Journal of Machine Learning Research 25 (2024) 1-41 Submitted 5/23; Revised 1/24; Published 2/24

# Tangential Wasserstein Projections

**Florian Gunsilius** FFG@UMICH.EDU
*Department of Economics*
*University of Michigan*
*Ann Arbor, MI 48109-1220, USA*

**Meng Hsuan Hsieh** REXHSIEH@UMICH.EDU
*Ross School of Business*
*University of Michigan*
*Ann Arbor, MI 48109-1234, USA*

**Myung Jin Lee** LEEMJIN@TAMU.EDU
*Mays Business School*
*Texas A&M University*
*College Station, TX 77843-0001, USA*

**Editor:** Quentin Berthet

## Abstract

We develop a notion of projections between sets of probability measures using the geometric properties of the 2-Wasserstein space. In contrast to existing methods, it is designed for multivariate probability measures that need not be regular, and is computationally efficient to implement via regression. The idea is to work on tangent cones of the Wasserstein space using generalized geodesics. Its structure and computational properties make the method applicable in a variety of settings where probability measures need not be regular, from causal inference to the analysis of object data. An application to estimating causal effects yields a generalization of the synthetic controls method for systems with general heterogeneity described via multivariate probability measures.

**Keywords:** Optimal Transport, Wasserstein distance, Generalized geodesics, Projection, Tangent Cone, Causal Inference, Synthetic Controls

## 1. Introduction

The concept of projections, that is, approximating a target quantity of interest by an optimally weighted combination of other quantities, is of fundamental relevance in mathematics, statistics, and machine learning. Statistical projections are generally defined between random variables in appropriately defined linear spaces (e.g. van der Vaart, 2000, chapter 11). In modern statistics and machine learning applications, the objects of interest are often probability measures themselves. Examples range from object- and functional data (e.g. Marron and Alonso, 2014) to causal inference with individual heterogeneity (e.g. Athey and Imbens, 2015).

A notion of projection between sets of probability measures should be applicable between any set of general probability measures, replicate geometric properties of the target measure, and possess good computational and statistical properties. We introduce such a notion of projection between sets of general probability measures supported on Euclidean spaces. It

also provides a global solution to the projection problem, which will in general be unique. To achieve this, we work in the 2-Wasserstein space, that is, the set of all probability measures with finite second moments equipped with the 2-Wasserstein distance.

Importantly, we focus on the multivariate setting, i.e. we consider the Wasserstein space over some Euclidean space $\mathbb{R}^d$, denoted by $\mathcal{W}_2(\mathbb{R}^d)$, where the dimension $d$ can be arbitrarily high. The multivariate setting poses challenges from a mathematical, computational, and statistical perspective. In particular, $\mathcal{W}_2$ is a positively curved metric space for $d > 1$ (e.g. Ambrosio et al., 2008; Kloeckner, 2010). Moreover, the 2-Wasserstein distance between two probability measures is defined as the value function of the Monge-Kantorovich optimal transportation problem (Villani, 2003, chapter 2), which does not have a closed-form solution in multivariate settings in general. This is coupled with a well-known statistical curse of dimensionality for general measures (Ajtai et al., 1984; Dudley, 1969; Fournier and Guillin, 2015; Talagrand, 1992, 1994; Weed and Bach, 2019).

### 1.1 Existing Approaches

These challenges have impeded the development of a method of projections between general and potentially high-dimensional probability measures. A focus so far has been on the univariate and low-dimensional setting. In particular, Chen et al. (2021), Ghodrati and Panaretos (2022), and Pegoraro and Beraha (2021) introduce frameworks for distribution-on-distribution regressions in the univariate setting for object data. Bigot et al. (2014) and Cazelles et al. (2017) develop principal component analyses on the space of univariate probability measures using geodesics on the Wasserstein space. Most recently, in the context of manifold learning, Cloninger et al. (2023) examine how to learn low-dimensional structures by approximating pairwise Wasserstein distances between data points, which are probability measures.

The most closely related works to ours are Wang et al. (2013), Kolouri et al. (2016), Bonneel et al. (2016), Mérigot et al. (2020), Werenski et al. (2022), and Fan and Alvarez-Melis (2023). Wang et al. (2013) and Kolouri et al. (2016) leverage generalized geodesics to construct linear Wasserstein embeddings, with an eye towards applications in computer vision. Bonneel et al. (2016) develop a regression approach in barycentric coordinates with applications in computer graphics as well as color and shape transport problems. Their method is defined directly on $\mathcal{W}_2$ and requires solving a bilevel optimization problem, which does not necessarily yield global solutions. This is not an issue for prediction problems such as color transport, which is their main focus. For causal inference, which is our main application, the weights obtained in the projection problem are of primary interest, however. Therefore, a method that obtains globally optimal weights is desirable.

Mérigot et al. (2020) introduce a linearization of the 2-Wasserstein space by lifting it to a $L^2$-space anchored at a measure that is absolutely continuous with respect to Lebesgue measure. Similarly, Fan and Alvarez-Melis (2023) introduce a method of projection along generalized geodesics similar to ours, but it requires the existence of optimal transport maps and hence a regular reference measure, which needs to be fixed; this also makes their practical implementation significantly more computationally involved. Since we focus on projections, the "anchor" in our case is naturally given as the target measure we want to project. This need not be absolutely continuous in practice. Allowing for general target

measures to project is important from a causal inference perspective, as many measures of interest as not continuous, such as treatment status.

Werenski et al. (2022) work with a tangential structure based on "Karcher means" (Karcher, 2014; Zemel and Panaretos, 2019). In particular, this means that their method requires all probability measures to be absolutely continuous with respect to the Lebesgue measure, their densities bounded away from zero, and with the target measure lying in the convex hull of the control measures, something that is as hard to check in practice as performing the projection.

### 1.2 Our Contributions

In contrast to the existing approaches, our method is applicable for general probability measures, allows for the target measure to be outside the generalized geodesic convex hull of the control measures, can be implemented by a standard constrained linear regression, and provides a global solution. Convexity of the projection in particular implies that solutions are unique conditional on fixing the optimal transport plans.

Specifically, our proposed method transforms the projection problem on the positively-curved Wasserstein space into a constrained regression problem in the geometric tangent cone. This problem takes the form of a deformable template (Boissard et al., 2015; Yuille, 1991), which connects our approach to this literature. Our method can be implemented in three steps: (i) obtain the general tangent cone structure at the target measure, (ii) construct a tangent space from the tangent cone via barycentric projections if it does not exist, and (iii) perform a regression constrained to the unit simplex to carry out the projection in the tangent space. This implementation of the projection approach via linear regression is computationally efficient, in particular compared to the existing methods in Bonneel et al. (2016), Fan and Alvarez-Melis (2023), and Werenski et al. (2022).

The challenging part of the implementation is lifting the problem to the tangential structure: this requires computing the corresponding optimal transport plans between the target and each measure used in the projection. Many methods have been developed for this, see for instance Benamou and Brenier (2000); Jacobs and Léger (2020); Makkuva et al. (2020); Peyré and Cuturi (2019); Ruthotto et al. (2020) and references therein. Other alternatives compute approximations of the optimal transport plans via regularized optimal transport problems (Peyré and Cuturi, 2019) such as entropy regularized optimal transport (Galichon and Salanié, 2010; Cuturi, 2013). The proposed projection approach is compatible with any such method, therefore its complexity scales with that of estimating optimal transport plans. As a statistical contribution, we provide results for the statistical consistency when estimating the measures via their empirical counterparts in practice.

To demonstrate the efficiency and utility of the proposed method, we apply it in different settings and compare it to existing benchmarks such as Werenski et al. (2022) where those are computationally feasible. Furthermore, we extend the classical synthetic control estimator (Abadie and Gardeazabal, 2003; Abadie et al., 2010), a fundamental approach for counterfactual prediction in causal inference, to settings with observed individual heterogeneity in multivariate outcomes. The synthetic controls estimator is a projection approach, where one tries to predict an aggregate outcome of a treated unit by an optimal convex combination of control units. The weights of this optimal combination is then used to construct

the counterfactual state of the treated unit had it not received treatment. The novelty of our application is that it lets us perform the synthetic control method on the joint distribution of several outcomes, which complements the recently introduced method in Gunsilius (2023) designed for univariate outcomes. The possibility to project entire probability measures allows us to disentangle treatment heterogeneity at the treatment unit level. The possibility of working with general probability measures is key in this setting, as many outcomes of interest are not regular. We illustrate this by applying our method to estimate the effects of a Medicaid expansion policy in Montana, where we consider—as outcome—non-regular probability measures in $d = 28$ dimensions.

All the code used to produce the synthetic experiment results and the application to synthetic control method can be found at the following GitHub repository: `https://github.com/menghsuanhsieh/tangential-wasserstein-projection`.

## 2. Methodology

### 2.1 The 2-Wasserstein Space $\mathcal{W}_2(\mathbb{R}^d)$

For probability measures $P_X, P_Y \in \mathscr{P}(\mathbb{R}^d)$ with supports $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, respectively, the 2-Wasserstein distance $W_2(P_X, P_Y)$ is defined as

$$W_2(P_X, P_Y) \triangleq \left( \min_{\gamma \in \Gamma(P_X, P_Y)} \int_{\mathcal{X} \times \mathcal{Y}} |x - y|^2 \, \mathrm{d}\gamma(x, y) \right)^{\frac{1}{2}}. \tag{1}$$

Here, $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$ and

$$\Gamma(P_X, P_Y) \triangleq \left\{ \gamma \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d) : (\pi_1)_{\#}\gamma = P_X, \ (\pi_2)_{\#}\gamma = P_Y \right\}$$

is the set of all couplings of $P_X$ and $P_Y$. The maps $\pi_1$ and $\pi_2$ are the projections onto the first and second coordinate, respectively, and $T_{\#}P$ denotes the pushforward measure of $P$ via $T$, i.e. for any measurable $A \subset \mathcal{Y}$, $T_{\#}P(A) \equiv P(T^{-1}(A))$. An optimal coupling $\gamma \in \Gamma(P_X, P_Y)$ solving the optimal transport problem (1) is an *optimal transport plan*. By Prokhorov's theorem, a solution always exists in our setting. When $P_X$ is regular, i.e. when it does not give mass to sets of lower Hausdorff dimension in its support, then the optimal transport plan $\gamma$ solving (1) is unique and takes the form $\gamma = (\mathrm{Id} \times \nabla\varphi)_{\#}P_X$, where Id is the identity map on $\mathbb{R}^d$ and $\nabla\varphi(x)$ is the gradient of some convex function. This result is known as Brenier's theorem (Brenier, 1991; McCann, 1997; Villani, 2003, Theorem 2.12). By definition, all measures that possess a density with respect to Lebesgue measure are regular. Our main contribution is to allow for general probability measures, where only optimal transport plans but no maps exist.

The 2-Wasserstein space $\mathcal{W}_2 \triangleq (\mathscr{P}_2(\mathbb{R}^d), W_2)$ is the metric space defined on the set $\mathscr{P}_2(\mathbb{R}^d)$ of all probability measures with finite second moments supported on $\mathbb{R}^d$, with the 2-Wasserstein distance as the metric. It is a geodesically complete space in the sense that between any two measures $P, P' \in \mathcal{W}_2$, one can define a geodesic $P_t : [0, 1] \to \mathcal{W}_2$ via the interpolation $P_t \triangleq (\pi_t)_{\#}\gamma$, where $\gamma$ is an optimal transport plan and $\pi_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is defined through $\pi_t(x, y) \triangleq (1 - t)x + ty$ (Ambrosio et al., 2008; McCann, 1997). Using this, it can be shown that $\mathcal{W}_2$ is a positively curved metric space when $d > 1$ (Ambrosio et al.,

2008, Theorem 7.3.2) and flat for $d = 1$ (Kloeckner, 2010), where curvature is defined in the sense of Aleksandrov (1951). This difference in the curvature properties is the main reason for why the multivariate setting requires different approaches compared to the established results for measures on the real line.

### 2.2 Tangent cone structure on $\mathcal{W}_2$

We exploit a tangential structure that can be defined for general measures on $\mathcal{W}_2$ (Ambrosio et al., 2008; Otto, 2001; Takatsu and Yokota, 2012). In particular, it allows us to circumvent solving a bilevel optimization problem as the one considered in Bonneel et al. (2016), which we review below.

The tangential structure relies on the fact that geodesics $P_t$ in $\mathcal{W}_2$ are linear in the transport plans $(\pi_t)_\# \gamma$. This implies a geometric tangent cone structure at each measure $P \in \mathcal{W}$ that can be defined as the closure in $\mathscr{P}_2(\mathbb{R}^d)$ of the set

$$\mathcal{G}(P) \triangleq \left\{ \gamma \in \mathscr{P}_2(\mathbb{R}^d \times \mathbb{R}^d) : (\pi_1)_\# \gamma = P \ , \ (\pi_1, \pi_1 + \varepsilon \pi_2)_\# \gamma \text{ is optimal for some } \varepsilon > 0 \right\} \tag{2}$$

with respect to the local distance

$$W_P^2(\gamma_{12}, \gamma_{13}) \triangleq \min \left\{ \int_{(\mathbb{R}^d)^3} |x_2 - x_3|^2 \, \mathrm{d}\gamma_{123} : \gamma_{123} \in \Gamma_1(\gamma_{12}, \gamma_{13}) \right\} \ , \tag{3}$$

where $\gamma_{12}$ and $\gamma_{13}$ are couplings between $P$ and some other measures $P_2$ and $P_3$, respectively, and $\Gamma_1(\gamma_{12}, \gamma_{13})$ is the set of all 3-couplings $\gamma_{123}$ such that the projection of $\gamma_{123}$ onto the first two elements is $\gamma_{12}$ and the projection onto the first and third element is $\gamma_{13}$ (Ambrosio et al., 2008, Appendix 12). The optimality requirement in $\mathcal{G}(P)$ is with respect to transport plans $\gamma$. We can then define the exponential map at $P$ with respect to some tangent element $\gamma \in \mathcal{G}(P)$ by

$$\exp_P(\gamma) = (\pi_1 + \pi_2)_\# \gamma \ .$$

This tangent cone can be constructed at every $P \in \mathcal{W}$, irrespective of its support; in particular, we do not assume that the corresponding measures are regular, i.e., give mass to subsets of $\mathbb{R}^d$ of lower Hausdorff dimension. In the case where $P$ is regular the tangent cone structure reduces to a tangent space (Ambrosio et al., 2008, Theorem 8.5.1).

### 2.3 Descriptions of Existing Approaches

One approach to defining projections on $\mathcal{W}_2$ is to work on the Wasserstein space directly. This leads to a bilevel optimization problem, based on the notion of barycenters in Wasserstein space (Agueh and Carlier, 2011; Carlier and Ekeland, 2010):

$$\bar{P}(\lambda) = \operatorname*{argmin}_{P \in \mathscr{P}_2(\mathbb{R}^d)} \sum_{j=1}^{J} \frac{\lambda_j}{2} W_2^2(P, P_j).$$

With this definition, and assuming that the barycenter $\bar{P}(\lambda)$ is unique for given $\lambda$, the bilevel projection problem reads:

$$\lambda^* \in \underset{\lambda \in \Delta^J}{\operatorname{argmin}} \; W_2(P_0, \bar{P}(\lambda)), \qquad \text{where} \quad \bar{P}(\lambda) = \underset{P \in \mathscr{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \sum_{j=1}^{J} \frac{\lambda_j}{2} W_2^2(P, P_j).$$

A version of this approach is used in Bonneel et al. (2016) to define a notion of regression between probability measures in low dimensions. The challenges here are mathematical and computational. Importantly, the optimal weights $\lambda^*$ need not be unique. This is not an issue for the applications considered in Bonneel et al. (2016), like color transport; however, it is important in statistical settings when the weights convey information used in further procedures, like causal inference via synthetic controls, where the optimal weights are used to introduced a counterfactual outcome of a treated unit had it not been treated (Abadie and Gardeazabal, 2003; Abadie et al., 2010; Abadie, 2021). Moreover, the bi-level optimization structure makes solving the problem prohibitively costly in higher dimensions. Bonneel et al. (2016) introduce a gradient descent approach based on an entropy-regularized analogue of $W_2$ (Cuturi, 2013; Peyré and Cuturi, 2019) that can be implemented in low-dimensional settings.

Other approaches like Werenski et al. (2022) introduce a tangential approach, but under strong assumptions on the involved measures: they need to be absolutely continuous with densities bounded away from zero on their support, and in particular the target measure must be known to lie inside the convex hull of the other measures, something that is as hard to check in practice as performing the projection. A starting point for this is to consider a characterization of the barycenter $\bar{P}(\lambda)$ for fixed weights of a set $\{P_j\}_{j \in [\![J]\!]}$ in regular tangent spaces. Agueh and Carlier (2011, Equation (3.10)) show that if at least one of the measures is absolutely continuous with respect to Lebesgue measure, then $\bar{P}(\lambda)$ can be characterized via

$$\sum_{j=1}^{J} \lambda_j \left( \nabla \tilde{\varphi}_j - \mathrm{Id} \right) = 0, \tag{4}$$

where $\{\tilde{\varphi}_j\}_{j \in [\![J]\!]}$ are the optimal transport maps from the barycenter to the respective measure $P_j$, i.e. $(\tilde{\varphi}_j)_{\#} \bar{P}(\lambda) = P_j$. Each term of the summand in (4) an element in $\mathcal{T}_{\bar{P}(\lambda)} \mathcal{W}_2(\mathbb{R}^d)$ by construction. We leave additional descriptions to Appendix A, including a projections result we can prove in this setting.

### 2.4 Tangential Wasserstein Projections

Our main contribution is to define a projection approach between general probability measures, where the target need not be regular. To define this notion of projection, we need to first define an appropriate notion of a geodesic convex hull. The novelty here is that we define this notion via generalized geodesics (Ambrosio et al., 2008, section 9.2) *centered at the target measure* $P_0$. For this, we extend the definition of $W_P$ to the multimarginal setting, by defining, for given couplings $\gamma_{0j} \in \Gamma(P_0, P_j)$, $j \in [\![J]\!]$

$$W^2_{P_0;\lambda}(\gamma_{01}, \gamma_{02}, \ldots, \gamma_{0J}) \triangleq \min \left\{ \int_{(\mathbb{R}^d)^{J+1}} \sum_{j=1}^{J} \lambda_j \left| x_j - x_0 \right|^2 \mathrm{d}\boldsymbol{\gamma} : \boldsymbol{\gamma} \in \Gamma_1(\gamma_{01}, \ldots, \gamma_{0J}) \right\}, \tag{5}$$

where $\Gamma_1(\gamma_{01}, \ldots, \gamma_{0J}) \subset \Gamma(P_0, P_1, \ldots, P_J)$ is the set of all $(J+1)$-couplings $\boldsymbol{\gamma}$ such that the projection of $\boldsymbol{\gamma}$ onto the first- and $j$-th element is $\gamma_{0j}$. Note that this definition is similar to the multimarginal definition of the 2-Wasserstein barycenter (Agueh and Carlier, 2011; Gangbo and Święch, 1998), but "centered" at $P_0$. Based on this, we define the generalized geodesic convex hull of measures $\{P_j\}_{j \in [\![J]\!]}$ with respect to the measure $P_0$ as

$$\mathfrak{Co}_{P_0}\left(\{P_j\}_{j=1}^J\right) \triangleq \left\{ P(\lambda) \in \mathscr{P}_2(\mathbb{R}^d) : P(\lambda) = \left(\sum_{j=1}^J \lambda_j \pi_{j+1}\right)_{\#} \boldsymbol{\gamma}, \right.$$

$$\left. \boldsymbol{\gamma} \text{ solves } W^2_{P_0;\lambda}(\gamma_{01}, \ldots, \gamma_{0J}), \ \gamma_{0j} \text{ is optimal in } \Gamma(P_0, P_j) \ \forall j \in [\![J]\!], \quad \lambda \in \Delta^J \right\} . \quad (6)$$

A direct application of our tangential projection idea would lead us to solving

$$\lambda^* \triangleq \operatorname*{argmin}_{\lambda \in \Delta^J} W^2_{P_0;\lambda}(\gamma_{01}, \ldots, \gamma_{0J}) \ , \quad (7)$$

which would be a computationally prohibitive bilevel optimization problem similar to the one in Bonneel et al. (2016). We therefore rely on barycentric projections to reduce the general cone structure to a regular tangent space which we denote by $\mathcal{T}_{P_0} \mathcal{W}_2$ (Ambrosio et al., 2008). In this structure the projection problem (7) is replaced by

$$\lambda^* \triangleq \operatorname*{argmin}_{\lambda \in \Delta^J} \left\| \sum_{j=1}^J \lambda_j \left( b_{\gamma_{0j}} - \mathrm{Id} \right) \right\|^2_{L^2(P_0)} , \quad \text{with} \quad b_{\gamma_{0j}}(x_1) \triangleq \int_{\mathbb{R}^d} x_2 \, \mathrm{d}\gamma_{0j,x_1}(x_2) \quad (8)$$

denoting the barycentric projections of optimal transport plans $\gamma_{0j}$ between $P_0$ and $P_j$. Here, $\gamma_{x_1}$ denotes the disintegration of the optimal transport plan $\gamma$ with respect to $P_0$.

This approach is a natural extension of the regular setting to general probability measures for two reasons. First, if the optimal transport plans $\gamma_{0j}$ are actually induced by some optimal transport map, then $b_{\gamma_{0j}}$ reduces to this optimal transport map; in this case the general tangent cone $\mathcal{G}(P_0)$ reduces to the regular tangent cone $\mathcal{T}_{P_0} \mathcal{W}_2$ (Ambrosio et al., 2008, Theorem 12.4.4). Second, by the definition of $b_\gamma$ and disintegrations in conjunction with Jensen's inequality it holds for all $\lambda \in \Delta^J$ that

$$\left\| \sum_{j=1}^J \lambda_j \left( b_{\gamma_{0j}} - \mathrm{Id} \right) \right\|^2_{L^2(P_0)} \leq W^2_{P_0;\lambda}(\gamma_{01}, \ldots, \gamma_{0J}) \ . \quad (9)$$

This implies that for general $P_0$ we can also define a convex hull based on barycentric projections, which is of the form

$$\widetilde{\mathfrak{Co}}_{P_0}\left(\{P_j\}_{j=1}^J\right) \triangleq \left\{ P(\lambda) \in \mathscr{P}_2(\mathbb{R}^d) : P(\lambda) = \left(\sum_{j=1}^J \lambda_j b_{\gamma_{0j}}\right)_{\#} P_0, \quad \lambda \in \Delta^J \right\} . \quad (10)$$

Using these definitions, the following defines our notion of projection for general $P_0$ and shows that it projects onto $\widetilde{\mathfrak{Co}}_{P_0}$.

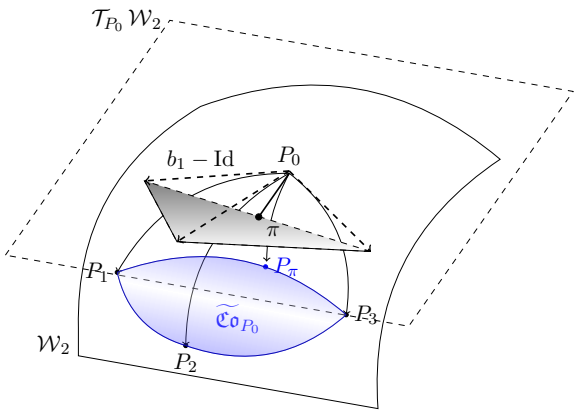



Figure 1: Tangential Wasserstein projection for a general target $P_0$.

$\mathcal{T}_{P_0}\mathcal{W}_2$ is the regular tangent space constructed by applying barycentric projection to $\mathcal{G}(P_0)$, the general tangent cone anchored at $P_0$. Thick dashed lines are tangent vectors $(b_j - \mathrm{Id})$ generated by the respective barycentric projections. The gray shaded region is their convex hull in this constructed tangent space and $\pi$ is the projection of Id onto this convex hull. $P_\pi \triangleq \exp_{P_0}(\pi)$ is the projection of $P_0$ onto the generalized geodesic convex hull $\widetilde{\mathfrak{Co}}_{P_0}\left(\{P_1, P_2, P_3\}\right) \subset \mathcal{W}_2$ (blue).

**Proposition 1** *Consider a general target measure $P_0$ and a set $\{P_j\}_{j\in[\![J]\!]}$ of general control measures. Construct the measure $\widetilde{P}_\pi$ as*

$$\widetilde{P}_\pi \triangleq \exp_{P_0}\left(\sum_{j=1}^{J} \lambda_j^* b_{\gamma_{0j}} - \mathrm{Id}\right) ,$$

*where the optimal weights $\lambda^* \in \Delta^J$ are obtained by solving* (8) *and $\gamma_{0j}$ are optimal plans transporting $P_0$ to $P_j$, respectively. Then for given optimal plans $\gamma_{0j}$, $\widetilde{P}_\pi$ is the unique metric projection of $P_0$ onto $\widetilde{\mathfrak{Co}}_{P_0}\left(\left\{P_j\right\}_{j=1}^J\right)$.*

The optimal plans $\gamma_{0j}$ transporting $P_0$ to $P_j$ need not be unique if $P_j$ lies outside the cut locus of $P_0$, i.e., when there is more than one optimal way to transport $P_0$ onto $P_j$. However, the projection for fixed $\gamma_{0j}$ is always unique by virtue of the linear regression.

## 3. Statistical Properties of the Weights and Projection

We now provide statistical consistency results for our method when the corresponding measures $\{P_j\}_{j\in[\![J]\!]}$ are estimated from data. We consider the case where the measures $P_j$ are replaced by their empirical counterparts

$$\mathbb{P}_{N_j}(A) \triangleq N_j^{-1} \sum_{n=1}^{N_j} \delta_{X_n}(A)$$

for every measurable set $A$ in the Borel $\sigma$-algebra on $\mathbb{R}^d$, where $\delta_x(A)$ is the Dirac measure and $\left(X_{1j}, \ldots, X_{N_j,j}\right)$ is an independent and identically distributed set of random variables

whose distribution is $P_j$. We explicitly allow for different sample sizes $\bigcup_{j=0}^{J} N_j = N$ for the different measures. To save on notation we write $\widehat{\varphi}_{N_j} \equiv \widehat{\varphi}_j$, $\widehat{b}_{0j} \equiv \widehat{b}_{\gamma_{0j},N_j}$ and $\widehat{\gamma}_{0j} \equiv \widehat{\gamma}_{N_j,N_0}$ in the following.

**Proposition 2 (Consistency of the optimal weights)** *Let* $\left\{\mathbb{P}_{N_j}\right\}_{j=0}^{J}$ *be the empirical measures corresponding to the data* $\left(X_{1j},\dots,X_{N_jj}\right)_{j=0}^{J}$ *which are independent and identical draws from* $P_j$*, respectively, and are supported on some common latent probability space* $(\Omega,\mathscr{A},P)$*. Assume all* $P_j$ *have finite second moments. As* $N_j \to \infty$ *for all* $j \in [\![J]\!]$*, the corresponding optimal weights* $\widehat{\lambda}_N^* = \left(\widehat{\lambda}_{N_1}^*,\dots,\widehat{\lambda}_{N_J}^*\right) \in \Delta^J$ *obtained via*

$$\widehat{\lambda}_N^* \triangleq \operatorname*{argmin}_{\lambda\in\Delta^J} \left\| \sum_{j=1}^{J} \lambda_j \left(\widehat{b}_{0j} - \mathrm{Id}\right) \right\|^2_{L^2(\mathbb{P}_{N_0})} , \tag{11}$$

*satisfy*

$$P\left(\left|\widehat{\lambda}_N^* - \lambda^*\right| > \varepsilon\right) \to 0 \qquad \text{for all } \varepsilon > 0 \ ,$$

*where* $\lambda^*$ *solve* (8)*.*

This consistency result directly implies consistency of the optimal weights in case the optimal transport problems between $\mathbb{P}_{N_0}$ and each $\mathbb{P}_{N_j}$ are achieved by optimal transport maps $\nabla\widehat{\varphi}_{N_j}$. We also have a consistency result for the empirical counterparts $\widetilde{\mathbb{P}}_{\pi,N}$ of the optimal projection $\widetilde{P}_\pi$.

**Corollary 3 (Consistency of the optimal projections)** *In the setting of Proposition 2, the estimated projections* $\widetilde{\mathbb{P}}_{\pi,N}$ *converge weakly in probability to the projection* $\widetilde{P}_\pi$ *as* $N_j \to \infty$ *for all* $j \in [\![J]\!]$*.*

Proposition 2 and Corollary 3 hold in all generality and without any assumptions on the corresponding measures $P_j$, except that they possess finite second moments. To get stronger results—for instance, parametric rates of convergences—we need to make strong regularity assumptions on the measures $P_j$, such as on the smoothness of their densities (provided they exist). Under such additional regularity conditions, the results for the asymptotic properties are standard (e.g. Andrews, 1989), because the proposed method reduces to a classical semiparametric estimation problem, as the weights $\lambda_j$ are finite-dimensional. Additionally, a recent work of Deb et al. (2021) examine the rates of convergence of estimating optimal transport maps via computing barycentric projections of optimal transport plans. Without additional regularity assumptions, the rate of convergence of optimal transport maps in terms of expected square loss is as slow as $n^{-2/d}$ (Hütter and Rigollet, 2021).

## 4. Illustrations

For all experiments in this section, we use the `POT` toolbox (Flamary et al., 2021) to compute optimal transport plans and free-support barycenters. To solve the regression problem

constrained to the unit simplex, we leverage the constrained optimization solver from the `CVXPY` toolbox (Diamond and Boyd, 2016).

As for the computational complexity, our proposed method consists of two steps: obtaining the optimal transport maps between the target and each of the $J$ control units, and then applying a regression constrained to the unit simplex to obtain the optimal weights. In the case where optimal maps do not exist, we estimate the optimal transport plans first, then obtain the optimal transport maps by applying barycentric projections to the estimated optimal transport plans. We included an analysis of the computational complexity of our proposed method, and a small simulation exercise, in Appendix E.

### 4.1 Image Experiment: MNIST

We compare our results to those from the experiment in Section 4.3 of Werenski et al. (2022). We follow the experimental procedure described therein, taking as experimental data the MNIST dataset of $28 \times 28$ pixel images of hand-written digits (LeCun, 1998). We show comparison to the test case with image occlusion and with salt and pepper noise. We treat the normalized matrix as probability measures supported on a $28 \times 28$ grid. Our experiment was run using 10 control images; these control images are shown in Appendix C. For each control unit, we record the relative weights the receive in the respective projection approach. 500 iterations of each of these exercises took 4 seconds to compute on an Apple M1 laptop with 8 cores and 16GB of working memory. In fact, computation takes 4 seconds on any exercise: any digit between 0 and 9, with either type of noise.

**Occlusion** Figures 2 shows our results. The occlusion for the "4" image is around 8% and for the "8" image it is around 25%.

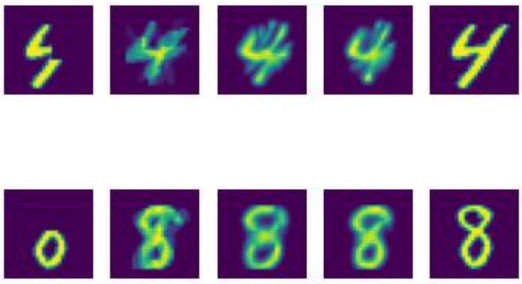


Figure 2: Left to right: occluded image; Euclidean projection; result from Werenski et al. (2022), using optimal weights obtained from their method; result from our approach, using optimal weights obtained from (8); target image.

**Salt and pepper noise** Figures 3 shows our results. The noisy pixels are chosen randomly and make up between 10% and 20% of all pixels.

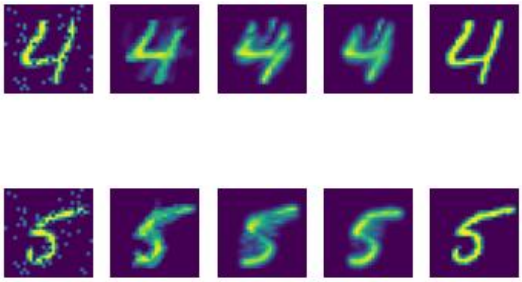
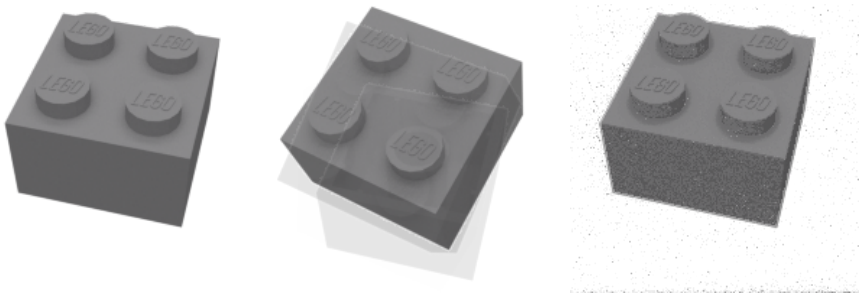


Figure 3: Left to right: image with salt and pepper noise; Euclidean projection; result from Werenski et al. (2022), using optimal weights obtained from their method; result from our approach, using optimal weights obtained from (8); target image.

### 4.2 Image Experiment: Lego Bricks

To examine the general properties of how our method obtains the optimal weights, we provide an application on replicating a target image of an object using images of the same object taken from different angles. We use the Lego Bricks dataset available from `Kaggle`, which contains approximately 12,700 images of 16 different Lego bricks in RGBA format. All images used have the resolution $200 \times 200$ pixels. We used 10 images as our controls, and these are illustrated in Figure 6. Figure 4 shows our results. Our method manages to replicate the target block rather well, while only using the information of control units that look sufficiently like the target (i.e. first row of Figure 6). In particular, in replication, our method places zero weights on every image from the underside of the Lego brick (see second row of Figure 6). In contrast, the Euclidean projection does not provide the correct rotation in the replication, and suffers from the standard blur induced by using a mixture of images.

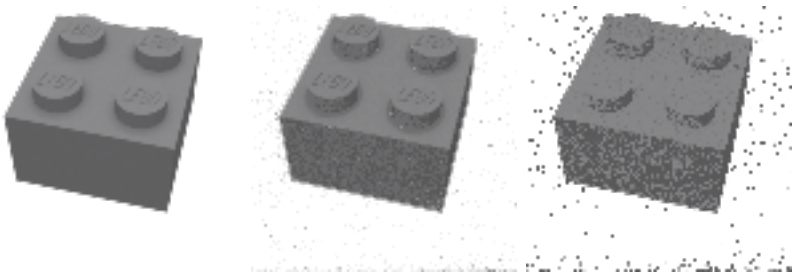


(a) (b) (c)

Figure 4: (a) target block, (b) Euclidean projection, and (c) projection from our method.

In Figure 5, we show a direct comparison of our method to that in Werenski et al. (2022) on a downsampled target image. We downsampled the target image to reduce the extensive computational time of the method in Werenski et al. (2022). Our projection retains fine details on the target block, while the method in Werenski et al. (2022) does not,

and it results in an image with a significant amount of noise. The 2-Wasserstein distances between Figure 5(a) and Figure 5(b) and Figure 5(a) and Figure 5(c) are 0.0068 and 0.0330, respectively. We also computed our proposed projection in 3 minutes on an Apple M1 laptop with 8 cores and 16GB of working memory, compared to 4 hours on a cluster computer with 36 cores and 180GB of working memory for the method of Werenski et al. (2022).



(a) (b) (c)

Figure 5: (a) target block, (b) projection from our method, and (c) barycenter obtained from Werenski et al. (2022).

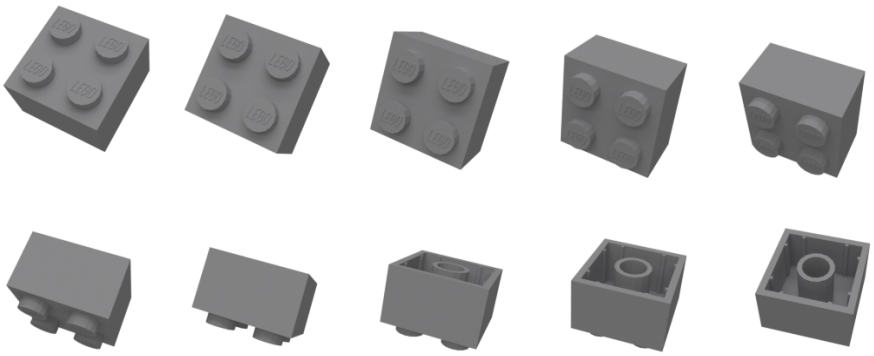


Figure 6: Control units used in simulation for Section 4.2.

### 4.3 Application to Color Transport

We further illustrate an application of our method to color transport as in Bonneel et al. (2016). We utilize the Berkeley DeepDrive dataset (Yu et al., 2020). Our goal is to take an image taken from the dashboard, at daytime, and show it as if the photo is taken during nighttime. The results are contained in Figure 7. The controls units used are contained in Figure 8. We note that our projection puts the most weights on control units that are both bright and situationally "close" to the target image (i.e. the center and right images in the first row of Figure 8). This allows us to achieve painting the target image as if it is nighttime. The images used for this example is downsampled to a size of $144 \times 216$ pixels from its original $720 \times 1080$ pixels to reduce computation time. Unfortunately, even the downsampled size requires a rather extreme amount of computing power and working memory for methods from Bonneel et al. (2016) and Werenski et al. (2022). We thus present results from our method only in comparison to the Euclidean projection. Our results were

computed on a cluster computer with 36 cores and 180GB of RAM within 1.5 hours. Lastly, our output is correctly classified as a nighttime image by a pre-trained night-and-day image classifier.

Figure 7: (a) target image, (b) Euclidean projection, and (c) projection from our method.

Figure 8: Control units used in simulation for Section 4.3.

## 5. Application to Causal Inference Via Synthetic Controls

When analyzing the causal effect of treatment on a unit that is observed over many time periods with an intervention in one time period, such as that of public policies or medical interventions on individuals, there is often no comparable control unit that can capture the treated unit's underlying characteristics and trend over time. The classical synthetic controls method (Abadie and Gardeazabal, 2003; Abadie et al., 2010) aims to create a suitable control unit by replicating the pre-treatment outcome trends of the treated unit, using some optimally chosen set of control units, the *synthetic control* unit. This is achieved by projecting the observed characteristics of the target unit onto the convex hull defined by the characteristics of control units in the pre-treatment periods. The optimal weights obtained by this projection, therefore, describe how much each control unit contributes to the target unit's counterfactual outcome in the post-treatment period (Abadie, 2021).

Current developments in the synthetic control method literature focus on vectors as the outcome of interest. The trend over time of the outcomes are assumed to be generated via latent factor models, and principal component analyses generate counterfactual outcomes (Agarwal et al., 2019; Athey et al., 2021; Bai and Ng, 2021). There is, separately, a growing literature that relates the synthetic control method for vector-valued outcomes to online learning and machine learning; see, for example, Chen (2023); Bottmer et al. (2023); Agarwal et al. (2023), and references therein. In contrast to the existing methods, our proposed

method of projection can be applied to define a synthetic controls estimator for outcomes that are *multivariate measures* instead of vectors; in that, it generalizes the recent univariate method of distributional synthetic controls introduced in Gunsilius (2023) and allows to nonparametrically disentangle multivariate heterogeneous treatment effects.

As demonstration, we study the effect of health insurance coverage following state-level Medicaid expansion in Montana in 2016. The variables of interest are Medicaid coverage, employment status, log wages, and log hours worked. For control units, we use the twelve states for which such expansion has never occurred; these are: Alabama, Florida, Georgia, Kansas, Mississippi, North Carolina, South Carolina, South Dakota, Tennessee, Texas, Wisconsin, Wyoming. Additional information can be found in Appendix F.

We estimate "synthetic Montana", i.e. Montana had it not adopted Medicaid expansion, by estimating the optimal weights $\lambda^*$ using data from 2010 to 2016, and solving (8) over the joint distribution of the four outcomes over the time period from 2010 to 2016, which generates measures in $d = 28$ dimensions. We note that we estimate one set of optimal weights—specifically, one for each control state—over the entire time period. We then estimate the counterfactual joint distribution using data from 2017 to 2020, by using the optimal weights $\lambda^*$ and computing the weighted barycenter (Agueh and Carlier, 2011) of the control states using these weights. Details of sample selection and estimating "synthetic Montana" are described in Appendix F. The results of the general causal effect of the Medicaid expansion policy in Montana averaged over the years $2017 - 2020$ are illustrated in Figures 9, 10, 11, 12.

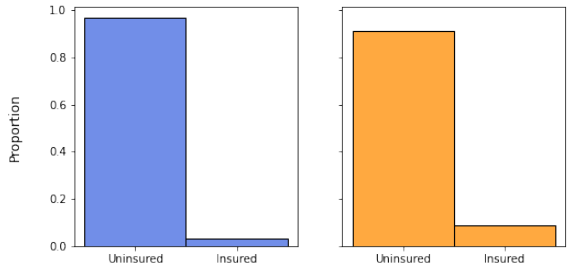



Figure 9: Counterfactual (blue) vs actual (orange) Medicaid coverage in Montana from 2017 to 2020.

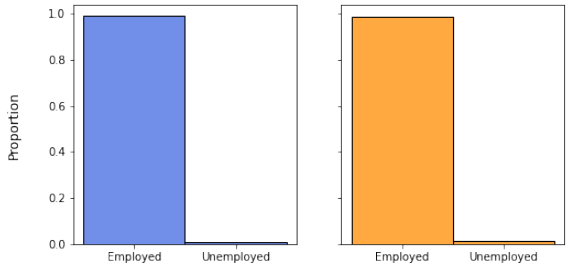



Figure 10: Counterfactual (blue) vs actual (orange) employment statuses in Montana from 2017 to 2020.

Consistent with findings in Courtemanche et al. (2017) and Mazurenko et al. (2018), we find significant first- and second order effects of Medicaid expansion. From Figures 9 and 10, "synthetic Montana" has much lower proportion of individuals insured under Medicaid, suggesting that expanding Medicaid eligibility directly affects the extensive margin of Medicaid enrollment. The disemployment effect is less pronounced in comparison to the enrollment effect we estimated, but nonetheless positive and nontrivial, consistent with the findings in, e.g., Peng et al. (2020), but inconsistent with those in, e.g., Gooptu et al. (2016). We do not find positive second-order effects, and the results we obtained are summarized in Figures 11 and 12. Additional details related to this application can be found in Appendix F.

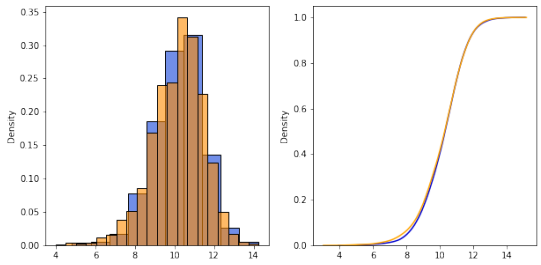



Figure 11: Counterfactual (blue) vs actual (orange) log wages in Montana from 2017 to 2020. Histogram of data distribution is shown on the left, and cumulative distribution function is shown on the right.

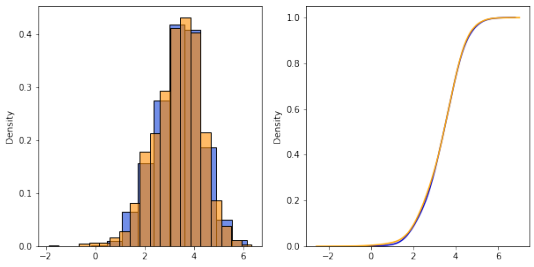



Figure 12: Counterfactual (blue) vs actual (orange) log labor hours supplied in Montana from 2017 to 2020. Histogram of data distribution is shown on the left, and cumulative distribution function is shown on the right.

## 6. Conclusion

We proposed a projection method between sets of probability measures supported on $\mathbb{R}^d$ based on the tangent cone structure of the 2-Wasserstein space. Our method seeks to best approximate some general target measure using some chosen set of control measures. In particular, it provides a global optimal solution that is unique conditional on fixing the optimal transport plans. It also demonstrably performs well compared to existing methods while being significantly more efficient in its implementation via a regression approach. Furthermore, it can be applied to general, that is, not necessarily regular, probability measures, in contrast to existing methods.

We derive statistical properties of the method when the respective measures are replaced with their empirical analogues. We also showcase the empirical performance of the method in several applications, the main one being in the field of causal inference where we generalize the concept of synthetic controls (Abadie et al., 2010; Abadie, 2021) to general multivariate probability measures. The proposed method still works without restricting optimal weights to be in the unit simplex, which would allow for extrapolation beyond the convex hull of the control units, providing a notion of tangential regression. It can also be extended to a continuum of measures, using established consistency results of barycenters (e.g. Le Gouic and Loubes, 2017).

## Acknowledgments

We would like to thank Sinho Chewi, Thibaut Le Gouic, and Philippe Rigollet for helpful discussions and comments. This research was supported in part by computational resources and services provided by Advanced Research Computing (ARC), a division of Information and Technology Services (ITS) at the University of Michigan, Ann Arbor. F. Gunsilius is supported by a MITRE grant from the University of Michigan.

## Appendix A. Existing Methods Based on Wasserstein Barycenters, and the Special Case of a Regular Target Measure

Existing methods such as Werenski et al. (2022) rely on the notion of "Karcher means" as mentioned in the main text. In that respect, the condition (4) is a sufficient condition for $\bar{P}(\lambda)$ to be a "Karcher mean" (Karcher, 2014) in $\mathcal{W}_2$ (Zemel and Panaretos, 2019). In fact, a "Karcher mean" of a set of measures $\{P_j\}_{j\in[\![J]\!]}$ is defined as the gradient of the Fréchet functional in $\mathcal{W}_2$ and is characterized through (4) holding $\bar{P}(\lambda)$-almost everywhere. (4) is a stronger condition because it is assumed to hold at every point in the support of $\bar{P}(\lambda)$, not just almost every point. Álvarez-Esteban et al. (2016) use this characterization to introduce a fixed-point approach to compute Wasserstein barycenters, and Werenski et al. (2022) use this structure to introduce a replication approach for absolutely continuous measures whose densities are bounded away from zero and whose target measure lies inside the convex hull of the control measures. Related is the recent definition of weak barycenters in Cazelles et al. (2021), where the authors replace the optimal transport maps from the classical optimal transport problem by the weak optimal transport problem introduced in Gozlan et al. (2017).

Heuristically, this characterization is that of a *deformable template*. A measure $P$ is a deformable template if there exists a set of deformations $\left\{\psi_j\right\}_{j=1,\dots,J}$ such that $\psi_{j\#}P = P_j$, in a way that their weighted average is "as close to the identity" as possible. In our setting $\psi_j \equiv \nabla\varphi_j - \mathrm{Id}$ (Anderes et al., 2015; Boissard et al., 2015; Yuille, 1991).

In our setting of interest, our tangential projection reduces to

$$\lambda^* \triangleq \underset{\lambda\in\Delta^J}{\operatorname{argmin}} \left\| \sum_{j=1}^{J} \lambda_j \left(\nabla\varphi_j - \mathrm{Id}\right) \right\|^2_{L^2(P_0)}, \tag{12}$$

where $\nabla\varphi_j$ are the optimal transport maps between the target $P_0$ and the control measures $P_j$, $j \in [\![J]\!]$. In contrast to Werenski et al. (2022) the target measure does not need to lie inside the convex hull of the other measures.

Based on these definitions we can show that our approach is a projection of the target $P_0$ onto $\mathfrak{Co}_{P_0}\left(\left\{P_j\right\}_{j=1}^J\right)$ in the case where $P_0$ is regular.

**Proposition 4** *Consider a regular target measure $P_0$ and a set $\{P_j\}_{j\in[\![J]\!]}$ of general control measures. Construct the measure $P_\pi$ as*

$$P_\pi \triangleq \exp_{P_0}\left(\sum_{j=1}^{J} \lambda_j^*(\nabla\varphi_j - \mathrm{Id})\right),$$

*where the optimal weights $\lambda^* \in \Delta^J$ are obtained by solving* (12) *and $\nabla\varphi_j$ are the optimal maps transporting $P_0$ to $P_j$, respectively. Then $P_\pi$ is the unique metric projection of $P_0$ onto $\mathfrak{Co}_{P_0}\left(\left\{P_j\right\}_{j=1}^J\right)$.*

## Appendix B. Proofs

**Proof** [Proof of Proposition 4] Define the following closed and convex subset $\mathcal{C} \subset L^2(P_0)$ for fixed optimal transportation maps between $P_0$ and $P_j$, denoted $\nabla\varphi_j$:

$$\mathcal{C} \triangleq \left\{ f \in L^2(P_0) : f = \sum_{j=1}^{J} \lambda_j \nabla\varphi_j \text{ for some } \lambda \in \Delta^J \right\} .$$

Recall that the transport maps $\nabla\varphi_j$ exist since $P_0$ is regular. Using $\mathcal{C}$, we can rewrite (12) as

$$\underset{\lambda\in\Delta^J}{\operatorname{argmin}} \left\| \sum_{j=1}^{J} \lambda_j \nabla\varphi_j - \mathrm{Id} \right\|^2_{L^2(P_0)} = \underset{f\in\mathcal{C}}{\operatorname{argmin}} \|f - \mathrm{Id}\|^2_{L^2(P_0)} \ ,$$

which by definition is the metric projection of Id onto $\mathcal{C}$. Since $\mathcal{C}$ is a non-empty closed and convex subset of the Hilbert space $L^2(P_0)$, this metric projection exists and is unique (Aliprantis and Border, 1999, Theorem 6.53). Moreover, if $\mathrm{Id} \in \mathcal{C}$, then $\pi_{\mathcal{C}} = \mathrm{Id}$; otherwise, $\pi_{\mathcal{C}} \in \partial\mathcal{C}$, where $\partial\mathcal{C}$ is the boundary of $\mathcal{C}$ (Aliprantis and Border, 1999, Lemma 6.54).

Since $P_0$ is regular, the exponential map is continuous. In fact, for every $j \neq k$,

$$W_2^2(P_j, P_k) = W_2^2((\nabla\varphi_j)_\# P_0, (\nabla\varphi_k)_\# P_0) \leq \int_{\mathbb{R}^d} \left|\nabla\varphi_j - \nabla\varphi_k\right|^2 \mathrm{d}P_0(x) \ .$$

In other words, the distance between $P_j$ and $P_k$ in $\mathcal{W}_2(\mathbb{R}^d)$ is smaller than that between corresponding elements $\nabla\varphi_j, \nabla\varphi_k$ in the tangent space. This implies continuity of the exponential map.

Furthermore, in this regular setting, the exponential map sends convex sets in $\mathcal{T}_{P_0}\mathcal{W}_2$ to generalized geodesically convex sets in $\mathcal{W}_2$. Mechanically, for any two (scaled) elements $t(\nabla\varphi_j - \mathrm{Id})$ and $s(\nabla\varphi_k - \mathrm{Id})$ in $\mathcal{T}_{P_0}\mathcal{W}_2$, and any $\rho \in [0,1]$,

$$\begin{aligned}
&\exp_{P_0}(\rho t(\nabla\varphi_j - \mathrm{Id}) + (1-\rho)s(\nabla\varphi_k - \mathrm{Id})) \\
&= \exp_{P_0}((\rho t\nabla\varphi_j + (1-\rho)s\nabla\varphi_k) - (\rho t + (1-\rho)s)\,\mathrm{Id}) \\
&= \exp_{P_0}\left(\widetilde{\rho}\left[\left[\frac{\rho t}{\widetilde{\rho}}\nabla\varphi_j + \frac{(1-\rho)s}{\widetilde{\rho}}\nabla\varphi_k\right] - \mathrm{Id}\right]\right) \\
&= \left(\left[\rho t\nabla\varphi_j + (1-\rho)s\nabla\varphi_k\right] + (1-\widetilde{\rho})\,\mathrm{Id}\right)_\# P_0 \\
&= \left(\left[\rho t(\nabla\varphi_j - \mathrm{Id}) + (1-\rho)s(\nabla\varphi_k - \mathrm{Id})\right] + \mathrm{Id}\right)_\# P_0 \ ,
\end{aligned}$$

where $\widetilde{\rho} \triangleq \rho t + (1-\rho)s$. This is a generalized geodesic connecting $P_j$ and $P_k$, via the optimal transport map between them and $P_0$ (Ambrosio et al., 2008, section 9.2). The same argument holds when extending generalized geodesics to generalized barycenters by taking convex combination of more measures than a binary interpolation with respect to $\rho$. Mechanically, for any $\lambda \in \Delta^J$ and $t_j > 0$ for all $j \in [\![J]\!]$,

$$\exp_{P_0}\left(\sum_{j=1}^{J} \lambda_j t_j(\nabla\varphi_j - \mathrm{Id})\right) = \exp_{P_0}\left(\sum_{j=1}^{J} \lambda_j t_j \nabla\varphi_j - \sum_{j=1}^{J} \lambda_j t_j\,\mathrm{Id}\right)$$

$$
\begin{aligned}
&= \exp_{P_0}\left(\widetilde{\rho}_J \left[\sum_{j=1}^{J} \widetilde{\rho}_J \nabla\varphi_j - \mathrm{Id}\right]\right) \\
&= \left(\left[\sum_{j=1}^{J} \lambda_j t_j \nabla\varphi_j\right] + (1-\widetilde{\rho}_J)\,\mathrm{Id}\right)_{\#} P_0 \\
&= \left(\left[\sum_{j=1}^{J} \lambda_j t_j (\nabla\varphi_j - \mathrm{Id})\right] + \mathrm{Id}\right)_{\#} P_0 \ ,
\end{aligned}
$$

where $\widetilde{\rho}_J \triangleq \sum_{j=1}^{J} \lambda_j t_j$. This proves the exponential map is generalized geodesically convex.

From above it follows that $P_\pi \triangleq \exp_{P_0}(\pi_C)$ is either in the interior of $\mathcal{C}$, which is the case if $\mathrm{Id} \in \mathcal{C}$, or on its boundary: since $\pi_C \in \partial C$, $\exp_{P_0}(\pi_C) \in \exp_{P_0}(\partial C)$. By continuity of the exponential map it follows that $\exp_{P_0}(\partial C) = \partial \exp_{P_0}(C)$. Combining all steps above show that $P_\pi$ is a *geodesic* metric projection of $P_0$ onto the geodesic convex hull of $\left\{P_j\right\}_{j=1}^{J}$. ■

**Proof** [Proof of Proposition 1] The result follows from the same argument as the proof of Proposition 4. Theorem 12.4.4 in Ambrosio et al. (2008) shows that $\mathcal{T}_{P_0}\mathcal{W}_2$ is the image of the barycentric projection of measures in the general tangent cone: $b_\gamma(x)$ is an optimal transport map if $\gamma$ is an optimal transport plan. But the exponential map satisfies

$$
\exp_{P_0}(v) = (v + \mathrm{Id})_{\#} P_0 \qquad \text{for all } v \in \mathcal{T}_{P_0}\mathcal{W}_2.
$$

This implies that

$$
\widetilde{P}_\pi \triangleq \exp_{P_0}\left(\sum_{j=1}^{J} \lambda_j^* b_{\gamma_{0j}} - \mathrm{Id}\right) = \left(\sum_{j=1}^{J} \lambda_j^* b_{\gamma_{0j}}\right)_{\#} P_0 \in \widetilde{\mathfrak{Co}}_{P_0}\left(\left\{P_j\right\}_{j=1}^{J}\right) \ ,
$$

since the convex combination of elements in the subgradients of convex functions lie in the subgradient of a convex function (provided the subgradient of each convex function is nonempty, which is the case here). Then the continuity and generalized convexity of the exponential map for elements in the regular tangent space $\mathcal{T}_{P_0}\mathcal{W}_2$ implies the result. ■

**Proof** [Proof of Proposition 2] We split the proof into two parts. In the first part we prove the convergence in probability of the family of objective functions (11) to their population counterparts (8) if the empirical measures $\mathbb{P}_{N_j}$ converge weakly in probability to the population measures $P_j$. In the second step we use the fact that $\widehat{\lambda}^*$ is a classical semiparametric estimator (Andrews, 1994; Newey and McFadden, 1994) to derive the convergence of the weights.

***Step 1: Convergence of the objective functions*** To show the convergence of the of the objective functions for obtaining the weights $\lambda^*$, we write

$$\begin{aligned}&\left|\left\|\sum_{j=1}^{J}\lambda_j b_{0j} - \mathrm{Id}\right\|^2_{L^2(P_0)} - \left\|\sum_{j=1}^{J}\lambda_j \widehat{b}_{0j} - \mathrm{Id}\right\|^2_{L^2(\mathbb{P}_{N_0})}\right| \\ &= \left|\int\left|\sum_{j=1}^{J}\lambda_j b_{0j}(x) - x\right|^2 \mathrm{d}P_0 - \int\left|\sum_{j=1}^{J}\lambda_j \widehat{b}_{0j}(x) - x\right|^2 \mathrm{d}\mathbb{P}_{N_0}\right| .\end{aligned}$$

We hence want to show that

$$\lim_{\bigwedge_j N_j\to\infty}\left|\int\left|\sum_{j=1}^{J}\lambda_j b_{0j}(x) - x\right|^2 \mathrm{d}P_0(x) - \int\left|\sum_{j=1}^{J}\lambda_j \widehat{b}_{0j}(x) - x\right|^2 \mathrm{d}\mathbb{P}_{N_0}(x)\right| = 0 ,$$

where $\bigwedge_j N_j \equiv \min\{N_0,\dots,N_J\}$.

We split the result into two parts. The first part shows that

$$\liminf_{\bigwedge_j N_j\to\infty}\int_{\mathbb{R}^d}\left|\sum_{j=1}^{J}\lambda_j \widehat{b}_{0j}(x_0) - x_0\right|^2 \mathrm{d}\mathbb{P}_{N_0}(x_0) \geq \int_{\mathbb{R}^d}\left|\sum_{j=1}^{J}\lambda_j b_{0j}(x_0) - x_0\right|^2 \mathrm{d}P_0(x_0).$$

In the second part we use the $L^2(P_0)$ convergence of the barycentric projections to prove that the limit exists and coincides with the limit inferior.

For the first part, we have

$$\begin{aligned}\liminf_{\bigwedge_j N_j\to\infty}\int_{\mathbb{R}^d}\left|\sum_{j=1}^{J}\lambda_j \widehat{b}_{0j}(x_0) - x_0\right|^2 \mathrm{d}\mathbb{P}_{N_0}(x_0) = \\ \liminf_{\bigwedge_j N_j\to\infty}\int_{(\mathbb{R}^d)^{J+1}}\left|\sum_{j=1}^{J}\lambda_j x_j - x_0\right|^2 \mathrm{d}\widehat{\boldsymbol{\gamma}}_N(x_0,x_1,\dots,x_J) ,\end{aligned}$$

where $\widehat{\boldsymbol{\gamma}}_N(x_0,x_1,\dots,x_J)$ is a measure that solves

$$\min\left\{\int_{(\mathbb{R}^d)^{J+1}}\sum_{j=1}^{J}\lambda_j\left|x_j - x_0\right|^2 \mathrm{d}\boldsymbol{\gamma} : \boldsymbol{\gamma}\in\Gamma_1(\widehat{\gamma}_{01},\dots,\widehat{\gamma}_{0J})\right\} ,$$

$\widehat{\gamma}_{0j}$ are the optimal couplings between $\mathbb{P}_{N_0}$ and $\widetilde{\mathbb{P}}_{N_j} \triangleq \left(\widehat{b}_{0j}\right)_{\#}\mathbb{P}_{N_0}$. Since all measures are defined on the complete and separable space $\mathbb{R}^d$, and by assumption of finite second moments, i.e.

$$\max_{j\in[\![J]\!]}\sup_{N_j}\int\left|x_j - x_0\right|^2 \mathrm{d}\widehat{\gamma}_{0j} < +\infty ,$$

it holds that each sequence $\widehat{\gamma}_{0j}$ is tight by Ulam's theorem (Dudley, 2018, Theorem 7.1.4). Using the fact that $\lambda \in \Delta^J$ and $\widehat{\boldsymbol{\gamma}}_N \in \Gamma_1\left(\widehat{\gamma}_{01}, \ldots, \widehat{\gamma}_{0J}\right)$, applying Jensen's inequality gives us

$$\max_{j \in [\![J]\!]} \sup_{N_j} \int_{\left(R^d\right)^{J+1}} \left| \sum_{j=1}^{J} \lambda_j x_j - x_0 \right|^2 \mathrm{d}\widehat{\boldsymbol{\gamma}}_N \leq \max_{j \in [\![J]\!]} \sup_{N_j} \sum_{j=1}^{J} \lambda_j \int_{\mathbb{R}^d} \left| x_j - x_0 \right|^2 \mathrm{d}\widehat{\gamma}_{0j} < +\infty \ ,$$

which implies that $\widehat{\boldsymbol{\gamma}}_N$ is tight. By Prokhorov's theorem, there exists a subsequence $\widehat{\boldsymbol{\gamma}}_{N_k}$ that weakly converges to a limit measure $\boldsymbol{\gamma}$. Therefore, by the continuity of the map $(x_0, x_j) \mapsto \sum_j \lambda_j x_j - x_0$, it follows from classical convergence results (Ambrosio et al., 2008, Lemma 5.1.12(d)) that

$$\liminf_{\bigwedge_j N_j \to \infty} \int_{(\mathbb{R}^d)^{J+1}} \left| \sum_{j=1}^{J} \lambda_j x_j - x_0 \right|^2 \mathrm{d}\widehat{\boldsymbol{\gamma}}_N(x_0, x_1, \ldots, x_J) = \int_{(\mathbb{R}^d)^{J+1}} \left| \sum_{j=1}^{J} \lambda_j x_j - x_0 \right|^2 \mathrm{d}\boldsymbol{\gamma}(x_0, \ldots, x_J) \ .$$

Furthermore, by the same argument via Jensen's inequality, i.e.,

$$\int_{(\mathbb{R}^d)^{J+1}} \left| \sum_{j=1}^{J} \lambda_j x_j - x_0 \right|^2 \mathrm{d}\boldsymbol{\gamma}(x_0, \ldots, x_J) \leq \sum_{j=1}^{J} \int_{(\mathbb{R}^d)^2} \left| \lambda_j x_j - x_0 \right|^2 \mathrm{d}\gamma_{0j}(x_0, x_j) < +\infty \ ,$$

it follows that the limit $\boldsymbol{\gamma} \in \Gamma_1\left(\gamma_{01}, \ldots, \gamma_{0J}\right)$ exists.

Now note that by the definition of disintegration it follows that (Ambrosio et al., 2008, Lemma 5.3.2)

$$\boldsymbol{\gamma} \in \Gamma_1(\gamma_{01}, \ldots, \gamma_{0J}) \qquad \Longleftrightarrow \qquad \boldsymbol{\gamma}_{x_0} \in \Gamma\left(\gamma_{1|x_0}, \ldots, \gamma_{J|x_0}\right) \ ,$$

where

$$\boldsymbol{\gamma} = \int \boldsymbol{\gamma}_{x_0} \, \mathrm{d}P_0(x_0) \qquad \text{and} \qquad \gamma_{0j} = \int \gamma_{j|x_0} \, \mathrm{d}P_0(x_0)$$

are the disintegrations of $\boldsymbol{\gamma}$ and $\gamma_{0j}$ with respect to $P_0$, respectively. Therefore, we have

$$\begin{aligned}
&\int_{(\mathbb{R}^d)^{J+1}} \left| \sum_{j=1}^{J} \lambda_j x_j - x_0 \right|^2 \mathrm{d}\boldsymbol{\gamma}(x_0, \ldots, x_J) \\
&= \int_{\mathbb{R}^d} \int_{\left(\mathbb{R}^d\right)^J} \left| \sum_{j=1}^{J} \lambda_j x_j - x_0 \right|^2 \mathrm{d}\boldsymbol{\gamma}_{x_0}(x_1, \ldots, x_J) \, \mathrm{d}P_0(x_0) \\
&\geq \int_{\mathbb{R}^d} \left| \int_{\left(\mathbb{R}^d\right)^J} \left( \sum_{j=1}^{J} \lambda_j x_j - x_0 \right) \mathrm{d}\boldsymbol{\gamma}_{x_0}(x_1, \ldots, x_J) \right|^2 \mathrm{d}P_0(x_0)
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^d} \left| \sum_{j=1}^{J} \lambda_j \int_{\left(\mathbb{R}^d\right)^J} x_j \, \mathrm{d}\boldsymbol{\gamma}_{x_0}(x_1, \ldots, x_J) - x_0 \right|^2 \mathrm{d}P_0(x_0) \\
&= \int_{\mathbb{R}^d} \left| \sum_{j=1}^{J} \lambda_j \int_{\mathbb{R}^d} x_j \, \mathrm{d}\gamma_{j|x_0}(x_j) - x_0 \right|^2 \mathrm{d}P_0(x_0) \\
&= \int_{\mathbb{R}^d} \left| \sum_{j=1}^{J} \lambda_j b_{0j}(x_0) - x_0 \right|^2 \mathrm{d}P_0(x_0) \ ,
\end{aligned}
$$

where the third lines follows from Jensen's inequality and the fifth line from the fact that $\boldsymbol{\gamma}_{x_0} \in \Gamma\left(\gamma_{1|x_0}, \ldots, \gamma_{J|x_0}\right)$. This shows the first part.

For the second part we use the fact that each barycentric projection $\widehat{b}_{0j}(x_1)$ is an optimal transport map between $\mathbb{P}_{N_0}$ and $\widetilde{\mathbb{P}}_{N_j}$ if $\widehat{\gamma}_{0j}$ is an optimal transport plan between $\mathbb{P}_{N_0}$ and $\mathbb{P}_{N_j}$, which follows from Theorem 12.4.4 in Ambrosio et al. (2008). As before, we know that $\left(\widehat{b}_{0j}\right)_{\#} \mathbb{P}_{N_0}$ is a tight sequence that converges to some $\widetilde{P}_j$. By definition and the fact that $\hat{b}_{0j}$ is the gradient of a convex function between $\mathbb{P}_{N_0}$ and $\widetilde{\mathbb{P}}_{N_j}$, $\widehat{b}_{0j}$ is the unique optimal transport map between $\mathbb{P}_{N_0}$ and $\widetilde{\mathbb{P}}_{N_j}$ for all $N_j$ and all $j$. Since the measures $P_j$ have finite second moments by assumption, we have

$$
\begin{aligned}
\limsup_{N_0 \wedge N_j \to \infty} \int_{\mathbb{R}^d} |x_j|^2 \, \mathrm{d}\widetilde{\mathbb{P}}_{N_j} &= \limsup_{N_0 \wedge N_j \to \infty} \int_{\mathbb{R}^d} \left| \widehat{b}_{0j}(x_0) \right|^2 \mathrm{d}\mathbb{P}_{N_0} \\
&= \limsup_{N_0 \wedge N_j \to \infty} \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} x_j \, \mathrm{d}\widehat{\gamma}_{j|x_0}(x_j) \right|^2 \mathrm{d}\mathbb{P}_{N_0} \\
&\leq \limsup_{N_0 \wedge N_j \to \infty} \int_{\left(\mathbb{R}^d\right)^2} \left| x_j \right|^2 \mathrm{d}\widehat{\gamma}_{0j}(x_0, x_j) \\
&= \int_{\left(\mathbb{R}^d\right)^2} \left| x_j \right|^2 \mathrm{d}\gamma_{0j}(x_0, x_j) < +\infty \ ,
\end{aligned}
$$

where the last equality follows from the tightness of $\widehat{\gamma}_{0j}$, as shown earlier. Therefore, by standard stability results for optimal transport maps (Segers, 2022; Panaretos and Zemel, 2020), it holds that $\widehat{b}_{0j}$ converges uniformly on every compact subset $K \subset \mathbb{R}^d$ in the support of the limit measure $\widetilde{P}_j$, that is

$$
\lim_{N_0 \wedge N_j \to \infty} \sup_{x_0 \in K} \left| \widehat{b}_{0j}(x_0) - v_j(x_0) \right| = 0 \ ,
$$

where $v_j$ is the optimal transport map between $P_0$ and $\widetilde{P}_j$.

We now show that $v_j = b_{0j}$ $P_0$-almost everywhere. From the local uniform convergence, we can then derive "strong $L^2$-convergence" (Ambrosio et al., 2008, Definition 5.4.3) of the potentials:

$$
\limsup_{N_0 \wedge N_j \to \infty} \left| \left\| \widehat{b}_{0j} \right\|_{L^2(\mathbb{P}_{N_0})} - \left\| v_j \right\|_{L^2(P_0)} \right|
$$

$$\leq \limsup_{N_0 \wedge N_j \to \infty} \left| \left\| \widehat{b}_{0j} \right\|_{L^2(\mathbb{P}_{N_0})} - \left\| v_j \right\|_{L^2(\mathbb{P}_{N_0})} \right| + \limsup_{N_0 \to \infty} \left| \left\| v_j \right\|_{L^2(\mathbb{P}_{N_0})} - \left\| v_j \right\|_{L^2(P_0)} \right|$$
$$\leq \limsup_{N_0 \wedge N_j \to \infty} \left\| \widehat{b}_{0j} - v_j \right\|_{L^2(\mathbb{P}_{N_0})} + \limsup_{N_0 \to \infty} \left| \left\| v_j \right\|_{L^2(\mathbb{P}_{N_0})} - \left\| v_j \right\|_{L^2(P_0)} \right|$$

Now the first term converges to zero by Hölder's inequality and the local uniform convergence of the optimal transport maps from above. The second term satisfies

$$\begin{aligned}
&\limsup_{N_0 \to \infty} \left| \left\| v_j \right\|_{L^2(\mathbb{P}_{N_0})} - \left\| v_j \right\|_{L^2(P_0)} \right| \\
=& \limsup_{N_0 \to \infty} \left| \left( \int_{\mathbb{R}^d} \left| v_j(x_0) \right|^2 \mathrm{d}\mathbb{P}_{N_0} \right)^{1/2} - \left( \int_{\mathbb{R}^d} \left| v_j(x_0) \right|^2 \mathrm{d}P_0 \right)^{1/2} \right| \\
\leq& \limsup_{N_0 \to \infty} \left| \int_{\mathbb{R}^d} \left| v_j(x_0) \right|^2 \mathrm{d}\mathbb{P}_{N_0} - \int_{\mathbb{R}^d} \left| v_j(x_0) \right|^2 \mathrm{d}P_0 \right|^{1/2} .
\end{aligned}$$

But since $P_0$ has finite second moments, it holds that this term also converges to zero.

Based on this we can show that $\widehat{\gamma}_{0j} \equiv \left( \mathrm{Id}, \widehat{b}_{0j} \right)$ converge weakly to $\gamma_{0j} \equiv \left( \mathrm{Id}, v_j \right)$. Indeed, if $\gamma_{0j}$ is a limit point of the sequence $\widehat{\gamma}_{0j}$, it holds that

$$\begin{aligned}
\int_{\left(\mathbb{R}^d\right)^2} \left| x_j \right|^2 \mathrm{d}\gamma_{0j}(x_0, x_j) &\leq \liminf_{N_0 \wedge N_j \to \infty} \int_{\left(\mathbb{R}^d\right)^2} \left| x_j \right|^2 \mathrm{d}\widehat{\gamma}_{0j}(x_0, x_j) \\
&\leq \limsup_{N_0 \wedge N_j \to \infty} \int_{\left(\mathbb{R}^d\right)^2} \left| x_j \right|^2 \mathrm{d}\widehat{\gamma}_{0j}(x_0, x_j) \\
&= \limsup_{N_0 \wedge N_j \to \infty} \int_{\mathbb{R}^d} \left| \widehat{b}_{0j}(x_0) \right|^2 \mathrm{d}\mathbb{P}_{N_0}(x_0) \\
&= \int_{\mathbb{R}^d} \left| v_j(x_0) \right|^2 \mathrm{d}P_0(x_0) .
\end{aligned}$$

Disintegrating the left-hand side with respect to $P_0$, and applying Jensen's inequality, gives

$$\begin{aligned}
\int_{\left(\mathbb{R}^d\right)^2} \left| x_j \right|^2 \mathrm{d}\gamma_{0j}(x_0, x_j) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left| x_j \right|^2 \mathrm{d}\gamma_{j|x_0}(x_j) \, \mathrm{d}P_0(x_0) \\
&\geq \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} x_j \, \mathrm{d}\gamma_{j|x_0}(x_j) \right|^2 \mathrm{d}P_0(x_0) = \int_{\mathbb{R}^d} \left| b_{0j}(x_0) \right|^2 \mathrm{d}P_0(x_0) ,
\end{aligned}$$

that is,

$$\int_{\mathbb{R}^d} \left| b_{0j}(x_0) \right|^2 \mathrm{d}P_0(x_0) \leq \int_{\mathbb{R}^d} \left| v_j(x_0) \right|^2 \mathrm{d}P_0(x_0) .$$

But since $v_j$ is an optimal transport map between $P_0$ and $\widetilde{P}_j$ by definition, it holds that

$$\int_{\mathbb{R}^d} \left| b_{0j}(x_0) \right|^2 \mathrm{d}P_0(x_0) \geq \int_{\mathbb{R}^d} \left| v_j(x_0) \right|^2 \mathrm{d}P_0(x_0) ,$$

which implies that equality holds and we have that

$$\int_{\mathbb{R}^d} \left[ \left| b_{0j}(x_0) \right|^2 - \left| v_j(x_0) \right|^2 \right] \mathrm{d}P_0(x_0) = 0 ,$$

which implies that $b_{0j} = v_j$ $P_0$-almost everywhere. We have hence shown that $\left(\mathrm{Id}, \widehat{b}_{0j}\right)_{\#} \mathbb{P}_{N_0}$ converges weakly to $\left(\mathrm{Id}, b_{0j}\right)_{\#} P_0$ for all $j$, where the barycentric projection $b_{0j}$ is the optimal transport map between $P_0$ and $\widetilde{P}_j$ (e.g. Villani, 2003, Theorem 2.12.(iii)).

Moreover, we have shown "strong $L^2$-convergence" of the barycentric projections in terms of Definition 5.4.3 in Ambrosio et al. (2008). Since this holds for all $j$, it also holds for their convex combination for fixed weights $\lambda \in \Delta^J$. Putting everything together, we then have that

$$\bigwedge_j \lim_{N_j \to \infty} \left\| \sum_{j=1}^J \lambda_j \widehat{b}_{0j} - \mathrm{Id} \right\|^2_{L^2(\mathbb{P}_{N_0})} = \left\| \sum_{j=1}^J \lambda_j b_{0j} - \mathrm{Id} \right\|^2_{L^2(P_0)} .$$

Since all observable measures $\mathbb{P}_j$ are empirical measures, they converge weakly in probability (Varadarajan, 1958), which implies that

$$\bigwedge_j \lim_{N_j \to \infty} P\left( \left| \left\| \sum_{j=1}^J \lambda_j \widehat{b}_{0j} - \mathrm{Id} \right\|^2_{L^2(\mathbb{P}_{N_0})} - \left\| \sum_{j=1}^J \lambda_j b_{0j} - \mathrm{Id} \right\|^2_{L^2(P_0)} \right| > \varepsilon \right) = 0 \quad \text{for all } \varepsilon > 0 .$$

This shows convergence in probability of the objective function for fixed $\lambda$.

***Step 2: Convergence of the optimal weights*** $\widehat{\lambda}^*_N$ The convergence of the optimal weights now follows from standard consistency results in semiparametric estimation. In particular, the objective functions are all convex for any $\lambda \in \mathbb{R}^J$, which implies that they converge uniformly on any compact set (Rockafellar, 1970, Theorem 10.8), so the objective function converges uniformly on $\Delta^J$. Now a standard consistency result like Theorem 2.1 in Newey and McFadden (1994) then implies that

$$\bigwedge_j \lim_{N_j \to \infty} P\left( \left| \widehat{\lambda}^*_N - \lambda^* \right| > \varepsilon \right) = 0 \qquad \text{for all } \varepsilon > 0 ,$$

which is what we wanted to show. Note that the result can also be shown if we allow the weights $\lambda$ to be negative, i.e., if we only require that $\sum_{j=1}^J \lambda_j = 1$. In this case, the fact that the objective functions are convex and coercive implies that an optimal $\lambda^*$ will be achieved at the interior of the extended Euclidean space, from which consistency follows by Theorem 2.7 in Newey and McFadden (1994). ■

**Proof** [Proof of Corollary 3] We want to show that $\left(\sum_{j=1}^J \hat{\lambda}^*_{N_j} \hat{b}_{0j}\right)_{\#} \mathbb{P}_{N_0}$ converges weakly in probability to $\left(\sum_{j=1}^J \lambda^*_j b_{0j}\right)_{\#} P_0$, where $\hat{\lambda}^*_N \triangleq \left(\hat{\lambda}^*_{N_1}, \ldots, \hat{\lambda}^*_{N_J}\right)$ are the optimal weights obtained in (11) and (8), respectively. The result follows by applying the extended continuous mapping theorem (van der Vaart and Wellner, 2013, Theorem 1.11.1) as follows.

As shown in the proof of Proposition 2 we have "strong $L^2$-convergence" of the maps $\sum_{j=1}^J \hat{\lambda}^*_{N_j} \hat{b}_{0j} - \mathrm{Id}$ to $\sum_{j=1}^J \lambda^*_j b_{0j} - \mathrm{Id}$. Therefore, by Theorem 5.4.4 (iii) in Ambrosio et al. (2008), it holds that

$$\lim_{\wedge_j N_j \to \infty} \int_{\mathbb{R}^d} f\left(x_0, \sum_{j=1}^J \hat{\lambda}^*_{N_j} \hat{b}_{0j}(x_0) - x_0\right) \mathrm{d}\mathbb{P}_{N_0}(x_0) = \int_{\mathbb{R}^d} f\left(x_0, \sum_{j=1}^J \lambda^*_j b_{0j}(x_0) - x_0\right) \mathrm{d}P_0(x_0)$$

for any continuous function such that $|f(x_0)| \leq C_1 + C_2 \left|\overline{x}_0 - x_0\right|^2$ for all $x_0$ in the support of $P_0$, where $C_1, C_2 < +\infty$ are some constants and $\overline{x}_0$ in some element in the support of $P_0$ (Ambrosio et al., 2008, equation (5.1.21)). In particular, this holds for any bounded and continuous function $f$, which implies that

$$\lim_{\wedge_j N_j \to \infty} \int_{\mathbb{R}^d} f\left(\sum_{j=1}^J \hat{\lambda}^*_{N_j} \hat{b}_{0j}(x_0)\right) \mathrm{d}\mathbb{P}_{N_0}(x_0) = \int_{\mathbb{R}^d} f\left(\sum_{j=1}^J \lambda^*_j b_{0j}(x_0)\right) \mathrm{d}P_0(x_0)$$

for any bounded and continuous function, which implies that $\left(\sum_{j=1}^J \hat{\lambda}^*_{N_j} \hat{b}_{0j}\right)_{\#} \mathbb{P}_{N_0}$ converges weakly to $\left(\sum_{j=1}^J \lambda^*_j b_{0j}\right)_{\#} P_0$ if $\mathbb{P}_{N_j}$ converge weakly to $P_j$, $j \in [\![J]\!]$.

Now we apply the extended continuous mapping theorem (van der Vaart and Wellner, 2013, Theorem 1.11.1). Equip $\mathscr{P}_2(\mathbb{R}^d)$ with any metric $\widetilde{d}(\cdot,\cdot)$ that metrizes weak convergence. We define the maps $g : \bigtimes_{j=0}^J \left(\mathscr{P}_2(\mathbb{R}^d), \widetilde{d}\right)_j \to \left(\mathscr{P}_2(\mathbb{R}^d), \widetilde{d}\right)$ by

$$g\left(P_0, \ldots, P_J\right) = \left(\sum_{j=1}^J \lambda^*_j b_{0j}\right)_{\#} P_0 \ ,$$

and analogously for their empirical counterparts $g_N$. Note that $g$ and $g_N$ are non-random functions if the measures $P_j$ and $\mathbb{P}_{N_j}$ are non-random themselves for all $j \in [\![J]\!]$. Moreover, by definition, $g$ and $g_N$ are continuous maps because $\sum_{j=1}^J \lambda^*_j b_{0j}$ are gradients of convex functions, which are continuous $P_0$-almost everywhere; the same thing holds for their empirical counterparts. Now from what we have shown above and in Proposition 2, it holds that

$$g_N\left(\mathbb{P}_{N_0}, \ldots, \mathbb{P}_{N_J}\right) \to g\left(P_0, \ldots, P_J\right)$$

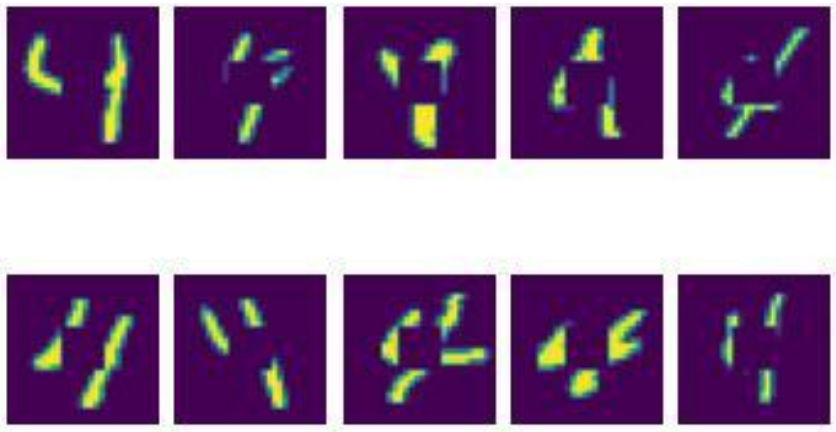
as $\mathbb{P}_{N_j}$ converge weakly to $P_j$. Since $\{\mathbb{P}_{N_j}\}_{j=1}^J$ here instead are the only random elements in $\bigtimes_{j=0}^J \left(\mathscr{P}_2(\mathbb{R}^d), \widetilde{d}\right)_j$, the extended continuous mapping theorem implies that

$$\lim_{\bigwedge_j N_j \to \infty} P\left(\widetilde{d}\left(g_N\left(\mathbb{P}_{N_0}, \ldots, \mathbb{P}_{N_J}\right), g\left(P_0, \ldots, P_J\right)\right) > \varepsilon\right) = 0 \qquad \text{for all } \varepsilon > 0 \ ,$$

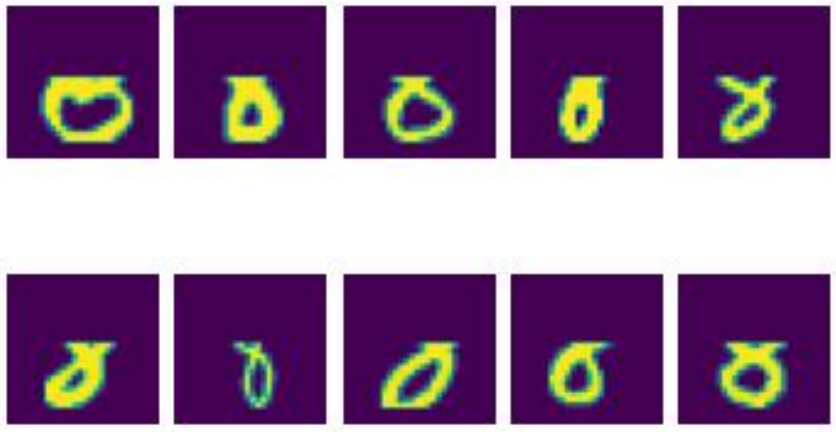
which is what we wanted to show. ■

## Appendix C. Additional Details of MNIST Experiment

We provide the control images used in the MNIST experiment described in Section 4.1 of the main text.



(a)



(b)

Figure 13: Control images used for MNIST experiment in Figure 2.

*Notes:* For Panel (a), the optimal weights from top left to bottom right are: (our method) (0.149, 0.025, 0, 0, 0.101, 0.369, 0.051, 0.101, 0.092, 0.109); (Werenski et al., 2022) (0.025, 0.015, 0, 0.103, 0.108, 0.260, 0, 0.268, 0, 0).
For panel (b), the optimal weights from top left to bottom right are: (our method) (0, 0.628, 0, 0, 0, 0, 0.142, 0.125, 0, 0.106); (Werenski et al., 2022) (0.284, 0.159, 0, 0, 0, 0.329, 0.248, 0.276, 0, 0).

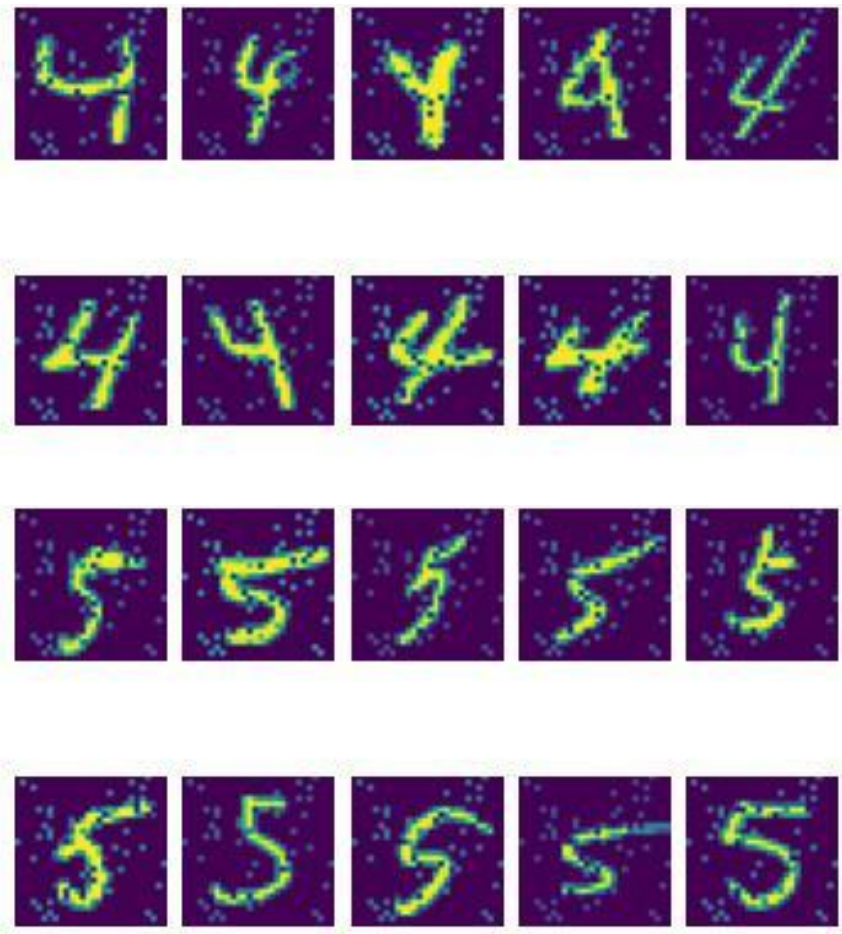


Figure 14: Control images used for MNIST experiment in Figure 3.

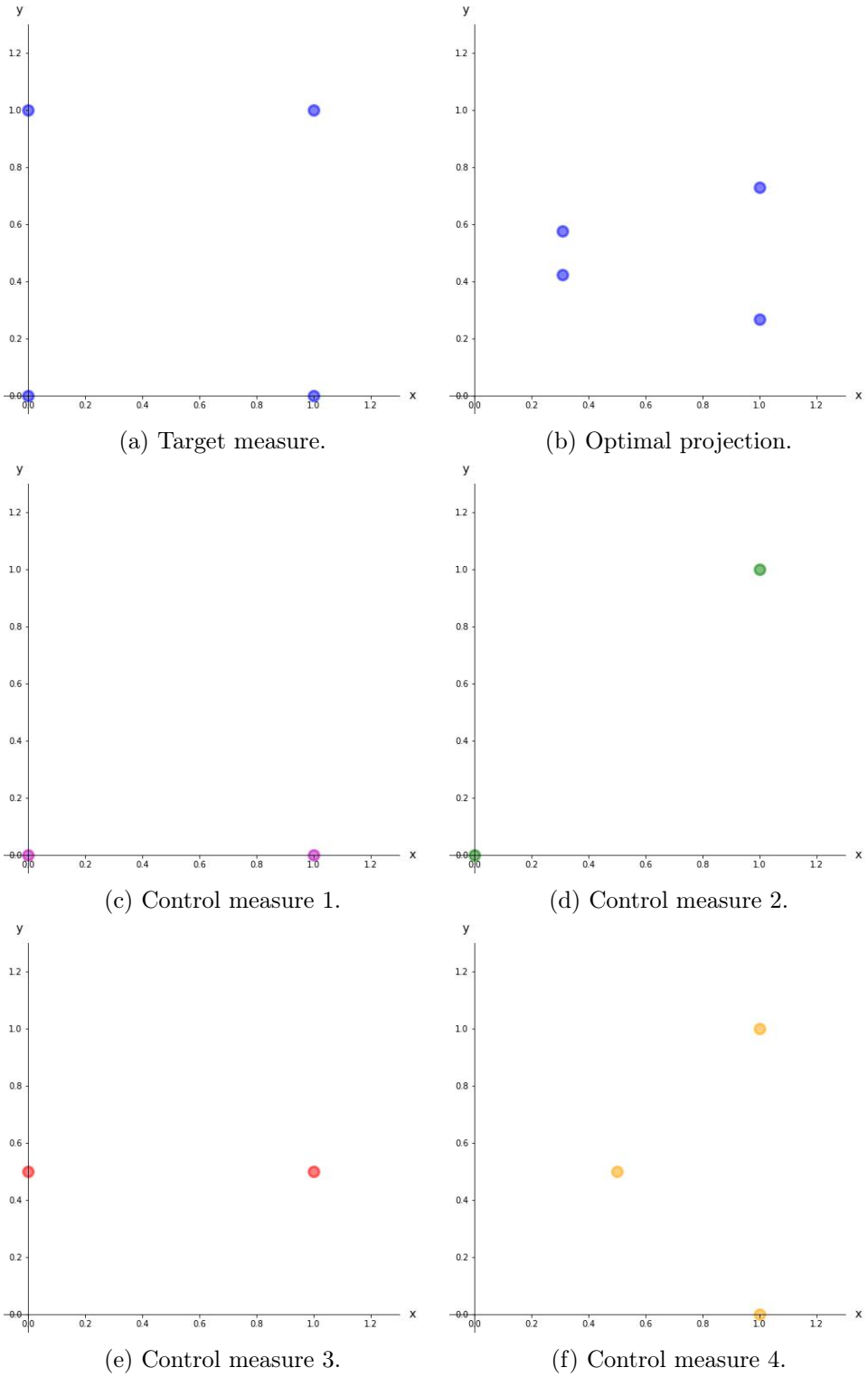
*Notes:* For Panel (a), the optimal weights from top left to bottom right are: (our method) (0.107, 0, 0, 0, 0.118, 0.388, 0.0340, 0.065, 0.1888, 0.095); (Werenski et al., 2022) (0.120, 0, 0, 0, 0, 0.657, 0, 0.223, 0, 0).
For panel (b), the optimal weights from top left to bottom right are: (our method) (0.109, 0.011, 0.057, 0.196, 0.058, 0.355, 0, 0.096, 0.082, 0.034); (Werenski et al., 2022) (0, 0.284, 0, 0.469, 0, 0.133, 0.113, 0, 0, 0)..

## Appendix D. Additional Simulation Results With $2D$ Empirical Distributions

To better illustrate how the projection method works, we illustrate an example involving very simply $2D$ empirical distributions, which consist of the following measures:

1. Target measure: supported with equal weights on 4 points in the unit cube $ABCD$, where $A = (0, 0)$, $B = (1, 0)$, $C = (1, 1)$, and $D = (0, 1)$.
2. Control measure 1: two-point measure supported on AB, with equal weights on $A$ and $B$.
3. Control measure 2: two point measure supported on $AC$, with equal weights on $A$ and $C$.
4. Control measure 3: two point measure supported on $\frac{1}{2}AD + \frac{1}{2}BC$, with equal weights on each of these midpoints.
5. Control measure 4: three point measure supported on one edge $BC$, and on the midpoint of the cube, with equal weights on $B$, $C$, and the said midpoint.

The target and control measures are illustrated in Figure 15. With this example, we illustrate the barycentric projection maps in Figure 16. This is an example where some optimal transport plans need not be unique. The method automatically picks one of them and the corresponding tangential projection is unique for these given transport plans.



(a) Target measure.

(b) Optimal projection.

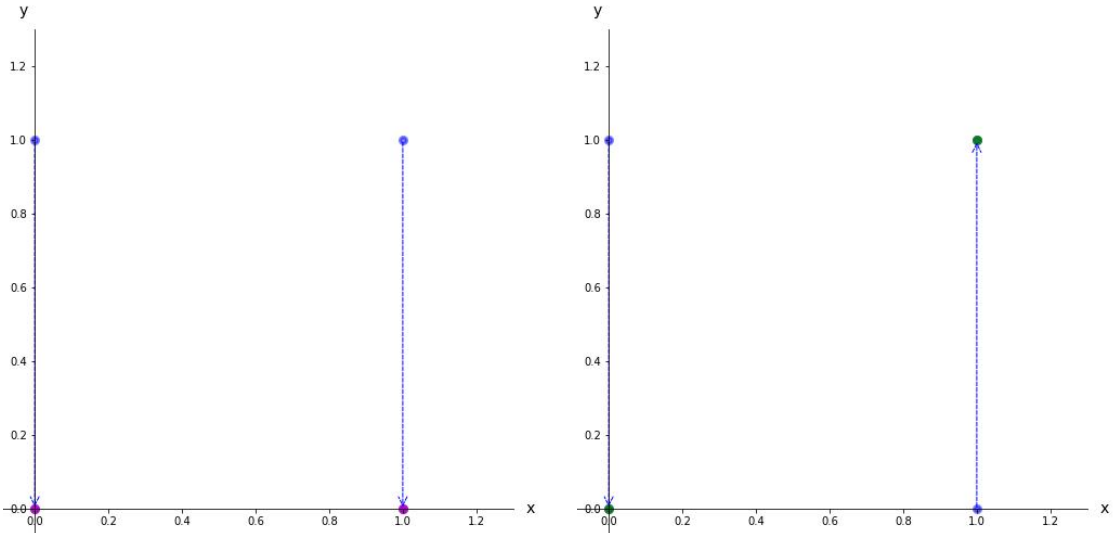
(c) Control measure 1.

(d) Control measure 2.

(e) Control measure 3.

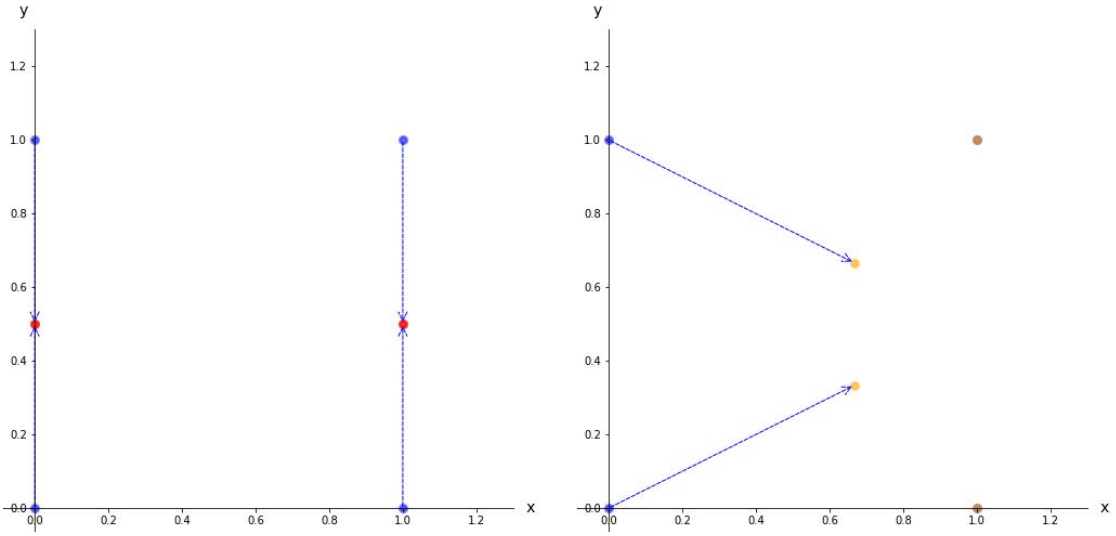
(f) Control measure 4.

Figure 15: Illustrations of target and control measures used in simulation, and the optimal projection obtained from our method.




(a) Barycentric projection associated with control measure 1, applied to the target measure.

(b) Barycentric projection associated with control measure 2, applied to the target measure.




(c) Barycentric projection associated with control measure 3, applied to the target measure.

(d) Barycentric projection associated with control measure 4, applied to the target measure.

Figure 16: Illustrations of barycentric maps associated with each control measure, applied to the target measure. The target and control measures of the experiment are illustrated in Figure 15.

## Appendix E. Runtime Complexity of the Proposed Method

As mentioned in the main text, the proposed method consists of two main steps: obtaining the optimal transport maps between the target and each of the $J$ control units, and then applying a regression constrained to the unit simplex to obtain the optimal weights. In the case where optimal maps do not exist, we estimate the optimal transport plans via barycentric projections, which are sums over columns of data matrices depending on the respective data samples of the measures. Assuming for simplicity that all measures have the same number of data points $N$, the complexity for this is $O(N^2 J)$. The exact solver of the optimal transport maps is of time complexity $O(JN^3 \log N)$ (Flamary et al., 2021).

The regression to obtain the optimal weights takes the form

$$\lambda^* = \underset{\lambda \in \Delta^J}{\operatorname{argmin}}\, f(\lambda) = \underset{\lambda \in \Delta^J}{\operatorname{argmin}} \sum_{n=1}^{N} \sum_{k=1}^{d} \left( \sum_{j=1}^{J} \lambda_j b_{\gamma_0 j,k}(x_n) - x_{nk} \right)^2 ,$$

There are many ways to solve this problem with different complexity dependencies. One way to solve it is via projected gradient descent. The gradient $\nabla f(\lambda)$ takes the form $\nabla f(\lambda) = \tilde{A}\lambda - \tilde{x}$, where $\tilde{A} \coloneqq (\tilde{a}_1, \cdots, \tilde{a}_J)^\top$ is the $J \times J$ matrix defined by the vectors $\tilde{a}'_k \coloneqq \big((A_1 \colon A_k), \ldots, (A_J \colon A_k)\big)$. Here, $A_j \coloneqq \left\{ \nabla \varphi^*_{jk}(x_n) \right\}_{k=1,\ldots,d, n=1,\ldots,N}$, is the $N \times d$ matrix for fixed $j$ and $A_j \colon A_k$ denotes the tensor double dot product $A_j \colon A_k \coloneqq \sum_{k=1}^{d} \sum_{n=1}^{N} A_j \circ A_k$, where $A_j \circ A_k$ denotes the Hadamard product between $A_j$ and $A_k$. Similarly, $\tilde{x}$ is the $J \times 1$ vector defined by stacking the scalar values $x_j \coloneqq X \colon A_j$, where $X \coloneqq \{x_{nk}\}_{n=1,\ldots,N, k=1,\ldots,d}$, is the $N \times d$ data matrix for $j = 0$.

Using the above approach, the complexity for the regression step consists of computing the matrix $\tilde{A}$ and the vector $\tilde{x}$ and solving the regression via projected gradient flow. The construction of the matrix requires $O(NdJ^2)$ computations. If solved via gradient descent, the complexity depends on the number of iterations $s > 0$. The overall complexity will be $O(NdJ^2 + J^2 s)$ in this case. If another numerical method is used to solve this regression problem, the complexity will be different.

To analyze runtime, we apply the method to an experiment with multivariate Gaussian distributions. The time complexity of depends on the number of observations in the target and control distributions (denoted $N$), and the number of controls in the control set (denoted $J$), since we work with the data points and not a grid approach. Thus we vary $N$ and $J$ while keeping the dimension $d$ fixed. We define $\mathbf{C}_{10} = I_{10} + 0.6 I_{-10}$, where $I_{10}$ denotes the $10 \times 10$ identity matrix and $I_{-10}$ denotes the $10 \times 10$ matrix that takes the value 1 everywhere *except* on the diagonal entries, and takes the value 0 on the diagonal. For the target and each of $J$ controls, we draw samples of size $N$ from the Gaussian distribution $\mathcal{N}(0, \mathbf{C}_{10})$.

Each entry in Table 1 is the runtime (in seconds) averaged over 2,000 iterations in seconds.

| | $J = 3$ | $J = 5$ | $J = 10$ |
|---|---|---|---|
| $N = 10$ | 0.003710 | 0.004332 | 0.006003 |
| $N = 100$ | 0.009927 | 0.014531 | 0.025691 |
| $N = 200$ | 0.024859 | 0.038397 | 0.072219 |
| $N = 300$ | 0.083083 | 0.133285 | 0.263616 |

Table 1: Runtimes (in seconds) averaged over 2,000 iterations.

## Appendix F. Details of Medicaid Expansion Application

Our implementation is available at the GitHub repository here. The images used can be found in the repository; the Medicaid data can be downloaded from the Dropbox folder here.

We use the ACS data with harmonized variables made available by IPUMS, a unified source of Census and survey data collected around the world. The data is at the household-person-year level. For our application, we select the household head and the spouse as our unit of analysis. The continuous outcomes are adjusted using the person-level sample weights available in the data.

We adopt the following sample restriction criteria: we included individuals

- of working age, i.e. between ages 18 and 65
- who has no missing outcomes (for those listed in the main text)
- who has no top-coded responses
- who are either household heads or their spouses

The sample size breakdown by states are follows:

| **State** | **Observations** |
|---|---|
| **Target** | |
| MT | 25,173 |
| **Control** | |
| AL | 106,464 |
| FL | 427,397 |
| GA | 227,659 |
| KS | 74,812 |
| MS | 61,505 |
| NC | 233,804 |
| SC | 107,905 |
| SD | 22,563 |
| TN | 152,470 |
| TX | 598,222 |
| WI | 157,410 |
| WY | 15,666 |

Table 2: Summary of the full data sample used to obtain $\lambda^*$.

We randomly select $N = 1500$ observations from each unit for estimating $\lambda^*$. In the Python implementation, if the entries of the target and control data are large enough, (8) becomes too large for `CVXPY` to compute an optimal solution numerically. Therefore, we introduce a stabilizing constant to prevent this. This stabilizing constant is determined by the mean value and dimensions of the target distribution, and the number of controls. The optimal weights we obtained are sparse and are displayed in Table 3.

| State | AL | FL | GA | KS | MS | NC | SC | SD | TN | TX | WI | WY |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Weight | 0.184 | 0 | 0 | 0 | 0.174 | 0 | 0.010 | 0.513 | 0 | 0 | 0.119 | 0 |

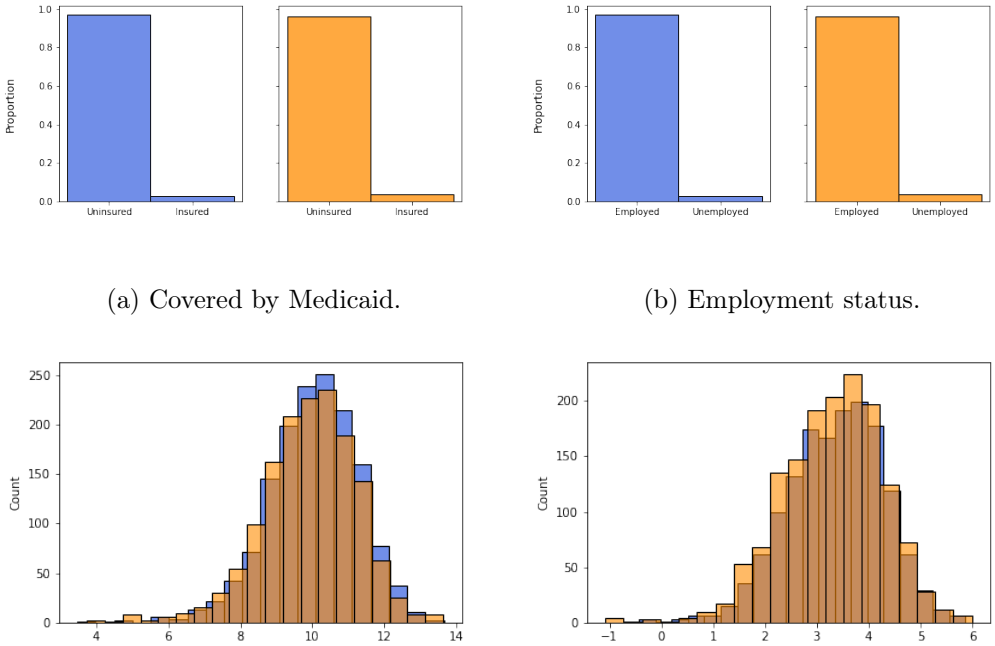
Table 3: Estimated Weights for Control States.

To obtain an estimate of a treatment effect, the replication of the target in pre-treatment periods and post-treatment periods needs to be comparable. Therefore, after obtaining the optimal weights $\lambda^*$, we compute the corresponding free barycenter

$$P(\lambda) = \underset{P \in \mathscr{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \sum_{j=1}^{J} \frac{\lambda_j^*}{2} W_2^2(P, P_j)$$

and check if this barycenter replicates the target distribution well. Implementation-wise, we computed the free-support barycenter, using the `POT` package (Flamary et al., 2021); this does not fix the support of the barycenter *a priori*, and allows it to be different from those of the control distributions. This is captured in Figures 17, and suggests that the barycenter constructed with these weights replicates the observed data well.

With the optimal weights $\lambda^*$, we estimate the counterfactual outcomes of interest for the four years after Medicaid expansion in Montana (namely, between 2017 and 2020). This involves solving the barycenter problem again, this time for post-intervention periods. We plot the densities and distributions of the counterfactual outcomes in Figures 11 and 12 of the main text.




(a) Covered by Medicaid.

(b) Employment status.

(c) Log wage.

(d) Log labor hours supplied.

Figure 17: Replicated (blue) vs actual (orange) Montana from 2010 to 2016.

| | Counterfactual | Actual |
|---|---|---|
| Log wages | 0.85 | 0.93 |
| Log hours worked | 1.55 | 1.74 |

Table 4: Variances of log labor and log hours worked across the post-intervention years.

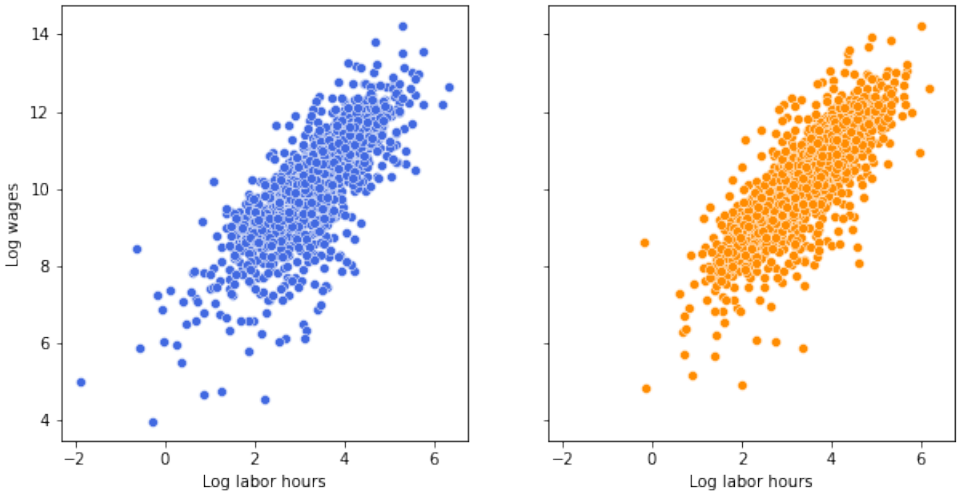
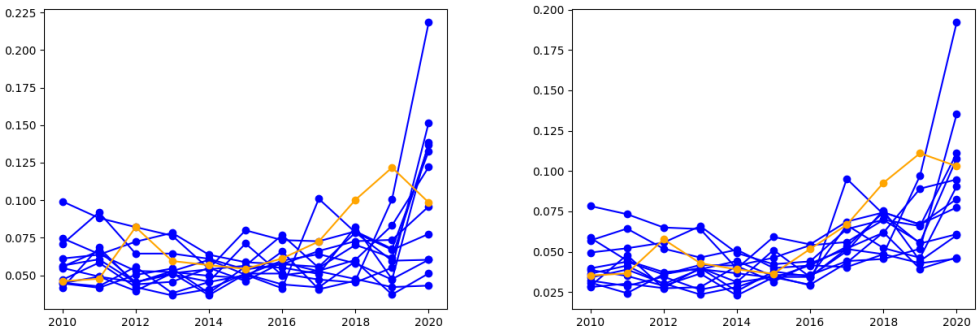



Figure 18: In blue: counterfactual Montana in the post-intervention periods. In orange: actual Montana in the post-intervention periods.

To perform inference on the estimated causal effect, we use a placebo permutation test in analogy to Abadie et al. (2010); Gunsilius (2023). The idea is to repeatedly apply the procedure described above to each control unit, pretending in turn each control unit is the treated unit. Post-intervention, if there exists an actual effect for the treated unit (Montana, in this application), then the estimated effect for the actual treatment unit should be among the largest.



(a) (b)

Figure 19: In orange: Montana. In blue: pretending each control state listed in Table 2 is a treated state. Panel (a) shows results obtained from using weights computed over all years in the pre-intervention period. Panel (b) shows results obtained from using weights computed from averaging weights obtained in each years in the pre-intervention period.

We plot the 2-Wasserstein distance between the treated, joint distribution of *all outcomes* and the pre-/post-intervention optimal projection (i.e. the barycenter problem with $\lambda^*$). We present two sets of results in Figure 19: in panel (a), the optimal projection is computed using $\lambda^*$ estimated using all years in the pre-intervention period; in panel (b), the $\lambda^*$ used is constructed from taking a simple average of weights estimated in each year of the pre-intervention period. Our results suggest that the estimated causal effect is valid in the post-intervention period, as we consistently observe the largest difference coming from Montana, especially from 2017-2019. The effect is less pronounced in 2020, however.

To accompany Figure 19, we also compute $p$-values, which we denote by and define as $p_t \triangleq \frac{r(d_{1t})}{J+1}$, where $d_{1t}$ is the 2-Wasserstein distance from the optimal projection to actual distribution when the target unit is Montana, $r(d_{1t})$ is the rank of $d_{1t}$ amongst $d_{jt}$s at given time $t$, and $J$ is the number of control units. Results are described in Table 5. A smaller $p_t$ value indicates larger treatment effect. We observed that $r(d_{1t}) = 1$ for 2018 and 2019, implying a nontrivial effect of the Medicaid expansion in Montana during these years. The values are $p_t$ are comparably higher in 2017 and 2020, which we attribute to the fact that it was the first year of the policy implementation, and the COVID-19 pandemic, respectively.

| Year ($t$) | $p_t$ (Weights Using All Years) | $p_t$ (Averaged Weights Over All Years) |
|---|---|---|
| 2017 | 0.231 | 0.308 |
| 2018 | 0.077 | 0.077 |
| 2019 | 0.077 | 0.077 |
| 2020 | 0.535 | 0.385 |

Table 5: Estimated $p_t \triangleq \frac{r(d_{1t})}{J+1}$ in the post-intervention period.

## References

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021.

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Anish Agarwal, Keegan Harris, Justin Whitehouse, and Zhiwei Steven Wu. Adaptive principal component regression with applications to panel data. *arXiv preprint arXiv:2307.01357*, 2023.

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein Space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, January 2011. ISSN 0036-1410, 1095-7154. doi: 10.1137/100805741. URL `http://epubs.siam.org/doi/10.1137/100805741`.

Miklós Ajtai, János Komlós, and Gábor Tusnády. On optimal matchings. *Combinatorica*, 4 (4):259–264, 1984.

Aleksandr Danilovich Aleksandrov. A theorem on triangles in a metric space and some of its applications. *Trudy Matematicheskogo Instituta imeni VA Steklova*, 38:5–23, 1951.

Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer, Berlin ; New York, 2nd, completely rev. and enl. ed edition, 1999. ISBN 978-3-540-65854-2.

Pedro C Álvarez-Esteban, E Del Barrio, JA 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.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in mathematics ETH Zürich. Birkhäuser, Basel, 2. ed edition, 2008. ISBN 978-3-7643-8721-1.

Ethan Anderes, Steffen Borgwardt, and Jacob Miller. Discrete Wasserstein Barycenters: Optimal Transport for Discrete Data. *arXiv:1507.07218 [math]*, August 2015. URL `http://arxiv.org/abs/1507.07218`. arXiv: 1507.07218.

Donald Andrews. Asymptotics for semiparametric econometric models: Iii. testing and examples. Technical report, Cowles Foundation for Research in Economics, Yale University, 1989.

Donald WK Andrews. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 43–72, 1994.

Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5):1–26, 2015.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

Jushan Bai and Serena Ng. Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763, 2021.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the Wasserstein space. *arXiv:1307.7721 [math, stat]*, October 2014. URL `http://arxiv.org/abs/1307.7721`. arXiv: 1307.7721.

Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. Distribution's template estimate with wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):1–10, July 2016. ISSN 0730-0301, 1557-7368. doi: 10.1145/2897824.2925918. URL `https://dl.acm.org/doi/10.1145/2897824.2925918`.

Lea Bottmer, Guido W Imbens, Jann Spiess, and Merrill Warnick. A design-based perspective on synthetic control methods. *Journal of Business & Economic Statistics*, pages 1–12, 2023.

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

Guillaume Carlier and Ivar Ekeland. Matching for teams. *Economic theory*, 42(2):397–418, 2010.

Elsa Cazelles, Vivien Seguy, Jérémie Bigot, Marco Cuturi, and Nicolas Papadakis. Log-pca versus geodesic pca of histograms in the wasserstein space. *arXiv preprint 1708.08143*, 2017.

Elsa Cazelles, Felipe Tobar, and Joaquin Fontbona. A novel notion of barycenter for probability distributions based on optimal weak mass transport. *Advances in Neural Information Processing Systems*, 34, 2021.

Jiafeng Chen. Synthetic control as online linear regression. *Econometrica*, 91(2):465–491, 2023.

Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, pages 1–14, 2021.

Alexander Cloninger, Keaton Hamm, Varun Khurana, and Caroline Moosmüller. Linearized wasserstein dimensionality reduction with approximation guarantees. *arXiv preprint arXiv:2302.07373*, 2023.

Charles Courtemanche, James Marton, Benjamin Ukert, Aaron Yelowitz, and Daniela Zapata. Early impacts of the affordable care act on health insurance coverage in medicaid expansion and non-expansion states. *Journal of Policy Analysis and Management*, 36(1): 178–210, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34, 2021.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Richard M Dudley. *Real analysis and probability*. CRC Press, 2018.

Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Jiaojiao Fan and David Alvarez-Melis. Generating synthetic datasets by interpolating along generalized geodesics. *arXiv preprint arXiv:2306.06866*, 2023.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Alfred Galichon and Bernard Salanié. Matching with trade-offs: Revealed preferences over competing characteristics, 2010. CEPR Discussion Paper No. DP7858.

Wilfrid Gangbo and Andrzej Święch. Optimal maps for the multidimensional monge-kantorovich problem. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 51(1):23–45, 1998.

Laya Ghodrati and Victor M Panaretos. Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 01 2022. asac005.

Angshuman Gooptu, Asako S Moriya, Kosali I Simon, and Benjamin D Sommers. Medicaid expansion did not result in significant employment changes or job reductions in 2014. *Health affairs*, 35(1):111–118, 2016.

Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, and Prasad Tetali. Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis*, 273(11): 3327–3405, 2017.

Florian F Gunsilius. Distributional synthetic controls. *Econometrica*, 91(3):1105–1117, 2023.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

Matt Jacobs and Flavien Léger. A fast approach to optimal transport: The back-and-forth method. *Numerische Mathematik*, 146(3):513–544, 2020.

Hermann Karcher. Riemannian center of mass and so called karcher mean. *arXiv preprint arXiv:1407.2087*, 2014.

Benoît Kloeckner. A geometric study of wasserstein spaces: Euclidean spaces. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 9(2):297–323, 2010.

Soheil Kolouri, Akif B Tosun, John A Ozolek, and Gustavo K Rohde. A continuous linear optimal transport approach for pattern analysis in image datasets. *Pattern recognition*, 51:453–462, 2016.

Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of wasserstein barycenters. *Probability Theory and Related Fields*, 168(3):901–917, 2017.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

J Steve Marron and Andrés M Alonso. Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753, 2014.

Olena Mazurenko, Casey P Balio, Rajender Agarwal, Aaron E Carroll, and Nir Menachemi. The effects of medicaid expansion under the aca: a systematic review. *Health Affairs*, 37 (6):944–950, 2018.

Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

Quentin Mérigot, Alex Delalande, and Frédéric Chazal. Quantitative stability of optimal transport maps and linearization of the 2-wasserstein space. In *International Conference on Artificial Intelligence and Statistics*, pages 3186–3196. PMLR, 2020.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

Victor M. Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. SpringerBriefs in Probability and Mathematical Statistics. SpringerOpen, Cham, 2020. ISBN 978-3-030-38437-1. doi: 10.1007/978-3-030-38438-8.

Matteo Pegoraro and Mario Beraha. Fast pca in 1-d wasserstein spaces via b-splines representation and metric projection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9342–9349, 2021.

Lizhong Peng, Xiaohui Guo, and Chad D Meyerhoefer. The effects of medicaid expansion on labor market outcomes: evidence from border counties. *Health economics*, 29(3):245–260, 2020.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

R Tyrrell Rockafellar. *Convex Analysis*, volume 36. Princeton University Press, 1970.

Lars Ruthotto, Stanley J Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung. A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences*, 117(17):9183–9193, 2020.

Johan Segers. Graphical and uniform consistency of estimated optimal transport plans. *arXiv preprint: 2208.02508*, 2022.

Asuka Takatsu and Takumi Yokota. Cone structure of $l2$-wasserstein spaces. *Journal of Topology and Analysis*, 4(02):237–253, 2012.

Michel Talagrand. Matching random samples in many dimensions. *The Annals of Applied Probability*, pages 846–856, 1992.

Michel Talagrand. The transportation cost from the uniform measure to the empirical measure in dimension $\geq 3$. *The Annals of Probability*, pages 919–959, 1994.

Aad van der Vaart and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Veeravalli S Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics*, 19(1):23–26, 1958.

Cédric Villani. *Topics in optimal transportation*. Number v. 58 in Graduate studies in mathematics. American Mathematical Society, Providence, RI, 2003. ISBN 978-0-8218-3312-4.

Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101:254–269, 2013.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Mattthew E Werenski, Ruijie Jiang, Abiy Tasissa, Shuchin Aeron, and James M Murphy. Measure estimation in the barycentric coding model. In *International Conference on Machine Learning*, pages 23781–23803. PMLR, 2022.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

Alan L. Yuille. Deformable Templates for Face Recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70, January 1991. ISSN 0898-929X, 1530-8898. doi: 10.1162/jocn.1991.3.1.59. URL `https://direct.mit.edu/jocn/article/3/1/59/3023/Deformable-Templates-for-Face-Recognition`.

Yoav Zemel and Victor M Panaretos. Fréchet means and procrustes analysis in wasserstein space. *Bernoulli*, 25(2):932–976, 2019.