TANGENTIAL WASSERSTEIN PROJECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a notion of projections between sets of probability measures using the geometric properties of the 2-Wasserstein space. It is designed for general multivariate probability measures, is computationally efficient to implement, and provides a unique solution in regular settings. The idea is to work on regular tangent cones of the Wasserstein space using generalized geodesics. Its structure and computational properties make the method applicable in a variety of settings, from causal inference to the analysis of object data. An application to estimating causal effects yields a generalization of the notion of synthetic controls for systems with general heterogeneity described via multivariate probability measures, as well as a way to estimate optimal weights jointly over all time periods.

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 statistics. 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 & Alonso, 2014) to causal inference with individual heterogeneity (e.g. Athey & Imbens, 2015).

We introduce a notion of projection between sets of probability measures supported on Euclidean spaces. The proposed definition is applicable between sets of general probability measures with different supports and possesses good computational and statistical properties. It also provides a unique solution to the projection problem under mild conditions and can replicate the geometric properties of the target measure, such as its shape and support. 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 $\mathcal{W}_2$, where the dimension d can be high. The multivariate setting poses particular 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. This is coupled with a well-known statistical curse of dimensionality for general measures (Ajtai et al., 1984, Dudley, 1969, Fournier & Guillin, 2015, Talagrand, 1992; 1994, Weed & Bach, 2019).

These challenges have impeded the development of a method of projections between 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 & Panaretos (2022), and Pegoraro & Beraha (2021) introduced frameworks for distribution-on-distribution regressions in the univariate setting for object data. Bigot et al. (2014), Cazelles et al. (2017) developed principal component analyses on the space of univariate probability measures using geodesics on the Wasserstein space.

The most closely related works to ours are Bonneel et al. (2016) and Werenski et al. (2022). The former develops a regression approach in barycentric coordinates with applications in computer graphics as well as color and shape transport problems. Their method requires solving a computationally costly bilevel optimization problem, which does not need to achieve a global solution.

The latter works on a tangential structure like we do, but is based on "Karcher means" (Karcher, 2014, Zemel & Panaretos, 2019). This implies that their method works between absolutely continuous measures with densities that are bounded away from zero, with the target measure lying in the convex hull of the control measures.

The notion of projection we propose in this article circumvents these challenges in the multivariate setting by lifting the projection problem to the regular tangent space to $\mathcal{W}_2$ at the target measure based on generalized geodesics. These tangent spaces exist when the target measure is regular, i.e., if it does not give mass to sets of lower Hausdorff dimension. For general measures, we construct a regular tangent space using barycentric projections (Ambrosio et al., 2008, appendix 12). In contrast to the existing approaches, our method works for general probability measures, allows for the target measure to be outside the generalized geodesic convex hull of the control measures, and can be implemented by a standard constrained linear regression. The method hence transforms the projection problem on the positively curved Wasserstein space into a linear optimization problem in the regular tangent space, which provides a unique solution to the projection problem under mild assumptions. This problem takes the form of a deformable template (Boissard et al., 2015, Yuille, 1991), which connects our approach to this literature. The method can be implemented in three steps: (i) obtain the general tangent cone structure at the target measure, (ii) construct a regular tangent space if it does not exist, and (iii) perform a linear regression to carry out the projection in the tangent space.

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 & Brenier (2000), Jacobs & Léger (2020), Makkuva et al. (2020), Peyré & Cuturi (2019), Ruthotto et al. (2020) and references therein. Other alternatives compute approximations of the optimal transport plans via regularized optimal transport problems (Peyré & Cuturi, 2019), such as entropy regularized optimal transport (Galichon & Salanié, 2010, Cuturi, 2013). The proposed projection approach is compatible with any such method. 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 extend the classical synthetic control estimator (Abadie & Gardeazabal, 2003, Abadie et al., 2010) to settings with observed individual heterogeneity in multivariate outcomes. This lets us perform the synthetic control method on the joint distribution of several outcomes and also estimate one set of optimal weights over all pre-intervention time periods, which complements the recently introduced method in Gunsilius (2022) designed for univariate outcomes. We apply this synthetic controls estimator to estimate the causal effect of Medicaid expansion on the population in individuals states and find that the policy induces nontrivial, positive effects on Medicaid enrolment, earnings, and labor supply, while its effect on employment agrees with some previous estimates, but less in magnitude compared to the other effects we estimated.

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 $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$, respectively, the 2-Wasserstein distance $W_2(P_X, P_Y)$ is defined as

$$W_2(P_X, P_Y) \triangleq \left(\min_{\gamma \in \Gamma(P_X, P_Y)} \int_{\mathcal{X} \times \mathcal{Y}} |x - y|^2 \, \mathrm{d}\gamma(x, y) \right)^{\frac{1}{2}}. \tag{2.1}$$

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 π_1 and π_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 $A \subseteq \mathcal{Y}$, $T_\# P(A) \equiv P(T^{-1}(A))$. An optimal coupling $\gamma \in \Gamma(P_X, P_Y)$ solving the

optimal transport problem equation 2.1 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 γ solving equation 2.1 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. In the following we will distinguish between regular and general measures.

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 γ 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 $d > 1$ (Ambrosio et al., 2008, Theorem 7.3.2) and flat for $d = 1$ (Kloeckner, 2010), where curvature is defined in the sense of Aleksandrov (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 Projections in $\mathcal{W}_2$, barycenters, and existing challenges

In linear spaces, the idea of projecting an element on a set of other elements is a linear or convex combination of these elements in the set. The analogues of averages in vector spaces are Fréchet means or barycenters in metric spaces. For Wasserstein spaces, this concept has been introduced in our setting in Agueh & Carlier (2011) and in a more abstract sense in Carlier & Ekeland (2010). For any collection of probability measures $\{P_j\}_{1 \leqslant j \leqslant J} \subseteq \mathcal{W}_2$, their weighted barycenter $\bar{P}(\lambda)$ for given weights $\lambda \equiv (\lambda_1, \ldots, \lambda_J) \in \Delta^J$ is a solution of the following minimization problem:

$$\bar{P}(\lambda) \in \operatorname*{arg\,min}_{P \in \mathscr{P}_2(\mathbb{R}^d)} \sum_{j=1}^{J} \frac{\lambda_j}{2} W_2^2(P, P_j). \tag{2.2}$$

The weights λ are defined to lie in the J-dimensional probability simplex Δ^J of nonnegative vectors in $\mathbb{R}^J$ that sum to unity. Prokhorov's theorem implies that a solution to equation 2.2 exists (Agueh & Carlier, 2011, Proposition 2.3).

Throughout, we focus on the case where $\lambda \in \Delta^J$, because this provides a natural notion of *interpolation* between the given measures P_j. We could also construct a linear projection that corresponds to a "geodesic *extrapolation*" by relaxing the requirement that all weights λ_j need to be non-negative. In the classical setting of random variables mapping to some Euclidean space and not random measures, this relaxation is the natural analogue to a linear regression, see Abadie et al. (2015) for a discussion. We focus on the interpolation setting because it is in line with the notion of a projection onto the convex hull spanned by other elements. All of our results can be extended to the extrapolation setting in principle.

One way to extend the notion of projection between a target probability measure P_0 and a set of control measures $\{P_j\}_{j=1,\ldots,J}$ in $\mathcal{W}_2$ would hence consist in finding the optimal weights $\lambda^* \in \Delta^J$ for which an induced barycenter $\bar{P}(\lambda^*)$ solving equation 2.2 is as close as possible to P_0. This would lead to the following bi-level optimization problem, assuming that the barycenter $\bar{P}(\lambda)$ is unique for given λ:

$$\lambda^* \in \operatorname*{arg\,min}_{\lambda \in \Delta^J} W_2(P_0, \bar{P}(\lambda)), \qquad \text{where} \quad \bar{P}(\lambda) = \operatorname*{arg\,min}_{P \in \mathscr{P}_2(\mathbb{R}^d)} \sum_{j=1}^{J} \frac{\lambda_j}{2} W_2^2(P, P_j). \tag{2.3}$$

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 with this approach are mathematical and computational. Importantly, the optimal weights λ^* 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 & 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 & Cuturi, 2019) that can be implemented in low-dimensional settings.

2.3 Tangential structures on $\mathcal{W}_2$

We circumvent these difficulties by exploiting a tangential structure that can be defined on $\mathcal{W}_2$ (Ambrosio et al., 2008, Otto, 2001). In particular, it allows us to entirely circumvent solving a bi-level optimization problem as the one in equation 2.3.

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

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

with respect to the local distance

$$W_P^2(\gamma_{12}, \gamma_{13}) \triangleq \min\left\{\int_{(\mathbb{R}^d)^3} |x_2 - x_3|^2 \,\mathrm{d}\gamma_{123} : \gamma_{123} \in \Gamma_1(\gamma_{12}, \gamma_{13})\right\}, \tag{2.4}$$

where γ_{12} and γ_{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 γ_{123} such that the projection of γ_{123} onto the first two elements is γ_{12} and the projection onto the first and third element is γ_{13} (Ambrosio et al., 2008, Appendix 12). 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 the case where P is absolutely continuous with respect to Lebesgue measure, the definition simplifies. The corresponding optimal transport plan between P and any other measure measure $Q \in \mathcal{W}_2$ is then supported on the graph of the gradient of a convex function $\nabla\varphi$ by Brenier's theorem. This allows the introduction of a regular tangent cone at P, defined via (Ambrosio et al., 2008, section 8.4)

$$\mathcal{T}_P\,\mathcal{W}_2(\mathbb{R}^d) \triangleq \overline{\left\{t(\nabla\,\varphi_j - \mathrm{Id}) : \left(\mathrm{Id} \times \nabla\,\varphi_j\right)_\#\,P \text{ is optimal in } \Gamma(P, (\nabla\varphi_j)_\#P), \;\; t > 0\right\}}^{L^2(P)}, \tag{2.5}$$

where $\overline{A}^{L^2(P)}$ defines the closure of the set A with respect to the distance induced by the L^2-norm on P. Interestingly, this tangent cone is actually a linear tangent space (Ambrosio et al., 2008, Theorem 8.5.1). In the following, we therefore call $\mathcal{T}_P\,\mathcal{W}_2$ the regular tangent space equation 2.5.

2.4 Tangential Wasserstein projections for regular target measures

We want to define a natural analogue to the barycenter projection problem equation 2.3 in the regular tangent space of the Wasserstein space. 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 & 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

$$\sum_{j=1}^{J} \lambda_j \left(\nabla\,\varphi_j - \mathrm{Id}\right) = 0, \tag{2.6}$$

where $\{\varphi_j\}_{j\in[\![J]\!]}$ are the optimal transport maps from the barycenter to the respective measure P_j, i.e. $(\varphi_j)_\#\bar{P}(\lambda) = P_j$. Each term of the summand in equation 2.6 is an element in $\mathcal{T}_{\bar{P}(\lambda)}\,\mathcal{W}_2(\mathbb{R}^d)$ by construction.

More generally, the condition equation 2.6 is a sufficient condition for $\bar{P}(\lambda)$ to be a "Karcher mean" (Karcher, 2014) in $\mathcal{W}_2$ (Zemel & 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 equation 2.6 holding $\bar{P}(\lambda)$-almost everywhere. equation 2.6 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 $\{\psi_j\}_{j=1,\dots,J}$ 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, we are given a target measure P_0 that we want to replicate given a set $\{P_j\}_{j\in[\![J]\!]}$. The key idea for this is to adapt the characterization equation 2.6 *with respect to the target measure* P_0. That is, we aim to find the optimal weights $\lambda^* \in \Delta^J$ that satisfy

$$\lambda^* \triangleq \underset{\lambda\in\Delta^J}{\arg\min} \left\| \sum_{j=1}^{J} \lambda_j \left(\nabla\varphi_j - \mathrm{Id}\right) \right\|^2_{L^2(P_0)}, \tag{2.7}$$

where $\nabla\varphi_j$ are the optimal transport maps between the target P_0 and the control measures P_j, $j \in [\![J]\!]$.

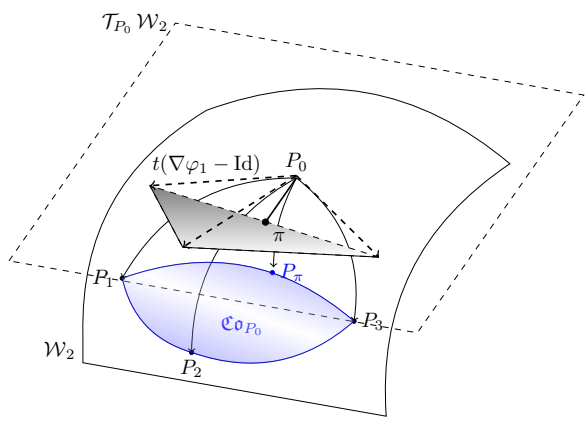


Figure 1: Tangential Wasserstein projection for a regular target P_0.

Thick dashed lines are tangent vectors $t(\nabla\varphi_j - \mathrm{Id})$. The gray shaded region is their convex hull in $\mathcal{T}_{P_0}\mathcal{W}_2$ and π 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 $\mathfrak{Co}_{P_0}\left(\{P_1, P_2, P_3\}\right) \subseteq \mathcal{W}_2$ (blue).

We now show that this approach is in fact a projection of P_0 onto the *generalized geodesic convex hull* of the control measures P_j with respect to P_0 as illustrated in Figure 1. To define this notion of convex hull, we extend the definition of generalized geodesics (Ambrosio et al., 2008, section 9.2), and in particular the definition of W_P to the multimarginal setting, by defining, for given couplings $\gamma_{0j} \in \Gamma(P_0, P_j)$, $j \in [\![J]\!]$

$$W^2_{P_0;\lambda}(\gamma_{01}, \gamma_{02}, \dots, \gamma_{0J}) \triangleq \min \left\{ \int_{(\mathbb{R}^d)^{J+1}} \sum_{j=1}^{J} \lambda_j \left|x_j - x_0\right|^2 \mathrm{d}\boldsymbol{\gamma} : \boldsymbol{\gamma} \in \Gamma_1(\gamma_{01}, \dots, \gamma_{0J}) \right\}, \tag{2.8}$$

where $\Gamma_1(\gamma_{01}, \dots, \gamma_{0J}) \subseteq \Gamma(P_0, P_1, \dots, P_J)$ 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 γ_{0j}. Note that this definition is similar to the multimarginal definition of the 2-Wasserstein barycenter (Agueh & Carlier, 2011, Gangbo & 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

$$\mathfrak{Co}_{P_0}\left(\left\{P_j\right\}_{j=1}^J\right) \triangleq \Bigg\{ P(\lambda) \in \mathscr{P}_2(\mathbb{R}^d) : P(\lambda) = \left(\sum_{j=1}^J \lambda_j \pi_{j+1}\right)_{\#} \boldsymbol{\gamma}, \\ \boldsymbol{\gamma} \text{ solves } W^2_{P_0;\lambda}(\gamma_{01}, \dots, \gamma_{0J}), \;\; \gamma_{0j} \text{ is optimal in } \Gamma(P_0, P_j) \;\; \forall j \in [\![J]\!], \quad \lambda \in \Delta^J \Bigg\}. \tag{2.9}$$

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

Proposition 2.1. *Consider a regular target measure P_0 and a set $\{P_j\}_{j\in[\![J]\!]}$ of general control measures. Construct the measure P_π 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 equation 2.7 and $\nabla\varphi_j$ are the optimal maps transporting P_0 to P_j, respectively. Then P_π is the unique metric projection of P_0 onto $\mathfrak{Co}_{P_0}\left(\{P_j\}_{j=1}^J\right)$.

2.5 Tangential Wasserstein projections for general target measures

Proposition 2.1 holds for a regular target P_0. In many practical settings, however, the target outcome is not a regular measure, as in our application to Medicaid expansion in section 4. In such settings, the corresponding optimal transport problems equation 2.1 between the target and the respective control measures P_j is only achieved via optimal transport plans γ_{0j}, not maps $\nabla\varphi_j$. In contrast to the regular setting, these transport maps also do not need to be unique. To apply our method in this more general setting would lead us to solving

$$\lambda^* \triangleq \operatorname*{arg\,min}_{\lambda\in\Delta^J} W^2_{P_0;\lambda}(\gamma_{01}, \dots, \gamma_{0J}), \tag{2.10}$$

where γ_{0j} are optimal transport plans between the target P_0 and the respective control measure P_j, which, again would be a computationally prohibitive bi-level optimization problem.

We therefore rely on barycentric projections to reduce the complex general setting to the regular tangent space and subsequently apply the projection equation 2.7. This is computationally inexpensive, as it amounts to computing

$$\lambda^* \triangleq \operatorname*{arg\,min}_{\lambda\in\Delta^J} \left\| \sum_{j=1}^J \lambda_j \left(b_{\gamma_{0j}} - \mathrm{Id}\right) \right\|^2_{L^2(P_0)}, \tag{2.11}$$

where

$$b_{\gamma_{0j}}(x_1) \triangleq \int_{\mathbb{R}^d} x_2 \, \mathrm{d}\gamma_{0j,x_1}(x_2)$$

are the barycentric projections of optimal transport plans γ_{0j} between P_0 and P_j. Here, γ_{x_1} denotes the disintegration of the optimal transport plan γ 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 γ_{0j} are actually induced by some optimal transport map

$\nabla\gamma_j$, 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_γ and disintegrations in conjunction with Jensen's inequality it holds for all $\lambda \in \Delta^J$ that

$$\left\| \sum_{j=1}^{J} \lambda_j \left(b_{\gamma_{0j}} - \mathrm{Id} \right) \right\|_{L^2(P_0)}^2 \leqslant W_{P_0;\lambda}^2(\gamma_{01}, \ldots, \gamma_{0J}). \tag{2.12}$$

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

$$\widetilde{\mathfrak{Co}}_{P_0}\left(\left\{P_j\right\}_{j=1}^J\right) \triangleq \left\{ P(\lambda) \in \mathscr{P}_2(\mathbb{R}^d) : P(\lambda) = \left(\sum_{j=1}^{J} \lambda_j b_{\gamma_{0j}} \right)_{\#} P_0, \quad \lambda \in \Delta^J \right\}. \tag{2.13}$$

Furthermore, the contraction property equation 2.12 implies that $\mathfrak{Co}_{P_0} \subseteq \widetilde{\mathfrak{Co}}_{P_0}$, with equality when all transport plans are achieved via maps $\nabla\varphi_j$. Using these definitions, the following defines our notion of projection for general P_0 and shows that it projects onto $\widetilde{\mathfrak{Co}}_{P_0}$.

Proposition 2.2. *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 equation 2.11 and γ_{0j} are optimal plans transporting P_0 to P_j, respectively. Then for given optimal plans γ_{0j}, $\widetilde{P}_\pi$ is the unique metric projection of P_0 onto $\widetilde{\mathfrak{Co}}_{P_0}\left(\left\{P_j\right\}_{j=1}^J\right)$.

In contrast to the regular case in Proposition 2.1, the optimal plans γ_{0j} transporting P_0 to P_j need not be unique. However, the projection for fixed γ_{0j} is unique.

3 Statistical properties of the weights and tangential 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 σ-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 3.1 (Consistency of the optimal weights)**.** *Let $\left\{\mathbb{P}_{N_j}\right\}_{j=0}^J$ be the empirical measures corresponding to the data $\left(X_{1j}, \ldots, X_{N_j j}\right)_{j=0}^J$ which are independent and identical draws from P_j, respectively, and are supported on some common latent probability space $(\Omega, \mathscr{A}, P)$. Assume P_j has finite second moment. As $N_j \to \infty$ for all $j \in [\![J]\!]$, the corresponding optimal weights $\widehat{\lambda}_N^* = \left(\widehat{\lambda}_{N_1}^*, \ldots, \widehat{\lambda}_{N_J}^*\right) \in \Delta^J$ obtained via*

$$\widehat{\lambda}_N^* \triangleq \underset{\lambda\in\Delta^J}{\arg\min} \left\| \sum_{j=1}^{J} \lambda_j \left(\widehat{b}_{0j} - \mathrm{Id} \right) \right\|_{L^2(\mathbb{P}_{N_0})}^2, \tag{3.1}$$

satisfy

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

where λ^* solve equation 2.11.

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 3.1 (Consistency of the optimal projections)**.** *In the setting of Proposition 3.1, 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 3.1 and Corollary 3.1 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 one needs to make strong regularity assumptions on the measures P_j. Without these, the rate of convergence of optimal transport maps in terms of expected square loss is as slow as $n^{-2/d}$ (Hütter & Rigollet, 2021). Under such additional regularity conditions, the results for the asymptotic properties are standard, because the proposed method reduces to a classical semiparametric estimation problem, as the weights λ_j are finite-dimensional.

4 Application to estimating causal effects via synthetic controls

In this section, we introduce a practical application of our method to causal inference by estimating the causal effect of Medicaid expansion in Montana in 2016 via our generalization of the classical synthetic controls estimator (Abadie & Gardeazabal, 2003, Abadie et al., 2010) to multivariate outcome distributions, complementing the univariate method in Gunsilius (2022). In particular, we can account for general heterogeneity in the multivariate outcomes by considering the entire distribution of several outcomes of interest jointly, something which has been out of reach of existing synthetic controls methods. We use the `POT` package (Flamary et al., 2021) to obtain the optimal transport plans, and `CVXPY` (Diamond & Boyd, 2016, Agrawal et al., 2018) to compute equation 2.7. Additional details of our applications as well as other simulations are contained in Appendix B.

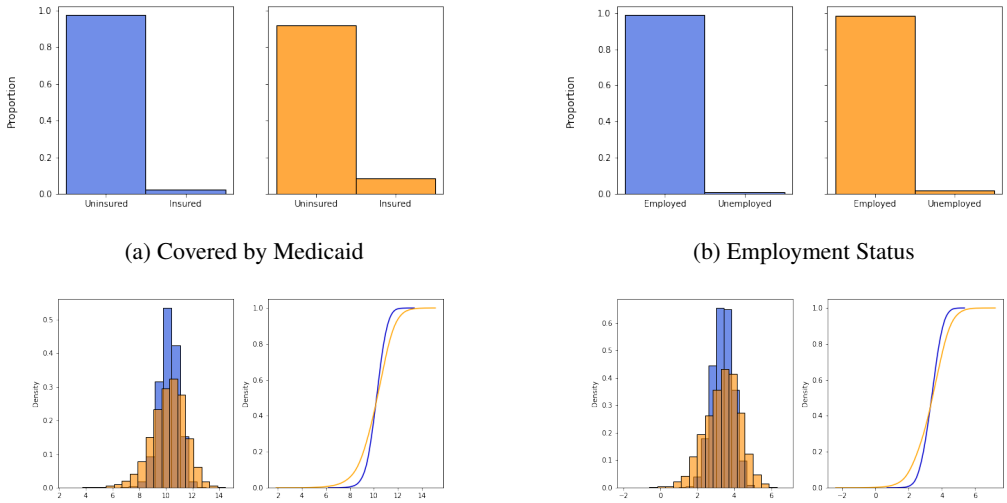
We study the effect of health insurance coverage following state-level Medicaid expansion in the United States. A provision within the Affordable Care Act (ACA) allows states to decide whether to expand Medicaid for low-income households. Some states decided to adopt such expansions early on, while others did not (and still have not done so). We investigate the economic and behavioral effects of Medicaid expansion. Specifically, we consider some first-order effects (i.e. the extensive margin of Medicaid enrollment post-expansion and disemployment effects) and second-order effects (i.e. income and labor supply effects) of expanded Medicaid access.

The observational data we use is the American Community Survey (ACS; additional details in Appendix B). From it, we collect the following variables:

1. HINSCAID, which indicates whether the individual interviewed is covered by Medicaid,
2. EMPSTAT, which indicates employment status (either employed or unemployed),
3. UHRSWORK, which indicates number of hours worked,
4. INCWAGE, which indicates previous-twelve-month labor income level.

We measure labor hours supplied and labor income in logs instead of levels. We consider Montana as the treated unit for this application. Montana adopted such expansion in 2016, which is the treated year. For control units, we use the twelve states for which such expansion has never occurred. As of 2022, the twelve states are: Alabama, Florida, Georgia, Kansas, Mississippi, North Carolina, South Carolina, South Dakota, Tennessee, Texas, Wisconsin, Wyoming. We use data from 2010 to 2016 to estimate optimal weights for the control states in order to create the "synthetic Montana", i.e. Montana had it not adopted Medicaid expansion. Once we obtain the optimal weights λ^*, we estimate the counterfactual outcomes of interest for the four years post-Medicaid expansion

in Montana. This involves solving equation 2.2 with the optimal weights λ^* obtained in the pre-intervention periods. Details of sample selection and estimating "synthetic Montana" are described in Appendix B.3. As we show there, the combination of control states replicating Montana is close to actual Montana in the pre-intervention period, implying that the observed differences in post-intervention periods can be attributed to the causal effect of the Medicaid expansion. Furthermore, therein, we also develop and show the results of a placebo permutation test in the sense of Abadie et al. (2010) to validate our findings.



(a) Covered by Medicaid

(b) Employment Status

(c) Log Wage

(d) Log Labor Hours Supplied

Figure 2: Counterfactual (blue) vs actual (orange) Montana from 2017 to 2020. In the bottom row, histograms of data distributions are shown on the left, and cumulative distribution functions are shown on the right.

Consistent with findings in Courtemanche et al. (2017), Mazurenko et al. (2018), we find significant first- and second order effects of Medicaid expansion, which are summarized in the top row and the bottom row of Figure 2. "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 also find positive second-order effects, summarized in the bottom row of Figure 2. Additional details are in Appendix B.

5 Conclusion

We have developed a projection method between sets of probability measures supported on $\mathbb{R}^d$ based on the tangential structure of the 2-Wasserstein space. Our method seeks to best approximate some target distribution that is potentially multivariate, using some chosen set of control distributions. We provide an implementation which gives unique, interpretable weights in a setting of regular probability measures. For general probability measures, we construct our projection by first creating a regular tangent space through applying barycentric projection to optimal transport plans. Our application to evaluating the first- and second-order effects of Medicaid expansion in Montana via an extension of the synthetic controls estimator (Abadie & Gardeazabal, 2003, Abadie et al., 2010) demonstrates the method's efficiency and the necessity to have a method that is applicable for general proabbility measures. The approach 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 & Loubes, 2017).

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.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

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

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

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.

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.

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

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*, pp. 2292–2300, 2013.

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.

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 Gunsilius. Distributional synthetic controls. *arXiv:2001.06118 [econ, stat]*, February 2022. URL `http://arxiv.org/abs/2001.06118`. arXiv: 2001.06118.

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.

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

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 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.

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.

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, pp. 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.

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

Michel Talagrand. The transportation cost from the uniform measure to the empirical measure in dimension $\geqslant 3$. *The Annals of Probability*, pp. 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.

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*, pp. 23781–23803. PMLR, 2022.

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.

APPENDIX

A PROOFS FROM THE MAIN TEXT

Proof of Proposition 2.1. Define the following closed and convex subset $\mathcal{C} \subseteq L^2(P_0)$ 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 equation 2.7 as

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

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 & 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 & Border, 1999, Lemma 6.54).

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

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

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 $\rho \in [0, 1]$,

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

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 ρ. Mechanically, for any $\lambda \in \Delta^J$ and $t_j > 0$ for all $j \in [\![J]\!]$,

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

$$= \left(\left[\sum_{j=1}^{J} \lambda_j t_j (\nabla \varphi_j - \mathrm{Id}) \right] + \mathrm{Id} \right)_{\#} P_0$$

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_π is a *geodesic* metric projection of P_0 onto the geodesic convex hull of $\{P_j\}_{j=1}^{J}$. ■

Proof of Proposition 2.2. The result follows from the same argument as the proof of Proposition 2.1. 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 γ 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

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

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 of Proposition 3.1. We split the proof into two parts. In the first part we prove the convergence in probability of the family of objective functions equation 3.1 to their population counterparts equation 2.11 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 & 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 λ^*, we write

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

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

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

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

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

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

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

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

$\widehat{\gamma}_{0j}$ 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

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

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.,

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

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

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 γ_{0j} with respect to P_0, respectively. Therefore, we have

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

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

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

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

where the last equality follows from the tightness of $\widehat{\gamma}_{0j}$, as shown earlier. Therefore, by standard stability results for optimal transport maps, in particular Proposition 6 in Zemel & Panaretos (2019), it holds that $\widehat{b}_{0j}$ converges uniformly on every compact subset $K \subseteq \mathbb{R}^d$ in the interior of 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:

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

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

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

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 γ_{0j} is a limit point of the sequence $\widehat{\gamma}_{0j}$, it holds that

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

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

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

that is,

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

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

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

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

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

This shows convergence in probability of the objective function for fixed λ.

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 Δ^J. Now a standard consistency result like Theorem 2.1 in Newey & 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 λ 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 λ^* will be achieved at the interior of the extended Euclidean space, from which consistency follows by Theorem 2.7 in Newey & McFadden (1994). ■

Proof of Corollary 3.1. 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 equation 3.1 and equation 2.11, respectively. The result follows by applying the extended continuous mapping theorem (van der Vaart & Wellner, 2013, Theorem 1.11.1) as follows.

As shown in the proof of Proposition 3.1 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

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

for any continuous function such that $|f(x_0)| \leqslant C_1 + C_2\,|\overline{x}_0 - x_0|^2$ 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

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

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 & 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 3.1, 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. ■

B Details from Simulations and Applications

B.1 Mixtures of Gaussian Simulation

As a simple first simulation to show that the method is applicable in higher dimensions, we consider mixtures of Gaussian in dimension $d = 20$. We draw from the following Gaussians:

$$\mathbf{X}_j \sim \mathcal{N}\left(\mu_j, \Sigma\right), \quad j = 0, 1, 2, 3\,,$$

where $\mu_0 = [10, 10, \ldots, 10]$, $\mu_1 = [50, 50, \ldots, 50]$, $\mu_2 = [200, 200, \ldots, 200]$, $\mu_3 = [-50, -50, \ldots, -50]$ and $\Sigma = \mathrm{Id}_{20} + 0.8\,\mathrm{Id}_{10}^{-}$, with Id_{10}^{-} the 10×10 matrix with zeros on the main diagonal and ones on all off-diagonal terms. We then define the following mixtures: $\mathbf{Y}_0$ as target, and $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$ as controls, where

$$\begin{aligned}
\mathbf{Y}_0 &= 0.3\mathbf{X}_0 + 0.6\mathbf{X}_1 + 0.1\mathbf{X}_2 \\
\mathbf{Y}_1 &= 0.8\mathbf{X}_0 + 0.1\mathbf{X}_1 + 0.1\mathbf{X}_2 \\
\mathbf{Y}_2 &= 0.2\mathbf{X}_1 + 0.7\mathbf{X}_2 + 0.1\mathbf{X}_3 \\
\mathbf{Y}_3 &= 0.2\mathbf{X}_0 + 0.2\mathbf{X}_2 + 0.6\mathbf{X}_3
\end{aligned}$$

We sample $N_0 = N_1 = N_2 = N_3 = 10000$ points from each mixture. The estimated optimal weights are $\lambda^* = [0.8204, 0.1796, 0]$. The results are as expected, since $\mathbf{Y}_1$ is defined by the same set of $\mathbf{X}$'s as those for $\mathbf{Y}_0$. In the case of $\mathbf{Y}_3$, it is not defined with $\mathbf{X}_1$, therefore we expected the associated weight to be very small. Table 1 suggests the projection is close to the target distribution when only considering the mean, despite only having three control units.

d	1	2	3	4	5	6	7	8	9	10
mean($\mathbf{Y}_0$)	53.0	53.0	53.0	53.0	53.0	53.0	53.0	53.0	53.0	53.0
mean($\sum_{j=1}^{3} \lambda_j^* \mathbf{Y}_j$)	53.1	53.1	53.1	53.1	53.1	53.1	53.1	53.1	53.1	53.1

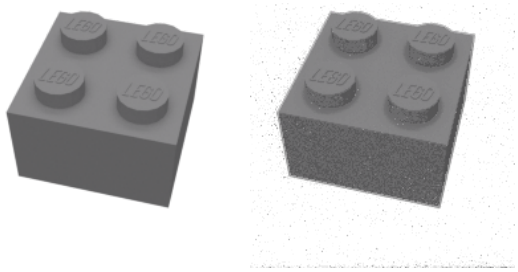
Table 1: Means of target and optimally-weighted controls $\mathbf{Y}_1$, $\mathbf{Y}_2$, $\mathbf{Y}_3$.

B.2 Simulation: Replicating Images

To demonstrate the property that our method provides sparse 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. These images are in RGBA format. For our application, the target image is contained in Figure 3(A), and the control images are in Figure 4.

The result, as shown in Figure 3, indicates the optimally-weighted projection replicates the target image well, even in a setting with only a few controls. We note two interesting findings. One, only the control images in the first row of Figure 4 received nontrivial weights; all others received

essentially zero weights. This suggests that our method use most information from controls which look sufficiently like the target and demonstrates that our method provides sparse weights. Two, even with few controls, the optimally weighted projection approximates the target well in this application.



(a) Target block (b) Projection

Figure 3: Target Lego block and projection.

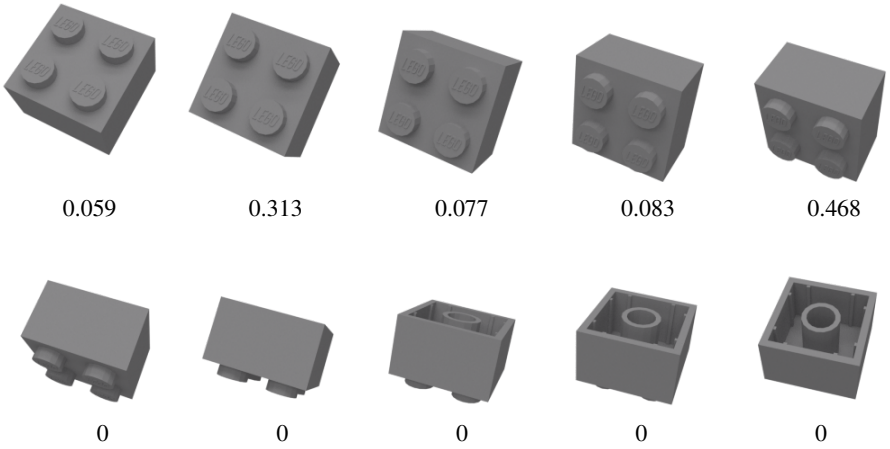


Figure 4: Control units used in simulation and their respective weights.

B.3 Details of Medicaid Expansion Application

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

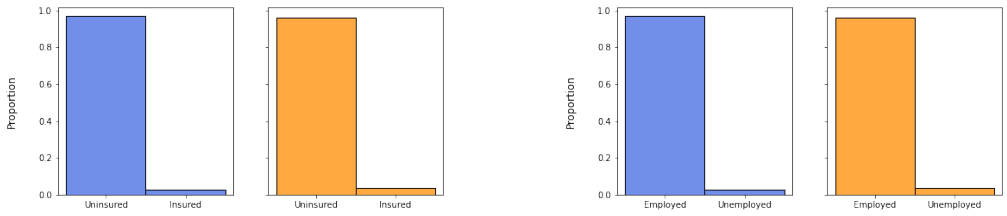
Table 2: Summary of the full data sample used to obtain λ^*.

We randomly select $N = 1500$ observations from each unit for estimating λ^*. In the Python implementation, we face a challenge where if the entries of the target and control data are large enough, equation 2.7 becomes too large for `CVXPY` to compute an optimal solution. 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 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.

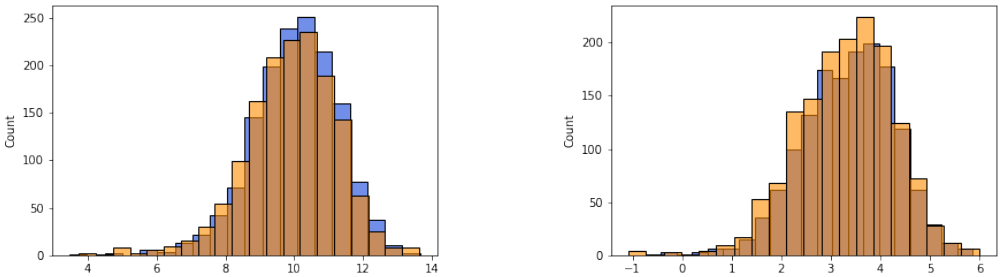
We check whether the obtained weights are fit for creating synthetic Montana by examining if they well-approximate actual Montana in the pre-treatment period. As seen in Figures 5 and 6, our projection is similar to the actual data.



(a) Covered by Medicaid

(b) Employment Status

Figure 5: Replicated (blue) vs actual (orange) Montana from 2010 to 2016.



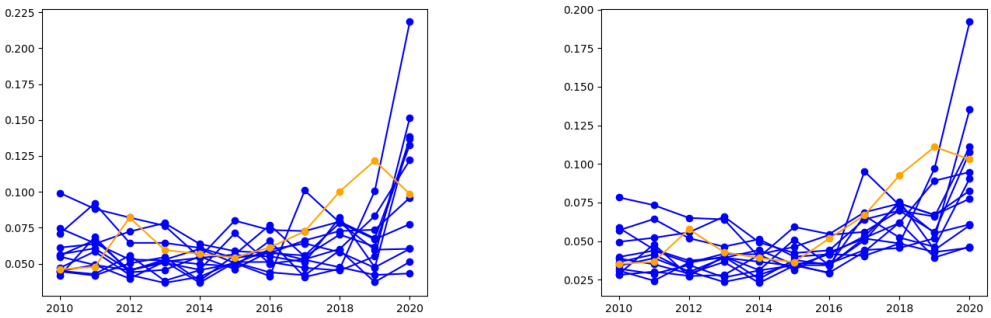
(a) Log Wage

(b) Log Labor Hours Supplied

Figure 6: Replicated (blue) vs actual (orange) Montana from 2010 to 2016. In each panel, histograms of data distributions are shown on the left, and cumulative distribution functions are shown on the right.

Once we obtain the optimal weights λ^*, we estimate the counterfactual outcomes of interest for the four years after Medicaid expansion in Montana (namely, between 2017 and 2020). This involves solving equation 2.2 with λ^* obtained from the pre-intervention period. Implementation-wise, we computed the free-support barycenter, using the `POT` package; this does not fix the support of the barycenter *a priori*, and allows it to be different from those of the control distributions. We plot the densities and distributions of the counterfactual outcomes in Figure 2 of the main text.

To perform inference on the estimated causal effect, we use a placebo permutation test in analogy to Abadie et al. (2010), Gunsilius (2022). 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 an actual treatment effect only appears in the treatment unit (Montana, in this application), then the estimated effect for the actual treatment unit should be among the largest.



(a) Weights Using All Years

(b) Weights Using Averaged Weights Over All Years

Figure 7: In orange: Montana. In blue: pretending each control state listed in Table 2 is a treated state.

We plot the 2-Wasserstein distance between the treated, joint distribution of *all outcomes* and the pre-/post-intervention optimal projection (i.e. equation 2.2 with λ^*). We present two sets of results in Figure 7: in panel (A), the optimal projection is computed using λ^* estimated using all years in the pre-intervention period; in panel (B), the λ^* 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 7, 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 4. 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 4: Estimated $p_t \triangleq \frac{r(d_{1t})}{J+1}$ in the post-intervention period.