in \mathbb{N}_0}$. The resulting sequences $(\widehat{\rho}_t)_{t \in \mathbb{N}_0}$ and $(\widehat{\mu}_t)_{t \in \mathbb{N}_0}$ are thus $(\mathcal{F}_t)_{t \in \mathbb{N}_0}$-adapted stochastic processes. In the next subsection, we will specify the choices of $(R_t)_{t \in \mathbb{N}_0}$, $(N_{t,k})_{k=1:K,\, t \in \mathbb{N}_0}$, and $(\Theta_{t,k})_{k=1:K,\, t \in \mathbb{N}_0}$ (which are $(\mathcal{F}_t)_{t \in \mathbb{N}_0}$-adapted stochastic processes) in order to achieve $\mathbb{P}$-almost sure convergence of the output process $(\widehat{\mu}_t)_{t \in \mathbb{N}_0}$ of Algorithm 2.

Remark 3.8 (Measurability)**.** *We would like to remark that the above description of the stochastic processes $(\widehat{\rho}_t)_{t \in \mathbb{N}_0}$ and $(\widehat{\mu}_t)_{t \in \mathbb{N}_0}$ has not yet rigorously justified the measurability of all the constructive operations. The rigorous justification of measurability will be carried out in Appendix A with specific choices of the plug-in OT map estimator $\widehat{T}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta)$.*

3.2. **Convergence analysis.**

The goal of this subsection is to develop sufficient conditions for the convergence of the output process $(\widehat{\mu}_t)_{t \in \mathbb{N}_0}$ in Algorithm 2. Let us begin by analyzing the decrements of the process $\big(V(\widehat{\mu}_t)\big)_{t \in \mathbb{N}_0}$. This will subsequently lead to sufficient conditions on the choices of $(R_t)_{t \in \mathbb{N}_0}$, $(N_{t,k})_{k=1:K,\, t \in \mathbb{N}_0}$, and $(\Theta_{t,k})_{k=1:K,\, t \in \mathbb{N}_0}$ to guarantee the convergence of $(\widehat{\mu}_t)_{t \in \mathbb{N}_0}$.

Proposition 3.9 (Decrement of the process $\big(V(\widehat{\mu}_t)\big)_{t \in \mathbb{N}_0}$)**.** *Let the inputs of Algorithm 2 satisfy Setting 3.5, let $\big(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t \in \mathbb{N}_0}\big)$ be the filtered probability space constructed by Algorithm 2, and let $(\widehat{\mu}_t)_{t \in \mathbb{N}_0}$ be the output of Algorithm 2. Moreover, let $V(\cdot)$ be the function defined in (1.2) and let $G(\cdot)$ be the operator defined in (1.3). Then, the sequence $\big(V(\widehat{\mu}_t)\big)_{t \in \mathbb{N}_0}$ satisfies*

$$
\begin{aligned}
V(\widehat{\mu}_{t+1}) - V(\widehat{\mu}_t) \le -\mathcal{W}_2\big(\widehat{\mu}_t, G(\widehat{\mu}_t)\big)^2 &+ \frac{2}{K}\sum_{k=1}^{K}\big\|\widehat{T}_{t+1,k} - T^{\widehat{\mu}_t}_{\nu_k}\big\|^2_{\mathcal{L}^2(\widehat{\mu}_t)} \\
&+ 2\mathcal{W}_2\Big(\big[\tfrac{1}{K}\textstyle\sum_{k=1}^{K}\widehat{T}_{t+1,k}\big]\sharp\widehat{\mu}_t, \widehat{\mu}_{t+1}\Big)^2 \qquad \forall t \in \mathbb{N}_0,\ \mathbb{P}\text{-a.s.}
\end{aligned}
\tag{3.2}
$$

In particular, taking expectations on both sides of (3.2) conditional on $\mathcal{F}_t$ yields

$$
\begin{aligned}
\mathbb{E}\big[V(\widehat{\mu}_{t+1})\big|\mathcal{F}_t\big] - V(\widehat{\mu}_t) \le -\mathcal{W}_2\big(\widehat{\mu}_t, G(\widehat{\mu}_t)\big)^2 &+ \frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\big\|\widehat{T}_{t+1,k} - T^{\widehat{\mu}_t}_{\nu_k}\big\|^2_{\mathcal{L}^2(\widehat{\mu}_t)}\Big|\mathcal{F}_t\Big] \\
&+ 2\mathbb{E}\Big[\mathcal{W}_2\Big(\big[\tfrac{1}{K}\textstyle\sum_{k=1}^{K}\widehat{T}_{t+1,k}\big]\sharp\widehat{\mu}_t, \widehat{\mu}_{t+1}\Big)^2\Big|\mathcal{F}_t\Big] \qquad \forall t \in \mathbb{N}_0,\ \mathbb{P}\text{-a.s.}
\end{aligned}
\tag{3.3}
$$

Proof of Proposition 3.9. Throughout this proof, let us fix an arbitrary $t \in \mathbb{N}_0$, denote $\bar{T}^{\widehat{\mu}_t} := \frac{1}{K}\sum_{k=1}^{K} T^{\widehat{\mu}_t}_{\nu_k}$, $\bar{T}_{t+1} := \frac{1}{K}\sum_{k=1}^{K}\widehat{T}_{t+1,k}$, and denote $\widetilde{\mu}_{t+1} := \bar{T}_{t+1}\sharp\widehat{\mu}_t$. As we have shown in the proof of Proposition 3.6, $\bar{T}_{t+1} : \mathbb{R}^d \to \mathbb{R}^d$ is a homeomorphism $\mathbb{P}$-almost surely. Subsequently, the change of variable formula for pushforward (see, e.g., [6, Lemma 5.5.3]) implies that $\widetilde{\mu}_{t+1} \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ $\mathbb{P}$-almost surely. Let $T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}} : \mathbb{R}^d \to \mathbb{R}^d$

denotes the OT map from $\widetilde{\mu}_{t+1}$ to $\widehat{\mu}_{t+1}$. In the remainder of this proof, all statements hold in $\mathbb{P}$-almost sure sense, and we will omit "$\mathbb{P}$-a.s." for simplicity.

Our proof below uses the following identity, which can be verified directly by expanding both sides:

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{y}-\boldsymbol{z}_k\|^2=\|\boldsymbol{y}-\bar{\boldsymbol{z}}\|^2+\frac{1}{K}\sum_{k=1}^{K}\|\bar{\boldsymbol{z}}-\boldsymbol{z}_k\|^2\\ &\qquad\qquad\text{where } \bar{\boldsymbol{z}}:=\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{z}_k \qquad \forall \boldsymbol{y},\boldsymbol{z}_1,\ldots,\boldsymbol{z}_k\in\mathbb{R}^d.\end{aligned}\tag{3.4}$$

For any $\boldsymbol{x}\in\mathbb{R}^d$, substituting $\boldsymbol{y}\leftarrow\boldsymbol{x}$ and $\boldsymbol{z}_k\leftarrow T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})$ in (3.4) gives us

$$\frac{1}{K}\sum_{k=1}^{K}\big\|\boldsymbol{x}-T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})\big\|^2=\big\|\boldsymbol{x}-\bar{T}^{\widehat{\mu}_t}(\boldsymbol{x})\big\|^2+\frac{1}{K}\sum_{k=1}^{K}\big\|\bar{T}^{\widehat{\mu}_t}(\boldsymbol{x})-T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})\big\|^2.\tag{3.5}$$

Moreover, substituting $\boldsymbol{y}\leftarrow T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}\circ\bar{T}_{t+1}(\boldsymbol{x})$ and $\boldsymbol{z}_k\leftarrow T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})$ in (3.4), we obtain

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\Big\|T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}\circ\bar{T}_{t+1}(\boldsymbol{x})-T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})\Big\|^2\\ &\quad=\Big\|T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}\circ\bar{T}_{t+1}(\boldsymbol{x})-\bar{T}^{\widehat{\mu}_t}(\boldsymbol{x})\Big\|^2+\frac{1}{K}\sum_{k=1}^{K}\big\|\bar{T}^{\widehat{\mu}_t}(\boldsymbol{x})-T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})\big\|^2.\end{aligned}\tag{3.6}$$

Combining (3.5) and (3.6) yields

$$\begin{aligned}&\left(\frac{1}{K}\sum_{k=1}^{K}\Big\|T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}\circ\bar{T}_{t+1}(\boldsymbol{x})-T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})\Big\|^2\right)-\left(\frac{1}{K}\sum_{k=1}^{K}\big\|\boldsymbol{x}-T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})\big\|^2\right)\\ &\quad=\Big\|T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}\circ\bar{T}_{t+1}(\boldsymbol{x})-\bar{T}^{\widehat{\mu}_t}(\boldsymbol{x})\Big\|^2-\big\|\boldsymbol{x}-\bar{T}^{\widehat{\mu}_t}(\boldsymbol{x})\big\|^2\\ &\quad\le 2\Big\|T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}\circ\bar{T}_{t+1}(\boldsymbol{x})-\bar{T}_{t+1}(\boldsymbol{x})\Big\|^2+2\big\|\bar{T}_{t+1}(\boldsymbol{x})-\bar{T}^{\widehat{\mu}_t}(\boldsymbol{x})\big\|^2-\big\|\boldsymbol{x}-\bar{T}^{\widehat{\mu}_t}(\boldsymbol{x})\big\|^2\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall\boldsymbol{x}\in\mathbb{R}^d.\end{aligned}\tag{3.7}$$

In the following, let us examine the integral of each term in (3.7) with respect to $\widehat{\mu}_t$. Firstly, for $k=1,\ldots,K$, let $\pi_k:=\Big[T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}\circ\bar{T}_{t+1},T^{\widehat{\mu}_t}_{\nu_k}\Big]\sharp\widehat{\mu}_t\in\mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d)$. Since $\Big(T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}\circ\bar{T}_{t+1}\Big)\sharp\widehat{\mu}_t=T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}\sharp\widetilde{\mu}_{t+1}=\widehat{\mu}_{t+1}$ and $T^{\widehat{\mu}_t}_{\nu_k}\sharp\widehat{\mu}_t=\nu_k$, it follows that $\pi_k\in\Pi(\widehat{\mu}_{t+1},\nu_k)$. Thus, we get

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\int_{\mathbb{R}^d}\Big\|T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}\circ\bar{T}_{t+1}(\boldsymbol{x})-T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})\Big\|^2\widehat{\mu}_t(\mathrm{d}\boldsymbol{x})&=\frac{1}{K}\sum_{k=1}^{K}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2\,\pi_k(\mathrm{d}\boldsymbol{x}_1,\mathrm{d}\boldsymbol{x}_2)\\ &\ge\frac{1}{K}\sum_{k=1}^{K}\mathcal{W}_2(\widehat{\mu}_{t+1},\nu_k)^2=V(\widehat{\mu}_{t+1}).\end{aligned}\tag{3.8}$$

Secondly, since $T^{\widehat{\mu}_t}_{\nu_k}$ is the OT map from $\widehat{\mu}_t$ to ν_k for $k=1,\ldots,K$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\int_{\mathbb{R}^d}\big\|\boldsymbol{x}-T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})\big\|^2\widehat{\mu}_t(\mathrm{d}\boldsymbol{x})=\frac{1}{K}\sum_{k=1}^{K}\mathcal{W}_2(\widehat{\mu}_t,\nu_k)^2=V(\widehat{\mu}_t).\tag{3.9}$$

Thirdly, it holds that

$$\int_{\mathbb{R}^d}\Big\|T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}\circ\bar{T}_{t+1}(\boldsymbol{x})-\bar{T}_{t+1}(\boldsymbol{x})\Big\|^2\widehat{\mu}_t(\mathrm{d}\boldsymbol{x})=\int_{\mathbb{R}^d}\Big\|T^{\widetilde{\mu}_{t+1}}_{\widehat{\mu}_{t+1}}(\boldsymbol{x})-\boldsymbol{x}\Big\|^2\widetilde{\mu}_{t+1}(\mathrm{d}\boldsymbol{x})=\mathcal{W}_2(\widetilde{\mu}_{t+1},\widehat{\mu}_{t+1})^2.\tag{3.10}$$

Fourthly, the convexity of $\mathbb{R}^d\ni\boldsymbol{z}\mapsto\|\boldsymbol{z}\|^2\in\mathbb{R}$ together with Jensen's inequality gives

$$\big\|\bar{T}_{t+1}(\boldsymbol{x})-\bar{T}^{\widehat{\mu}_t}(\boldsymbol{x})\big\|^2=\left\|\frac{1}{K}\sum_{k=1}^{K}\big(\widehat{T}_{t+1,k}(\boldsymbol{x})-T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})\big)\right\|^2\le\frac{1}{K}\sum_{k=1}^{K}\big\|\widehat{T}_{t+1,k}(\boldsymbol{x})-T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})\big\|^2\quad\forall\boldsymbol{x}\in\mathbb{R}^d,$$

which results in

$$\begin{aligned}\int_{\mathbb{R}^d} \big\|\bar{T}_{t+1}(\boldsymbol{x}) - \bar{T}^{\widehat{\mu}_t}(\boldsymbol{x})\big\|^2\,\widehat{\mu}_t(\mathrm{d}\boldsymbol{x}) &\le \frac{1}{K}\sum_{k=1}^{K}\int_{\mathbb{R}^d}\big\|\widehat{T}_{t+1,k}(\boldsymbol{x}) - T^{\widehat{\mu}_t}_{\nu_k}(\boldsymbol{x})\big\|^2\,\widehat{\mu}_t(\mathrm{d}\boldsymbol{x})\\ &= \frac{1}{K}\sum_{k=1}^{K}\big\|\widehat{T}_{t+1,k} - T^{\widehat{\mu}_t}_{\nu_k}\big\|^2_{\mathcal{L}^2(\widehat{\mu}_t)}.\end{aligned} \tag{3.11}$$

Lastly, for $k = 1,\ldots,K$, let $\varphi^{\widehat{\mu}_t}_{\nu_k} \in \mathfrak{C}_{\underline{\lambda}_k,\overline{\lambda}_k}(\mathbb{R}^d)$ denote the optimal Brenier potential from $\widehat{\mu}_t$ to ν_k, where $0 < \underline{\lambda}_k \le \overline{\lambda}_k < \infty$ (see Lemma 3.2). It follows that $\bar{T}^{\widehat{\mu}_t}$ is the gradient of the continuously differentiable convex function $\frac{1}{K}\sum_{k=1}^{K}\varphi^{\widehat{\mu}_t}_{\nu_k}$. Thus, Brenier's theorem (Theorem 2.4) implies that $\bar{T}^{\widehat{\mu}_t}$ is the OT map from $\widehat{\mu}_t$ to $\bar{T}^{\widehat{\mu}_t}\sharp\widehat{\mu}_t = G(\widehat{\mu}_t)$, resulting in

$$\int_{\mathbb{R}^d}\big\|\boldsymbol{x} - \bar{T}^{\widehat{\mu}_t}(\boldsymbol{x})\big\|^2\,\widehat{\mu}_t(\mathrm{d}\boldsymbol{x}) = \mathcal{W}_2\big(\widehat{\mu}_t, G(\widehat{\mu}_t)\big)^2. \tag{3.12}$$

Now, integrating both sides of (3.7) with respect to $\widehat{\mu}_t$ and then combining it with (3.8)–(3.12) completes the proof of (3.2). Finally, taking conditional expectations with respect to $\mathcal{F}_t$ on both sides of (3.2) proves (3.3). The proof is now complete. □

Remark 3.10. *In [5, Proposition 3.3], the decrement of the sequence $\big(V(\mu_t)\big)_{t\in\mathbb{N}_0}$ in the deterministic fixed-point iteration $\mu_{t+1} \leftarrow G(\mu_t)$ $\forall t \in \mathbb{N}_0$ is controlled through the inequality:*

$$V(\mu_{t+1}) - V(\mu_t) \le -\mathcal{W}_2(\mu_t, G(\mu_t))^2 \qquad \forall t \in \mathbb{N}_0. \tag{3.13}$$

Compared to (3.13), the stochastic decrement (3.2) in Proposition 3.9 has two additional terms on the right-hand side:

- *the term $\frac{2}{K}\sum_{k=1}^{K}\big\|\widehat{T}_{t+1,k} - T^{\widehat{\mu}_t}_{\nu_k}\big\|^2_{\mathcal{L}^2(\widehat{\mu}_t)}$ comes from the inexactness when approximating the true OT map $T^{\widehat{\mu}_t}_{\nu_k}$ by the plug-in OT map estimator $\widehat{T}_{t+1,k}$, i.e., from the approximation in Line 6 of Conceptual Algorithm 1;*
- *the term $2\mathcal{W}_2\big(\big[\frac{1}{K}\sum_{k=1}^{K}\widehat{T}_{t+1,k}\big]\sharp\widehat{\mu}_t, \widehat{\mu}_{t+1}\big)^2$ comes from the inexactness when approximating $\big[\frac{1}{K}\sum_{k=1}^{K}\widehat{T}_{t+1,k}\big]\sharp\widehat{\mu}_t$ by $\widehat{\mu}_{t+1} = \Big(\big[\frac{1}{K}\sum_{k=1}^{K}\widehat{T}_{t+1,k}\big]\sharp\widehat{\rho}_t\Big)\Big|_{\mathcal{X}_{R_{t+1}}}$, i.e., from the approximation in Line 7 of Conceptual Algorithm 1.*

To guarantee the convergence of $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$, we aim to control the two error terms $\frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\big\|\widehat{T}_{t+1,k} - T^{\widehat{\mu}_t}_{\nu_k}\big\|^2_{\mathcal{L}^2(\widehat{\mu}_t)}\Big|\mathcal{F}_t\Big]$ and $2\mathbb{E}\Big[\mathcal{W}_2\big(\big[\frac{1}{K}\sum_{k=1}^{K}\widehat{T}_{t+1,k}\big]\sharp\widehat{\mu}_t, \widehat{\mu}_{t+1}\big)^2\Big|\mathcal{F}_t\Big]$ on the right-hand side of (3.3) to be arbitrarily close to 0. Before presenting our concrete setting of Algorithm 2 that guarantees the convergence of $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$, let us first establish an intermediate result about choosing the truncation set $\mathcal{X}_{R_t}$ on Line 3 presented in the lemma below.

Lemma 3.11 (Choice of the truncation set)**.** *Let $\nu_1,\ldots,\nu_K \in \mathcal{M}$ and let $\rho \in \mathcal{M}_{\mathrm{full}}$. Moreover, let $G(\cdot)$ be the operator defined in (1.3), let $(\mathcal{X}_r)_{r\in\mathbb{N}}$ be a family of increasing sets satisfying Assumption 3.3, and let $\widehat{T}^{\mu,m}_{\nu,n}(\,\cdot\,;\theta)$ be a plug-in OT map estimator satisfying Assumption 3.4. Then, for any $\epsilon > 0$, there exist two numbers $\overline{r}_1(\rho,\epsilon), \overline{r}_2(\rho,\nu_1,\ldots,\nu_K,\epsilon) \in \mathbb{N}$ where $\overline{r}_1(\rho,\epsilon)$ depends on ρ,ϵ and $\overline{r}_2(\rho,\nu_1,\ldots,\nu_K,\epsilon)$ depends on $\rho,\nu_1,\ldots,\nu_K,\epsilon$ such that the following statements hold.*

(i) For all $r \ge \overline{r}_1(\rho,\epsilon)$, $\mu := \rho|_{\mathcal{X}_r}$ satisfies $\mathcal{W}_2(\mu,\rho)^2 \le \epsilon$.

(ii) For all $r \ge \overline{r}_2(\rho,\nu_1,\ldots,\nu_K,\epsilon)$ and for any $m,n,\theta \in \mathbb{N}$, $\mu := \rho|_{\mathcal{X}_r}$ satisfies

$$\mathbb{E}\Big[\mathcal{W}_2\big(\widehat{T}^{\mu,m}_{\nu_k,n}(\,\cdot\,;\theta)\sharp\mu, \widehat{T}^{\mu,m}_{\nu_k,n}(\,\cdot\,;\theta)\sharp\rho\big)^2\Big] \le \epsilon \qquad \forall 1 \le k \le K.$$

Moreover, in this case, $\bar{T} := \frac{1}{K}\sum_{k=1}^{K}\widehat{T}^{\mu,m}_{\nu_k,n}(\,\cdot\,;\theta)$ satisfies $\mathbb{E}\Big[\mathcal{W}_2\big(\bar{T}\sharp\mu, \bar{T}\sharp\rho\big)^2\Big] \le \epsilon$.

Proof of Lemma 3.11. Let us first prove statement (i). For every $r \in \mathbb{N}$, let us define $\widehat{\mu}_r := \rho|_{\mathcal{X}_r}$ and define $\breve{\mu}_r := \rho|_{\mathcal{X}_r^c}$, where $\mathcal{X}_r^c := \mathbb{R}^d \setminus \mathcal{X}_r$. Notice that $\rho = \rho(\mathcal{X}_r)\widehat{\mu}_r + (1-\rho(\mathcal{X}_r))\breve{\mu}_r$ for all $r \in \mathbb{N}$. Let $\pi_{r,1} := [I_d, I_d]\sharp\widehat{\mu}_r$ where $I_d : \mathbb{R}^d \to \mathbb{R}^d$ denotes the identity mapping on $\mathbb{R}^d$, let $\pi_{r,2} \in \Pi(\widehat{\mu}_r, \breve{\mu}_r)$ be arbitrary, and let $\pi_r := \rho(\mathcal{X}_r)\pi_{r,1} + (1-\rho(\mathcal{X}_r))\pi_{r,2} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$. One may check that $\pi_r \in \Pi(\hat{\mu}_r, \rho)$ for all $r \in \mathbb{N}$. Subsequently, it holds for all $r \in \mathbb{N}$ that

$$\begin{aligned}
\mathcal{W}_2(\hat{\mu}_r,\rho)^2 &\le \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\boldsymbol{x}-\boldsymbol{y}\|^2\,\pi_r(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{y})\\
&= \rho(\mathcal{X}_r)\int_{\mathbb{R}^d} \|\boldsymbol{x}-\boldsymbol{x}\|^2\,\hat{\mu}_r(\mathrm{d}\boldsymbol{x}) + (1-\rho(\mathcal{X}_r))\int_{\mathbb{R}^d\times\mathbb{R}^d}\|\boldsymbol{x}-\boldsymbol{y}\|^2\,\pi_{r,2}(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{y})\\
&\le (1-\rho(\mathcal{X}_r))\int_{\mathbb{R}^d\times\mathbb{R}^d} 2\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{y}\|^2\,\pi_{r,2}(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{y})\\
&= (1-\rho(\mathcal{X}_r))\int_{\mathbb{R}^d} 2\|\boldsymbol{x}\|^2\,\hat{\mu}_r(\mathrm{d}\boldsymbol{x}) + (1-\rho(\mathcal{X}_r))\int_{\mathbb{R}^d} 2\|\boldsymbol{y}\|^2\,\breve{\mu}_r(\mathrm{d}\boldsymbol{y})\\
&\le \frac{1-\rho(\mathcal{X}_r)}{\rho(\mathcal{X}_r)}\int_{\mathbb{R}^d} 2\|\boldsymbol{x}\|^2\,\rho(\mathrm{d}\boldsymbol{x}) + \int_{\mathbb{R}^d} 2\|\boldsymbol{y}\|^2\mathbb{1}_{\mathcal{X}_r^c}(\boldsymbol{y})\,\rho(\mathrm{d}\boldsymbol{y}).
\end{aligned}$$

Since $\bigcup_{r\in\mathbb{N}_0}\mathcal{X}_r = \mathbb{R}^d$ and $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ by assumption, it follows from Lebesgue's dominated convergence theorem that

$$\limsup_{r\to\infty}\mathcal{W}_2(\hat{\mu}_r,\rho)^2 \le \limsup_{r\to\infty}\frac{1-\rho(\mathcal{X}_r)}{\rho(\mathcal{X}_r)}\int_{\mathbb{R}^d} 2\|\boldsymbol{x}\|^2\,\rho(\mathrm{d}\boldsymbol{x}) + \limsup_{r\to\infty}\int_{\mathbb{R}^d} 2\|\boldsymbol{y}\|^2\mathbb{1}_{\mathcal{X}_r^c}(\boldsymbol{y})\,\rho(\mathrm{d}\boldsymbol{y}) = 0.$$

Therefore, for any $\epsilon > 0$, there exists $\overline{r}_1(\rho,\epsilon) \in \mathbb{N}$ such that $\mathcal{W}_2(\hat{\mu}_r,\rho)^2 \le \epsilon$ for all $r \ge \overline{r}_1(\rho,\epsilon)$. This proves statement (i).

Next, let us use the growth condition in Assumption 3.4(ii) to prove statement (ii). For every $r \in \mathbb{N}$, let $\hat{\mu}_r$, $\breve{\mu}_r$, $\pi_{r,1}$, $\pi_{r,2}$, and π_r be defined as in the proof of statement (i). Recall that $\pi_r \in \Pi(\hat{\mu}_r,\rho)$. Moreover, let $m,n,\theta \in \mathbb{N}$ be arbitrary and denote $\dot{T}_{k,r} := \widehat{T}^{\hat{\mu}_r,m}_{\nu_k,n}(\,\cdot\,;\theta)$ for $k = 1,\ldots,K$, $\dot{T}_r := \frac{1}{K}\sum_{k=1}^K \widehat{T}^{\hat{\mu}_r,m}_{\nu_k,n}(\,\cdot\,;\theta)$ for notational simplicity. Furthermore, for $k=1,\ldots,K$, we denote by $\dot{T}_{k,r}\otimes\dot{T}_{k,r}$ the function $\mathbb{R}^d\times\mathbb{R}^d \ni (\boldsymbol{x},\boldsymbol{y}) \mapsto \dot{T}_{k,r}\otimes\dot{T}_{k,r}(\boldsymbol{x},\boldsymbol{y}) := \big(\dot{T}_{k,r}(\boldsymbol{x}),\dot{T}_{k,r}(\boldsymbol{y})\big) \in \mathbb{R}^d\times\mathbb{R}^d$. In addition, we denote by $\dot{T}_r\otimes\dot{T}_r$ the function $\mathbb{R}^d\times\mathbb{R}^d \ni (\boldsymbol{x},\boldsymbol{y}) \mapsto \dot{T}_r\otimes\dot{T}_r(\boldsymbol{x},\boldsymbol{y}) := \big(\dot{T}_r(\boldsymbol{x}),\dot{T}_r(\boldsymbol{y})\big) \in \mathbb{R}^d\times\mathbb{R}^d$. It holds that $\big[\dot{T}_{k,r}\otimes\dot{T}_{k,r}\big]\sharp\pi_r \in \Pi\big(\dot{T}_{k,r}\sharp\hat{\mu}_r,\dot{T}_{k,r}\sharp\rho\big)$ for $k=1,\ldots,K$, and $\big[\dot{T}_r\otimes\dot{T}_r\big]\sharp\pi_r \in \Pi\big(\dot{T}_r\sharp\hat{\mu}_r,\dot{T}_r\sharp\rho\big)$. Therefore, for $k=1,\ldots,K$, we are able to bound $\mathcal{W}_2\big(\dot{T}_{k,r}\sharp\hat{\mu}_r,\dot{T}_{k,r}\sharp\rho\big)^2$ by

$$\begin{aligned}
\mathcal{W}_2\big(\dot{T}_{k,r}\sharp\hat{\mu}_r,\dot{T}_{k,r}\sharp\rho\big)^2 &\le \int_{\mathbb{R}^d\times\mathbb{R}^d}\big\|\boldsymbol{x}-\boldsymbol{y}\big\|^2\,\big[\dot{T}_{k,r}\otimes\dot{T}_{k,r}\big]\sharp\pi_r(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{y})\\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d}\big\|\dot{T}_{k,r}(\boldsymbol{x})-\dot{T}_{k,r}(\boldsymbol{y})\big\|^2\,\pi_r(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{y})\\
&= \rho(\mathcal{X}_r)\int_{\mathbb{R}^d}\big\|\dot{T}_{k,r}(\boldsymbol{x})-\dot{T}_{k,r}(\boldsymbol{x})\big\|^2\,\hat{\mu}_r(\mathrm{d}\boldsymbol{x})\\
&\qquad + (1-\rho(\mathcal{X}_r))\int_{\mathbb{R}^d\times\mathbb{R}^d}\big\|\dot{T}_{k,r}(\boldsymbol{x})-\dot{T}_{k,r}(\boldsymbol{y})\big\|^2\,\pi_{r,2}(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{y})\\
&\le (1-\rho(\mathcal{X}_r))\int_{\mathbb{R}^d\times\mathbb{R}^d} 2\big\|\dot{T}_{k,r}(\boldsymbol{x})\big\|^2 + 2\big\|\dot{T}_{k,r}(\boldsymbol{y})\big\|^2\,\pi_{r,2}(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{y})\\
&= (1-\rho(\mathcal{X}_r))\int_{\mathbb{R}^d} 2\big\|\dot{T}_{k,r}(\boldsymbol{x})\big\|^2\,\hat{\mu}_r(\mathrm{d}\boldsymbol{x})\\
&\qquad + (1-\rho(\mathcal{X}_r))\int_{\mathbb{R}^d} 2\big\|\dot{T}_{k,r}(\boldsymbol{y})\big\|^2\,\breve{\mu}_r(\mathrm{d}\boldsymbol{y})\\
&\le \frac{1-\rho(\mathcal{X}_r)}{\rho(\mathcal{X}_r)}\int_{\mathbb{R}^d} 2\big\|\dot{T}_{k,r}(\boldsymbol{x})\big\|^2\,\rho(\mathrm{d}\boldsymbol{x}) + \int_{\mathbb{R}^d} 2\big\|\dot{T}_{k,r}(\boldsymbol{y})\big\|^2\mathbb{1}_{\mathcal{X}_r^c}(\boldsymbol{y})\,\rho(\mathrm{d}\boldsymbol{y}).
\end{aligned} \tag{3.14}$$

For $k=1,\ldots,K$, observe that the growth condition of $\dot{T}_{k,r}$ guarantees $\mathbb{E}\Big[\big\|\dot{T}_{k,r}(\boldsymbol{x})\big\|^2\Big] \le u_1(\nu_k)+u_2(\nu_k)\|\boldsymbol{x}\|^2$ for all $\boldsymbol{x}\in\mathbb{R}^d$, where $u_1(\nu_k)\in\mathbb{R}_+$ and $u_2(\nu_k)\in\mathbb{R}_+$ are constants that only depend on ν_k. Let $\overline{u}_1 := \max_{1\le k\le K}\big\{u_1(\nu_k)\big\}$ and $\overline{u}_2 := \max_{1\le k\le K}\big\{u_2(\nu_k)\big\}$. It thus holds that

$$\mathbb{E}\Big[\big\|\dot{T}_{k,r}(\boldsymbol{x})\big\|^2\Big] \le \overline{u}_1 + \overline{u}_2\|\boldsymbol{x}\|^2 \qquad \forall \boldsymbol{x}\in\mathbb{R}^d,\ \forall 1\le k\le K. \tag{3.15}$$

Taking expectations on both sides of (3.14) then applying Fubini's theorem and (3.15) yields

$$\begin{aligned}\mathbb{E}\Big[\mathcal{W}_2\big(\dot{T}_{k,r}\sharp\widehat{\mu}_r,\dot{T}_{k,r}\sharp\rho\big)^2\Big]&\le\frac{1-\rho(\mathcal{X}_r)}{\rho(\mathcal{X}_r)}\int_{\mathbb{R}^d}2\mathbb{E}\Big[\big\|\dot{T}_{k,r}(\boldsymbol{x})\big\|^2\Big]\,\rho(\mathrm{d}\boldsymbol{x})+\int_{\mathbb{R}^d}2\mathbb{E}\Big[\big\|\dot{T}_{k,r}(\boldsymbol{y})\big\|^2\Big]\mathbb{1}_{\mathcal{X}_r^c}(\boldsymbol{y})\,\rho(\mathrm{d}\boldsymbol{y})\\&\le\frac{1-\rho(\mathcal{X}_r)}{\rho(\mathcal{X}_r)}\int_{\mathbb{R}^d}2\big(\overline{u}_1+\overline{u}_2\|\boldsymbol{x}\|^2\big)\,\rho(\mathrm{d}\boldsymbol{x})+\int_{\mathbb{R}^d}2\big(\overline{u}_1+\overline{u}_2\|\boldsymbol{y}\|^2\big)\mathbb{1}_{\mathcal{X}_r^c}(\boldsymbol{y})\,\rho(\mathrm{d}\boldsymbol{y})\\&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall 1\le k\le K.\end{aligned}$$

Same as in the proof of statement (i), since $\bigcup_{r\in\mathbb{N}_0}\mathcal{X}_r=\mathbb{R}^d$ and $\rho\in\mathcal{P}_2(\mathbb{R}^d)$ by assumption, it follows from Lebesgue's dominated convergence theorem that

$$\begin{aligned}\limsup_{r\to\infty}\mathbb{E}\Big[\mathcal{W}_2\big(\dot{T}_{k,r}\sharp\widehat{\mu}_r,\dot{T}_{k,r}\sharp\rho\big)^2\Big]&\le\limsup_{r\to\infty}\frac{1-\rho(\mathcal{X}_r)}{\rho(\mathcal{X}_r)}\int_{\mathbb{R}^d}2\big(\overline{u}_1+\overline{u}_2\|\boldsymbol{x}\|^2\big)\,\rho(\mathrm{d}\boldsymbol{x})\\&\qquad+\limsup_{r\to\infty}\int_{\mathbb{R}^d}2\big(\overline{u}_1+\overline{u}_2\|\boldsymbol{y}\|^2\big)\mathbb{1}_{\mathcal{X}_r^c}(\boldsymbol{y})\,\rho(\mathrm{d}\boldsymbol{y})\\&=0\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall 1\le k\le K.\end{aligned}$$

Therefore, for any $\epsilon>0$, there exists $\overline{r}_2(\rho,\nu_1,\ldots,\nu_K,\epsilon)\in\mathbb{N}$ such that for any $m,n,\theta\in\mathbb{N}$, the inequality $\mathbb{E}\Big[\mathcal{W}_2\big(\dot{T}_{k,r}\sharp\widehat{\mu}_r,\dot{T}_{k,r}\sharp\rho\big)^2\Big]\le\epsilon$ holds for all $k=1,\ldots,K$ and all $r\ge\overline{r}_2(\rho,\nu_1,\ldots,\nu_K,\epsilon)$. Furthermore, repeating the same derivation in (3.14) with $\dot{T}_{k,r}$ replaced by $\dot{T}_r$ yields

$$\mathcal{W}_2\big(\dot{T}_r\sharp\widehat{\mu}_r,\dot{T}_r\sharp\rho\big)^2\le\frac{1-\rho(\mathcal{X}_r)}{\rho(\mathcal{X}_r)}\int_{\mathbb{R}^d}2\big\|\dot{T}_r(\boldsymbol{x})\big\|^2\,\rho(\mathrm{d}\boldsymbol{x})+\int_{\mathbb{R}^d}2\big\|\dot{T}_r(\boldsymbol{y})\big\|^2\mathbb{1}_{\mathcal{X}_r^c}(\boldsymbol{y})\,\rho(\mathrm{d}\boldsymbol{y}).\tag{3.16}$$

Since the convexity of $\mathbb{R}^d\ni\boldsymbol{z}\mapsto\|\boldsymbol{z}\|^2\in\mathbb{R}$ and Jensen's inequality imply

$$\mathbb{E}\Big[\big\|\dot{T}_r(\boldsymbol{x})\big\|^2\Big]\le\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\big\|\dot{T}_{k,r}(\boldsymbol{x};\theta)\big\|^2\Big]\le\overline{u}_1+\overline{u}_2\|\boldsymbol{x}\|^2\qquad\forall\boldsymbol{x}\in\mathbb{R}^d,\tag{3.17}$$

taking expectations on both sides of (3.16) then applying Fubini's theorem and (3.17) leads to

$$\mathbb{E}\Big[\mathcal{W}_2\big(\dot{T}_r\sharp\widehat{\mu}_r,\dot{T}_r\sharp\rho\big)^2\Big]\le\frac{1-\rho(\mathcal{X}_r)}{\rho(\mathcal{X}_r)}\int_{\mathbb{R}^d}2\big(\overline{u}_1+\overline{u}_2\|\boldsymbol{x}\|^2\big)\,\rho(\mathrm{d}\boldsymbol{x})+\int_{\mathbb{R}^d}2\big(\overline{u}_1+\overline{u}_2\|\boldsymbol{y}\|^2\big)\mathbb{1}_{\mathcal{X}_r^c}(\boldsymbol{y})\,\rho(\mathrm{d}\boldsymbol{y}).$$

Consequently, it follows from the same argument as above that $\mathbb{E}\Big[\mathcal{W}_2\big(\dot{T}_r\sharp\widehat{\mu}_r,\dot{T}_r\sharp\rho\big)^2\Big]\le\epsilon$ holds for all $r\ge\overline{r}_2(\rho,\nu_1,\ldots,\nu_K,\epsilon)$. The proof is now complete. □

The results in Proposition 3.9 and Lemma 3.11 suggest the following sufficient conditions for the convergence of Algorithm 2.

Setting 3.12 (Convergence conditions for Algorithm 2)**.** *Let $\beta>0$. In addition to Setting 3.5, let the $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$-adapted stochastic processes $(R_t)_{t\in\mathbb{N}_0}$, $(N_{t,k})_{t\in\mathbb{N}_0,\,k=1:K}$, and $(\Theta_{t,k})_{t\in\mathbb{N}_0,\,k=1:K}$ in Algorithm 2 be chosen as follows.*

(a) For every $t\in\mathbb{N}_0$, let R_t be set as follows:

$$\begin{aligned}R_0&:=\overline{r}_2(\widehat{\rho}_0,\nu_1,\ldots,\nu_K,1),\\R_t&:=\max\Big\{\overline{r}_1\big(\widehat{\rho}_t,t^{-(1+\beta)}\big),\overline{r}_2\big(\widehat{\rho}_t,\nu_1,\ldots,\nu_K,(t+1)^{-2(1+\beta)}\big)\Big\}\qquad\forall t\ge1,\end{aligned}$$

where $\overline{r}_1(\,\cdot\,,\cdot\,)$ and $\overline{r}_2(\,\cdot\,,\ldots,\cdot\,)$ are given by Lemma 3.11. Note that R_t is $\mathcal{F}_t$-measurable for all $t\in\mathbb{N}_0$.

(b) For every $t\in\mathbb{N}_0$ and for $k=1,\ldots,K$, let $N_{t,k}:=\overline{n}\big(\widehat{\mu}_t,\nu_k,(t+1)^{-2(1+\beta)}\big)$, $\Theta_{t,k}:=\overline{\theta}\big(\widehat{\mu}_t,\nu_k,N_{t,k},N_{t,k},(t+1)^{-2(1+\beta)}\big)$, where $\overline{n}(\,\cdot\,,\cdot\,,\cdot\,)$ and $\overline{\theta}(\,\cdot\,,\cdot\,,\cdot\,,\cdot\,,\cdot\,)$ are given by Assumption 3.4(iii). Note that $N_{t,k}$ and $\Theta_{t,k}$ are $\mathcal{F}_t$-measurable for all $t\in\mathbb{N}_0$.

We are now ready to present our main convergence result.

Theorem 3.13 (Convergence of Algorithm 2)**.** *Let the inputs of Algorithm 2 satisfy Setting 3.5 and let $\big(\Omega,\mathcal{F},\mathbb{P},(\mathcal{F}_t)_{t\in\mathbb{N}_0}\big)$ be the filtered probability space constructed by Algorithm 2. Let the $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$-adapted stochastic processes $(R_t)_{t\in\mathbb{N}_0}$, $(N_{t,k})_{t\in\mathbb{N}_0,\,k=1:K}$, and $(\Theta_{t,k})_{t\in\mathbb{N}_0,\,k=1:K}$ in Algorithm 2 be specified by Setting 3.12, and let $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$ be the output of Algorithm 2. Then, the following statements hold.*

(i) It holds $\mathbb{P}$-almost surely that $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$ is a tight sequence of probability measures, and that every accumulation point of $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$ with respect to the $\mathcal{W}_2$ metric is a fixed-point of G.

(ii) In particular, if G has a unique fixed-point, then $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$ converges $\mathbb{P}$-almost surely in $\mathcal{W}_2$ to the Wasserstein barycenter of $\nu_1,\dots,\nu_K$.

Proof of Theorem 3.13. Let us denote $\bar{T}_{t+1} := \frac{1}{K}\sum_{k=1}^{K}\widehat{T}_{t+1,k}$. Recall that $\widehat{\rho}_{t+1} := \bar{T}_{t+1}\sharp\widehat{\rho}_t$ by Line 10 of Algorithm 2. As implied by Setting 3.12 and the properties of $\overline{r}_1(\cdot,\cdot)$, $\overline{r}_2(\cdot,\dots,\cdot)$, $\overline{n}(\cdot,\cdot,\cdot)$, $\overline{\theta}(\cdot,\cdot,\cdot,\cdot,\cdot)$ in Lemma 3.11 and Assumption 3.4(iii), the following inequalities hold $\mathbb{P}$-almost surely:

$$\mathcal{W}_2(\widehat{\mu}_{t+1},\widehat{\rho}_{t+1})^2 \le (t+1)^{-(1+\beta)} \qquad \forall t\in\mathbb{N}_0, \tag{3.18}$$

$$\mathbb{E}\Big[\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\mu}_t,\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t\big)^2\Big|\mathcal{F}_t\Big] \le (t+1)^{-2(1+\beta)} \qquad \forall 1\le k\le K,\ \forall t\in\mathbb{N}_0, \tag{3.19}$$

$$\mathbb{E}\Big[\mathcal{W}_2\big(\bar{T}_{t+1}\sharp\widehat{\mu}_t,\widehat{\rho}_{t+1}\big)^2\Big|\mathcal{F}_t\Big] \le (t+1)^{-2(1+\beta)} \qquad \forall t\in\mathbb{N}_0, \tag{3.20}$$

$$\mathbb{E}\Big[\big\|\widehat{T}_{t+1,k}-T_{\nu_k}^{\widehat{\mu}_t}\big\|^2_{\mathcal{L}^2(\widehat{\mu}_t)}\Big|\mathcal{F}_t\Big] \le (t+1)^{-2(1+\beta)} \qquad \forall 1\le k\le K,\ \forall t\in\mathbb{N}_0, \tag{3.21}$$

where $\beta>0$. The proof of statement (i) is divided into five steps.

Step 1: showing that $\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t \xrightarrow[t\to\infty]{\mathcal{W}_2} \nu_k$ $\mathbb{P}$-almost surely for $k=1,\dots,K$. Notice that, for each $t\in\mathbb{N}_0$ and for $k=1,\dots,K$, it holds that $\big[\widehat{T}_{t+1,k},T_{\nu_k}^{\widehat{\mu}_t}\big]\sharp\widehat{\mu}_t \in \Pi\big(\widehat{T}_{t+1,k}\sharp\widehat{\mu}_t,\nu_k\big)$. Thus, we have

$$\begin{aligned}
\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t,\nu_k\big)^2 &\le \Big(\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\mu}_t,\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t\big)+\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\mu}_t,\nu_k\big)\Big)^2\\
&\le 2\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\mu}_t,\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t\big)^2+2\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\mu}_t,\nu_k\big)^2\\
&\le 2\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\mu}_t,\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t\big)^2+2\int_{\mathbb{R}^d}\big\|\widehat{T}_{t+1,k}(\boldsymbol{x})-T_{\nu_k}^{\widehat{\mu}_t}(\boldsymbol{x})\big\|^2\,\widehat{\mu}_t(\boldsymbol{x})\\
&= 2\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\mu}_t,\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t\big)^2+2\big\|\widehat{T}_{t+1,k}-T_{\nu_k}^{\widehat{\mu}_t}\big\|^2_{\mathcal{L}^2(\widehat{\mu}_t)}.
\end{aligned}$$

Taking expectations on both sides conditional on $\mathcal{F}_t$ and then applying (3.19) and (3.21) yields

$$\mathbb{E}\Big[\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t,\nu_k\big)^2\Big|\mathcal{F}_t\Big] \le 4(t+1)^{-2(1+\beta)} \qquad \forall 1\le k\le K,\ \forall t\in\mathbb{N}_0.$$

Applying the law of total expectation and Markov's inequality then gives

$$\mathbb{P}\Big[\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t,\nu_k\big)^2 \ge (t+1)^{-(1+\beta)}\Big] \le (t+1)^{1+\beta}\mathbb{E}\Big[\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t,\nu_k\big)^2\Big] \le 4(t+1)^{-(1+\beta)}$$
$$\forall 1\le k\le K,\ \forall t\in\mathbb{N}_0.$$

Subsequently, since $\sum_{t\in\mathbb{N}_0}4(t+1)^{-(1+\beta)}<\infty$, we conclude by the Borel–Cantelli lemma that, $\mathbb{P}$-almost surely, $\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t,\nu_k\big)^2 \le (t+1)^{-(1+\beta)}$ holds for all but finitely many $t\in\mathbb{N}_0$, and therefore $\lim_{t\to\infty}\mathcal{W}_2\big(\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t,\nu_k\big)=0$ $\mathbb{P}$-almost surely for $k=1,\dots,K$.

Step 2: showing that $(\widehat{\rho}_t)_{t\in\mathbb{N}_0}$ is tight $\mathbb{P}$-almost surely. By Prokhorov's theorem, for $k=1,\dots,K$, $\big(\widehat{T}_{t+1,k}\sharp\widehat{\rho}_t\big)_{t\in\mathbb{N}_0}$ is $\mathbb{P}$-almost surely a tight sequence of probability measures since it is $\mathbb{P}$-almost surely convergent. Let $\eta_t := \big[\widehat{T}_{t+1,1},\dots,\widehat{T}_{t+1,K}\big]\sharp\widehat{\rho}_t \in \mathcal{P}(\underbrace{\mathbb{R}^d\times\cdots\times\mathbb{R}^d}_{K\text{ copies}})$ for $t\in\mathbb{N}_0$. It hence holds $\mathbb{P}$-almost surely that each marginal of each probability measure in $(\eta_t)_{t\in\mathbb{N}_0}$ (on each copy of $\mathbb{R}^d$) belongs to a tight set of probability measures on $\mathbb{R}^d$, and it thus follows from a multi-marginal generalization of [96, Lemma 4.4] that $(\eta_t)_{t\in\mathbb{N}_0}$ is a tight set of probability measures on $(\mathbb{R}^d)^K$. Let A denote the mapping $(\mathbb{R}^d)^K\ni(\boldsymbol{x}_1,\dots,\boldsymbol{x}_K)\mapsto\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}_k\in\mathbb{R}^d$. Hence, we have $\widehat{\rho}_{t+1}=\bar{T}_{t+1}\sharp\widehat{\rho}_t=A\sharp\eta_t$ for all $t\in\mathbb{N}_0$. Consequently, the tightness of $(\eta_t)_{t\in\mathbb{N}_0}$ and the continuity of the mapping A imply the tightness of $(\widehat{\rho}_t)_{t\in\mathbb{N}_0}$ in the $\mathbb{P}$-almost sure sense.

Step 3: showing that $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$ is tight $\mathbb{P}$-almost surely. It follows from (3.18) that $\lim_{t\to\infty}\mathcal{W}_2(\widehat{\mu}_t,\widehat{\rho}_t)=0$ $\mathbb{P}$-almost surely, and hence $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$ is $\mathbb{P}$-almost surely sequentially precompact due to the $\mathbb{P}$-almost sure tightness of $(\widehat{\rho}_t)_{t\in\mathbb{N}_0}$ and Prokhorov's theorem. Applying Prokhorov's theorem again then yields the $\mathbb{P}$-almost sure tightness of $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$.

Step 4: constructing a subset $\widetilde{\Omega}\subseteq\Omega$ with $\mathbb{P}[\widetilde{\Omega}]=1$ in which the convergence is analyzed. Similar to the argument used in Step 1, applications of the law of total expectation together with Markov's inequality to (3.20) and (3.21) lead to

$$\mathbb{P}\Big[\mathcal{W}_2\big(\bar{T}_{t+1}\sharp\widehat{\mu}_t,\widehat{\rho}_{t+1}\big)^2\geq(t+1)^{-(1+\beta)}\Big]\leq(t+1)^{-(1+\beta)}\qquad\qquad\forall t\in\mathbb{N}_0,$$
$$\mathbb{P}\Big[\big\|\widehat{T}_{t+1,k}-T^{\widehat{\mu}_t}_{\nu_k}\big\|^2_{\mathcal{L}^2(\widehat{\mu}_t)}\geq(t+1)^{-(1+\beta)}\Big]\leq(t+1)^{-(1+\beta)}\qquad\forall 1\leq k\leq K,\ \forall t\in\mathbb{N}_0.$$

Since $\sum_{t\in\mathbb{N}_0}(t+1)^{-(1+\beta)}<\infty$, we use the Borel–Cantelli lemma again to show that, $\mathbb{P}$-almost surely, $\mathcal{W}_2\big(\bar{T}_{t+1}\sharp\widehat{\mu}_t,\widehat{\rho}_{t+1}\big)^2\leq(t+1)^{-(1+\beta)}$ and $\big\|\widehat{T}_{t+1,k}-T^{\widehat{\mu}_t}_{\nu_k}\big\|^2_{\mathcal{L}^2(\widehat{\mu}_t)}\leq(t+1)^{-(1+\beta)}$ $\forall 1\leq k\leq K$ hold for all but finitely many $t\in\mathbb{N}_0$. In the following, for every $\omega\in\Omega$, let us use the notations $\widehat{\rho}^{(\omega)}_t$, $\widehat{\mu}^{(\omega)}_t$, $\widehat{T}^{(\omega)}_{t+1,k}$, $\bar{T}^{(\omega)}_{t+1}$ to explicitly express the dependence of $\widehat{\rho}_t$, $\widehat{\mu}_t$, $\widehat{T}_{t+1,k}$, $\bar{T}_{t+1}$ on ω. The above analyses have shown the existence of an $\mathcal{F}$-measurable set $\widetilde{\Omega}\subseteq\Omega$ with $\mathbb{P}[\widetilde{\Omega}]=1$, which satisfies:

$$\forall\omega\in\widetilde{\Omega},\ \exists\bar{t}^{(\omega)}\in\mathbb{N},\ \begin{cases}\big(\widehat{\mu}^{(\omega)}_t\big)_{t\in\mathbb{N}_0}\text{ is tight,}\\ \mathcal{W}_2\big(\widehat{\mu}^{(\omega)}_{t+1},\widehat{\rho}^{(\omega)}_{t+1}\big)^2\leq(t+1)^{-(1+\beta)} & \forall t\geq\bar{t}^{(\omega)},\\ \mathcal{W}_2\big(\bar{T}^{(\omega)}_{t+1}\sharp\widehat{\mu}^{(\omega)}_t,\widehat{\rho}^{(\omega)}_{t+1}\big)^2\leq(t+1)^{-(1+\beta)} & \forall t\geq\bar{t}^{(\omega)},\\ \Big\|\widehat{T}^{(\omega)}_{t+1,k}-T^{\widehat{\mu}^{(\omega)}_t}_{\nu_k}\Big\|^2_{\mathcal{L}^2\left(\widehat{\mu}^{(\omega)}_t\right)}\leq(t+1)^{-(1+\beta)} & \forall t\geq\bar{t}^{(\omega)},\ \forall 1\leq k\leq K.\end{cases}\tag{3.22}$$

Step 5: showing that for every $\omega\in\widetilde{\Omega}$, every accumulation point of $\big(\widehat{\mu}^{(\omega)}_t\big)_{t\in\mathbb{N}_0}$ is a fixed-point of G. Let us fix an arbitrary $\omega\in\widetilde{\Omega}$ and suppose there is a subsequence $(t_i)_{i\in\mathbb{N}_0}$ such that $\widehat{\mu}^{(\omega)}_{t_i}\xrightarrow[i\to\infty]{\mathcal{W}_2}\widehat{\mu}^{(\omega)}_\infty\in\mathcal{P}_2(\mathbb{R}^d)$. The continuity of $V(\cdot)$ on $\mathcal{P}_2(\mathbb{R}^d)$ then implies that $\lim_{i\to\infty}V\big(\widehat{\mu}^{(\omega)}_{t_i}\big)=V\big(\widehat{\mu}^{(\omega)}_\infty\big)$. Removing finitely many initial terms from $(t_i)_{i\in\mathbb{N}_0}$ if necessary, we assume without loss of generality that $t_0\geq\bar{t}^{(\omega)}$. For each $i\in\mathbb{N}_0$, summing (3.2) over $s=t_i,t_i+1,\ldots,t_{i+1}-1$, using the inequality $\mathcal{W}_2\big(\bar{T}^{(\omega)}_{s+1}\sharp\widehat{\mu}^{(\omega)}_s,\widehat{\mu}^{(\omega)}_{s+1}\big)^2\leq 2\mathcal{W}_2\big(\bar{T}^{(\omega)}_{s+1}\sharp\widehat{\mu}^{(\omega)}_s,\widehat{\rho}^{(\omega)}_{s+1}\big)^2+2\mathcal{W}_2\big(\widehat{\mu}^{(\omega)}_{s+1},\widehat{\rho}^{(\omega)}_{s+1}\big)^2$, and using the properties in (3.22) show that

$$\begin{aligned}V\big(\widehat{\mu}^{(\omega)}_{t_{i+1}}\big)-V\big(\widehat{\mu}^{(\omega)}_{t_i}\big)&=\sum_{s=t_i}^{t_{i+1}-1}V\big(\widehat{\mu}^{(\omega)}_{s+1}\big)-V\big(\widehat{\mu}^{(\omega)}_s\big)\\&\leq-\left(\sum_{s=t_i}^{t_{i+1}-1}\mathcal{W}_2\Big(\widehat{\mu}^{(\omega)}_s,G\big(\widehat{\mu}^{(\omega)}_s\big)\Big)^2\right)+\left(\sum_{s=t_i}^{t_{i+1}-1}\frac{2}{K}\sum_{k=1}^{K}\Big\|\widehat{T}^{(\omega)}_{s+1,k}-T^{\widehat{\mu}^{(\omega)}_s}_{\nu_k}\Big\|^2_{\mathcal{L}^2\left(\widehat{\mu}^{(\omega)}_s\right)}\right)\\&\quad+\left(\sum_{s=t_i}^{t_{i+1}-1}4\mathcal{W}_2\big(\bar{T}^{(\omega)}_{s+1}\sharp\widehat{\mu}^{(\omega)}_s,\widehat{\rho}^{(\omega)}_{s+1}\big)^2\right)+\left(\sum_{s=t_i}^{t_{i+1}-1}4\mathcal{W}_2\big(\widehat{\mu}^{(\omega)}_{s+1},\widehat{\rho}^{(\omega)}_{s+1}\big)^2\right)\\&\leq-\left(\sum_{s=t_i}^{t_{i+1}-1}\mathcal{W}_2\Big(\widehat{\mu}^{(\omega)}_s,G\big(\widehat{\mu}^{(\omega)}_s\big)\Big)^2\right)+\left(\sum_{s=t_i}^{t_{i+1}-1}10(s+1)^{-(1+\beta)}\right)\\&\leq-\mathcal{W}_2\Big(\widehat{\mu}^{(\omega)}_{t_i},G\big(\widehat{\mu}^{(\omega)}_{t_i}\big)\Big)^2+\left(\sum_{s=t_i}^{\infty}10(s+1)^{-(1+\beta)}\right)\qquad\forall i\in\mathbb{N}_0.\end{aligned}$$

Rearranging the terms above leads to

$$\mathcal{W}_2\Big(\widehat{\mu}^{(\omega)}_{t_i},G\big(\widehat{\mu}^{(\omega)}_{t_i}\big)\Big)^2\leq\Big|V\big(\widehat{\mu}^{(\omega)}_{t_{i+1}}\big)-V\big(\widehat{\mu}^{(\omega)}_{t_i}\big)\Big|+\left(\sum_{s=t_i}^{\infty}10(s+1)^{1+\beta}\right)\qquad\forall i\in\mathbb{N}_0.\tag{3.23}$$

Since $\sum_{s=0}^{\infty}(s+1)^{-(1+\beta)}$ is a convergent series, (3.23) implies that

$$\limsup_{i\to\infty}\mathcal{W}_2\Big(\widehat{\mu}^{(\omega)}_{t_i},G\big(\widehat{\mu}^{(\omega)}_{t_i}\big)\Big)^2\leq\limsup_{i\to\infty}\Big|V\big(\widehat{\mu}^{(\omega)}_{t_{i+1}}\big)-V\big(\widehat{\mu}^{(\omega)}_{t_i}\big)\Big|+\limsup_{i\to\infty}\left(\sum_{s=t_i}^{\infty}10(s+1)^{1+\beta}\right)=0.$$

This shows that $G\big(\widehat{\mu}^{(\omega)}_{t_i}\big)\xrightarrow[i\to\infty]{\mathcal{W}_2}\widehat{\mu}^{(\omega)}_\infty$. Moreover, for any $\mu\in\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, the analysis in [5, Remark 3.2] demonstrates that the density function $f_{G(\mu)}$ of $G(\mu)\in\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ satisfies $\sup_{\boldsymbol{x}\in\mathbb{R}^d}\big\{f_{G(\mu)}(\boldsymbol{x})\big\}\leq$

$K^d \sup_{\boldsymbol{x}\in\mathrm{supp}(\nu_1)}\big\{f_{\nu_1}(\boldsymbol{x})\big\} < \infty$, where f_{ν_1} denotes the density function of ν_1. Consequently, it holds for every open set $E \subseteq \mathbb{R}^d$ that $\widehat{\mu}_\infty^{(\omega)}(E) \le \liminf_{i\to\infty} G\big(\widehat{\mu}_{t_i}^{(\omega)}\big)(E) \le K^d \sup_{\boldsymbol{x}\in\mathrm{supp}(\nu_1)}\big\{f_{\nu_1}(\boldsymbol{x})\big\}\mathscr{L}(E)$, where $\mathscr{L}$ denotes the Lebesgue measure on $\mathbb{R}^d$. It thus follows that $\widehat{\mu}_\infty^{(\omega)} \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. Now, the continuity of the mapping $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \ni \mu \mapsto \mathcal{W}_2\big(\mu, G(\mu)\big)^2 \in \mathbb{R}_+$ in Theorem 1.2(i) implies that $\mathcal{W}_2\Big(\widehat{\mu}_\infty^{(\omega)}, G\big(\widehat{\mu}_\infty^{(\omega)}\big)\Big)^2 = \lim_{i\to\infty} \mathcal{W}_2\Big(\widehat{\mu}_{t_i}^{(\omega)}, G\big(\widehat{\mu}_{t_i}^{(\omega)}\big)\Big)^2 = 0$, which shows that $\widehat{\mu}_\infty^{(\omega)}$ is a fixed-point of G. Since $\mathbb{P}[\widetilde{\Omega}] = 1$, it holds $\mathbb{P}$-almost surely that every accumulation point of $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$ is a fixed-point of G. We have thus completed the proof of statement (i).

Finally, if G has a unique fixed-point $\bar{\mu} \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, then statement (i) implies that, $\mathbb{P}$-almost surely, every accumulation point of $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$ is equal to $\bar{\mu}$. Therefore, $(\widehat{\mu}_t)_{t\in\mathbb{N}_0}$ converges $\mathbb{P}$-almost surely to $\bar{\mu}$, which is the unique Wasserstein barycenter of $\nu_1, \dots, \nu_K$ by Theorem 1.2(ii). The proof is now complete. □

Remark 3.14 (Computational tractability of Algorithm 2)**.** *Assume that: (i) independent random samples from $\nu_1, \dots, \nu_K$, and ρ_0 can be efficiently generated; (ii) the plug-in OT map estimator $\widehat{T}_{\nu,n}^{\mu,m}(\,\cdot\,;\theta)$ can be tractably computed and can be tractably evaluated at any point $\boldsymbol{x} \in \mathbb{R}^d$; (iii) for all $r \in \mathbb{N}$, checking whether a point $\boldsymbol{x} \in \mathbb{R}^d$ belongs to $\mathcal{X}_r$ is computationally tractable. Then, Algorithm 2 is computationally tractable. Indeed, for $t \in \mathbb{N}$, a random sample from $\widehat{\mu}_t$ can be generated by rejection sampling. Specifically, one first generates a random sample $\boldsymbol{X} \in \mathbb{R}^d$ from ρ_0 and evaluates the composition $\widehat{\boldsymbol{X}} := \big[\sum_{k=1}^K \widehat{T}_{t,k}\big] \circ \cdots \circ \big[\sum_{k=1}^K \widehat{T}_{1,k}\big](\boldsymbol{X})$. This sample $\widehat{\boldsymbol{X}}$ is subsequently accepted if $\widehat{\boldsymbol{X}} \in \mathcal{X}_{R_t}$. Otherwise, this process repeatedly generates $\widehat{\boldsymbol{X}}$ until $\widehat{\boldsymbol{X}}$ is accepted. The computational tractability of the plug-in OT map estimator $\widehat{T}_{\nu,n}^{\mu,m}(\,\cdot\,;\theta)$ is discussed in Remark 4.9 and Remark 4.11 in Section 4.*

Remark 3.15. *We would like to remark that the operator G in (1.3) does not always have a unique fixed-point for general input probability measures $\nu_1, \dots, \nu_K \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$; see, e.g., Example 3.1 of [5] for a concrete counterexample. It is known that G has a unique fixed-point when $\nu_1, \dots, \nu_K$ belong to the same parametric family of elliptical distributions [5, Section 4], e.g., Gaussian distributions. However, to the best of our knowledge, sufficient conditions to guarantee the uniqueness of the fixed-point of G for non-parametric $\nu_1, \dots, \nu_K$ is still an open problem.*

Remark 3.16. *Same as the deterministic fixed-point iterative scheme of Álvarez-Esteban et al. [5], our stochastic extension in Algorithm 2 does not provide any non-asymptotic rate of convergence. [To discuss: add discussions about empirical evidence?]*

4. Concrete Examples of Plug-In OT Map Estimators

As stated in Setting 3.12, the convergence of Algorithm 2 depends crucially on the plug-in OT map estimator $\widehat{T}_{\nu,n}^{\mu,m}(\,\cdot\,;\theta)$, specifically on its shape, growth, and consistency properties required by Assumption 3.4. In this section, we consider two admissible compactly supported probability measures $\mu, \nu \in \mathcal{M}$ and introduce two concrete examples of plug-in OT map estimators that satisfy Assumption 3.4. For the sake of notational simplicity, we will omit μ, ν, m, n in the notations for estimators in this section. Nonetheless, m and n will always be understood as the numbers of samples from μ and ν, respectively.

Both of our examples are inspired by the estimation error bound of plug-in OT map estimators developed by Manole et al. [59] as well as the shape-constrained convex least squares regression and shape-constrained convex interpolation methods developed by Taylor [91], Taylor et al. [92]. We will first introduce these preliminary results in Section 4.1. Subsequently, we will introduce our kernel-smoothed OT map estimator $\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)$ in Section 4.2 and introduce our barrier-based map estimator $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta)$ in Section 4.3.

4.1. Preliminaries on plug-in OT map estimators and shaped-constrained interpolation.

Both of our proposed estimators are based on the following shape-constrained convex least squares OT map estimator, defined as follows.

Definition 4.1 (Shape-constrained convex least squares OT map estimator)**.** *Let $\mu, \nu \in \mathcal{M}$ (recall Definition 3.1) with $\mathbf{0}_d \in \mathrm{supp}(\mu)$, let T_ν^μ be the OT map from μ to ν, and let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\underline{\lambda} \equiv \underline{\lambda}(\mu,\nu)$ and $\overline{\lambda} \equiv \overline{\lambda}(\mu,\nu)$ be chosen based on μ and ν such that $0 < \underline{\lambda} < \overline{\lambda} < \infty$ satisfy $\underline{\lambda}\mathbf{I}_d \preceq \nabla T_\nu^\mu(\boldsymbol{x}) \preceq \overline{\lambda}\mathbf{I}_d$ for all $\boldsymbol{x} \in \mathrm{supp}(\mu)$ (recall Lemma 3.2). For $m \in \mathbb{N}$ independent random samples*

$\boldsymbol{X}_1, \dots, \boldsymbol{X}_m : \Omega \to \mathbb{R}^d$ *from* μ *and* $n \in \mathbb{N}$ *independent random samples* $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n : \Omega \to \mathbb{R}^d$ *from* ν, *let* $(\widehat{\pi}^{\star}_{i,j})_{i=1:m,\, j=1:n}$ *be an optimal solution of the following linear programming (LP) problem:*

$$\begin{aligned} &\underset{(\widehat{\pi}_{i,j})}{\text{minimize}} && \sum_{i=1}^{m}\sum_{j=1}^{n} \widehat{\pi}_{i,j} \|\boldsymbol{X}_i - \boldsymbol{Y}_j\|^2 \\ &\textit{subject to} && \sum_{j=1}^{n} \widehat{\pi}_{i,j} = \frac{1}{m} && \forall 1 \le i \le m, \\ &&& \sum_{i=1}^{m} \widehat{\pi}_{i,j} = \frac{1}{n} && \forall 1 \le j \le n, \\ &&& \widehat{\pi}_{i,j} \ge 0 \qquad \forall 1 \le i \le m,\ \forall 1 \le j \le n. \end{aligned} \tag{4.1}$$

Subsequently, let $(\widetilde{\varphi}^{\star}_i)_{i=1:m}$, $(\widetilde{\boldsymbol{g}}^{\star}_i)_{i=1:m}$ *be an optimal solution of the following quadratically constrained quadratic programming (QCQP) problem:*

$$\begin{aligned} &\underset{(\widetilde{\varphi}_i),\,(\widetilde{\boldsymbol{g}}_i)}{\text{minimize}} && \sum_{i=1}^{m}\sum_{j=1}^{n} \widehat{\pi}^{\star}_{i,j} \|\widetilde{\boldsymbol{g}}_i + \underline{\lambda}\boldsymbol{X}_i - \boldsymbol{Y}_j\|^2 \\ &\textit{subject to} && \widetilde{\varphi}_j \ge \widetilde{\varphi}_i + \langle \widetilde{\boldsymbol{g}}_i, \boldsymbol{X}_j - \boldsymbol{X}_i \rangle + \tfrac{1}{2(\bar{\lambda}-\underline{\lambda})} \|\widetilde{\boldsymbol{g}}_i - \widetilde{\boldsymbol{g}}_j\|^2 \quad \forall 1 \le i \le m,\ \forall 1 \le j \le m, \\ &&& \|\widetilde{\boldsymbol{g}}_i + \underline{\lambda}\boldsymbol{X}_i\|^2 \le \overline{u}_0(\nu)^2 \qquad \forall 1 \le i \le m, \end{aligned} \tag{4.2}$$

where $\overline{u}_0(\nu) := \inf\big\{r \in \mathbb{R}_+ : \operatorname{supp}(\nu) \subseteq \bar{B}(\boldsymbol{0}, r)\big\}$. *Let* $\varphi^{\star}_i := \widetilde{\varphi}^{\star}_i + \frac{\underline{\lambda}}{2}\|\boldsymbol{X}_i\|^2$, $\boldsymbol{g}^{\star}_i := \widetilde{\boldsymbol{g}}^{\star}_i + \underline{\lambda}\boldsymbol{X}_i$ *for* $i = 1, \dots, m$. *We call* $\widehat{T}_{\mathrm{CLS}} : \mathbb{R}^d \to \mathbb{R}^d$ *a shaped-constrained convex least squares OT map estimator of* T^{μ}_{ν} *if there exists* $\widehat{\varphi}_{\mathrm{CLS}} \in \mathfrak{C}_{\underline{\lambda},\bar{\lambda}}(\mathbb{R}^d)$ *such that* $\widehat{T}_{\mathrm{CLS}} = \nabla\widehat{\varphi}_{\mathrm{CLS}}$ *and* $\widehat{\varphi}_{\mathrm{CLS}}(\boldsymbol{X}_i) = \varphi^{\star}_i$, $\widehat{T}_{\mathrm{CLS}}(\boldsymbol{X}_i) = \boldsymbol{g}^{\star}_i$ *for* $i = 1, \dots, m$.

Note that the linear programming problem (4.1) computes an optimal coupling of the empirical measures $\dot{\mu}_m := \frac{1}{m}\sum_{i=1}^{m} \delta_{\boldsymbol{X}_i} \in \mathcal{P}_2(\mathbb{R}^d)$ and $\dot{\nu}_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{\boldsymbol{Y}_j} \in \mathcal{P}_2(\mathbb{R}^d)$, i.e., $\widehat{\pi}^{\star} := \sum_{i=1}^{m}\sum_{j=1}^{n} \widehat{\pi}^{\star}_{i,j}\delta_{(\boldsymbol{X}_i, \boldsymbol{Y}_j)} \in \Pi(\dot{\mu}_m, \dot{\nu}_n)$ satisfies $\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \,\widehat{\pi}^{\star}(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) = \mathcal{W}_2(\dot{\mu}_m, \dot{\nu}_n)^2$. The QCQP problem (4.2) utilizes the tractable formulation of smoothness and strong convexity constraints developed by Taylor [91]. Below we present a version of their results adapted to our settings.

Theorem 4.2 (Formulation of shape constraints; adapted from [91, Theorem 3.8])**.** *Under the settings of Definition 4.1,* $\widehat{T}_{\mathrm{CLS}} : \mathbb{R}^d \to \mathbb{R}^d$ *is a shaped-constrained convex least squares OT map estimator of* T^{μ}_{ν} *if and only if* $\widehat{T}_{\mathrm{CLS}} = \nabla\widehat{\varphi}_{\mathrm{CLS}}$ *and* $\widehat{\varphi}_{\mathrm{CLS}}$ *is an optimizer of the following minimization problem over the set of all convex functions on* $\mathbb{R}^d$ *subject to shape constraints:*

$$\begin{aligned} &\underset{\varphi}{\text{minimize}} && \sum_{i=1}^{m}\sum_{j=1}^{n} \widehat{\pi}^{\star}_{i,j} \big\|\nabla\varphi(\boldsymbol{X}_i) - \boldsymbol{Y}_j\big\|^2 \\ &\textit{subject to} && \varphi \in \mathfrak{C}_{\underline{\lambda},\bar{\lambda}}(\mathbb{R}^d),\ \big\|\nabla\varphi(\boldsymbol{X}_i)\big\| \le \overline{u}_0(\nu) \quad \forall 1 \le i \le m. \end{aligned} \tag{4.3}$$

Let us first state the following result from [91] as a lemma, which will be used in the proofs in this section.

Lemma 4.3 ([91, Theorem 3.8])**.** *For any* $0 \le \underline{l} < \bar{l} \le \infty$, *we call* $(\boldsymbol{x}_i, \boldsymbol{g}_i, \varphi_i)_{i=1:m}$ $\mathfrak{C}_{\underline{l},\bar{l}}(\mathbb{R}^d)$*-interpolable, if there exists* $\varphi \in \mathfrak{C}_{\underline{l},\bar{l}}(\mathbb{R}^d)$ *such that* $\varphi(\boldsymbol{x}_i) = \varphi_i$ *and* $\boldsymbol{g}_i \in \partial\varphi(\boldsymbol{x}_i)$ *for* $i = 1, \dots, m$. *In this case, we say* $(\boldsymbol{x}_i, \boldsymbol{g}_i, \varphi_i)_{i=1:m}$ $\mathfrak{C}_{\underline{l},\bar{l}}(\mathbb{R}^d)$ *is interpolated by* φ. *Then, the following statements are equivalent:*

(a) $\big(\frac{\bar{l}}{\bar{l}-\underline{l}}\boldsymbol{x}_i - \frac{1}{\bar{l}-\underline{l}}\boldsymbol{g}_i, \boldsymbol{g}_i - \underline{l}\boldsymbol{x}_i, \varphi_i + \frac{\underline{l}}{\bar{l}-\underline{l}}\langle \boldsymbol{g}_i, \boldsymbol{x}_i\rangle - \frac{1}{2(\bar{l}-\underline{l})}\|\boldsymbol{g}_i\|^2 - \frac{\underline{l}\bar{l}}{2(\bar{l}-\underline{l})}\|\boldsymbol{x}_i\|^2\big)_{i=1:m}$ *is interpolated by* $\varphi \in \mathfrak{C}_{0,\infty}(\mathbb{R}^d)$.

(b) $\big(\boldsymbol{g}_i - \underline{l}\boldsymbol{x}_i, \frac{\bar{l}}{\bar{l}-\underline{l}}\boldsymbol{x}_i - \frac{1}{\bar{l}-\underline{l}}\boldsymbol{g}_i, \frac{\bar{l}}{\bar{l}-\underline{l}}\langle \boldsymbol{g}_i, \boldsymbol{x}_i\rangle - \varphi_i - \frac{1}{2(\bar{l}-\underline{l})}\|\boldsymbol{g}_i\|^2 - \frac{\underline{l}\bar{l}}{2(\bar{l}-\underline{l})}\|\boldsymbol{x}_i\|^2\big)_{i=1:m}$ *is interpolated by* $\varphi^* \in \mathfrak{C}_{0,\infty}(\mathbb{R}^d)$;

(c) $\big(\boldsymbol{g}_i - \underline{l}\boldsymbol{x}_i, \boldsymbol{x}_i, \langle \boldsymbol{g}_i, \boldsymbol{x}_i\rangle - \varphi_i - \frac{\underline{l}}{2}\|\boldsymbol{x}_i\|^2\big)_{i=1:m}$ *is interpolated by* $\varphi^* + \frac{1}{2(\bar{l}-\underline{l})}\|\cdot\|^2 \in \mathfrak{C}_{\frac{1}{\bar{l}-\underline{l}},\infty}(\mathbb{R}^d)$;

(d) $\big(\boldsymbol{x}_i, \boldsymbol{g}_i - \underline{l}\boldsymbol{x}_i, \varphi_i - \frac{\underline{l}}{2}\|\boldsymbol{x}_i\|^2\big)_{i=1:m}$ *is interpolated by* $\big(\varphi^* + \frac{1}{2(\bar{l}-\underline{l})}\|\cdot\|^2\big)^* \in \mathfrak{C}_{0,\bar{l}-\underline{l}}(\mathbb{R}^d)$;

(e) $(\boldsymbol{x}_i, \boldsymbol{g}_i, \varphi_i)_{i=1:m}$ *is interpolated by* $\big(\varphi^* + \frac{1}{2(\overline{l}-\underline{l})}\|\cdot\|^2\big)^* + \frac{l}{2}\|\cdot\|^2 \in \mathfrak{C}_{\underline{l},\overline{l}}(\mathbb{R}^d)$*;*

Proof of Theorem 4.2. It holds by the equivalence of (e) and (d) in Lemma 4.3 and the statement of [91, Theorem 3.8] that:

$$
\begin{aligned}
& \exists \varphi \in \mathfrak{C}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d),\ \varphi(\boldsymbol{X}_i) = \varphi_i,\ \nabla\varphi(\boldsymbol{X}_i) = \boldsymbol{g}_i\ \forall 1 \le i \le m \\
\Leftrightarrow \quad & \exists \widetilde{\varphi} \in \mathfrak{C}_{0,\overline{\lambda}-\underline{\lambda}}(\mathbb{R}^d),\ \widetilde{\varphi}(\boldsymbol{X}_i) = \widetilde{\varphi}_i := \varphi_i - \tfrac{\underline{\lambda}}{2}\|\boldsymbol{X}_i\|^2,\ \nabla\widetilde{\varphi}(\boldsymbol{X}_i) = \widetilde{\boldsymbol{g}}_i := \boldsymbol{g}_i - \underline{\lambda}\boldsymbol{x}_i\ \forall 1 \le i \le m \\
\Leftrightarrow \quad & \widetilde{\varphi}_j \ge \widetilde{\varphi}_i + \langle \widetilde{\boldsymbol{g}}_i, \boldsymbol{X}_j - \boldsymbol{X}_i \rangle + \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}_i - \widetilde{\boldsymbol{g}}_j\|^2\ \forall 1 \le i \le m,\ \forall 1 \le j \le m.
\end{aligned}
$$

Thus, the two optimization problems (4.2) and (4.3) are equivalent under the reparametrization of the decision variables: $\varphi(\boldsymbol{X}_i) \leftrightarrow \widetilde{\varphi}_i + \frac{\underline{\lambda}}{2}\|\boldsymbol{X}_i\|^2$, $\nabla\varphi(\boldsymbol{X}_i) \leftrightarrow \widetilde{\boldsymbol{g}}_i + \underline{\lambda}\boldsymbol{X}_i$ for $i = 1, \ldots, m$. This completes the proof. □

Remark 4.4. *Since solving the QCQP in* (4.2) *suffers from prohibitive computational complexity when the sample size and the dimension grow, we provide in Appendix B a paralleled implementation of it using the first-order alternating direction method of multipliers (ADMM). The algorithm is adapted from Simonetto [80] by exploiting the underlying decomposable structure in* (4.2)*, and similar techniques for efficient shape-constrained convex regression have been considered in literature; see, e.g., Aybat and Wang [9] and Mazumder, Choudhury, Iyengar, and Sen [60].*

In the following, we adapt the results of Manole et al. [59] to derive the estimation error bound of any shape-constrained convex least squares OT map estimator in Definition 4.1.

Theorem 4.5 (Estimation error bound of $\widehat{T}_{\mathrm{CLS}}$; adapted from [59, Proposition 15])**.** *Under the settings of Definition 4.1, let* $\widehat{T}_{\mathrm{CLS}} : \mathbb{R}^d \to \mathbb{R}^d$ *be a shape-constrained convex least squares OT map estimator of* T^μ_ν *based on* $m \in \mathbb{N}$ *independent random samples* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m : \Omega \to \mathbb{R}^d$ *from* μ *and* $n \in \mathbb{N}$ *independent random samples* $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n : \Omega \to \mathbb{R}^d$ *from* ν*. Then, there exists a constant* $C(\mu, \nu, \underline{\lambda}, \overline{\lambda}) > 0$ *that depends on* μ*,* ν *and the choices of* $\underline{\lambda} \equiv \underline{\lambda}(\mu,\nu)$*,* $\overline{\lambda} \equiv \overline{\lambda}(\mu,\nu)$*, such that*

$$
\mathbb{E}\Big[\big\|\widehat{T}_{\mathrm{CLS}} - T^\mu_\nu\big\|^2_{\mathcal{L}^2(\mu)}\Big] \le C(\mu, \nu, \underline{\lambda}, \overline{\lambda}) \log(m)^2 \kappa\big(\min\{m, n\}\big),
$$

where

$$
\kappa(q) := \begin{cases} q^{-\frac{1}{2}} & d \le 3, \\ q^{-\frac{1}{2}}\log(q) & d = 4, \\ q^{-\frac{2}{d}} & d \ge 5 \end{cases} \qquad \forall q \in \mathbb{N}.
$$

Proof of Theorem 4.5. Before applying [59, Proposition 15], let us show that the support of μ and ν satisfy the premises of [59, Proposition 15]. Let $\mathscr{L}$ denote the Lebesgue measure on $\mathbb{R}^d$. Specifically, we will prove the following claim: for every $\mathcal{X} \in \mathcal{S}$, there exist $\epsilon_0 > 0$ and $\delta_0 > 0$ such that $\mathscr{L}\big(\mathcal{X} \cap B(\boldsymbol{x}, \epsilon)\big) \ge \delta_0 \mathscr{L}\big(B(\boldsymbol{x}, \epsilon)\big)$ for all $\boldsymbol{x} \in \mathcal{X}$ and for all $\epsilon \in (0, \epsilon_0)$. To that end, let us fix an arbitrary $\mathcal{X} \in \mathcal{S}$ and let $r > 0$ satisfy $B(\boldsymbol{0}, r) \supset \mathcal{X}$. Subsequently, let $\epsilon_0 := 2r$ and let $\delta_0 := \frac{\mathscr{L}(\mathcal{X})}{\mathscr{L}(B(\boldsymbol{0},\epsilon_0))} > 0$. Then, for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, it holds that $\|\boldsymbol{x} - \boldsymbol{x}'\| \le \|\boldsymbol{x}\| + \|\boldsymbol{x}'\| \le 2r = \epsilon_0$, which implies that $\mathcal{X} \subset B(\boldsymbol{x}, \epsilon_0)$. Therefore, $\mathscr{L}\big(\mathcal{X} \cap B(\boldsymbol{x}, \epsilon_0)\big) = \mathscr{L}(\mathcal{X}) = \delta_0 \mathscr{L}\big(B(\boldsymbol{0}, \epsilon_0)\big) = \delta_0 \mathscr{L}\big(B(\boldsymbol{x}, \epsilon_0)\big)$. Next, for any $\epsilon \in (0, \epsilon_0)$ and for any $\boldsymbol{y} \in \mathcal{X} \cap B(\boldsymbol{x}, \epsilon_0)$, note that $\boldsymbol{x} + \frac{\epsilon}{\epsilon_0}(\boldsymbol{y} - \boldsymbol{x}) \in \mathcal{X} \cap B(\boldsymbol{x}, \epsilon)$ due to the convexity of $\mathcal{X}$. This implies that

$$
\begin{aligned}
\mathscr{L}\big(\mathcal{X} \cap B(\boldsymbol{x}, \epsilon)\big) &\ge \mathscr{L}\big(\big\{\boldsymbol{x} + \tfrac{\epsilon}{\epsilon_0}(\boldsymbol{y} - \boldsymbol{x}) : \boldsymbol{y} \in \mathcal{X} \cap B(\boldsymbol{x}, \epsilon_0)\big\}\big) \\
&= \mathscr{L}\big(\tfrac{\epsilon_0 - \epsilon}{\epsilon_0}\boldsymbol{x} + \tfrac{\epsilon}{\epsilon_0}\big(\mathcal{X} \cap B(\boldsymbol{x}, \epsilon_0)\big)\big) \\
&= \big(\tfrac{\epsilon}{\epsilon_0}\big)^d \mathscr{L}\big(\mathcal{X} \cap B(\boldsymbol{x}, \epsilon_0)\big) \\
&= \delta_0 \big(\tfrac{\epsilon}{\epsilon_0}\big)^d \mathscr{L}\big(B(\boldsymbol{x}, \epsilon_0)\big) \\
&= \delta_0 \mathscr{L}\big(B(\boldsymbol{x}, \epsilon)\big) \qquad\qquad \forall \epsilon \in (0, \epsilon_0).
\end{aligned}
$$

This proves the claim, and hence the assumption (S2) of [59, Proposition 15] holds.

Moreover, let $\dot{\mu}_m := \frac{1}{m}\sum_{i=1}^m \delta_{\boldsymbol{X}_i} \in \mathcal{P}_2(\mathbb{R}^d)$. We have by the constraints in the linear programming problem (4.1) that

$$\begin{aligned}\big\|\widehat{T}_{\mathrm{CLS}} - T^{\mu}_{\nu}\big\|^2_{\mathcal{L}^2(\dot{\mu}_m)} &= \frac{1}{m}\sum_{i=1}^{m}\big\|\widehat{T}_{\mathrm{CLS}}(\boldsymbol{X}_i) - T^{\mu}_{\nu}(\boldsymbol{X}_i)\big\|^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{\pi}^{\star}_{i,j}\big\|\widehat{T}_{\mathrm{CLS}}(\boldsymbol{X}_i) - T^{\mu}_{\nu}(\boldsymbol{X}_i)\big\|^2 \\ &\le 2\sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{\pi}^{\star}_{i,j}\Big(\big\|\widehat{T}_{\mathrm{CLS}}(\boldsymbol{X}_i) - \boldsymbol{Y}_j\big\|^2 + \big\|T^{\mu}_{\nu}(\boldsymbol{X}_i) - \boldsymbol{Y}_j\big\|^2\Big).\end{aligned} \tag{4.4}$$

Notice that $T^{\mu}_{\nu} = \nabla\varphi^{\mu}_{\nu}$, and by Lemma 3.2 we can assume without loss of generality that $\varphi^{\mu}_{\nu} \in \mathfrak{C}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$. Furthermore, since $(\boldsymbol{X}_i)_{i=1:m} \subset \mathrm{supp}(\mu)$ $\mathbb{P}$-almost surely and $\nabla\varphi^{\mu}_{\nu}(\mathrm{supp}(\mu)) = \mathrm{supp}(\nu) \subseteq \bar{B}\big(\mathbf{0}, \overline{u}_0(\nu)\big)$, the inequality $\big\|\nabla\varphi^{\mu}_{\nu}(\boldsymbol{X}_i)\big\| \le \overline{u}_0(\nu)$ holds $\mathbb{P}$-almost surely for $i = 1, \ldots, m$. Consequently, φ^{μ}_{ν} is $\mathbb{P}$-almost surely feasible for the optimization problem (4.3). Since $\widehat{T}_{\mathrm{CLS}} = \nabla\widehat{\varphi}_{\mathrm{CLS}}$ where $\widehat{\varphi}_{\mathrm{CLS}}$ is optimal for (4.3) by Theorem 4.2, we get

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{\pi}^{\star}_{i,j}\big\|\widehat{T}_{\mathrm{CLS}}(\boldsymbol{X}_i) - \boldsymbol{Y}_j\big\|^2 \le \sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{\pi}^{\star}_{i,j}\big\|T^{\mu}_{\nu}(\boldsymbol{X}_i) - \boldsymbol{Y}_j\big\|^2 \qquad \mathbb{P}\text{-a.s.}$$

Substituting this into (4.4) yields

$$\big\|\widehat{T}_{\mathrm{CLS}} - T^{\mu}_{\nu}\big\|^2_{\mathcal{L}^2(\dot{\mu}_m)} \le 4\sum_{i=1}^{m}\sum_{j=1}^{n}\widehat{\pi}^{\star}_{i,j}\big\|T^{\mu}_{\nu}(\boldsymbol{X}_i) - \boldsymbol{Y}_j\big\|^2 \qquad \mathbb{P}\text{-a.s.}$$

The rest of the proof is then identical to the proof of Proposition 15 in [59], up to rescaling $\mathrm{supp}(\mu)$ and $\mathrm{supp}(\nu)$ to be contained in $[0,1]^d$ as well as extending to the case where $\mathrm{supp}(\mu) \neq \mathrm{supp}(\nu)$. □

Remark 4.6 (Choice of $\underline{\lambda}(\mu,\nu)$ and $\overline{\lambda}(\mu,\nu)$)**.** *In Theorem 4.5, the dependence of the constant term $C(\mu,\nu,\underline{\lambda},\overline{\lambda})$ on the choices of $\underline{\lambda} \equiv \underline{\lambda}(\mu,\nu)$, $\overline{\lambda} \equiv \overline{\lambda}(\mu,\nu)$ is explicitly stated. Even though any choices of $0 < \underline{\lambda}(\mu,\nu) < \overline{\lambda}(\mu,\nu) < \infty$ that satisfy $\underline{\lambda}(\mu,\nu)\mathbf{I}_d \preceq \nabla T^{\mu}_{\nu}(\boldsymbol{x}) \preceq \overline{\lambda}(\mu,\nu)\mathbf{I}_d$ for all $\boldsymbol{x} \in \mathrm{supp}(\mu)$ would be valid, both a decrease in $\underline{\lambda}(\mu,\nu)$ and an increase in $\overline{\lambda}(\mu,\nu)$ will lead to an increase of $C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)$. Despite this, we assume that $\underline{\lambda} \equiv \underline{\lambda}(\mu,\nu)$, $\overline{\lambda} \equiv \overline{\lambda}(\mu,\nu)$ can be unambiguously chosen given any $\mu,\nu \in \mathcal{M}$.*

Given $(\varphi^{\star}_i)_{i=1:m}$, $(\boldsymbol{g}^{\star}_i)_{i=1:m}$ that are constructed via Definition 4.1, the functions $\widehat{\varphi}_{\mathrm{CLS}} \in \mathfrak{C}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$ and $\widehat{T}_{\mathrm{CLS}} = \nabla\widehat{\varphi}_{\mathrm{CLS}}$ are only specified at the sample points $(\boldsymbol{X}_i)_{i=1:m}$. In fact, interpolation (and extrapolation) of $\widehat{\varphi}_{\mathrm{CLS}}$ and $\widehat{T}_{\mathrm{CLS}}$ to the entire $\mathbb{R}^d$ is non-unique. Thus, let us introduce the following quadratic programming (QP) formulation for computing the smallest of such interpolation functions, which is a simplified version of the formulation developed by Taylor [91].

Theorem 4.7 (Smallest shape-constrained interpolation function [91, Theorem 3.14])**.** *Under the settings of Definition 4.1, let us define the following terms:*

$$\begin{aligned}\Delta &:= \Big\{\boldsymbol{w} = (w_1,\ldots,w_m)^{\mathsf{T}} : \textstyle\sum_{i=1}^m w_i = 1,\ w_i \ge 0\ \forall 1 \le i \le m\Big\} \subset \mathbb{R}^m, \\ \widetilde{\mathbf{G}}^{\star} &:= \begin{pmatrix} | & | & & | \\ \widetilde{\boldsymbol{g}}^{\star}_1 & \widetilde{\boldsymbol{g}}^{\star}_2 & \cdots & \widetilde{\boldsymbol{g}}^{\star}_m \\ | & | & & | \end{pmatrix} \in \mathbb{R}^{d\times m}, \\ v_i &:= \varphi^{\star}_i + \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\boldsymbol{g}^{\star}_i\|^2 + \tfrac{\underline{\lambda}\overline{\lambda}}{2(\overline{\lambda}-\underline{\lambda})}\|\boldsymbol{X}_i\|^2 - \tfrac{\overline{\lambda}}{\overline{\lambda}-\underline{\lambda}}\langle\boldsymbol{g}^{\star}_i, \boldsymbol{X}_i\rangle \in \mathbb{R} \qquad \forall 1 \le i \le m, \\ \boldsymbol{v} &:= (v_1,\ldots,v_m)^{\mathsf{T}} \in \mathbb{R}^m.\end{aligned} \tag{4.5}$$

Let $\widehat{\varphi}_{\text{CLS-LB}}:\mathbb{R}^d\to\mathbb{R}$ *and* $\widehat{T}_{\text{CLS-LB}}:\mathbb{R}^d\to\mathbb{R}^d$ *be defined as follows:*

$$\widehat{\varphi}_{\text{CLS-LB}}(\boldsymbol{x}) := \tfrac{\underline{\lambda}}{2}\|\boldsymbol{x}\|^2 + \sup_{\boldsymbol{w}\in\Delta}\Big\{\langle\widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x}+\boldsymbol{v},\boldsymbol{w}\rangle - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\mathbf{G}}^{\star}\boldsymbol{w}\|^2\Big\} \qquad \forall\boldsymbol{x}\in\mathbb{R}^d, \tag{4.6}$$

$$\begin{aligned}\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x}) &:= \underline{\lambda}\boldsymbol{x} + \widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x}),\\ &\textit{where } \widehat{\boldsymbol{w}}(\boldsymbol{x}) \in \operatorname*{arg\,max}_{\boldsymbol{w}\in\Delta}\Big\{\langle\widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x}+\boldsymbol{v},\boldsymbol{w}\rangle - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\mathbf{G}}^{\star}\boldsymbol{w}\|^2\Big\} \qquad \forall\boldsymbol{x}\in\mathbb{R}^d.\end{aligned} \tag{4.7}$$

Then, the following statements hold.

(i) $\widehat{\varphi}_{\text{CLS-LB}}\in\mathfrak{C}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$ *and* $\widehat{\varphi}_{\text{CLS-LB}}$ *is the smallest interpolation function of* $(\varphi_i^\star,\boldsymbol{g}_i^\star)_{i=1:m}$ *in* $\mathfrak{C}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$*; i.e., for any* $\varphi\in\mathfrak{C}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$ *satisfying* $\varphi(\boldsymbol{X}_i)=\varphi_i^\star$ *and* $\nabla\varphi(\boldsymbol{X}_i)=\boldsymbol{g}_i^\star$ *for* $i=1,\ldots,m$*, it holds that* $\widehat{\varphi}_{\text{CLS-LB}}(\boldsymbol{x})\le\varphi(\boldsymbol{x})\ \forall\boldsymbol{x}\in\mathbb{R}^d$.

(ii) $\widehat{T}_{\text{CLS-LB}}$ *is uniquely defined by (4.7) for every* $\boldsymbol{x}\in\mathbb{R}^d$ *and* $\widehat{T}_{\text{CLS-LB}}=\nabla\widehat{\varphi}_{\text{CLS-LB}}$.

Since Taylor [91] did not provide a detailed proof of this result, we will present a detailed derivation for the sake of completeness.

Proof of Theorem 4.7. To begin, for arbitrary $(\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{g}}_i,\bar{\varphi}_i)_{i=1:m}$ that is $\mathfrak{C}_{0,\infty}(\mathbb{R}^d)$-interpolable, the smallest function $\varphi_{\text{LB}}\in\mathfrak{C}_{0,\infty}(\mathbb{R}^d)$ that interpolates $(\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{g}}_i,\bar{\varphi}_i)_{i=1:m}$ is given by

$$\varphi_{\text{LB}}(\boldsymbol{x}) = \max_{1\le i\le m}\{\bar{\varphi}_i+\langle\bar{\boldsymbol{g}}_i,\boldsymbol{x}-\bar{\boldsymbol{x}}_i\rangle\} \qquad \forall\boldsymbol{x}\in\mathbb{R}^d. \tag{4.8}$$

It holds that $\varphi_{\text{LB}}(\boldsymbol{x})\le\varphi(\boldsymbol{x})$ for all $\boldsymbol{x}\in\mathbb{R}^d$ for any other $\varphi\in\mathfrak{C}_{0,\infty}(\mathbb{R}^d)$ that interpolates $(\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{g}}_i,\bar{\varphi}_i)_{i=1:m}$; see, e.g., [91, Remark 3.5].

To prove statement (i), let us first use (4.8) to derive the smallest interpolation function in Lemma 4.3(a) with $\boldsymbol{x}_i\leftarrow\boldsymbol{X}_i$, $\boldsymbol{g}_i\leftarrow\boldsymbol{g}_i^\star$, $\varphi_i\leftarrow\varphi_i^\star$ for $i=1,\ldots,m$, $\underline{l}\leftarrow\underline{\lambda}$, $\bar{l}\leftarrow\overline{\lambda}$. Let us define $\varphi_{\text{LB}}^{(a)}:\mathbb{R}^d\to\mathbb{R}$ as follows:

$$\begin{aligned}\varphi_{\text{LB}}^{(a)}(\boldsymbol{x}) &:= \sup_{1\le i\le m}\Big\{\varphi_i^\star+\tfrac{\underline{\lambda}}{\overline{\lambda}-\underline{\lambda}}\langle\boldsymbol{g}_i^\star,\boldsymbol{X}_i\rangle - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\boldsymbol{g}_i^\star\|^2 - \tfrac{\underline{\lambda}\overline{\lambda}}{2(\overline{\lambda}-\underline{\lambda})}\|\boldsymbol{X}_i\|^2\\ &\qquad\qquad + \big\langle\boldsymbol{g}_i^\star-\underline{\lambda}\boldsymbol{X}_i,\boldsymbol{x}-\tfrac{\overline{\lambda}}{\overline{\lambda}-\underline{\lambda}}\boldsymbol{X}_i+\tfrac{1}{\overline{\lambda}-\underline{\lambda}}\boldsymbol{g}_i^\star\big\rangle\Big\}\\ &= \sup_{1\le i\le m}\Big\{\langle\boldsymbol{g}_i^\star-\underline{\lambda}\boldsymbol{X}_i,\boldsymbol{x}\rangle+\varphi_i^\star+\tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\boldsymbol{g}_i^\star\|^2+\tfrac{\underline{\lambda}\overline{\lambda}}{2(\overline{\lambda}-\underline{\lambda})}\|\boldsymbol{X}_i\|^2-\tfrac{\overline{\lambda}}{\overline{\lambda}-\underline{\lambda}}\langle\boldsymbol{g}_i^\star,\boldsymbol{X}_i\rangle\Big\}\\ &= \sup_{1\le i\le m}\big\{\langle\widetilde{\boldsymbol{g}}_i^\star,\boldsymbol{x}\rangle+v_i\big\} \qquad \forall\boldsymbol{x}\in\mathbb{R}^d.\end{aligned}$$

Next, we transform $\varphi_{\text{LB}}^{(a)}$ using the equivalence between Lemma 4.3(a) and Lemma 4.3(b) by taking the convex conjugate of $\varphi_{\text{LB}}^{(a)}$. We define $\varphi_{\text{UB}}^{(b)}:\mathbb{R}^d\to\mathbb{R}\cup\{\infty\}$ as follows:

$$\begin{aligned}\varphi_{\text{UB}}^{(b)}(\widetilde{\boldsymbol{g}}) &:= \sup_{\boldsymbol{x}\in\mathbb{R}^d}\big\{\langle\widetilde{\boldsymbol{g}},\boldsymbol{x}\rangle-\varphi_{\text{LB}}^{(a)}(\boldsymbol{x})\big\}\\ &= \sup_{\boldsymbol{x}\in\mathbb{R}^d}\Big\{\inf_{1\le i\le m}\Big\{\langle\widetilde{\boldsymbol{g}}-\widetilde{\boldsymbol{g}}_i^\star,\boldsymbol{x}\rangle-v_i\Big\}\Big\}\\ &= \sup_{\boldsymbol{x}\in\mathbb{R}^d}\left\{\inf_{(w_i)_{i=1:n}\in\Delta}\left\{\sum_{i=1}^{n}w_i\big(\langle\widetilde{\boldsymbol{g}}-\widetilde{\boldsymbol{g}}_i^\star,\boldsymbol{x}\rangle-v_i\big)\right\}\right\}\\ &= \sup_{\boldsymbol{x}\in\mathbb{R}^d}\Big\{\inf_{\boldsymbol{w}\in\Delta}\big\{\langle\widetilde{\boldsymbol{g}}-\widetilde{\mathbf{G}}^\star\boldsymbol{w},\boldsymbol{x}\rangle-\langle\boldsymbol{v},\boldsymbol{w}\rangle\big\}\Big\} \qquad \forall\widetilde{\boldsymbol{g}}\in\mathbb{R}^d.\end{aligned}$$

Since $\mathbb{R}^d \times \Delta \ni (\boldsymbol{x}, \boldsymbol{w}) \mapsto \langle \widetilde{\boldsymbol{g}} - \widetilde{\mathbf{G}}^\star \boldsymbol{w}, \boldsymbol{x} \rangle - \langle \boldsymbol{v}, \boldsymbol{w} \rangle \in \mathbb{R}$ is bilinear in $\boldsymbol{x}$ and $\boldsymbol{w}$, it follows from the compactness of Δ and Sion's minimax theorem [82] that

$$\begin{aligned}\varphi^{(b)}_{\mathrm{UB}}(\widetilde{\boldsymbol{g}}) &= \sup_{\boldsymbol{x}\in\mathbb{R}^d} \Big\{ \inf_{\boldsymbol{w}\in\Delta} \big\{ \langle \widetilde{\boldsymbol{g}} - \widetilde{\mathbf{G}}^\star \boldsymbol{w}, \boldsymbol{x} \rangle - \langle \boldsymbol{v}, \boldsymbol{w} \rangle \big\} \Big\} \\ &= \inf_{\boldsymbol{w}\in\Delta} \Big\{ \sup_{\boldsymbol{x}\in\mathbb{R}^d} \Big\{ \langle \widetilde{\boldsymbol{g}} - \widetilde{\mathbf{G}}^\star \boldsymbol{w}, \boldsymbol{x} \rangle \Big\} - \langle \boldsymbol{v}, \boldsymbol{w} \rangle \Big\} \\ &= - \sup_{\boldsymbol{w}\in\Delta,\, \widetilde{\mathbf{G}}^\star\boldsymbol{w} = \widetilde{\boldsymbol{g}}} \Big\{ \langle \boldsymbol{v}, \boldsymbol{w} \rangle \Big\} \qquad \forall \widetilde{\boldsymbol{g}} \in \mathbb{R}^d.\end{aligned}$$

The order reversing property of convex conjugation implies that $\varphi^{(b)}_{\mathrm{UB}}$ is the largest function in $\mathfrak{C}_{0,\infty}(\mathbb{R}^d)$ that interpolates $\big(\boldsymbol{g}^\star_i - \underline{\lambda}\boldsymbol{X}_i, \frac{\overline{\lambda}}{\overline{\lambda}-\underline{\lambda}}\boldsymbol{X}_i - \frac{1}{\overline{\lambda}-\underline{\lambda}}\boldsymbol{g}^\star_i, \frac{\overline{\lambda}}{\overline{\lambda}-\underline{\lambda}}\langle \boldsymbol{g}^\star_i, \boldsymbol{X}_i\rangle - \varphi^\star_i - \frac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\boldsymbol{g}^\star_i\|^2 - \frac{\underline{\lambda}\overline{\lambda}}{2(\overline{\lambda}-\underline{\lambda})}\|\boldsymbol{X}_i\|^2\big)_{i=1:m}$. Using the equivalence between Lemma 4.3(b) and Lemma 4.3(c), we add $\widetilde{\boldsymbol{g}} \mapsto \frac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}\|^2$ to $\varphi^{(b)}_{\mathrm{UB}}$ and define $\varphi^{(c)}_{\mathrm{UB}} : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ as follows:

$$\varphi^{(c)}_{\mathrm{UB}}(\widetilde{\boldsymbol{g}}) := \varphi^{(b)}_{\mathrm{UB}}(\widetilde{\boldsymbol{g}}) + \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}\|^2 = \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}\|^2 - \sup_{\boldsymbol{w}\in\Delta,\, \widetilde{\mathbf{G}}^\star\boldsymbol{w} = \widetilde{\boldsymbol{g}}} \Big\{ \langle \boldsymbol{v}, \boldsymbol{w} \rangle \Big\} \qquad \forall \widetilde{\boldsymbol{g}} \in \mathbb{R}^d.$$

Thus, $\varphi^{(c)}_{\mathrm{UB}}$ is the largest function in $\mathfrak{C}_{\frac{1}{\overline{\lambda}-\underline{\lambda}},\infty}(\mathbb{R}^d)$ that interpolates $\big(\boldsymbol{g}^\star_i - \underline{\lambda}\boldsymbol{X}_i, \boldsymbol{X}_i, \langle \boldsymbol{g}^\star_i, \boldsymbol{X}_i\rangle - \varphi^\star_i - \frac{\underline{\lambda}}{2}\|\boldsymbol{X}_i\|^2\big)_{i=1:m}$. Subsequently, we use the equivalence between Lemma 4.3(c) and Lemma 4.3(d) and define $\varphi^{(d)}_{\mathrm{LB}} : \mathbb{R}^d \to \mathbb{R}$ by taking the convex conjugate of $\varphi^{(c)}_{\mathrm{UB}}$:

$$\begin{aligned}\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x}) &:= \sup_{\widetilde{\boldsymbol{g}}\in\mathbb{R}^d} \big\{ \langle \widetilde{\boldsymbol{g}}, \boldsymbol{x} \rangle - \varphi^{(c)}_{\mathrm{UB}}(\widetilde{\boldsymbol{g}}) \big\} \\ &= \sup_{\widetilde{\boldsymbol{g}}\in\mathbb{R}^d,\, \boldsymbol{w}\in\Delta,\, \widetilde{\mathbf{G}}^\star\boldsymbol{w} = \widetilde{\boldsymbol{g}}} \Big\{ \langle \widetilde{\boldsymbol{g}}, \boldsymbol{x} \rangle - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}\|^2 + \langle \boldsymbol{v}, \boldsymbol{w} \rangle \Big\} \\ &= \sup_{\boldsymbol{w}\in\Delta} \Big\{ \langle \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} + \boldsymbol{v}, \boldsymbol{w} \rangle - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\mathbf{G}}^\star\boldsymbol{w}\|^2 \Big\} \qquad \forall \boldsymbol{x} \in \mathbb{R}^d.\end{aligned}$$

It follows again from the order reversing property of convex conjugation that $\varphi^{(d)}_{\mathrm{LB}}$ is the smallest function in $\mathfrak{C}_{0,\overline{\lambda}-\underline{\lambda}}(\mathbb{R}^d)$ that interpolates $\big(\boldsymbol{X}_i, \boldsymbol{g}^\star_i - \underline{\lambda}\boldsymbol{X}_i, \varphi^\star_i - \frac{\underline{\lambda}}{2}\|\boldsymbol{X}_i\|^2\big)_{i=1:m}$. Finally, using the equivalence between Lemma 4.3(d) and Lemma 4.3(e), we add $\boldsymbol{x} \mapsto \frac{\underline{\lambda}}{2}\|\boldsymbol{x}\|^2$ to $\varphi^{(d)}_{\mathrm{LB}}$ and define $\varphi^{(e)}_{\mathrm{LB}} : \mathbb{R}^d \to \mathbb{R}$ as follows:

$$\varphi^{(e)}_{\mathrm{LB}}(\boldsymbol{x}) := \varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x}) + \tfrac{\underline{\lambda}}{2}\|\boldsymbol{x}\|^2 = \tfrac{\underline{\lambda}}{2}\|\boldsymbol{x}\|^2 + \sup_{\boldsymbol{w}\in\Delta}\Big\{ \langle \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} + \boldsymbol{v}, \boldsymbol{w} \rangle - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\mathbf{G}}^\star\boldsymbol{w}\|^2 \Big\} \qquad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Therefore, $\varphi^{(e)}_{\mathrm{LB}}$ is the smallest function in $\mathfrak{C}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$ that interpolates $\big(\boldsymbol{X}_i, \boldsymbol{g}^\star_i, \varphi^\star_i\big)_{i=1:m}$. Since $\widehat{\varphi}_{\text{CLS-LB}} = \varphi^{(e)}_{\mathrm{LB}}$, we have completed the proof of statement (i).

To prove statement (ii), observe that $\widehat{T}_{\text{CLS-LB}}$ can be equivalently expressed as follows:

$$\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x}) = \underline{\lambda}\boldsymbol{x} + \mathop{\arg\max}_{\widetilde{\boldsymbol{g}}\in\mathbb{R}^d}\Big\{ \langle \boldsymbol{x}, \widetilde{\boldsymbol{g}} \rangle - \varphi^{(b)}_{\mathrm{UB}}(\widetilde{\boldsymbol{g}}) - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}\|^2 \Big\} \qquad \forall \boldsymbol{x} \in \mathbb{R}^d. \tag{4.9}$$

For every $\boldsymbol{x} \in \mathbb{R}^d$, since the maximization in (4.9) is equivalent to the minimization of a $\frac{1}{\overline{\lambda}-\underline{\lambda}}$-strongly convex function over Δ, a unique optimizer $\widetilde{\boldsymbol{g}}^\star \in \mathbb{R}^d$ is attained. This $\widetilde{\boldsymbol{g}}^\star$ also satisfies $\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x}) = \langle \widetilde{\boldsymbol{g}}^\star, \boldsymbol{x} \rangle - \varphi^{(c)}_{\mathrm{UB}}(\widetilde{\boldsymbol{g}}^\star)$, which implies that $\widetilde{\boldsymbol{g}}^\star = \nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})$ (recall that $\varphi^{(d)}_{\mathrm{LB}} \in \mathfrak{C}_{0,\overline{\lambda}-\underline{\lambda}}(\mathbb{R}^d)$ is the convex conjugate of $\varphi^{(c)}_{\mathrm{UB}}$). Consequently, $\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x}) = \underline{\lambda}\boldsymbol{x} + \widetilde{\boldsymbol{g}}^\star = \nabla\big(\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x}) + \frac{\underline{\lambda}}{2}\|\boldsymbol{x}\|^2\big) = \nabla\varphi^{(e)}_{\mathrm{LB}}(\boldsymbol{x}) = \nabla\widehat{\varphi}_{\text{CLS-LB}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$. The proof is now complete. □

In summary, Definition 4.1 and Theorem 4.7 have provided a computationally tractable plug-in OT map estimator $\widehat{T}_{\text{CLS-LB}}$, and Theorem 4.5 has shown the consistency of $\widehat{T}_{\text{CLS-LB}}$ as well as its estimation error bound. However, $\widehat{T}_{\text{CLS-LB}}$ does not satisfy Assumption 3.4 since the condition $\widehat{\varphi}_{\text{CLS-LB}} \in \mathcal{C}^{2,\alpha}_{\mathrm{loc}}(\mathbb{R}^d)$ does not hold in general. In the next two subsections, we will introduce two smoothing procedures for $\widehat{T}_{\text{CLS-LB}}$ to build plug-in OT map estimators that satisfy Assumption 3.4.

4.2. **Kernel-smoothed OT map estimator.** Let us now introduce our first concrete example of plug-in OT map estimators named kernel-smoothed convex least squares estimator denoted by $\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)$, which satisfies Assumption 3.4. The idea of $\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)$ is to convolute $\widehat{T}_{\text{CLS-LB}}$ in Theorem 4.7 with an infinitely differentiable kernel in order to make it satisfy the differentiability condition in Assumption 3.4(i). [To be discussed: theorem or proposition?]

Theorem 4.8 (Kernel-smoothed convex least squares OT map estimator)**.** *Let the settings of Definition 4.1 hold. Replacing $\underline{\lambda}(\mu,\nu)$ with $\min\big\{\underline{\lambda}(\mu,\nu),1\big\}$ if necessary, let us assume without loss of generality that $\underline{\lambda}\equiv\underline{\lambda}(\mu,\nu)\le 1$. Let $(\widehat{\pi}^{\star}_{i,j})_{i=1:m,\,j=1:n}$, $(\widetilde{\varphi}^{\star}_{i})_{i=1:m}$, and $(\widetilde{\boldsymbol{g}}^{\star}_{i})_{i=1:m}$ be defined via (4.1) and (4.2), and let $\overline{u}_0(\nu):=\inf\big\{r\in\mathbb{R}_+ : \operatorname{supp}(\nu)\subseteq\bar{B}(\mathbf{0},r)\big\}$. Moreover, let Δ, $\widetilde{\mathbf{G}}^{\star}$, and $\boldsymbol{v}$ be defined via (4.5) and let $\widehat{\varphi}_{\text{CLS-LB}}$, $\widehat{T}_{\text{CLS-LB}}$ be defined as in Theorem 4.7. For every $\theta\in\mathbb{N}$, let $\Psi_\theta\in\mathcal{C}^{\infty}(\mathbb{R}^d)$ be the density function of a d-dimensional probability measure with mean $\mathbf{0}_d$ and covariance matrix $\frac{1}{\theta^2}\mathbf{I}_d$, and let $\widehat{\varphi}_{\mathrm{kern}}(\,\cdot\,;\theta):\mathbb{R}^d\to\mathbb{R}$ and $\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta):\mathbb{R}^d\to\mathbb{R}^d$ be defined as follows (the $\mathbb{R}^d$-valued integral in (4.11) is evaluated entry-wise):*

$$\widehat{\varphi}_{\mathrm{kern}}(\boldsymbol{x};\theta):=\int_{\mathbb{R}^d}\Psi_\theta(\boldsymbol{\eta})\widehat{\varphi}_{\text{CLS-LB}}(\boldsymbol{x}-\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}\qquad\forall\boldsymbol{x}\in\mathbb{R}^d,\tag{4.10}$$

$$\widehat{T}_{\mathrm{kern}}(\boldsymbol{x};\theta):=\int_{\mathbb{R}^d}\Psi_\theta(\boldsymbol{\eta})\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x}-\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}\qquad\forall\boldsymbol{x}\in\mathbb{R}^d.\tag{4.11}$$

Then, the following statements hold.

(i) For all $m,n,\theta\in\mathbb{N}$, $\widehat{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)=\nabla\widehat{\varphi}_{\mathrm{kern}}(\boldsymbol{x};\theta)$ for all $\boldsymbol{x}\in\mathbb{R}^d$ and $\widehat{\varphi}_{\mathrm{kern}}(\,\cdot\,;\theta)\in\mathfrak{C}^{\infty}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$.

(ii) For all $m,n,\theta\in\mathbb{N}$, it holds $\mathbb{P}$-almost surely that $\big\|\widehat{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)\big\|^2\le 18\overline{u}_0(\nu)^2+2\|\boldsymbol{x}\|^2$ for all $\boldsymbol{x}\in\mathbb{R}^d$.

(iii) For all $m,n,\theta\in\mathbb{N}$, it holds $\mathbb{P}$-almost surely that $\big\|\widehat{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)-\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x})\big\|^2\le\frac{(\overline{\lambda}-\underline{\lambda})^2d}{\theta^2}$ for all $\boldsymbol{x}\in\mathbb{R}^d$.

(iv) In particular, $\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)$ satisfies Assumption 3.4 with respect to $u_1(\nu):=18\overline{u}_0(\nu)^2$, $u_2(\nu):=2$, $\overline{n}(\mu,\nu,\epsilon):=\min\big\{n\in\mathbb{N} : C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log(n)^2\kappa(n)\le\frac{\epsilon}{4}\big\}$, and $\overline{\theta}(\mu,\nu,m,n,\epsilon):=\Big\lceil\big(\frac{4d}{\epsilon}\big)^{\frac{1}{2}}\big(\overline{\lambda}(\mu,\nu)-\underline{\lambda}(\mu,\nu)\big)\Big\rceil$, where $C(\,\cdot\,,\cdot\,,\cdot\,,\cdot\,)$ and $\kappa(\cdot)$ are given by Theorem 4.5.

Proof of Theorem 4.8. Throughout this proof, let us fix arbitrary $m,n,\theta\in\mathbb{N}$ and let us denote $\widetilde{\varphi}_{\text{CLS-LB}}(\boldsymbol{x}):=\widehat{\varphi}_{\text{CLS-LB}}(\boldsymbol{x})-\frac{\underline{\lambda}}{2}\|\boldsymbol{x}\|^2$, $\widetilde{T}_{\text{CLS-LB}}(\boldsymbol{x}):=\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x})-\underline{\lambda}\boldsymbol{x}$, $\widetilde{\varphi}_{\mathrm{kern}}(\boldsymbol{x};\theta):=\int_{\mathbb{R}^d}\Psi_\theta(\boldsymbol{\eta})\widetilde{\varphi}_{\text{CLS-LB}}(\boldsymbol{x}-\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}$, $\widetilde{T}_{\mathrm{kern}}(\boldsymbol{x};\theta):=\int_{\mathbb{R}^d}\Psi_\theta(\boldsymbol{\eta})\widetilde{T}_{\text{CLS-LB}}(\boldsymbol{x}-\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}$ for all $\boldsymbol{x}\in\mathbb{R}^d$. Observe that $\widehat{\varphi}_{\mathrm{kern}}(\boldsymbol{x};\theta)=\widetilde{\varphi}_{\mathrm{kern}}(\boldsymbol{x};\theta)+\frac{\underline{\lambda}}{2}\|\boldsymbol{x}\|^2+\frac{\underline{\lambda}d}{2\theta^2}$ and $\widehat{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)=\widetilde{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)+\underline{\lambda}\boldsymbol{x}$ for all $\boldsymbol{x}\in\mathbb{R}^d$.

To prove statement (i), note that the proof of Theorem 4.7 has shown that $\widetilde{T}_{\text{CLS-LB}}=\nabla\widetilde{\varphi}_{\text{CLS-LB}}$ and that $\widetilde{T}_{\text{CLS-LB}}$ is $(\overline{\lambda}-\underline{\lambda})$-Lipschitz continuous. Consequently, we may apply the Leibniz integral rule to exchange integration and differentiation as follows:

$$\nabla\widetilde{\varphi}_{\mathrm{kern}}(\boldsymbol{x};\theta)=\int_{\mathbb{R}^d}\Psi_\theta(\boldsymbol{\eta})\nabla\widetilde{\varphi}_{\text{CLS-LB}}(\boldsymbol{x}-\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}=\int_{\mathbb{R}^d}\Psi_\theta(\boldsymbol{\eta})\widetilde{T}_{\text{CLS-LB}}(\boldsymbol{x}-\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}=\widetilde{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)\quad\forall\boldsymbol{x}\in\mathbb{R}^d.$$

Thus, it holds that $\widehat{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)=\widetilde{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)+\underline{\lambda}\boldsymbol{x}=\nabla\big(\widetilde{\varphi}_{\mathrm{kern}}(\boldsymbol{x};\theta)+\frac{\underline{\lambda}}{2}\|\boldsymbol{x}\|^2+\frac{\underline{\lambda}d}{2\theta^2}\big)=\nabla\widehat{\varphi}_{\mathrm{kern}}(\boldsymbol{x};\theta)$. Moreover, since $\widetilde{\varphi}_{\mathrm{kern}}$ is the convolution of $\widetilde{\varphi}_{\text{CLS-LB}}$ with $\Psi_\theta\in\mathcal{C}^{\infty}(\mathbb{R}^d)$, it holds that $\widetilde{\varphi}_{\mathrm{kern}}\in\mathcal{C}^{\infty}(\mathbb{R}^d)$. Furthermore, observe that $\widetilde{\varphi}_{\mathrm{kern}}$ is convex due to the convexity of $\widetilde{\varphi}_{\text{CLS-LB}}$, and that

$$\big\|\widetilde{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)-\widetilde{T}_{\mathrm{kern}}(\boldsymbol{y};\theta)\big\|\le\int_{\mathbb{R}^d}\big\|\widetilde{T}_{\text{CLS-LB}}(\boldsymbol{x})-\widetilde{T}_{\text{CLS-LB}}(\boldsymbol{y})\big\|\Psi_\theta(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}\le(\overline{\lambda}-\underline{\lambda})\|\boldsymbol{x}-\boldsymbol{y}\|\quad\forall\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d,$$

which show that $\widetilde{\varphi}_{\mathrm{kern}}\in\mathfrak{C}^{\infty}_{0,\overline{\lambda}-\underline{\lambda}}(\mathbb{R}^d)$. We thus conclude that $\widehat{\varphi}_{\mathrm{kern}}\in\mathfrak{C}^{\infty}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$. The proof of statement (i) is now complete.

To prove statement (ii), observe from (4.7) that

$$\widetilde{T}_{\text{CLS-LB}}(\boldsymbol{x})\in\operatorname{conv}\big(\{\widetilde{\boldsymbol{g}}^{\star}_1,\dots,\widetilde{\boldsymbol{g}}^{\star}_m\}\big)\qquad\forall\boldsymbol{x}\in\mathbb{R}^d.\tag{4.12}$$

Moreover, since $(\widetilde{\varphi}^{\star}_i)_{i=1:m}$, $(\widetilde{\boldsymbol{g}}^{\star}_i)_{i=1:m}$ is an optimizer of the QCQP problem (4.2), it follows from the constraints in (4.2) that

$$\max_{1\le i\le m}\big\{\|\widetilde{\boldsymbol{g}}^{\star}_i\|\big\}\le\max_{1\le i\le m}\big\{\|\widetilde{\boldsymbol{g}}^{\star}_i+\underline{\lambda}\boldsymbol{X}_i\|+\underline{\lambda}\|\boldsymbol{X}_i\|\big\}\le\overline{u}_0(\nu)+\underline{\lambda}\max_{1\le i\le m}\big\{\|\boldsymbol{X}_i\|\big\}.$$

Thus, we have

$$\max_{1\le i\le m}\big\{\|\widetilde{\boldsymbol{g}}_i^\star\|\big\} \le \overline{u}_0(\nu) + \underline{\lambda} \sup_{\boldsymbol{y}\in\operatorname{supp}(\mu)}\big\{\|\boldsymbol{y}\|\big\} \qquad \mathbb{P}\text{-a.s.} \tag{4.13}$$

Furthermore, since φ_ν^μ is $\underline{\lambda}$-strongly convex by assumption, it holds for all $\boldsymbol{y}\in\operatorname{supp}(\mu)$ that

$$\begin{aligned}\varphi_\nu^\mu(\boldsymbol{y}) - \varphi_\nu^\mu(\mathbf{0}_d) - \langle T_\nu^\mu(\mathbf{0}_d), \boldsymbol{y}\rangle &\ge \frac{\underline{\lambda}}{2}\|\boldsymbol{y}\|^2,\\ \varphi_\nu^\mu(\mathbf{0}_d) - \varphi_\nu^\mu(\boldsymbol{y}) + \langle T_\nu^\mu(\boldsymbol{y}), \boldsymbol{y}\rangle &\ge \frac{\underline{\lambda}}{2}\|\boldsymbol{y}\|^2.\end{aligned}$$

Adding both sides of the two inequalities, applying the Cauchy–Schwarz inequality, and using the assumption $\mathbf{0}_d\in\operatorname{supp}(\mu)$ in Definition 4.1 yields

$$\underline{\lambda}\|\boldsymbol{y}\| \le \big\|T_\nu^\mu(\boldsymbol{y}) - T_\nu^\mu(\mathbf{0}_d)\big\| \le 2\sup_{\boldsymbol{z}\in\operatorname{supp}(\nu)}\big\{\|\boldsymbol{z}\|\big\} \le 2\overline{u}_0(\nu) \qquad \forall \boldsymbol{y}\in\operatorname{supp}(\mu). \tag{4.14}$$

Combining (4.13) and (4.14) yields

$$\max_{1\le i\le m}\big\{\|\widetilde{\boldsymbol{g}}_i^\star\|\big\} \le 3\overline{u}_0(\nu) \qquad \mathbb{P}\text{-a.s.} \tag{4.15}$$

Subsequently, Jensen's inequality together with the convexity of $\mathbb{R}^d\ni\boldsymbol{x}\mapsto\|\boldsymbol{x}\|^2\in\mathbb{R}$, (4.12), (4.15), and the assumption $\underline{\lambda}\le 1$ imply that

$$\begin{aligned}\big\|\widehat{T}_{\text{kern}}(\boldsymbol{x};\theta)\big\|^2 &\le 2\big\|\widetilde{T}_{\text{kern}}(\boldsymbol{x},\theta)\big\|^2 + 2\underline{\lambda}^2\|\boldsymbol{x}\|^2\\ &\le 2\int_{\mathbb{R}^d}\big\|\widetilde{T}_{\text{CLS-LB}}(\boldsymbol{x}-\boldsymbol{\eta})\big\|^2\Psi_\theta(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta} + 2\underline{\lambda}^2\|\boldsymbol{x}\|^2\\ &\le 2\max_{1\le i\le m}\big\{\|\widetilde{\boldsymbol{g}}_i^\star\|\big\}^2 + 2\|\boldsymbol{x}\|^2\\ &\le 18\overline{u}_0(\nu)^2 + 2\|\boldsymbol{x}\|^2 \qquad \mathbb{P}\text{-a.s.}\end{aligned}$$

We have thus completed the proof of statement (ii).

Statement (iii) follows from the convexity of $\mathbb{R}^d\ni\boldsymbol{x}\mapsto\|\boldsymbol{x}\|^2\in\mathbb{R}$, Jensen's inequality, the $(\overline{\lambda}-\underline{\lambda})$-Lipschitz continuity of $\widetilde{T}_{\text{CLS-LB}}$, and the assumption that Ψ_θ is the density function of a distribution with mean $\mathbf{0}_d$ and covariance $\frac{1}{\theta^2}\mathbf{I}_d$:

$$\begin{aligned}\big\|\widehat{T}_{\text{kern}}(\boldsymbol{x};\theta) - \widehat{T}_{\text{CLS-LB}}(\boldsymbol{x})\big\|^2 &= \big\|\widetilde{T}_{\text{kern}}(\boldsymbol{x};\theta) - \widetilde{T}_{\text{CLS-LB}}(\boldsymbol{x})\big\|^2\\ &= \left\|\int_{\mathbb{R}^d}\big(\widetilde{T}_{\text{CLS-LB}}(\boldsymbol{x}-\boldsymbol{\eta}) - \widetilde{T}_{\text{CLS-LB}}(\boldsymbol{x})\big)\Psi_\theta(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}\right\|^2\\ &\le \int_{\mathbb{R}^d}\big\|\widetilde{T}_{\text{CLS-LB}}(\boldsymbol{x}-\boldsymbol{\eta}) - \widetilde{T}_{\text{CLS-LB}}(\boldsymbol{x})\big\|^2\Psi_\theta(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}\\ &\le \int_{\mathbb{R}^d}(\overline{\lambda}-\underline{\lambda})^2\|\boldsymbol{\eta}\|^2\Psi_\theta(\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\eta}\\ &= \frac{(\overline{\lambda}-\underline{\lambda})^2 d}{\theta^2} \qquad \forall\boldsymbol{x}\in\mathbb{R}^d.\end{aligned}$$

Finally, observe that statement (i) shows that $\widehat{T}_{\text{kern}}(\,\cdot\,;\theta)$ satisfies the shape condition in Assumption 3.4(i). Moreover, statement (ii) implies that $\mathbb{E}\Big[\big\|\widehat{T}_{\text{kern}}(\boldsymbol{x};\theta)\big\|^2\Big] \le 18\overline{u}_0(\nu)^2 + 2\|\boldsymbol{x}\|^2$ for all $\boldsymbol{x}$, which shows that $\widehat{T}_{\text{kern}}(\,\cdot\,;\theta)$ satisfies the growth condition in Assumption 3.4(ii) with respect to $u_1(\nu) := 18\overline{u}_0(\nu)^2$, $u_2(\nu) := 2$. Furthermore, statement (iii) shows that

$$\begin{aligned}\big\|\widehat{T}_{\text{kern}}(\boldsymbol{x};\theta) - T_\nu^\mu(\boldsymbol{x})\big\|^2 &\le 2\big\|\widehat{T}_{\text{kern}}(\boldsymbol{x};\theta) - \widehat{T}_{\text{CLS-LB}}(\boldsymbol{x})\big\|^2 + 2\big\|\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x}) - T_\nu^\mu(\boldsymbol{x})\big\|^2\\ &\le \frac{2(\overline{\lambda}(\mu,\nu)-\underline{\lambda}(\mu,\nu))^2 d}{\theta^2} + 2\big\|\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x}) - T_\nu^\mu(\boldsymbol{x})\big\|^2 \qquad \forall\boldsymbol{x}\in\mathbb{R}^d.\end{aligned}$$

Combining this with Theorem 4.7, Theorem 4.2, and Theorem 4.5 leads to

$$\mathbb{E}\Big[\big\|\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)-T^{\mu}_{\nu}\big\|^2_{\mathcal{L}^2(\mu)}\Big]\le\frac{2(\overline{\lambda}(\mu,\nu)-\underline{\lambda}(\mu,\nu))^2d}{\theta^2}\\ +2C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log(m)^2\kappa\big(\min\{m,n\}\big).$$

Consequently, for any $\epsilon>0$, with $\overline{n}(\mu,\nu,\epsilon):=\min\big\{n\in\mathbb{N}:C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log(n)^2\kappa(n)\le\frac{\epsilon}{4}\big\}$, $\overline{\theta}(\mu,\nu,m,n,\epsilon):=\Big\lceil\big(\frac{4d}{\epsilon}\big)^{\frac{1}{2}}(\overline{\lambda}(\mu,\nu)-\underline{\lambda}(\mu,\nu))\Big\rceil$, it holds for all $m\ge\overline{n}(\mu,\nu,\epsilon)$, $n\ge\overline{n}(\mu,\nu,\epsilon)$ and all $\theta\ge\overline{\theta}(\mu,\nu,m,n,\epsilon)$ that

$$\begin{aligned}&\mathbb{E}\Big[\big\|\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)-T^{\mu}_{\nu}\big\|^2_{\mathcal{L}^2(\mu)}\Big]\\&\quad\le\frac{2(\overline{\lambda}(\mu,\nu)-\underline{\lambda}(\mu,\nu))^2d}{\theta^2}+2C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log(m)^2\kappa\big(\min\{m,n\}\big)\\&\quad\le\frac{2(\overline{\lambda}(\mu,\nu)-\underline{\lambda}(\mu,\nu))^2d}{\frac{4d}{\epsilon}(\overline{\lambda}(\mu,\nu)-\underline{\lambda}(\mu,\nu))^2}+2C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log\big(\overline{n}(\mu,\nu,\epsilon)\big)^2\kappa\big(\overline{n}(\mu,\nu,\epsilon)\big)\\&\quad\le\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon.\end{aligned}$$

We have thus shown that $\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)$ satisfies the consistency condition in Assumption 3.4(iii) with respect to $\overline{n}(\mu,\nu,\epsilon):=\min\big\{n\in\mathbb{N}:C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log(n)^2\kappa(n)\le\frac{\epsilon}{4}\big\}$ and $\overline{\theta}(\mu,\nu,m,n,\epsilon):=\Big\lceil\big(\frac{4d}{\epsilon}\big)^{\frac{1}{2}}(\overline{\lambda}(\mu,\nu)-\underline{\lambda}(\mu,\nu))\Big\rceil$. The proof is now complete. □

Remark 4.9 (Computational tractability of $\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)$)**.** *The computation of $\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)$ is done in two phases. In the first phase, given the m samples $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_m$ from μ and the n samples $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n$ from ν, one computes $(\widetilde{\varphi}^{\star}_i)_{i=1:m}$ and $(\widetilde{\boldsymbol{g}}^{\star}_i)_{i=1:m}$ by first solving the LP problem (4.1) and then solving the QCQP problem (4.2); see Definition 4.1. These two problems can be solved by state-of-the-art LP and QCQP solvers such as Gurobi [43]. Subsequently, in the second phase, for every $\boldsymbol{x}\in\mathbb{R}^d$ at which $\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)$ needs to be evaluated, one can compute $\widehat{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)$ by Monte Carlo approximation. Specifically, one generates $S\in\mathbb{N}$ independent random samples $\{\boldsymbol{\eta}_{[s]}\}_{s=1:S}$ from the probability distribution whose density function is equal to Ψ_θ, and then approximates $\widehat{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)$ by $\widehat{T}_{\mathrm{kern}}(\boldsymbol{x};\theta)\approx\frac{1}{S}\sum_{s=1}^{S}\widehat{T}_{\mathrm{CLS\text{-}LB}}(\boldsymbol{x}-\boldsymbol{\eta}_{[s]})$, where each $\widehat{T}_{\mathrm{CLS\text{-}LB}}(\boldsymbol{x}-\boldsymbol{\eta}_{[s]})$ is computed by solving the QP problem in (4.7). This can be done using state-of-the-art QP solvers, which is also provided by Gurobi [43]. Consequently, using $\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)$ as the plug-in OT map estimator in Algorithm 2 results in a computationally tractable algorithm for $\mathcal{W}_2$-barycenter that is also provably convergent.*

4.3. Barrier-based OT map estimator.

In this subsection, we introduce the following barrier-based convex least squares estimator $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta)$ inspired by the smoothing technique of Nesterov [66]. [To discuss: theorem or proposition?]

Theorem 4.10 (Barrier-based convex least squares OT map estimator)**.** *Let the settings of Definition 4.1 hold. Replacing $\underline{\lambda}(\mu,\nu)$ with $\min\big\{\underline{\lambda}(\mu,\nu),1\big\}$ if necessary, let us assume without loss of generality that $\underline{\lambda}\equiv\underline{\lambda}(\mu,\nu)\le1$. Let $(\widetilde{\pi}^{\star}_{i,j})_{i=1:m,j=1:n}$, $(\widetilde{\varphi}^{\star}_i)_{i=1:m}$, and $(\widetilde{\boldsymbol{g}}^{\star}_i)_{i=1:m}$ be defined via (4.1) and (4.2), and let $\overline{u}_0(\nu):=\inf\big\{r\in\mathbb{R}_+:\operatorname{supp}(\nu)\subseteq\bar{B}(\boldsymbol{0},r)\big\}$. Moreover, let Δ, $\widetilde{\mathbf{G}}^{\star}$, and $\boldsymbol{v}$ be defined via (4.5) and let $\eta:\Delta\to\mathbb{R}_+$ be given by $\eta(w_1,\ldots,w_m):=-\sum_{i=1}^{m}\log(w_i)$. For every $\theta\in\mathbb{N}$, let $\widehat{\varphi}_{\mathrm{barr}}(\,\cdot\,;\theta):\mathbb{R}^d\to\mathbb{R}$ and $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta):\mathbb{R}^d\to\mathbb{R}^d$ be defined as follows:*

$$\widehat{\varphi}_{\mathrm{barr}}(\boldsymbol{x};\theta):=\tfrac{\underline{\lambda}}{2}\|\boldsymbol{x}\|^2+\sup_{\boldsymbol{w}\in\Delta}\Big\{\langle\widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x}+\boldsymbol{v},\boldsymbol{w}\rangle-\tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\mathbf{G}}^{\star}\boldsymbol{w}\|^2-\tfrac{\eta(\boldsymbol{w})}{\theta}\Big\}\qquad\forall\boldsymbol{x}\in\mathbb{R}^d,\tag{4.16}$$

$$\begin{aligned}&\widehat{T}_{\mathrm{barr}}(\boldsymbol{x};\theta):=\underline{\lambda}\boldsymbol{x}+\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta),\\&\qquad\textit{where }\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta):=\operatorname*{arg\,max}_{\boldsymbol{w}\in\Delta}\Big\{\langle\widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x}+\boldsymbol{v},\boldsymbol{w}\rangle-\tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\mathbf{G}}^{\star}\boldsymbol{w}\|^2-\tfrac{\eta(\boldsymbol{w})}{\theta}\Big\}\qquad\forall\boldsymbol{x}\in\mathbb{R}^d.\end{aligned}\tag{4.17}$$

Then, the following statements hold.

(i) For all $m,n,\theta\in\mathbb{N}$, $\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)$ in (4.17) is uniquely defined and $\widehat{T}_{\mathrm{barr}}(\boldsymbol{x};\theta)=\nabla\widehat{\varphi}_{\mathrm{barr}}(\boldsymbol{x};\theta)$ for all $\boldsymbol{x}\in\mathbb{R}^d$. Moreover, it holds that $\widehat{\varphi}_{\mathrm{barr}}(\,\cdot\,;\theta)\in\mathfrak{C}^{\infty}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$.

(ii) *For all $m,n,\theta\in\mathbb{N}$, it holds $\mathbb{P}$-almost surely that $\big\|\widehat{T}_{\mathrm{barr}}(\boldsymbol{x};\theta)\big\|^2\le 18\overline{u}_0(\nu)^2+2\|\boldsymbol{x}\|^2$ for all $\boldsymbol{x}\in\mathbb{R}^d$.*

(iii) *For all $m,n,\theta\in\mathbb{N}$, it holds $\mathbb{P}$-almost surely that $\big\|\widehat{T}_{\mathrm{barr}}(\boldsymbol{x};\theta)-\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x})\big\|^2\le\frac{2m(\overline{\lambda}-\underline{\lambda})}{\theta}$ for all $\boldsymbol{x}\in\mathbb{R}^d$, where $\widehat{T}_{\text{CLS-LB}}$ is defined in Theorem 4.7.*

(iv) *In particular, $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta)$ satisfies Assumption 3.4 with respect to $u_1(\nu):=18\overline{u}_0(\nu)^2$, $u_2(\nu):=2$, $\overline{n}(\mu,\nu,\epsilon):=\min\big\{n\in\mathbb{N}:C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log(n)^2\kappa(n)\le\frac{\epsilon}{4}\big\}$, and $\overline{\theta}(\mu,\nu,m,n,\epsilon):=\big\lceil\frac{8m}{\epsilon}\big(\overline{\lambda}(\mu,\nu)-\underline{\lambda}(\mu,\nu)\big)\big\rceil$, where $C(\,\cdot\,,\cdot\,,\cdot\,,\cdot\,)$ and $\kappa(\cdot)$ are given by Theorem 4.5.*

Proof of Theorem 4.10. Throughout this proof, let us fix arbitrary $m,n,\theta\in\mathbb{N}$. To begin, observe that we can extend the definition of η to $\mathbb{R}^d$ such that $\eta\in\mathfrak{C}_{1,\infty}(\mathbb{R}^d)$. To prove statement (i), let us define $\psi_\theta^{(a)}:\mathbb{R}^d\to\mathbb{R}$ and $\psi_\theta^{(b)}:\mathbb{R}^d\to\mathbb{R}\cup\{\infty\}$ as follows:

$$\begin{aligned}\psi_\theta^{(a)}(\boldsymbol{x})&:=\sup_{\boldsymbol{w}\in\Delta}\big\{\langle\widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x}+\boldsymbol{v},\boldsymbol{w}\rangle-\tfrac{\eta(\boldsymbol{w})}{\theta}\big\} && \forall\boldsymbol{x}\in\mathbb{R}^d,\\ \psi_\theta^{(b)}(\widetilde{\boldsymbol{g}})&:=\sup_{\boldsymbol{x}\in\mathbb{R}^d}\big\{\langle\widetilde{\boldsymbol{g}},\boldsymbol{x}\rangle-\psi_\theta^{(a)}(\boldsymbol{x})\big\}\\ &=\sup_{\boldsymbol{x}\in\mathbb{R}^d}\Big\{\inf_{\boldsymbol{w}\in\Delta}\Big\{\langle\widetilde{\boldsymbol{g}}-\widetilde{\mathbf{G}}^{\star}\boldsymbol{w},\boldsymbol{x}\rangle-\langle\boldsymbol{v},\boldsymbol{w}\rangle+\tfrac{\eta(\boldsymbol{w})}{\theta}\Big\}\Big\} && \forall\widetilde{\boldsymbol{g}}\in\mathbb{R}^d.\end{aligned}$$

Since $\mathbb{R}^d\times\Delta\ni(\boldsymbol{x},\boldsymbol{w})\mapsto\langle\widetilde{\boldsymbol{g}}-\widetilde{\mathbf{G}}^{\star}\boldsymbol{w},\boldsymbol{x}\rangle-\langle\boldsymbol{v},\boldsymbol{w}\rangle+\frac{\eta(\boldsymbol{w})}{\theta}\in\mathbb{R}$ is concave in $\boldsymbol{x}$ for every $\boldsymbol{w}\in\Delta$ and convex in $\boldsymbol{w}$ for every $\boldsymbol{x}\in\mathbb{R}^d$, it follows from the compactness of Δ and Sion's minimax theorem [82] that

$$\begin{aligned}\psi_\theta^{(b)}(\widetilde{\boldsymbol{g}})&=\sup_{\boldsymbol{x}\in\mathbb{R}^d}\Big\{\inf_{\boldsymbol{w}\in\Delta}\Big\{\langle\widetilde{\boldsymbol{g}}-\widetilde{\mathbf{G}}^{\star}\boldsymbol{w},\boldsymbol{x}\rangle-\langle\boldsymbol{v},\boldsymbol{w}\rangle+\tfrac{\eta(\boldsymbol{w})}{\theta}\Big\}\Big\}\\ &=\inf_{\boldsymbol{w}\in\Delta}\left\{\sup_{\boldsymbol{x}\in\mathbb{R}^d}\big\{\langle\widetilde{\boldsymbol{g}}-\widetilde{\mathbf{G}}^{\star}\boldsymbol{w},\boldsymbol{x}\rangle\big\}-\langle\boldsymbol{v},\boldsymbol{w}\rangle+\tfrac{\eta(\boldsymbol{w})}{\theta}\right\}\\ &=-\sup_{\boldsymbol{w}\in\Delta,\,\widetilde{\mathbf{G}}^{\star}\boldsymbol{w}=\widetilde{\boldsymbol{g}}}\Big\{\langle\boldsymbol{v},\boldsymbol{w}\rangle-\tfrac{\eta(\boldsymbol{w})}{\theta}\Big\} \qquad \forall\widetilde{\boldsymbol{g}}\in\mathbb{R}^d.\end{aligned}$$

We have $\psi_\theta^{(b)}\in\mathfrak{C}_{0,\infty}(\mathbb{R}^d)$ by its definition. Subsequently, let us define $\psi_\theta^{(c)}:\mathbb{R}^d\to\mathbb{R}\cup\{\infty\}$ as follows:

$$\psi_\theta^{(c)}(\widetilde{\boldsymbol{g}}):=\psi_\theta^{(b)}(\widetilde{\boldsymbol{g}})+\tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}\|^2=\tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}\|^2-\sup_{\boldsymbol{w}\in\Delta,\,\widetilde{\mathbf{G}}^{\star}\boldsymbol{w}=\widetilde{\boldsymbol{g}}}\Big\{\langle\boldsymbol{v},\boldsymbol{w}\rangle-\tfrac{\eta(\boldsymbol{w})}{\theta}\Big\} \qquad \forall\widetilde{\boldsymbol{g}}\in\mathbb{R}^d.$$

Thus, $\psi_\theta^{(c)}\in\mathfrak{C}_{\frac{1}{\overline{\lambda}-\underline{\lambda}},\infty}(\mathbb{R}^d)$. Continuing on, let us define $\psi_\theta^{(d)}:\mathbb{R}^d\to\mathbb{R}$ as the convex conjugate of $\psi_\theta^{(c)}$:

$$\begin{aligned}\psi_\theta^{(d)}(\boldsymbol{x})&:=\sup_{\widetilde{\boldsymbol{g}}\in\mathbb{R}^d}\big\{\langle\widetilde{\boldsymbol{g}},\boldsymbol{x}\rangle-\psi_\theta^{(c)}(\widetilde{\boldsymbol{g}})\big\}\\ &=\sup_{\widetilde{\boldsymbol{g}}\in\mathbb{R}^d,\,\boldsymbol{w}\in\Delta,\,\widetilde{\mathbf{G}}^{\star}\boldsymbol{w}=\widetilde{\boldsymbol{g}}}\Big\{\langle\widetilde{\boldsymbol{g}},\boldsymbol{x}\rangle-\tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}\|^2+\langle\boldsymbol{v},\boldsymbol{w}\rangle-\tfrac{\eta(\boldsymbol{w})}{\theta}\Big\}\\ &=\sup_{\boldsymbol{w}\in\Delta}\Big\{\langle\widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x}+\boldsymbol{v},\boldsymbol{w}\rangle-\tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\mathbf{G}}^{\star}\boldsymbol{w}\|^2-\tfrac{\eta(\boldsymbol{w})}{\theta}\Big\} \qquad \forall\boldsymbol{x}\in\mathbb{R}^d.\end{aligned}\tag{4.18}$$

It hence holds by the well-known duality between smooth convex functions and strongly convex functions (see, e.g., [77, Proposition 12.60]) that $\psi_\theta^{(d)}\in\mathfrak{C}_{0,\overline{\lambda}-\underline{\lambda}}(\mathbb{R}^d)$. Lastly, since $\widehat{\varphi}_{\mathrm{barr}}(\,\cdot\,;\theta)=\psi_\theta^{(d)}(\cdot)+\frac{\underline{\lambda}}{2}\|\cdot\|^2$, we get $\widehat{\varphi}_{\mathrm{barr}}(\,\cdot\,;\theta)\in\mathfrak{C}_{\underline{\lambda},\overline{\lambda}}(\mathbb{R}^d)$. Moreover, for every $\boldsymbol{x}\in\mathbb{R}^d$, the maximization in (4.17) is equivalent to the minimization of a $\frac{1}{\theta}$-strongly convex and l.s.c. function over the compact set Δ, and is hence always attained at a unique maximizer $\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)$. Observe from (4.18) that $\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)$ is a maximizer of $\sup_{\widetilde{\boldsymbol{g}}\in\mathbb{R}^d}\big\{\langle\widetilde{\boldsymbol{g}},\boldsymbol{x}\rangle-\psi_\theta^{(c)}(\widetilde{\boldsymbol{g}})\big\}$, and thus $\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\in\partial\psi_\theta^{(d)}(\boldsymbol{x})$. Since $\psi_\theta^{(d)}\in\mathfrak{C}_{0,\overline{\lambda}-\underline{\lambda}}(\mathbb{R}^d)$ and $\overline{\lambda}-\underline{\lambda}<\infty$, it holds that $\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)=\nabla\psi_\theta^{(d)}(\boldsymbol{x})$. Consequently, we get $\widehat{T}_{\mathrm{barr}}(\boldsymbol{x};\theta):=\underline{\lambda}\boldsymbol{x}+\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)=\underline{\lambda}\boldsymbol{x}+\nabla\psi_\theta^{(d)}(\boldsymbol{x})=\nabla\widehat{\varphi}_{\mathrm{barr}}(\boldsymbol{x};\theta)$, which shows that $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta)=\nabla\widehat{\varphi}_{\mathrm{barr}}(\,\cdot\,;\theta)$.

To complete the proof of statement (i), it remains to show that $\widehat{\varphi}_{\mathrm{barr}}(\,\cdot\,;\theta)\in\mathcal{C}^\infty(\mathbb{R}^d)$. By denoting the entry-wise inverse of $\boldsymbol{w}$ by $\boldsymbol{w}^{\circ-1}\in\mathbb{R}^m$ and introducing Lagrange multiplier variables $\boldsymbol{\zeta}\in\mathbb{R}^m_+$ and $\xi\in\mathbb{R}$, the Karush–Kuhn–Tucker (KKT) optimality conditions associated with the maximization problem in (4.17) are given by:

$$\begin{cases} -\frac{1}{\theta}\boldsymbol{w}^{\circ-1} + \frac{1}{\overline{\lambda}-\underline{\lambda}}\widetilde{\mathbf{G}}^{\star\mathsf{T}}\widetilde{\mathbf{G}}^{\star}\boldsymbol{w} - \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} - \boldsymbol{v} - \boldsymbol{\zeta} - \xi\mathbf{1}_m = \mathbf{0}_m, \\ \boldsymbol{w} \ge \mathbf{0}_m, \\ \langle \mathbf{1}_m, \boldsymbol{w}\rangle = 1, \\ \boldsymbol{\zeta} \ge \mathbf{0}_m, \\ \langle \boldsymbol{\zeta}, \boldsymbol{w}\rangle = 0, \end{cases}$$

where $\mathbf{1}_m$ denotes the vector in $\mathbb{R}^m$ with all entries equal to 1. For every $\boldsymbol{x} \in \mathbb{R}^d$, the log-barrier $\eta(\cdot)$ guarantees that the maximizer $\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)$ of the maximization problem in (4.17) must be in the relative interior of Δ. Hence, $\boldsymbol{\zeta} = \mathbf{0}_m$ holds necessarily and the above KKT conditions can be simplified into the following system of equations: $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w},\xi) = \mathbf{0}_{m+1}$, where the vector-valued function $\boldsymbol{F} : \mathbb{R}^d \times (0,1)^m \times \mathbb{R} \to \mathbb{R}^{m+1}$ is defined by

$$\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w},\xi) := \begin{pmatrix} -\frac{1}{\theta}\boldsymbol{w}^{\circ-1} + \frac{1}{\overline{\lambda}-\underline{\lambda}}\widetilde{\mathbf{G}}^{\star\mathsf{T}}\widetilde{\mathbf{G}}^{\star}\boldsymbol{w} - \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} - \boldsymbol{v} - \xi\mathbf{1}_m \\ \langle \mathbf{1}_m, \boldsymbol{w}\rangle - 1 \end{pmatrix} \in \mathbb{R}^{m+1} \tag{4.19}$$
$$\forall \boldsymbol{x} \in \mathbb{R}^d,\ \forall \boldsymbol{w} \in (0,1)^m,\ \forall \xi \in \mathbb{R}.$$

Note that $\boldsymbol{F}$ is infinitely differentiable on $\mathbb{R}^d \times (0,1)^m \times \mathbb{R}$. Since we have already shown that the maximizer $\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)$ in (4.17) is unique for every $\boldsymbol{x} \in \mathbb{R}^d$, and that $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w},\xi) = \mathbf{0}_{m+1}$ are the sufficient and necessary optimality conditions, the function $\widehat{\xi}(\,\cdot\,;\theta) : \mathbb{R}^d \to \mathbb{R}$ defined below satisfies $F\big(\boldsymbol{x}, \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta), \widehat{\xi}(\boldsymbol{x};\theta)\big) = \mathbf{0}_{m+1}$ for all $\boldsymbol{x} \in \mathbb{R}^d$:

$$\widehat{\xi}(\boldsymbol{x};\theta) = -\tfrac{1}{m\theta}\langle \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)^{\circ-1}, \mathbf{1}_m\rangle + \tfrac{1}{m(\overline{\lambda}-\underline{\lambda})}\langle \widetilde{\mathbf{G}}^{\star\mathsf{T}}\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta), \mathbf{1}_m\rangle - \tfrac{1}{m}\langle \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} + \boldsymbol{v}, \mathbf{1}_m\rangle \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Let $\nabla_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w},\xi)$ denote the partial derivative of $\boldsymbol{F}$ with respect to $\boldsymbol{x}$ and let $\nabla_{\boldsymbol{w},\xi}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w},\xi)$ denote the partial derivative of $\boldsymbol{F}$ with respect to $(\boldsymbol{w},\xi)$. They are given by the following matrix-valued functions:

$$\nabla_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w},\xi) = -\begin{pmatrix} \widetilde{\mathbf{G}}^{\star} \\ \mathbf{0}_d^{\mathsf{T}} \end{pmatrix} \in \mathbb{R}^{(m+1)\times d} \qquad \forall \boldsymbol{x} \in \mathbb{R}^d,\ \forall \boldsymbol{w} \in (0,1)^m,\ \forall \xi \in \mathbb{R},$$

$$\nabla_{\boldsymbol{w},\xi}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w},\xi) = \begin{pmatrix} \frac{1}{\theta}\mathrm{diag}(\boldsymbol{w})^{-2} + \frac{1}{\overline{\lambda}-\underline{\lambda}}\widetilde{\mathbf{G}}^{\star\mathsf{T}}\widetilde{\mathbf{G}}^{\star} & -\mathbf{1}_m \\ \mathbf{1}_m^{\mathsf{T}} & 0 \end{pmatrix} \in \mathbb{R}^{(m+1)\times(m+1)}$$
$$\forall \boldsymbol{x} \in \mathbb{R}^d,\ \forall \boldsymbol{w} \in (0,1)^m,\ \forall \xi \in \mathbb{R}.$$

Note that since $\frac{1}{\theta}\mathrm{diag}(\boldsymbol{w})^{-2} + \frac{1}{\overline{\lambda}-\underline{\lambda}}\widetilde{\mathbf{G}}^{\star\mathsf{T}}\widetilde{\mathbf{G}}^{\star} \succeq \frac{1}{\theta}\mathbf{I}_m$ for every $\boldsymbol{w} \in (0,1)^m$, we have $\det\Big(\frac{1}{\theta}\mathrm{diag}(\boldsymbol{w})^{-2} + \frac{1}{\overline{\lambda}-\underline{\lambda}}\widetilde{\mathbf{G}}^{\star\mathsf{T}}\widetilde{\mathbf{G}}^{\star}\Big) \ge \frac{1}{\theta^m}$. Consequently, we get via the determinant formula of block matrices that

$$\begin{aligned} \det\big(\nabla_{\boldsymbol{w},\xi}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w},\xi)\big) &= \det\Big(\tfrac{1}{\theta}\mathrm{diag}(\boldsymbol{w})^{-2} + \tfrac{1}{\overline{\lambda}-\underline{\lambda}}\widetilde{\mathbf{G}}^{\star\mathsf{T}}\widetilde{\mathbf{G}}^{\star}\Big)\det\Big(\mathbf{1}_m^{\mathsf{T}}\Big(\tfrac{1}{\theta}\mathrm{diag}(\boldsymbol{w})^{-2} + \tfrac{1}{\overline{\lambda}-\underline{\lambda}}\widetilde{\mathbf{G}}^{\star\mathsf{T}}\widetilde{\mathbf{G}}^{\star}\Big)\mathbf{1}_m\Big) \\ &\ge \tfrac{m}{\theta^{m+1}} > 0 \qquad\qquad \forall \boldsymbol{x} \in \mathbb{R}^d,\ \forall \boldsymbol{w} \in (0,1)^m,\ \forall \xi \in \mathbb{R}. \end{aligned}$$

It thus follows from the implicit function theorem (see, e.g., [34, Theorem 1B.1]) that $\widehat{\boldsymbol{w}}(\,\cdot\,;\theta)$ and $\widehat{\xi}(\,\cdot\,;\theta)$ are both continuous, and that

$$\nabla\begin{pmatrix} \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) \\ \widehat{\xi}(\boldsymbol{x};\theta) \end{pmatrix} = -\big[\nabla_{\boldsymbol{w},\xi}\boldsymbol{F}\big(\boldsymbol{x}, \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta), \widehat{\xi}(\boldsymbol{x};\theta)\big)\big]^{-1}\nabla_{\boldsymbol{x}}\boldsymbol{F}\big(\boldsymbol{x}, \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta), \widehat{\xi}(\boldsymbol{x};\theta)\big) \in \mathbb{R}^{(m+1)\times d}$$
$$\forall \boldsymbol{x} \in \mathbb{R}^d,\ \forall \boldsymbol{w} \in (0,1)^m,\ \forall \xi \in \mathbb{R}.$$

Therefore, it follows from the chain rule of differentiation, the infinite differentiability of $\boldsymbol{F}$, the continuity of $\widehat{\boldsymbol{w}}(\,\cdot\,;\theta)$, $\widehat{\xi}(\,\cdot\,;\theta)$, and an inductive argument that $\nabla\widehat{\boldsymbol{w}}(\,\cdot\,;\theta)$ is infinitely differentiable. Since $\nabla\widehat{\varphi}_{\mathrm{barr}}(\boldsymbol{x};\theta) = \widehat{T}_{\mathrm{barr}}(\boldsymbol{x};\theta) = \widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) + \underline{\lambda}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^d$, it holds that $\widehat{\varphi}_{\mathrm{barr}}(\,\cdot\,;\theta) \in \mathcal{C}^{\infty}(\mathbb{R}^d)$. The proof of statement (i) is now complete.

Next, we fix an arbitrary $\boldsymbol{x} \in \mathbb{R}^d$ and prove statement (ii). Since $\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) \in \Delta$, it holds that $\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) \in \operatorname{conv}\big(\{\widetilde{\boldsymbol{g}}_1^{\star},\dots,\widetilde{\boldsymbol{g}}_n^{\star}\}\big)$. Hence, using (4.15) and the assumption $\underline{\lambda} \le 1$, we get

$$\begin{aligned}\big\|\widehat{T}_{\mathrm{barr}}(\boldsymbol{x};\theta)\big\|^2 = \big\|\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) + \underline{\lambda}\boldsymbol{x}\big\|^2 &\le 2\big\|\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\big\|^2 + 2\underline{\lambda}^2\|\boldsymbol{x}\|^2\\ &\le 2\max_{1\le i\le m}\big\{\|\widetilde{\boldsymbol{g}}_i^{\star}\|\big\}^2 + 2\|\boldsymbol{x}\|^2\\ &\le 18\overline{u}_0(\nu)^2 + 2\|\boldsymbol{x}\|^2 \qquad \mathbb{P}\text{-a.s.}\end{aligned}$$

This proves statement (ii).

Now, let us turn to the proof of statement (iii). Let us fix an arbitrary $\boldsymbol{x} \in \mathbb{R}^d$ and let $\varphi^{(a)}_{\mathrm{LB}}(\cdot)$, $\varphi^{(b)}_{\mathrm{UB}}(\cdot)$, $\varphi^{(c)}_{\mathrm{UB}}(\cdot)$, $\varphi^{(d)}_{\mathrm{LB}}(\cdot)$, and $\varphi^{(e)}_{\mathrm{LB}}(\cdot)$ be defined as in the proof of Theorem 4.7. We will focus on the maximization problem characterizing $\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})$:

$$p^{\star} := \varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x}) = \sup_{\boldsymbol{w}\in\Delta}\Big\{\langle \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} + \boldsymbol{v}, \boldsymbol{w}\rangle - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\mathbf{G}}^{\star}\boldsymbol{w}\|^2\Big\}. \tag{4.20}$$

By introducing the Lagrange multiplier variables $\boldsymbol{\zeta} \in \mathbb{R}^m_+$ and $\xi \in \mathbb{R}$, the Lagrangian $L : \mathbb{R}^m \times \mathbb{R}^m_+ \times \mathbb{R} \to \mathbb{R}$ associated with (4.20) is given by:

$$\begin{aligned}L(\boldsymbol{w},\boldsymbol{\zeta},\xi) := \langle \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} + \boldsymbol{v}, \boldsymbol{w}\rangle - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\mathbf{G}}^{\star}\boldsymbol{w}\|^2 + \langle \boldsymbol{\zeta}, \boldsymbol{w}\rangle + \xi(\langle \mathbf{1}_m, \boldsymbol{w}\rangle - 1)&\\ \forall \boldsymbol{w}\in\mathbb{R}^m,\ \forall\boldsymbol{\zeta}\in\mathbb{R}^m_+,\ \forall\xi\in\mathbb{R}.&\end{aligned}$$

Let $p : \mathbb{R}^m \to \mathbb{R}\cup\{-\infty\}$ and $q : \mathbb{R}^m_+ \times \mathbb{R} \to \mathbb{R}$ be defined as follows:

$$\begin{aligned} p(\boldsymbol{w}) &:= \inf_{\boldsymbol{\zeta}\in\mathbb{R}^m_+,\,\xi\in\mathbb{R}}\big\{L(\boldsymbol{w},\boldsymbol{\zeta},\xi)\big\} && \forall\boldsymbol{w}\in\mathbb{R}^m,\\ q(\boldsymbol{\zeta},\xi) &:= \sup_{\boldsymbol{w}\in\mathbb{R}^m}\big\{L(\boldsymbol{w},\boldsymbol{\zeta},\xi)\big\} && \forall\boldsymbol{\zeta}\in\mathbb{R}^m_+,\ \forall\xi\in\mathbb{R}.\end{aligned}$$

In particular, we have $p^{\star} = \sup_{\boldsymbol{w}\in\mathbb{R}^m}\big\{p(\boldsymbol{w})\big\}$ and

$$q(\boldsymbol{\zeta},\xi) \ge \sup_{\boldsymbol{w}\in\mathbb{R}^m}\big\{p(\boldsymbol{w})\big\} = p^{\star} \qquad \forall\boldsymbol{\zeta}\in\mathbb{R}^m_+,\ \forall\xi\in\mathbb{R}. \tag{4.21}$$

Moreover, let $\widehat{\xi}(\boldsymbol{x};\theta)$ be defined as in the proof of statement (i). Recall that we have shown by the KKT optimality conditions in (4.19) that

$$\begin{aligned} -\tfrac{1}{\theta}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)^{\circ-1} + \tfrac{1}{\overline{\lambda}-\underline{\lambda}}\widetilde{\mathbf{G}}^{\star\mathsf{T}}\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) - \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} - \boldsymbol{v} - \widehat{\xi}(\boldsymbol{x};\theta)\mathbf{1}_m = \mathbf{0}_m,&\\ \langle \mathbf{1}_m, \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\rangle - 1 = 0.&\end{aligned} \tag{4.22}$$

Subsequently, let $\bar{\boldsymbol{\zeta}} := \frac{1}{\theta}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)^{\circ-1} \in \mathbb{R}^m_+$ and let $\bar{\xi} := \widehat{\xi}(\boldsymbol{x};\theta) \in \mathbb{R}$. Let $\nabla_{\boldsymbol{w}}L(\boldsymbol{w},\boldsymbol{\zeta},\xi)$ denote the partial derivative of L with respect to $\boldsymbol{w}$. Observe that (4.22) implies that

$$\nabla_{\boldsymbol{w}}L\big(\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta),\bar{\boldsymbol{\zeta}},\bar{\xi}\big) = \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} + \boldsymbol{v} - \tfrac{1}{\overline{\lambda}-\underline{\lambda}}\widetilde{\mathbf{G}}^{\star\mathsf{T}}\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) + \tfrac{1}{\theta}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)^{\circ-1} + \widehat{\xi}(\boldsymbol{x};\theta)\mathbf{1}_m = \mathbf{0}_m. \tag{4.23}$$

Since $L(\,\cdot\,,\bar{\boldsymbol{\zeta}},\bar{\xi})$ is concave, (4.23) shows that $\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) \in \arg\max_{\boldsymbol{w}\in\mathbb{R}^d}\big\{L(\boldsymbol{w},\bar{\boldsymbol{\zeta}},\bar{\xi})\big\}$. Thus, we have by (4.21) and (4.22) that

$$\begin{aligned} p^{\star} &\le q(\bar{\boldsymbol{\zeta}},\bar{\xi})\\ &= L\big(\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta),\bar{\boldsymbol{\zeta}},\bar{\xi}\big)\\ &= \langle \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} + \boldsymbol{v}, \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\rangle - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\big\|\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\big\|^2 + \tfrac{1}{\theta}\langle \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)^{\circ-1}, \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\rangle\\ &\qquad + \widehat{\xi}(\boldsymbol{x};\theta)(\langle \mathbf{1}_m, \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\rangle - 1)\\ &= \langle \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} + \boldsymbol{v}, \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\rangle - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\big\|\widetilde{\mathbf{G}}^{\star}\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\big\|^2 + \tfrac{m}{\theta}.\end{aligned} \tag{4.24}$$

The definition of $p^{\star}$ in (4.20) then shows that $\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)$ is a $\frac{m}{\theta}$-optimal solution of (4.20). Next, observe that

$$p^{\star} = \varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x}) = \sup_{\widetilde{\boldsymbol{g}}\in\mathbb{R}^d}\Big\{\langle \widetilde{\boldsymbol{g}}, \boldsymbol{x}\rangle - \varphi^{(c)}_{\mathrm{UB}}(\widetilde{\boldsymbol{g}})\Big\} = \sup_{\widetilde{\boldsymbol{g}}\in\mathbb{R}^d}\Big\{\langle \widetilde{\boldsymbol{g}}, \boldsymbol{x}\rangle - \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}\|^2 - \varphi^{(b)}_{\mathrm{UB}}(\widetilde{\boldsymbol{g}})\Big\}. \tag{4.25}$$

Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be defined by $h(\widetilde{\boldsymbol{g}}) := \frac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}\|^2 - \langle \widetilde{\boldsymbol{g}}, \boldsymbol{x}\rangle + \varphi^{(b)}_{\mathrm{LB}}(\widetilde{\boldsymbol{g}})$ for all $\widetilde{\boldsymbol{g}} \in \mathbb{R}^d$. We hence have $h \in \mathfrak{C}_{\frac{1}{\overline{\lambda}-\underline{\lambda}},\infty}(\mathbb{R}^d)$. It follows from (4.25) that $\nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})$ is the unique minimizer of h. We can subsequently derive from (4.25) and (4.24) that

$$\begin{aligned} h\big(\nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})\big) &= \inf_{\widetilde{\boldsymbol{g}}\in\mathbb{R}^d} \Big\{ \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}\|^2 - \langle \widetilde{\boldsymbol{g}}, \boldsymbol{x}\rangle + \varphi^{(b)}_{\mathrm{UB}}(\widetilde{\boldsymbol{g}}) \Big\} \\ &= -p^\star \\ &\geq \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\big\|\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\big\|^2 - \langle \widetilde{\mathbf{G}}^{\star\mathsf{T}}\boldsymbol{x} + \boldsymbol{v}, \widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\rangle - \tfrac{m}{\theta} \\ &\geq \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\big\|\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\big\|^2 - \langle \widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta), \boldsymbol{x}\rangle - \sup_{\boldsymbol{w}\in\Delta,\, \widetilde{\mathbf{G}}^\star\boldsymbol{w}=\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)} \big\{\langle \boldsymbol{v}, \boldsymbol{w}\rangle\big\} - \tfrac{m}{\theta} \\ &= \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\big\|\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\big\|^2 - \langle \widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta), \boldsymbol{x}\rangle + \varphi^{(b)}_{\mathrm{UB}}\big(\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\big) - \tfrac{m}{\theta} \\ &= h\big(\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\big) - \tfrac{m}{\theta}. \end{aligned} \tag{4.26}$$

Moreover, it follows from the first-order optimality condition that $\mathbf{0} \in \partial h\big(\nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})\big)$, and hence the $\frac{1}{\overline{\lambda}-\underline{\lambda}}$-strong convexity of h implies that

$$\begin{aligned} h\big(\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta)\big) &\geq h\big(\nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})\big) + \big\langle \mathbf{0}, \widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) - \nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})\big\rangle \\ &\qquad + \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\big\|\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) - \nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})\big\|^2 \\ &= h\big(\nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})\big) + \tfrac{1}{2(\overline{\lambda}-\underline{\lambda})}\big\|\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) - \nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})\big\|^2. \end{aligned} \tag{4.27}$$

Subsequently, combining (4.26) and (4.27) yields $\big\|\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) - \nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})\big\|^2 \leq \frac{2m(\overline{\lambda}-\underline{\lambda})}{\theta}$. Lastly, recall from (4.9) that $\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x}) = \underline{\lambda}\boldsymbol{x} + \nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})$. It hence follows from (4.17) that

$$\begin{aligned} \big\|\widehat{T}_{\mathrm{barr}}(\boldsymbol{x};\theta) - \widehat{T}_{\text{CLS-LB}}(\boldsymbol{x})\big\|^2 &= \big\|\big(\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) + \underline{\lambda}\boldsymbol{x}\big) - \big(\nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x}) + \underline{\lambda}\boldsymbol{x}\big)\big\|^2 \\ &= \big\|\widetilde{\mathbf{G}}^\star\widehat{\boldsymbol{w}}(\boldsymbol{x};\theta) - \nabla\varphi^{(d)}_{\mathrm{LB}}(\boldsymbol{x})\big\|^2 \\ &\leq \frac{2m(\overline{\lambda}-\underline{\lambda})}{\theta}. \end{aligned}$$

This completes the proof of statement (iii).

Finally, observe that statement (i) shows that $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta)$ satisfies the shape condition in Assumption 3.4(i). Moreover, statement (ii) implies that $\mathbb{E}\Big[\big\|\widehat{T}_{\mathrm{barr}}(\boldsymbol{x};\theta)\big\|^2\Big] \leq 18\overline{u}_0(\nu)^2 + 2\|\boldsymbol{x}\|^2$ for all $\boldsymbol{x}$, which shows that $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta)$ satisfies the growth condition in Assumption 3.4(ii) with respect to $u_1(\nu) := 18\overline{u}_0(\nu)^2$, $u_2(\nu) := 2$. Furthermore, statement (iii) shows that

$$\begin{aligned} \big\|\widehat{T}_{\mathrm{barr}}(\boldsymbol{x};\theta) - T^\mu_\nu(\boldsymbol{x})\big\|^2 &\leq 2\big\|\widehat{T}_{\mathrm{barr}}(\boldsymbol{x};\theta) - \widehat{T}_{\text{CLS-LB}}(\boldsymbol{x})\big\|^2 + 2\big\|\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x}) - T^\mu_\nu(\boldsymbol{x})\big\|^2 \\ &\leq \frac{4m(\overline{\lambda}(\mu,\nu) - \underline{\lambda}(\mu,\nu))}{\theta} + 2\big\|\widehat{T}_{\text{CLS-LB}}(\boldsymbol{x}) - T^\mu_\nu(\boldsymbol{x})\big\|^2 \qquad \forall \boldsymbol{x} \in \mathbb{R}^d. \end{aligned}$$

Combining this with Theorem 4.7, Theorem 4.2, and Theorem 4.5 leads to

$$\begin{aligned} \mathbb{E}\Big[\big\|\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta) - T^\mu_\nu\big\|^2_{\mathcal{L}^2(\mu)}\Big] &\leq \frac{4m(\overline{\lambda}(\mu,\nu) - \underline{\lambda}(\mu,\nu))}{\theta} \\ &\qquad + 2C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log(m)^2\kappa\big(\min\{m,n\}\big). \end{aligned}$$

Consequently, for any $\epsilon > 0$, with $\overline{n}(\mu,\nu,\epsilon) := \min\big\{n \in \mathbb{N} : C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log(n)^2\kappa(n) \leq \frac{\epsilon}{4}\big\}$ and $\overline{\theta}(\mu,\nu,m,n,\epsilon) := \big\lceil \frac{8m}{\epsilon}(\overline{\lambda}(\mu,\nu) - \underline{\lambda}(\mu,\nu))\big\rceil$, it holds for all $m \geq \overline{n}(\mu,\nu,\epsilon)$, $n \geq \overline{n}(\mu,\nu,\epsilon)$ and all

$\theta \geq \overline{\theta}(\mu,\nu,m,n,\epsilon)$ that

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta) - T^{\mu}_{\nu}\big\|^2_{\mathcal{L}^2(\mu)}\Big]\\
&\quad\leq \frac{4m(\overline{\lambda}(\mu,\nu) - \underline{\lambda}(\mu,\nu))}{\theta} + 2C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log(m)^2\kappa\big(\min\{m,n\}\big)\\
&\quad\leq \frac{4m(\overline{\lambda}(\mu,\nu) - \underline{\lambda}(\mu,\nu))}{\frac{8}{\epsilon}m(\overline{\lambda}(\mu,\nu) - \underline{\lambda}(\mu,\nu))} + 2C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log\big(\overline{n}(\mu,\nu,\epsilon)\big)^2\kappa\big(\overline{n}(\mu,\nu,\epsilon)\big)\\
&\quad\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}
$$

We have thus shown that $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta)$ satisfies the consistency condition in Assumption 3.4(iii) with respect to $\overline{n}(\mu,\nu,\epsilon) := \min\big\{n \in \mathbb{N} : C\big(\mu,\nu,\underline{\lambda}(\mu,\nu),\overline{\lambda}(\mu,\nu)\big)\log(n)^2\kappa(n) \leq \frac{\epsilon}{4}\big\}$ and $\overline{\theta}(\mu,\nu,m,n,\epsilon) := \big\lceil\frac{8m}{\epsilon}(\overline{\lambda}(\mu,\nu) - \underline{\lambda}(\mu,\nu))\big\rceil$. The proof is now complete. □

Remark 4.11 (Computational tractability of $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta)$)**.** *Same as $\widehat{T}_{\mathrm{kern}}(\,\cdot\,;\theta)$ discussed in Remark 4.9, the computation of $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta)$ is also done in two phases. In the first phase, given the m samples $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_m$ from μ and the n samples $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n$ from ν, one can compute $(\widetilde{\varphi}^{\star}_i)_{i=1:m}$ and $(\widetilde{\boldsymbol{g}}^{\star}_i)_{i=1:m}$ via state-of-the-art LP and QCQP solvers such as Gurobi [43], as discussed in Remark 4.9. Subsequently, in the second phase, for every $\boldsymbol{x} \in \mathbb{R}^d$ at which $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta)$ needs to be evaluated, one solves the convex optimization problem in (4.17). This can be numerically implemented by, for example, Newton's method with equality constraints; see, e.g., [17, Section 10.2]. Consequently, using $\widehat{T}_{\mathrm{barr}}(\,\cdot\,;\theta)$ as the plug-in OT map estimator in Algorithm 2 results in a computationally tractable algorithm for $\mathcal{W}_2$-barycenter that is also provably convergent.*

5. Numerical experiments

[To be done]

Appendix A. Discussion and proof related to measurability issues

[To be discussed]

Appendix B. A Paralleled implementation for soving (4.2) via ADMM

As mentioned in Remark (4.4), we provide the setting and the implementation details of our paralleled algorithm for solving the quadratically constrained quadratic program (QCQP) in (4.2) via the alternating direction method of multipliers (ADMM), adapted from Simonetto [80, Section 3.C].

Let us consider the complete digraph $\mathcal{G} := (\mathcal{V},\mathcal{E})$ where $\mathcal{V} := \{1,2,\ldots,m\}$ and $\mathcal{E} := \{e(i \to j) : i,j \in \mathcal{V}, i \neq j\}$. For any $e(i \to j) \in \mathcal{E}$, we refer i as the *source node* and j as the *target node*. For $i = 1,\ldots,m$, besides the existing node variables $\widetilde{\varphi}_i$ and $\widetilde{\boldsymbol{g}}_i$ in (4.2), we additionally assign on each edge $e(i \to j) \in \mathcal{E}$ decision variables $\widetilde{\varphi}^{(s)}_{i,j} \in \mathbb{R}$ and $\widetilde{\boldsymbol{g}}^{(s)}_{i,j} \in \mathbb{R}^d$. Meanwhile, when i serves as a target node, we correspondingly assign $\widetilde{\varphi}^{(t)}_{i,j} \in \mathbb{R}$ and $\widetilde{\boldsymbol{g}}^{(t)}_{i,j} \in \mathbb{R}^d$ for each $e(j \to i) \in \mathcal{E}$. We are therefore able to define, for each $e(i \to j) \in \mathcal{E}$, a concatenated decision variable $\boldsymbol{\xi}_{i\to j} := \big(\widetilde{\varphi}^{(s)}_{i,j}, \widetilde{\boldsymbol{g}}^{(s)\mathsf{T}}_{i,j}, \widetilde{\varphi}^{(t)}_{j,i}, \widetilde{\boldsymbol{g}}^{(t)\mathsf{T}}_{j,i}\big)^{\mathsf{T}} \in \mathbb{R}^{2d+2}$.

We now show that the original QCQP in (4.2) can now be reformulated in a separable form in cater for the application of ADMM. Indeed, by denoting $\mathcal{C}_{i\to j} := \big\{\boldsymbol{\xi}_{i\to j} : \widetilde{\varphi}^{(t)}_{j.i} \geq \widetilde{\varphi}^{(s)}_{i,j} + \langle\widetilde{\boldsymbol{g}}^{(s)}_{i,j}, \boldsymbol{X}_j - \boldsymbol{X}_i\rangle + \frac{1}{2(\overline{\lambda}-\underline{\lambda})}\|\widetilde{\boldsymbol{g}}^{(s)}_{i,j} - \widetilde{\boldsymbol{g}}^{(t)}_{j,i}\|^2\big\}$ whenever $i \neq j$ and denoting $\delta_{\mathcal{S}}$ as the indicator function for $\mathcal{S} \subseteq \mathbb{R}^d$, one can rewrite (4.2) in the following equivalent form:

$$
\begin{aligned}
\underset{(\widetilde{\varphi}_i),\,(\widetilde{\boldsymbol{g}}_i),\,(\boldsymbol{\xi}_{i\to j})}{\text{minimize}}\quad & \sum_{i=1}^{m}\sum_{j=1}^{m}\widehat{\pi}^{\star}_{i,j}\big\|\widetilde{\boldsymbol{g}}_i + \underline{\lambda}\boldsymbol{X}_i - \boldsymbol{Y}_j\big\|^2 + \sum_{i=1}^{m}\delta_{\bar{B}(-\underline{\lambda}\boldsymbol{X}_i,\overline{u}_0(\nu))}(\widetilde{\boldsymbol{g}}_i) + \sum_{i=1}^{m}\sum_{j=1}^{m}\delta_{\mathcal{C}_{i\to j}}(\boldsymbol{\xi}_{i\to j})\\
\text{subject to}\quad & \widetilde{\varphi}^{(s)}_{i,j} = \widetilde{\varphi}^{(t)}_{i,j} = \widetilde{\varphi}_i \quad \forall i = 1:m,\ \forall j = 1:m,\ i \neq j\\
& \widetilde{\boldsymbol{g}}^{(s)}_{i,j} = \widetilde{\boldsymbol{g}}^{(t)}_{i,j} = \widetilde{\boldsymbol{g}}_i \quad \forall i = 1:m,\ \forall j = 1:m,\ i \neq j
\end{aligned}
\tag{B.1}
$$

Since (B.1) contains a separable objective function and equality constraints, it can be tackled by ADMM; see, e.g., [16, Section 3]. For every $e(i \to j) \in \mathcal{E}$, we define the dual scaled variables $u_{i,j}^{(s)}, u_{j,i}^{(t)} \in \mathbb{R}$ and $\boldsymbol{v}_{i,j}^{(s)}, \boldsymbol{v}_{j,i}^{(t)} \in \mathbb{R}^d$, and denote $\boldsymbol{\vartheta}_{i\to j} := \big(u_{i,j}^{(s)}, \boldsymbol{v}_{i,j}^{(s)\mathsf{T}}, u_{j,i}^{(t)}, \boldsymbol{v}_{j,i}^{(t)\mathsf{T}}\big) \in \mathbb{R}^{2d+2}$. The associated augmented Lagrangian of (B.1) with a penalty term $\rho > 0$ is formed as

$$
\begin{aligned}
L_\rho\big((\widetilde{\varphi}_i), (\widetilde{\boldsymbol{g}}_i), (\boldsymbol{\xi}_{i\to j}), (\boldsymbol{\vartheta}_{i\to j})\big) :=& \sum_{i=1}^m \sum_{j=1}^m \widehat{\pi}_{i,j}^\star \big\|\widetilde{\boldsymbol{g}}_i + \underline{\lambda}\boldsymbol{X}_i - \boldsymbol{Y}_j\big\|^2 + \sum_{i=1}^m \delta_{\bar{B}(-\underline{\lambda}\boldsymbol{X}_i, \overline{u}_0(\nu))}(\widetilde{\boldsymbol{g}}_i) \\
&+ \sum_{i=1}^m \sum_{j=1, j\neq i}^m \delta_{\mathcal{C}_{i\to j}}(\boldsymbol{\xi}_{i\to j}) \\
&+ \frac{\rho}{2} \sum_{i=1}^m \sum_{j=1, j\neq i}^m \Big[\big\|\widetilde{\boldsymbol{g}}_{i,j}^{(s)} - \widetilde{\boldsymbol{g}}_i + \boldsymbol{v}_{i,j}^{(s)}\big\|^2 + \big\|\widetilde{\boldsymbol{g}}_{j,i}^{(t)} - \widetilde{\boldsymbol{g}}_j + \boldsymbol{v}_{j,i}^{(t)}\big\|^2 \\
&+ \big(\widetilde{\varphi}_{i,j}^{(s)} - \widetilde{\varphi}_i + u_{i,j}^{(s)}\big)^2 + \big(\widetilde{\varphi}_{j,i}^{(t)} - \widetilde{\varphi}_j + u_{j,i}^{(t)}\big)^2 \\
&- \big\|\boldsymbol{v}_{i,j}^{(s)}\big\|^2 - \big\|\boldsymbol{v}_{j,i}^{(t)}\big\|^2 - \big(u_{i,j}^{(s)}\big)^2 - \big(u_{j,i}^{(t)}\big)^2\Big]
\end{aligned}
$$

As such, with certain initializations $(\widetilde{\varphi}_i^0), (\widetilde{\boldsymbol{g}}_i^0), (\boldsymbol{\xi}_{i\to j}^0), (\boldsymbol{\vartheta}_{i\to j}^0)$, the decomposable blockwise ADMM updating the k-th iteration can be performed in the following procedure:

(i) **(node variables update)** We solve for each $i = 1, \dots, m$:

$$
\begin{aligned}
\widetilde{\boldsymbol{g}}_i^k := \mathop{\arg\min}_{\boldsymbol{g}\in \bar{B}(-\underline{\lambda}\boldsymbol{X}_i, \overline{u}_0(\nu))} & \sum_{j=1}^m \widehat{\pi}_{ij}^\star \big\|\boldsymbol{g} + \underline{\lambda}\boldsymbol{X}_i - \boldsymbol{Y}_j\big\|^2 + \frac{\rho}{2} \sum_{j=1, j\neq i}^m \Big(\big\|\widetilde{\boldsymbol{g}}_{i,j}^{(s),k-1} + \boldsymbol{v}_{i,j}^{(s),k-1} - \boldsymbol{g}\big\|^2 \\
& + \big\|\widetilde{\boldsymbol{g}}_{i,j}^{(t),k-1} + \boldsymbol{v}_{i,j}^{(t),k-1} - \boldsymbol{g}\big\|^2\Big).
\end{aligned} \tag{B.2}
$$

$$
\widetilde{\varphi}_i^k := \mathop{\arg\min}_{\varphi\in\mathbb{R}} \sum_{j=1, j\neq i}^m \Big(\big(\widetilde{\varphi}_{i,j}^{(s),k-1} + u_{i,j}^{(s),k-1} - \varphi\big)^2 + \big(\widetilde{\varphi}_{i,j}^{(t),k-1} + u_{i,j}^{(t),k-1} - \varphi\big)^2\Big) \tag{B.3}
$$

(ii) **(edge variables update)** We solve for each $e(i \to j) \in \mathcal{E}$:

$$
\begin{aligned}
\boldsymbol{\xi}_{i\to j}^k := \mathop{\arg\min}_{\boldsymbol{\xi}\in\mathcal{C}_{i\to j}} \Big(& \big\|\big(\ \boldsymbol{I}_{d+1} \ \big| \ \mathit{0}_{(d+1)\times(d+1)} \ \big)\boldsymbol{\xi} - (\widetilde{\varphi}_i^{k-1}, \widetilde{\boldsymbol{g}}_i^{k-1\,\mathsf{T}})^\mathsf{T} + (u_{i,j}^{(s),k-1}, \boldsymbol{v}_{i,j}^{(s),k-1\,\mathsf{T}})^\mathsf{T}\big\|^2 \\
& + \big\|\big(\ \mathit{0}_{(d+1)\times(d+1)} \ \big| \ \boldsymbol{I}_{d+1} \ \big)\boldsymbol{\xi} - (\widetilde{\varphi}_j^{k-1}, \widetilde{\boldsymbol{g}}_j^{k-1\,\mathsf{T}})^\mathsf{T} + (u_{j,i}^{(t),k-1}, \boldsymbol{v}_{j,i}^{(t),k-1\,\mathsf{T}})^\mathsf{T}\big\|^2\Big)
\end{aligned} \tag{B.4}
$$

where $\mathit{0}_{(d+1)\times(d+1)}$ denotes the all-zero $(d+1)\times(d+1)$ square matrix.

(iii) **(dual variables update)** We compute for $i, j = 1, \dots, m$ with $i \neq j$:

$$
u_{i,j}^{(s),k} := u_{i,j}^{(s),k-1} + \widetilde{\varphi}_{i,j}^{(s),k} - \widetilde{\varphi}_i^k, \qquad \boldsymbol{v}_{i,j}^{(s),k} := \boldsymbol{v}_{i,j}^{(s),k-1} + \widetilde{\boldsymbol{g}}_{i,j}^{(s),k-1} - \widetilde{\boldsymbol{g}}_i^{k-1}. \tag{B.5}
$$

$$
u_{i,j}^{(t),k} := u_{i,j}^{(t),k-1} + \widetilde{\varphi}_{i,j}^{(t),k} - \widetilde{\varphi}_i^k, \qquad \boldsymbol{v}_{i,j}^{(t),k} := \boldsymbol{v}_{i,j}^{(t),k-1} + \widetilde{\boldsymbol{g}}_{i,j}^{(t),k-1} - \widetilde{\boldsymbol{g}}_i^{k-1}. \tag{B.6}
$$

We now delve into each block and solve the optimization problems therein for alternating updates on nodes and edges. One can observe that both (B.2) and (B.3) attain closed-form solutions. Indeed, we have

$$
\begin{aligned}
\widetilde{\boldsymbol{g}}_i^k &= \mathop{\arg\min}_{\boldsymbol{g}\in \bar{B}(-\underline{\lambda}\boldsymbol{X}_i, \overline{u}_0(\nu))} \ \big[\tfrac{1}{m} + \rho(m-1)\big]\|\boldsymbol{g}\|^2 + 2\textstyle\sum_{j=1}^m \big[\widehat{\pi}_{i,j}^\star(\underline{\lambda}\boldsymbol{X}_i - \boldsymbol{Y}_j)\big]^\mathsf{T}\boldsymbol{g} \\
&\qquad\qquad - \rho\textstyle\sum_{j=1, j\neq i}^m \big(\widetilde{\boldsymbol{g}}_{i,j}^{(s),k-1} + \boldsymbol{v}_{i,j}^{(s),k-1} + \widetilde{\boldsymbol{g}}_{i,j}^{(t),k-1} + \boldsymbol{v}_{i,j}^{(t),k-1}\big)^\mathsf{T}\boldsymbol{g} \\
&= \mathop{\arg\min}_{\boldsymbol{g}\in \bar{B}(-\underline{\lambda}\boldsymbol{X}_i, \overline{u}_0(\nu))} \ \|\boldsymbol{g} - \boldsymbol{z}_i\|^2
\end{aligned} \tag{B.7}
$$

where

$$
\boldsymbol{z}_i := -\frac{1}{\frac{1}{m} + \rho(m-1)}\Big[\textstyle\sum_{j=1}^m \widehat{\pi}_{i,j}^\star(\underline{\lambda}\boldsymbol{X}_i - \boldsymbol{Y}_j) - \frac{\rho}{2}\sum_{j=1, j\neq i}^m \big(\widetilde{\boldsymbol{g}}_{i,j}^{(s),k-1} + \boldsymbol{v}_{i,j}^{(s),k-1} + \widetilde{\boldsymbol{g}}_{i,j}^{(t),k-1} + \boldsymbol{v}_{i,j}^{(t),k-1}\big)\Big].
$$

Therefore,

$$\widetilde{\boldsymbol{g}}_i^k = \begin{cases} \boldsymbol{z}_i, & \boldsymbol{z}_i \in \bar{B}(-\underline{\lambda}\boldsymbol{X}_i, \overline{u}_0(\nu)) \\ -\underline{\lambda}\boldsymbol{X}_i + \frac{\overline{u}_0(\nu)}{\|\boldsymbol{z}_i + \underline{\lambda}\boldsymbol{X}_i\|}(\boldsymbol{z}_i + \underline{\lambda}\boldsymbol{X}_i), & \boldsymbol{z}_i \notin \bar{B}(-\underline{\lambda}\boldsymbol{X}_i, \overline{u}_0(\nu)) \end{cases} \tag{B.8}$$

In terms of (B.3), it is a standard least square problem, thus the solution is

$$\widetilde{\varphi}_i^k = \frac{1}{2m-2} \sum_{j=1, j\neq i}^{m} \left(\widetilde{\varphi}_{i,j}^{(s),k-1} + u_{i,j}^{(s),k-1} + \widetilde{\varphi}_{i,j}^{(t),k-1} + u_{i,j}^{(t),k-1} \right) \tag{B.9}$$

It is left to solve (B.4) which is itself a QCQP with a single scalar constraint. To this end, we rewrite it in the following standard from of QCQP as elucidated in [80]:

$$\begin{aligned} \underset{\boldsymbol{\xi} \in \mathbb{R}^{2d+2}}{\text{minimize}} \quad & \boldsymbol{\xi}^\mathsf{T}\boldsymbol{\xi} + \boldsymbol{q}_0^\mathsf{T}\boldsymbol{\xi} \\ \text{subject to} \quad & \boldsymbol{\xi}^\mathsf{T}\boldsymbol{Q}\boldsymbol{\xi} + \boldsymbol{q}_1^\mathsf{T}\boldsymbol{\xi} \leq 0 \end{aligned} \tag{B.10}$$

where

$$\boldsymbol{q}_0 := 2\left(u_{i,j}^{(s),k-1} - \widetilde{\varphi}_i^{k-1}, \boldsymbol{v}_{i,j}^{(s),k-1\,\mathsf{T}} - \widetilde{\boldsymbol{g}}_i^{k-1\,\mathsf{T}}, u_{j,i}^{(t),k-1} - \widetilde{\varphi}_j^{k-1}, \boldsymbol{v}_{j,i}^{(t),k-1\,\mathsf{T}} - \widetilde{\boldsymbol{g}}_j^{k-1\,\mathsf{T}}\right)^\mathsf{T} \in \mathbb{R}^{2d+2}$$

$$\boldsymbol{Q} := \frac{1}{2(\overline{\lambda} - \underline{\lambda})} \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & \cdots & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & \cdots & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & \cdots & 0 & 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{(2d+2)\times(2d+2)}$$

$$\boldsymbol{q}_1 := \left(1, (\boldsymbol{X}_j - \boldsymbol{X}_i)^\mathsf{T}, -1, \boldsymbol{0}_d^\mathsf{T}\right)^\mathsf{T} \in \mathbb{R}^{2d+2}$$

With the dual variable $\gamma \geq 0$, the Lagrange dual function associated to (B.10) is

$$\begin{aligned} \phi(\gamma) &:= \inf_{\boldsymbol{\xi}} \boldsymbol{L}(\boldsymbol{\xi}, \gamma) = \min_{\boldsymbol{\xi}} \left\{ \boldsymbol{\xi}^\mathsf{T}\boldsymbol{\xi} + \boldsymbol{q}_0^\mathsf{T}\boldsymbol{\xi} + \gamma(\boldsymbol{\xi}^\mathsf{T}\boldsymbol{Q}\boldsymbol{\xi} + \boldsymbol{q}_1^\mathsf{T}\boldsymbol{\xi}) \right\} \\ &= -\frac{1}{4}(\boldsymbol{q}_0 + \gamma\boldsymbol{q}_1)^\mathsf{T}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}(\boldsymbol{q}_0 + \gamma\boldsymbol{q}_1). \end{aligned} \tag{B.11}$$

Defining $\gamma^\star := \arg\max_{\gamma \geq 0} \phi(\gamma)$, one can retrieve $\boldsymbol{\xi}_{i\to j}^k = -\frac{1}{2}(\boldsymbol{I} + \gamma^\star\boldsymbol{Q})^{-1}(\boldsymbol{q}_0 + \gamma^\star\boldsymbol{q}_1)$.

We apply Newton's method to solve $\gamma^\star$. The gradient and the hessian of ϕ in (B.11) with respect to γ is evaluated as follows:

$$\nabla\phi(\gamma) = -\tfrac{1}{2}\boldsymbol{q}_1^\mathsf{T}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}(\boldsymbol{q}_0 + \gamma\boldsymbol{q}_1) + \tfrac{1}{4}(\boldsymbol{q}_0 + \gamma\boldsymbol{q}_1)^\mathsf{T}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}\boldsymbol{Q}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}(\boldsymbol{q}_0 + \gamma\boldsymbol{q}_1) \tag{B.12}$$

$$\begin{aligned} \nabla^2\phi(\gamma) = &-\tfrac{1}{2}\boldsymbol{q}_1^\mathsf{T}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}\boldsymbol{q}_1 + \boldsymbol{q}_1^\mathsf{T}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}\boldsymbol{Q}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}(\boldsymbol{q}_0 + \gamma\boldsymbol{q}_1) \\ &- \tfrac{1}{2}(\boldsymbol{q}_0 + \gamma\boldsymbol{q}_1)^\mathsf{T}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}\boldsymbol{Q}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}\boldsymbol{Q}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}(\boldsymbol{q}_0 + \gamma\boldsymbol{q}_1) \end{aligned} \tag{B.13}$$

Specifically, we consider the eigen-decomposition $\boldsymbol{Q} = \boldsymbol{U}^\mathsf{T}\boldsymbol{D}\boldsymbol{U}$ where $\boldsymbol{D}$ is the diagonal matrix containing eigenvalues. Then, we have

$$\begin{aligned} (\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1} &= \gamma^{-1}\boldsymbol{U}^\mathsf{T}(\gamma^{-1}\boldsymbol{I} + \boldsymbol{D})^{-1}\boldsymbol{U}, \\ (\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}\boldsymbol{Q}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1} &= \gamma^{-2}\boldsymbol{U}^\mathsf{T}(\gamma^{-1}\boldsymbol{I} + \boldsymbol{D})^{-2}\boldsymbol{D}\boldsymbol{U}, \\ (\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}\boldsymbol{Q}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1}\boldsymbol{Q}(\boldsymbol{I} + \gamma\boldsymbol{Q})^{-1} &= \gamma^{-3}\boldsymbol{U}^\mathsf{T}(\gamma^{-1}\boldsymbol{I} + \boldsymbol{D})^{-3}\boldsymbol{D}^2\boldsymbol{U}. \end{aligned} \tag{B.14}$$

Appendix C. Generation details of synthetic input measures

References

[1] M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] J. M. Altschuler and E. Boix-Adsera. Wasserstein barycenters can be computed in polynomial time in fixed dimension. *Journal of Machine Learning Research*, 22(44):1–19, 2021.

[3] J. M. Altschuler and E. Boix-Adsera. Wasserstein barycenters are np-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.

[4] J. M. Altschuler and E. Boix-Adserá. Polynomial-time algorithms for multimarginal optimal transport problems with structure. *Mathematical Programming*, 199:1107–1178, 2023. doi: 10.1007/s10107-022-01868-7.

[5] P. C. Álvarez-Esteban, E. Del Barrio, J. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[6] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[7] E. Anderes, S. Borgwardt, and J. Miller. Discrete wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84:389–409, 2016.

[8] M. A. Arias-Serna, J.-M. Loubes, and F. J. Caro-Lopera. Risk measures estimation under wasserstein barycenter. *arXiv preprint arXiv:2008.05824*, 2020.

[9] N. S. Aybat and Z. Wang. A parallel method for large scale convex regression problems. In *53rd IEEE Conference on Decision and Control*, pages 5710–5717. IEEE, 2014.

[10] J. Backhoff-Veraguas, J. Fontbona, G. Rios, and F. Tobar. Bayesian learning with wasserstein barycenters. *ESAIM: Probability and Statistics*, 26:436–472, 2022.

[11] R. Bardenet, A. Doucet, and C. Holmes. On markov chain monte carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017.

[12] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[13] D. Bertsekas and J. Tsitsiklis. *Parallel and distributed computation: numerical methods*. Athena Scientific, 2015.

[14] G. Blekherman, P. A. Parrilo, and R. R. Thomas. *Semidefinite Optimization and Convex Algebraic Geometry*. Society for Industrial and Applied Mathematics, USA, 2012. ISBN 1611972280.

[15] S. Borgwardt. An lp-based, strongly-polynomial 2-approximation algorithm for sparse wasserstein barycenters. *Operational Research*, 22(2):1511–1551, 2022.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[17] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[18] L. Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[19] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[20] L. A. Caffarelli. A localization property of viscosity solutions to the monge-ampere equation and their strict convexity. *Annals of mathematics*, 131(1):129–134, 1990.

[21] L. A. Caffarelli. Some regularity properties of solutions of monge amp re equation. *Communications on pure and applied mathematics*, 44(8-9):965–969, 1991.

[22] L. A. Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

[23] L. A. Caffarelli. Boundary regularity of maps with convex potentials–ii. *Annals of mathematics*, 144(3):453–496, 1996.

[24] Y. Chen, J. Fan, and A. Taghvaei. Scalable computations of wasserstein barycenter via input convex neural networks. In *International Conference on Machine Learning*, pages 1571–1581. PMLR, 2021.

[25] S. Chewi, T. Maunu, P. Rigollet, and A. J. Stromme. Gradient descent algorithms for bures-wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.

[26] J. Chi, Z. Yang, X. Li, J. Ouyang, and R. Guan. Variational wasserstein barycenters with c-cyclical monotonicity regularization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7157–7165, 2023. doi: 10.1609/aaai.v37i6.25873.

[27] S. Claici, E. Chien, and J. Solomon. Stochastic wasserstein barycenters. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2018.

[28] R. T. Clemen and R. L. Winkler. Combining probability distributions from experts in risk analysis. *Risk analysis*, 19:187–203, 1999.

[29] S. Cohen, M. Arbel, and M. P. Deisenroth. Estimating barycenters of measures in high dimensions. *arXiv preprint arXiv:2007.07105*, 2020.

[30] M. Curmei and G. Hall. Shape-constrained regression using sum of squares polynomials. *Operations Research*, 2023.

[31] M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

[32] N. Deb, P. Ghosal, and B. Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753, 2021.

[33] P. Dognin, I. Melnyk, Y. Mroueh, J. Ross, C. D. Santos, and T. Sercu. Wasserstein barycenter model ensembling. *arXiv preprint arXiv:1902.04999*, 2019.

[34] A. L. Dontchev and R. T. Rockafellar. *Implicit functions and solution mappings: A view from variational analysis*. Springer Monographs in Mathematics. Springer, Dordrecht, 2009.

[35] P. Dvurechenskii, D. Dvinskikh, A. Gasnikov, C. Uribe, and A. Nedich. Decentralize and randomize: Faster algorithm for wasserstein barycenters. *Advances in Neural Information Processing Systems*, 31, 2018.

[36] D. Estrin, R. Govindan, J. Heidemann, and S. Kumar. Next century challenges: Scalable coordination in sensor networks. In *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*, pages 263–270, 1999.

[37] L. C. Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2010.

[38] D. Ge, H. Wang, Z. Xiong, and Y. Ye. Interior-point methods strike back: Solving the wasserstein barycenter problem. *Advances in neural information processing systems*, 32, 2019.

[39] N. Gigli. On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proc. Edinb. Math. Soc. (2)*, 54(2):401–409, 2011.

[40] A. González-Sanz, L. De Lara, L. Béthune, and J.-M. Loubes. Gan estimation of lipschitz optimal transport maps. *arXiv preprint arXiv:2202.07965*, 2022.

[41] F. F. Gunsilius. On the convergence rate of potentials of brenier maps. *Econometric Theory*, 38(2): 381–417, 2022.

[42] W. Guo, N. Ho, and M. Jordan. Fast algorithms for computational optimal transport and wasserstein barycenter. In *International Conference on Artificial Intelligence and Statistics*, pages 2088–2097. PMLR, 2020.

[43] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024. URL http://www.gurobi.com.

[44] F. Heinemann, A. Munk, and Y. Zemel. Randomized wasserstein barycenter computation: resampling with statistical guarantees. *SIAM Journal on Mathematics of Data Science*, 4(1):229–259, 2022.

[45] J.-C. Hütter and P. Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

[46] C. Intanagonwiwat, D. Estrin, R. Govindan, and J. Heidemann. Impact of network density on data aggregation in wireless sensor networks. In *Proceedings 22nd international conference on distributed computing systems*, pages 457–458. IEEE, 2002.

[47] Z. Izzo, S. Silwal, and S. Zhou. Dimensionality reduction for wasserstein barycenter. *Advances in neural information processing systems*, 34:15582–15594, 2021.

[48] S. N. Jagarlapudi and P. K. Jawanpuria. Statistical optimal transport posed as learning kernel embedding. *Advances in Neural Information Processing Systems*, 33:17334–17345, 2020.

[49] L. V. Kantorovich. On a problem of monge. *CR (Doklady) Acad. Sci. URSS (NS)*, 3:225–226, 1948.

[50] M. Knott and C. S. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49, 1984.

[51] A. Korotin, L. Li, J. Solomon, and E. Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. In *International Conference on Learning Representations*, 2021.

[52] A. Korotin, V. Egiazarian, L. Li, and E. Burnaev. Wasserstein iterative networks for barycenter estimation. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[53] S. Kum, M. H. Duong, Y. Lim, and S. Yun. A gpm-based algorithm for solving regularized wasserstein barycenter problems in some spaces of probability measures. *Journal of Computational and Applied Mathematics*, 416:114588, 2022.

[54] L. Li, A. Genevay, M. Yurochkin, and J. M. Solomon. Continuous regularized wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33:17755–17765, 2020.

[55] E. Lim and P. W. Glynn. Consistency of multidimensional convex regression. *Operations Research*, 60 (1):196–208, 2012.

[56] T. Lin, N. Ho, X. Chen, M. Cuturi, and M. Jordan. Fixed-support wasserstein barycenters: Computational hardness and fast algorithm. *Advances in neural information processing systems*, 33:5368–5380, 2020.

[57] G. Luise, S. Salzo, M. Pontil, and C. Ciliberto. Sinkhorn barycenters with free support via frank-wolfe algorithm. *Advances in neural information processing systems*, 32, 2019.

[58] A. Makkuva, A. Taghvaei, S. Oh, and J. Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

[59] T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman. Plugin estimation of smooth optimal transport maps. *Preprint, arXiv:2107.12364v2*, 2021.

[60] R. Mazumder, A. Choudhury, G. Iyengar, and B. Sen. A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, 114(525):318–331, 2019.

[61] S. Minsker, S. Srivastava, L. Lin, and D. B. Dunson. Robust and scalable Bayes via a median of subset posterior measures. *The Journal of Machine Learning Research*, 18(1):4488–4527, 2017.

[62] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[63] B. Muzellec and M. Cuturi. Generalizing point embeddings using the wasserstein space of elliptical distributions. *Advances in Neural Information Processing Systems*, 31, 2018.

[64] B. Muzellec, A. Vacher, F. Bach, F.-X. Vialard, and A. Rudi. Near-optimal estimation of smooth transport maps with kernel sums-of-squares. *arXiv preprint arXiv:2112.01907*, 2021.

[65] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2004.

[66] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

[67] A. Neufeld and Q. Xiang. Numerical method for feasible and approximately optimal solutions of multi-marginal optimal transport beyond discrete measures. *arXiv preprint arXiv:2203.01633*, 2022.

[68] A. Neufeld and Q. Xiang. Feasible approximation of matching equilibria for large-scale matching for teams problems. *arXiv preprint arXiv:2308.03550*, 2023.

[69] A. O'Hagan, C. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Chichester, UK, 2006.

[70] G. I. Papayiannis. Static hedging of freight risk under model uncertainty. *arXiv preprint arXiv:2207.00862*, 2022.

[71] F.-P. Paty, A. d'Aspremont, and M. Cuturi. Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1222–1232. PMLR, 2020.

[72] M. Perrot, N. Courty, R. Flamary, and A. Habrard. Mapping estimation for discrete optimal transport. *Advances in Neural Information Processing Systems*, 29, 2016.

[73] E. V. Petracou, A. Xepapadeas, and A. N. Yannacopoulos. Decision making under model uncertainty: Fréchet–wasserstein mean preferences. *Management Science*, 68(2):1195–1211, 2022.

[74] G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[75] A.-A. Pooladian and J. Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

[76] J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer, 2012.

[77] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences].* Springer-Verlag, Berlin, 1998. ISBN 3-540-62772-3.

[78] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[79] E. Seijo and B. Sen. Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(6):3126–3157, 2011.

[80] A. Simonetto. Smooth strongly convex regression. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 2130–2134. IEEE, 2021.

[81] S. P. Singh, A. Hug, A. Dieuleveut, and M. Jaggi. Context mover's distance & barycenters: Optimal transport of contexts for building representations. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 3437–3449. PMLR, 2020.

[82] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 03 1958.

[83] J. Solomon, F. De Goes, G. Peyre, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

[84] A. Spelta and N. Pecora. Wasserstein barycenter for link prediction in temporal networks. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187(1):180–208, 2024.

[85] S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson. Wasp: Scalable bayes via barycenters of subset posteriors. In *Artificial intelligence and statistics*, pages 912–920. PMLR, 2015.

[86] S. Srivastava, C. Li, and D. B. Dunson. Scalable bayes via barycenter in wasserstein space. *Journal of Machine Learning Research*, 19(8):1–35, 2018.

[87] M. Staib, S. Claici, J. M. Solomon, and S. Jegelka. Parallel streaming wasserstein barycenters. *Advances in Neural Information Processing Systems*, 30, 2017.

[88] B. Taşkesen, S. Shafieezadeh-Abadeh, and D. Kuhn. Semi-discrete optimal transport: hardness, regularization and numerical solution. *Mathematical Programming*, 199:1033–1106, 2023. doi: 10.1007/s10107-022-01856-x.

[89] Y. Takezawa, R. Sato, Z. Kozareva, S. Ravi, and M. Yamada. Fixed support tree-sliced wasserstein barycenter. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151, pages 1120–1137. PMLR, 2022.

[90] B. Taşkesen, S. Shafieezadeh-Abadeh, D. Kuhn, and K. Natarajan. Discrete optimal transport with independent marginals is# p-hard. *SIAM Journal on Optimization*, 33(2):589–614, 2023.

[91] A. B. Taylor. *Convex interpolation and performance estimation of first-order methods for convex optimization.* PhD thesis, Catholic University of Louvain, Louvain-la-Neuve, Belgium, 2017.

[92] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2017.

[93] D. Tiapkin, A. Gasnikov, and P. Dvurechensky. Stochastic saddle-point optimization for the wasserstein barycenter problem. *Optimization Letters*, 16(7):2145–2175, 2022.

[94] A. Vacher, B. Muzellec, A. Rudi, F. Bach, and F.-X. Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. In *Conference on Learning Theory*, pages 4143–4173. PMLR, 2021.

[95] C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2003.

[96] C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[97] J. von Lindheim. Approximative algorithms for multi-marginal optimal transport and free-support wasserstein barycenters. *arXiv preprint arXiv:2202.00954*, 2022.

[98] R. L. Winkler. The consensus of subjective probability distributions. *Management science*, 15(2):B–61, 1968.

[99] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[100] J. Ye, P. Wu, J. Z. Wang, and J. Li. Fast discrete distribution clustering using wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, 2017.

[101] Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, 1st edition, 2012. doi: 10.1201/b12207.

DIVISION OF MATHEMATICAL SCIENCES, NANYANG TECHNOLOGICAL UNIVERSITY, 21 NANYANG LINK, 637371 SINGAPORE
Email address: chen1417@e.ntu.edu.sg

DIVISION OF MATHEMATICAL SCIENCES, NANYANG TECHNOLOGICAL UNIVERSITY, 21 NANYANG LINK, 637371 SINGAPORE
Email address: ariel.neufeld@ntu.edu.sg

DIVISION OF MATHEMATICAL SCIENCES, NANYANG TECHNOLOGICAL UNIVERSITY, 21 NANYANG LINK, 637371 SINGAPORE
Email address: qikun.xiang@ntu.edu.sg