Machine Learning (2021) 110:3135–3155
https://doi.org/10.1007/s10994-021-06060-0

Data driven conditional optimal transport

Esteban G. Tabak[1] · **Giulio Trigila**[2] · **Wenjun Zhao**[1]

Received: 10 December 2019 / Revised: 20 May 2021 / Accepted: 5 August 2021 /
Published online: 1 October 2021

Abstract
A data-driven procedure is developed to compute the optimal map between two conditional probabilities $\rho(x|z_1,\dots,z_L)$ and $\mu(y|z_1,\dots,z_L)$, known only through samples and depending on a set of covariates z_l. The procedure is tested on synthetic data from the ACIC Data Analysis Challenge 2017 and it is applied to non-uniform lightness transfer between images. Exactly solvable examples and simulations are performed to highlight the differences with ordinary optimal transport.

Keywords Optimal transport · Conditional average treatment effect · Uncertainty quantification · Color transfer · Image restoration

1 Introduction

Optimal transport seeks the mass preserving map T between two probability distributions that minimizes the expected value of a given cost function, the *transportation cost* between a point and its image under T (Villani et al., 2003). The corresponding minimal cost defines a metric in the space of probability distributions, the *Wasserstein distance* for cost functions of the form $c(x,y) = \|y-x\|^p$. Beyond providing a metric, the optimal map T itself has broad applicability, which this article extends through the development of conditional optimal transport.

Consider as a specific example the evaluation of the effects of a long-term medical treatment (alternatively of a habit, such as smoking or dieting). Optimal transport can be used to quantify changes in the probability distribution of quantities that characterize the health

Editor: Pradeep Ravikumar.

✉ Giulio Trigila
giulio.trigila@baruch.cuny.edu

Esteban G. Tabak
tabak@cims.nyu.edu

Wenjun Zhao
wenjun@cims.nyu.edu

[1] Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA

[2] Baruch College, CUNY, 55 Lexington Avenue, New York, NY 10035, USA

state of a person (blood pressure, blood sugar level, heart beat rate) between scenarios with and without treatment. Data typically consist of independent measurements of these quantities in treated and untreated populations. Yet more often than not, the distribution of these quantities depends on many covariates beyond the presence or absence of treatment, such as age, weight, sex, habits. Hence one should refine the search, looking for the effect of the treatment as a function of these covariates.

Motivated by this and similar applications, this article develops a data driven procedure to compute the optimal map $T(x, z)$ between two conditional probability densities $\rho(x|z_1, \dots, z_L)$ and $\mu(y|z_1, \dots, z_L)$, with covariates z_l. In the example above, $y = T(x, z)$ estimates the value y that the quantity of interest would have under treatment if, without treatment, its value were x, under specific values of the covariates z_l. The procedure is data driven, as it uses only samples $\left\{x^i, z_1^i, \dots, z_L^i\right\}$ and $\left\{y^j, z_1^j, \dots, z_L^j\right\}$ from ρ and μ. Notice that we do not seek a pairwise matching between $\left\{x^i, z_1^i, \dots, z_L^i\right\}$ and $\left\{y^j, z_1^j, \dots, z_L^j\right\}$: typically these two data sets do not even have the same cardinality. Instead, we work under the hypothesis that these samples are drawn from smooth conditional densities $\rho(x|z) = \rho(x, z)/\gamma^x(z)$, $\mu(y|z) = \mu(y, z)/\gamma^y(z)$ and covariate distributions $\gamma^x(z)$ and $\gamma^y(z)$, and hence we seek a map $y = T(x, z)$ that is a smooth function of its arguments.

The need for conditional optimal transport is particularly apparent when the distributions for the covariates z for the source and target distributions are unbalanced, i.e. when γ^x and γ^y differ. Consider as a particularly telling example a situation when the treatment has no effect, i.e. $\rho(x|z) = \mu(x|z)$, so the "true" answer should be $y = x$, yet the covariates are unbalanced, i.e. $\gamma^x \neq \gamma^y$. For concreteness, suppose that

$$\rho(x|z) = \mu(y|z) = N(z, 1), \quad \gamma^x(z) = N(-1, 1), \quad \gamma^y(z) = N(1, 1),$$

where $N(a, b)$ denotes the 1d normal distribution with mean a and variance b. Then

$$\rho(x) = \int \rho(x|z)\gamma^x(z)\, dz = N(-1, 2), \quad \mu(y) = \int \mu(y|z)\gamma^y(z)\, dz = N(1, 2).$$

It follows that, if one would not look at the covariate z, one would infer incorrectly that $y = x + 2$, i.e. that the treatment does have a significant effect. We will see in Sect. 4.1 an instance of this phenomenon appearing in the more complex setting of a biomedical application, where conditional transport provides critical aid.

Conditional transport provides a very flexible toolbox for data analysis, as the choice of which variables are conditioned to which others is left at the discretion of the analyst. In anticipation of the application of this principle to color transfer problems in Sect. 4.3, we illustrate it here with a simple example. Consider a covariate $z \sim N(0, 1)$ and two dependent variables $x \sim N(z, 1)$ and $y \sim N(-z, 1)$ (see Fig. 1 for a sketch related to this problem). Since the marginals $\rho(x)$ and $\mu(y)$ are identical, performing optimal transport between them yields the identity map $y = x$, while conditioning to z yields $y = x - 2z$, effectively rotating the joint distribution $\rho(x, z)$ clockwise. For a third alternative, consider performing regular two dimensional transport between $\rho(x, z)$ and $\mu(y, z)$, which yields an irrotational map (Villani et al., 2003). Finally, if in a thought experiment we would identify x and y and switch the roles of dependent and independent variables, conditioning the transport in z-space to x, we would obtain $z_2 = z_1 - 2x$, effectively rotating the joint distribution $\rho(x, z)$ counter-clockwise. Here $z_{1,2}$ denote the variable z when attached to x and y respectively.

Conditional optimal transport shares similarities with normalizing flows, introduced in Tabak and Vanden-Eijnden (2010); Tabak and Turner (2013) and further developed, for

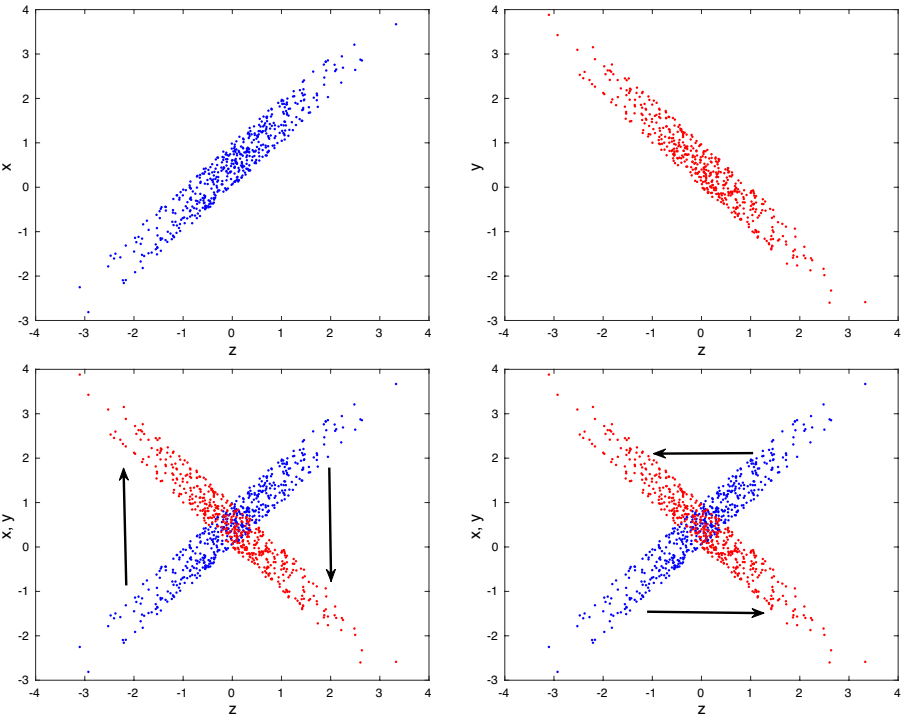


Fig. 1 Upper row: source (left) and target (right) distributions. Lower left: optimal transport of x conditioned on z. The arrows indicate that the lower left branch and the upper right branch of the source distribution are mapped respectively to the upper left branch and the lower right branch of the target distribution. Lower right: optimal transport of z conditioned on x. In this case, it is the upper right branch of the source distribution that is mapped to the upper left branch of the target distribution

instance, in Rezende and Mohamed (2015), Dinh et al. (2016), with autoregressive networks (Oliva et al., 2018), and with Generative Adversarial Networks (Grover et al., 2017; Finlay et al., 2020). One of the main differences with these procedures is the objective function being optimized. Optimal transport minimizes a user-defined, pairwise transportation cost, while the other methodologies, by and large, have objective functions, such as the relative entropy, concerned only with the fulfillment of the push-forward condition. In the specific kind of applications that we consider, we interpret the pairwise cost as a measure of data deformation, and thus seek the solution that minimally deforms the data, for instance among those maps that best characterize the changes in a person's health between scenarios with and without treatment. In this sense, the work presented here is closer to Trigila and Tabak (2016), which uses a normalizing flow that at the same time minimizes the transportation cost. The other main difference, of course, is the consideration of factors, which distinguishes conditional from regular optimal transport.

The plan of this article is as follows: after this introduction, Sect. 2 formulates the conditional optimal transport problem in an adversarial framework conducive to effective computation. This minimax formulation involves two players: one with strategy $T(x, z)$, the cost-minimizing map, and one with strategy $g(y, z)$, a test function that discriminates between the target $\mu(y|z)$ and the push-forward $T\#\rho(x|z)$ via a variational formulation of the relative

entropy between the two. Section 3 describes some alternative ways to parameterize these two strategies. Section 4 illustrates the procedure through real examples of practical and conceptual relevance: the determination of the effect of a medical treatment, lightness transfer: given two photographs of different objects under different lightness conditions, render the first under the lightness condition of the second, and "automatic restoration": given a painting that has deteriorated over time and one that has not, "restore" the first to its likely original condition. Finally, some conclusions and directions of further research are summarized in Sect. 5.

2 Conditional optimal transport

Conditional optimal transport between two conditional distributions $\rho(x|z)$ and $\mu(y|z)$ can be defined simply as the map $T(x, z)$ that performs optimal transport between them for each value of z:

$$\min_{T(:,z)} \int c(T(x,z),x)\rho(x|z)dx \quad \text{s.t.} \quad T\#\rho(:|z) = \mu(:|z), \tag{1}$$

where $c(x, y)$ represents the cost of moving a unit of mass from x to y and the symbol # indicates the push forward of probability measures, i.e. if x has distribution $\rho(x|z)$ then $y = T(x, z)$ has distribution $\mu(y|z) = T\#\rho(:|z)$. Since $T(:, z)$ decouples under different values of z, we can multiply the cost by the distribution $\gamma^x(z) \geq 0$ of the covariates z in the source and integrate over z, yielding

$$\min_{T(:,z)} \int c(T(x,z),x)\ \rho(x,z)\ dxdz \quad \text{s.t.} \quad T\#\rho(:|z) = \mu(:|z)\forall z, \tag{2}$$

where $\rho(x, z) = \rho(x|z)\gamma^x(z)$ is the joint distribution of x and z.

We need to reformulate this problem in a way that is implementable in terms of samples $\left\{x^i, z_x^i\right\}$ and $\left\{y^j, z_y^j\right\}$. As it stands in (2), two immediate problems emerge: there are not enough samples for each value of z, typically none or one for continuous covariates, to characterize the corresponding conditional distributions, and it is not clear how to enforce the push forward condition. The first problem is at the very heart of the need for conditional optimal transport: even though the objective functions for each value of z decouple, one assumes a commonality across z that makes samples from each conditional distribution be informative on the others. In the case of continuous covariates z, this can be posed as a smoothness (in z) condition on $\rho(x|z)$.

In order to address the second problem, we interpret the push forward condition in terms of relative entropy. Recall that the relative entropy between two distributions ρ_1 and ρ_2 is given by

$$D_{KL}\left(\rho_1||\rho_2\right) = \int \rho_1(x) \log \frac{\rho_1(x)}{\rho_2(x)} dx \geq 0,$$

which vanishes only when ρ_1 and ρ_2 agree almost everywhere. Hence the push forward condition $T\#\rho = \mu$ can be restated as $D_{KL}(\mu||T\#\rho) = 0$. Yet the relative entropy is not a robust quantifier of the difference between distributions, as it is bounded only when the first distribution is absolutely continuous with respect to the second. To resolve this issue, we replace the second distribution by an interpolation between $T\#\rho$ and μ:

$$D_{KL}\left(\mu\middle|\middle|\frac{1}{2}(T\#\rho+\mu)\right)=0,$$

so that the absolute continuity requirement is automatically satisfied.

In order to incorporate the dependence on the covariate z, we replace the relative entropy by its conditional counterpart:

$$D_{KL}(\rho_1(x|z)||\rho_2(x|z))=\int\gamma_1(z)\int\log\left(\frac{\rho_1(x|z)}{\rho_2(x|z)}\right)\rho_1(x|z)\,dxdz,$$

the conditional Kullback–Leibler divergence between ρ_1 and ρ_2 (Cover & Thomas, 2012). Since this is non-negative, we can rewrite the problem in (2) as

$$\min_{T(\cdot,z)}\max_{\lambda\geq 0}\left[\int c(T(x,z),x)\rho(x,z)\,dxdz+\lambda D_{KL}\left(\mu(x|z)\middle|\middle|\frac{1}{2}(T\#\rho(x|z)+\mu(x|z))\right)\right].$$

Instead of maximizing over λ, it will be convenient to fix a value of λ large enough that the push forward condition can be considered satisfied for all practical purposes. (It is straightforward to prove that, as $\lambda\to\infty$, the solution with fixed λ converges to the true minimax solution. In our implementation below, λ grows at each step of the algorithm.) Then the problem above becomes

$$\min_{T}\left[\int c(T(x,z),x)\rho(x,z)\,dxdz+\lambda\,D_{KL}\left(\mu(x|z)\middle|\middle|\frac{1}{2}(T\#\rho(x|z)+\mu(x|z))\right)\right],$$
$$\lambda\gg 1.$$

For any $\rho_1(x,z)=\gamma_1(z)\rho_1(x|z)$ and $\rho_2(x,z)=\gamma_2(z)\rho_2(x|z)$, the following "chain rule" for relative entropy holds (Cover & Thomas, 2012):

$$D_{KL}(\rho_1(x|z)||\rho_2(x|z))=D_{KL}(\rho_1(x,z)||\rho_2(x,z))-D_{KL}(\gamma_1(z)||\gamma_2(z)).$$

Since the map T acts only on x, it has no effect on the last term of this expression, so we can write

$$\min_{T}\left[\int c(T(x,z),x)\rho(x,z)\,dxdz+\lambda D_{KL}\left(\mu(x,z)||\frac{1}{2}(T\#\rho(x,z)+\mu(x,z))\right)\right].$$

This formulation improves over the one in (1) by consolidating an infinite set of problems, one for every value of z, into a single one. Yet it is not clear yet how to enforce the push forward condition in terms of samples, as the definition of the relative entropy involves evaluating logarithms of ρ and μ. To address this, we invoke a variational formulation of the relative entropy between two distributions (Donsker & Varadhan, 1975):

$$D_{KL}(\rho_1||\rho_2)=\max_{g}\left[\int g(x,z)\rho_1(x,z)dxdz-\log\left(\int e^{g(x,z)}\rho_2(x,z)\,dxdz\right)\right],\qquad(3)$$

which involves ρ_1 and ρ_2 only in the calculation of the expected values of g and e^g, with a natural sample-based interpretation as empirical means. Then our problem becomes

$$\min_T \max_g \int c(T(x,z),x)\rho(x,z)dxdz + \lambda\Bigg[\int g(y,z)\mu(y,z)dydz - \log\left(\frac{1}{2}\int e^{g(y,z)}\mu(y,z)dydz + \frac{1}{2}\int e^{g(T(x,z),z)}\rho(x,z)dxdz\right)\Bigg] \tag{4}$$

or, in terms of samples,

$$\min_T \max_g \frac{1}{N}\sum_{i=1}^{N} c(T(x^i,z_x^i),x^i) + \lambda\Bigg[\frac{1}{M}\sum_{j=1}^{M} g(y^j,z_y^j) - \log\left(\frac{1}{2M}\sum_{j=1}^{M} e^{g(y^j,z_y^j)} + \frac{1}{2N}\sum_{i=1}^{N} e^{g(T(x^i,z_x^i),z_x^i)}\right)\Bigg].$$

This adversarial formulation has two players with strategies T and g, one minimizing the cost and the other enforcing the push forward condition, providing an adaptive "lens" that identifies those places where the push-forward condition does not hold: for any T, the optimal g in (4) is given by

$$g = \log\left(\frac{\mu(y,z)}{\frac{\mu(y,z)+T\#\rho(x,z)}{2}}\right) = \log\left(\frac{(1+w(z))\mu(y|z)}{w(z)\mu(y|z)+T\#\rho(x|z)}\right) + \log\left(\frac{2w(z)}{1+w(z)}\right),$$

where $w(z) = \gamma^y(z)/\gamma^x(z)$, and the first term is furthest from zero in those places where $T\#\rho(x|z)$ and $\mu(y|z)$ differ the most.

It is interesting to notice a feature in the solution $g(x, z)$ to the variational formulation (3) for the relative entropy involving conditional distributions. The optimal g is given by

$$g(x,z) = \log\left(\frac{\rho_1(x,z)}{\rho_2(x,z)}\right) = \log\left(\frac{\rho_1(x|z)\gamma_1(z)}{\rho_2(x|z)\gamma_2(z)}\right).$$

Consider a situation where we have already performed conditional optimal transport, so that $\rho_1(x|z) = \rho_2(x|z)$. If the distributions for z in source and target are unbalanced, the corresponding optimal g will be a nonzero function of z alone:

$$g(x,z) = \log\left(\frac{\gamma_1(z)}{\gamma_2(z)}\right) = w(z), \tag{5}$$

in contrast to the situation in regular optimal transport between the joint distributions $\rho_{1,2}(x, z)$, where the final optimal $g(x, z)$ equals zero. As a consequence of this, we must not expect the penalizing term on the entropy to necessarily vanish in the solution of (4). The final value of the penalization term depends on the ratio between the possibly unbalanced distributions for z in the source and target distributions. Other choices to impose the push forward condition can be made, for instance by adopting the work in Nowozin et al. (2016) where the notion of f-divergence family is introduced.

3 Parametrization of the flows

In order to complete the problem formulation in (5), we need to specify the family of functions over which the map $T(x, z)$ and the test-function $g(y, z)$ are optimized. These families should satisfy some general properties:

1. Be rich enough that g can capture all significant differences between $\rho(x|z)$ and $\mu(y|z)$ and T can resolve them.
2. Not be so rich as to overfit the sample points $\{x^i, z_x^i\}$, $\{y^j, z_y^j\}$. For instance, a g with arbitrarily small bandwidth would force the sets $\{T(x^i, z_x^i), z_x^i\}$, $\{y^j, z_y^j\}$ to agree pointwise, an extreme case of overfitting that is not only undesirable but also unattainable when their cardinality differs. More generally, the dependence of the functions on z should be such that, with a finite number of samples, it should still capture the assumed smoothness of $\rho(x|z)$: functions that are too localized in z space effectively decouple the transport problems for every value of z, for which there are not enough available sample points.
3. Be well-balanced: if one of the two players has a much richer toolbox than the other, the game would be "unfair", leading to a waste of computational resources and possibly to instability and/or inaccuracy.

These conditions leave space for many proposals. For instance, we could define both T and g through neural networks, as done in Yang and Tabak (2019) in the context of optimal transport-based factor discovery. Instead, the examples in this article are solved with the two implementations detailed below. Both share the feature that T is built on map composition: at each step n of the mini-maximization algorithm, an elementary map E^n is applied not to the original sample points $\{x^i\}$, but to their current images:

$$T^n(x^i, z_x^i) = E^n\left(T^{n-1}(x^i, z_x^i), z_x^i\right).$$

This way, simple elementary maps E depending on only a handful of parameters can give rise through map composition to rich global maps T. The two proposals differ in that one builds nonlinear richness through evolving Gaussian mixtures, while the other builds complex z-dependence through an extra compositional step. In this article, the first method is applied to a lightness transfer problem, and the second to the effect of a medical treatment, as the latter is linear in x but has complex, nonlinear dependence on many covariates z. We close this section with a proposition that specifies the third point above. The goal is to motivate mathematically the choice of the potential and the test function that are made in the following two subsections.

Proposition 1 *The test component of the objective function*,

$$L_{test} = \int g(y,z)\mu(y,z)dydz + \\ - \log\left(\frac{1}{2}\int e^{g(y,z)}\mu(y,z)dydz + \frac{1}{2}\int e^{g(T(x,z),z)}\rho(x,z)dxdz\right),$$

necessarily decreases along the direction

$$dT = \nabla_y g(y, z)$$

when $dT \neq 0$. *Moreover, if* $dT = 0$ *and the push-forward condition*

$$\forall z,\ T\#\rho(\cdot|z) = \mu(\cdot|z)$$

is not satisfied, then g has not yet achieved its optimal value.

Proof The variational derivative of L_{test} with respect to the map $T(x, z)$ is given by

$$\frac{\delta}{\delta T} L_{test} = -\frac{e^{g(T(x,z),z)}\rho(x,z)}{\int e^{g(y,z)}\mu(y,z)dydz + \int e^{g(T(x,z),z)}\rho(x,z)dxdz} \nabla_y g(y,z)\Big|_{y=T(x,z)},$$

a strictly negative quantity times dT. Hence $dT = \nabla_y g$ is a direction of descent, unless it vanishes identically. But maximizing L over g yields

$$g(\cdot, z) = \log\left(\frac{\mu(\cdot|z)}{T\#\rho(\cdot|z)}\right),$$

which is constant in its first argument only when $T\#\rho(\cdot|z) = \mu(\cdot|z)$ (in which case $g(\cdot|z) = 0$.) □

3.1 Evolving Gaussian mixtures

We adopt as elementary map the gradient of a potential function convex in x: $E(x, z) = \nabla_x \boldsymbol{\Phi}(x, z)$, chosen from a family that includes the identity map and where $x \in \mathbb{R}^{d_x}$ and $z \in \mathbb{R}^{d_z}$. Strict convexity of the potential guarantees that the resulting elementary map is one-to-one. The potential $\boldsymbol{\Phi}$ is built from a quadratic form in x with coefficients that depend on z, plus a linear combination of K Gaussians in (x, z) space, and similarly for the test function g. The idea underlying this choice, introduced and discussed in more details in Tabak and Vanden-Eijnden (2010), Tabak and Turner (2013), Trigila and Tabak (2016), can be summarized by saying that the quadratic part of the potential is responsible for the matching of the mean and the covariance of ρ and μ while the Gaussian mixture add a non linearity of the elementary map to enforce that higher order moments of the push forward of ρ are mapped to the corresponding moements of μ . In addition we allow the centers of these Gaussians to evolve so to be able to approximate quite general functions $T = \nabla\boldsymbol{\Phi}$ and g.

In order to guarantee the convexity of $\boldsymbol{\Phi}$, notice that the gradient with respect to x of a radial basis function kernel with bandwidth d,

$$G_d(x, x') = \exp\left(-\frac{||x - x'||^2}{2d^2}\right),$$

is bounded by $\pm\frac{1}{d\exp(1/2)}$, and its second order derivatives by $\frac{2}{d^2\exp(3/2)} < \frac{1}{2d^2}$. It follows that $\frac{1}{2d^2}\frac{||\mathbf{x}||_2^2}{2} \pm G_d([\mathbf{z}, \mathbf{x}], [\mathbf{m}_{z_i}, \mathbf{m}_i])$ is convex, so we propose

$$\boldsymbol{\Phi}(\mathbf{x},\mathbf{z}) = (\mathbf{c}_0^T + \mathbf{z}^T\mathbf{c}_1)x + \frac{1}{2}\mathbf{x}^T\mathbf{C}_2(\mathbf{z})\mathbf{x} + \sum_{i=1}^{K} a_i^2\left(\frac{||\mathbf{x}||_2^2}{4d^2} - G_d([\mathbf{z},\mathbf{x}],[\mathbf{m}_{z_i},\mathbf{m}_i])\right)$$
$$+ \sum_{i=1}^{K} b_i^2\left(\frac{||\mathbf{x}||_2^2}{4d^2} + G_d([\mathbf{z},\mathbf{x}],[\mathbf{m}_{z_i},\mathbf{m}_i])\right),\ \mathbf{C}_2(\mathbf{z}) = \mathbf{C}_{2,0}^T\mathbf{C}_{2,0} + \mathbf{z}^T\mathbf{C}_{2,1}^T\mathbf{C}_{2,1}\mathbf{z},$$

with parameters of the model $\mathbf{C}_{2,0} \in R^{d_x \times d_x}$, $\mathbf{C}_{2,1} \in R^{d_x \times d_z}$ lower triangular, $\mathbf{c}_0$ and $c_1 \in R^{d_z}$, $m_i \in \mathbb{R}^{d_x}$, $m_{z_i} \in \mathbb{R}^{d_z}$, and a_i, b_i and $d \in \mathbb{R}$. Notice that, if $a_i = b_i$, the Gaussians cancel each other, and we are left with a purely quadratic potential. Therefore, in order to start the map at every step at the identity, the initialization must satisfy

$$\mathbf{C}_{2,0}(i,i)^2 + \sum_{i=1}^{K} \frac{1}{4d^2}(a_i^2 + b_i^2) = 1, \quad a_i^2 = b_i^2,$$

so we propose

$$a_i^2 = b_i^2 = \frac{4d^2\delta}{2K}, \quad \mathbf{C}_{2,0}(i,i) = \sqrt{1-\delta}, \quad \delta = \frac{1}{2},$$

with all other parameters starting from zero. The bandwidth d is chosen via $d = quantile(pdist([y;z]), 1/K)$, where *pdist* is the pairwise distance function. With this choice there are approximately $1/K$ points in the effective support of each Gaussian.

An illustration of how the map warps the source density with both x and $z \in \mathbb{R}$ is shown in Fig. 2. The joint distribution of x and z is a standard normal distribution. With only a linear term in z and no Gaussian kernels ($K = 0$), the results display a global distortion (upper right panel). When one kernel ($K = 1$) is used, it results in a local shrink (lower left panel) or stretch (lower right panel), depending on whether $a_1^2 < b_1^2$ or $a_1^2 > b_1^2$.

The choice of kernels is flexible in general and should have the following properties (1) they are smooth and differentiable with bounded second derivatives, so to enforce convexity of the potential through suitable parameterization; (2) users should be able to control the scale of effect, i.e. the kernels should have a free parameter being (or related to) the bandwidth. Except the radial basis function kernel mentioned above, for example, one can also use the rational quadratic kernel: $\mathcal{K}(x,x') = 1 - ||x-x'||^2/\left(||x-x'||^2 + c\right)$ with a free parameter c or the Cauchy kernel: $\mathcal{K}(x,x') = 1/\left(1 + \frac{||x-x'||^2}{h^2}\right)$ with parameter h.

For the test function, we propose

$$g(\mathbf{x},\mathbf{z}) = \sum_{i=1}^{K} \alpha_i G_d([\mathbf{z},\mathbf{x}],[\mathbf{m}_{z_i},\mathbf{m}_i]) + (\boldsymbol{\beta_0}^T + \mathbf{z}^T\boldsymbol{\beta_1})\mathbf{x} + \mathbf{x}^T(\boldsymbol{\beta_2} + \mathbf{z}^T\boldsymbol{\beta_3})\mathbf{x},$$

where $\beta_0 \in \mathbb{R}^{d_x}$, $\beta_1 \in \mathbb{R}^{d_z \times d_x}$, $\beta_2 \in \mathbb{R}^{d_x \times d_x}$ and $\beta_3 \in \mathbb{R}^{d_z \times d_x \times d_x}$ and with each iteration starting at the parameter values from the previous step. As for the potential $\boldsymbol{\Phi}$, the test function g is given by the sum of kernel functions and a quadratic function in $\mathbf{x}$ with coefficients that depend linearly from $\mathbf{z}$. The rationale behind the quadratic form in $\mathbf{x}$ is to check that the centers and covariance of $\mu(\mathbf{y}|\mathbf{z})$ and $\rho(\mathbf{x}|\mathbf{z})$ agree. The Gaussian centers are treated differently in the test function g, where they are extra parameters to ascend, and in the potential $\boldsymbol{\Phi}$, where they are fixed at their values from g in the prior step. The underlying notion is that g locates those areas where the distributions do not agree, and then T corrects them. The rationale for the parameterizations of the test function g and the potential $\boldsymbol{\Phi}$ to have the same form and, moreover, share the same centers, find its justification in Proposition 1 at

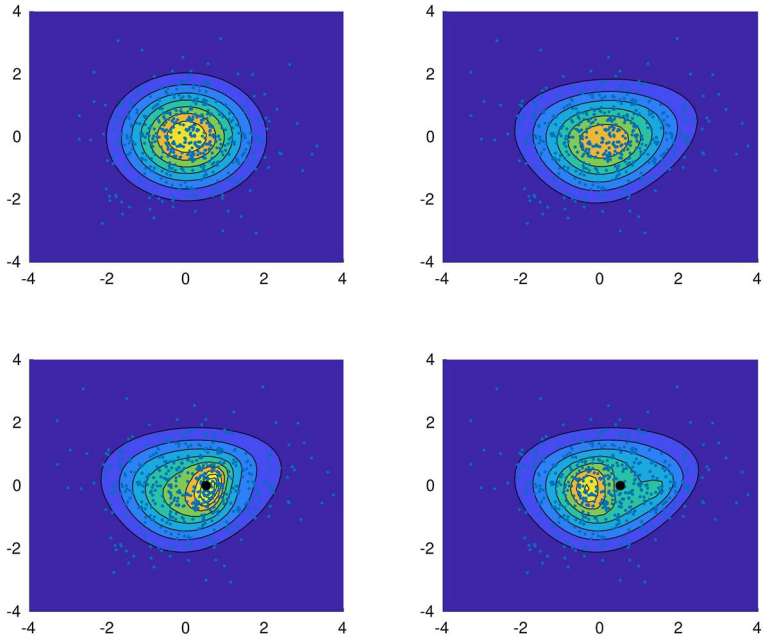


Fig. 2 An illustration of how different choices of the elementary map warp an isotropic standard Gaussian and a set of points (both displayed in the upper left panel) in the two dimensional (x, z) space. Upper right: warp with a global linear term in z. The panels on the lower row display a warp involving a Gaussian kernel ($K = 1$) centered at the black dot. Lower left panel: warp with the same linear term in z as in the upper right panel and $a_1^2 < b_1^2$. Lower right panel: warp with the same linear term in z as in the upper right and $a_1^2 > b_1^2$. The color scale is the same for all the four plots (Color figure online)

the beginning of the section saying that for any given test function $g(y, z)$, its gradient $\nabla_y g$ is a direction of descent for T—that is, a direction along which T lowers the test component of the objective function.

3.2 Extended map composition

This second methodology considers maps given by rigid translations and test functions that capture the conditional mean $\bar{x}(z)$:

$$T(x, z) = x + U(z), \quad g(y, z) = V(z)y + W(z), \quad x \in \mathbb{R},$$

with general, nonlinear dependence on $z \in \mathbb{R}^{d_z}$. Notice that the y independent function $W(z)$ is required, from (5), to handle a possible unbalance between $\gamma^x(z)$ and $\gamma^y(z)$. We will build U, V and W through generalized flows (Tabak & Vanden-Eijnden, 2010; Tabak & Turner, 2013) in z space , through the composition of function of the form

$$F(a, z, v, u) = \left(a_0^1 + \sum_{i=1}^{d_z} a_i^1 z_i + a_{L+1}^1 u\right) + \left(a_0^2 + \sum_{i=1}^{d_z} a_i^2 z_i + a_{L+1}^2 u\right) v.$$

Then we define the map $T = x + U(z)$ at time $n + 1$ via the recursion

$$T^{n+1}(T^n, z) = T^n + u^{n+1}, \quad u^{n+1} = F(\boldsymbol{\alpha}, z, u^n, v^n).$$

and the test function via

$$g^{n+1}(y, z) = v^{n+1}y + w^{n+1}, \quad v^{n+1} = F(\boldsymbol{\beta}, z, v^n, u^n), \quad w^{n+1} = F(\boldsymbol{\eta}, z, w^n, 0).$$

Notice that the T-independent function $W(z)$ evolves on its own, while $U(z)$ and $V(z)$ depend on the prior values of each other, as they compete through the minimax formulation. Also notice that this parameterization is such that it includes the maps $T = \nabla_y g$ that, according to Proposition 1, guarantee descent of F_{test}.

These maps are initialized at $u^0 = v^0 = w^0 = 0$. Before each step, α is set to 0 (as T is reinitialized every step to the identity), and so are β and η, except for $\beta_0^2 = \eta_0^2 = 1$, which makes g evolve from its value at the previous step.

4 Examples

We illustrate the procedure with two applications: determination of the effect of a medical treatment and lightness transfer. In order to solve the mini-maximization problem () we use the general procedure described in Essid et al. (2019).

4.1 Effect of a treatment

We apply conditional optimal transport to determine the response to a treatment of a diagnostic variable $x \in \mathbb{R}$ in terms of covariates z. As described in the introduction, given a set of available samples from the treated and untreated populations, we seek to infer the effect of the treatment. We propose to model this as a map $y = T(x, z)$ yielding the state y under treatment of a patient that, with covariates z, would have state x without treatment.

The data created for the ACIC data analysis challenge 2017 (Hahn et al., 2017) (https://arxiv.org/pdf/1905.09515.pdf) is particularly well-suited to test our approach. For concreteness, we consider the first of their 32 generating models, which includes 8 covariates: 6 binary and 2 continuous. We divide the data set into two groups: the untreated (x) and treated (y) patients, with samples drawn from distributions $\rho(x, z) = \gamma^x(z)\rho(x|z)$ and $\mu(y, z) = \gamma^y(z)\mu(y|z)$, having the property that

$$\mu(y|z) = \rho(y - \tau(z)|z), \quad \gamma^x(z) \neq \gamma^y(z).$$

The function $\tau(z)$ represents the Conditional Average Treatment Effect (CATE). It will be important for the analysis below to know that, in the model under consideration, τ depends only on the binary covariates, but the marginals $\gamma(z)$ depend also on the continuous ones (Hahn et al., 2017). The data is provided in 250 batches, each referring to the same 4302 patients, i.e. the same values of z_i under different realizations of the noise. We use only the first of these batches to compute the optimal map $T(x, z)$, reserving the other 249 to validate our results. In this first batch there is no repeated patient, so each patient is either treated or not-treated. This invalidates the use of regular regression, which would require pairs (x, y) for the same patient with and without treatment. Our distribution-based methodology, on the other hand, does not require the availability of such pairs.

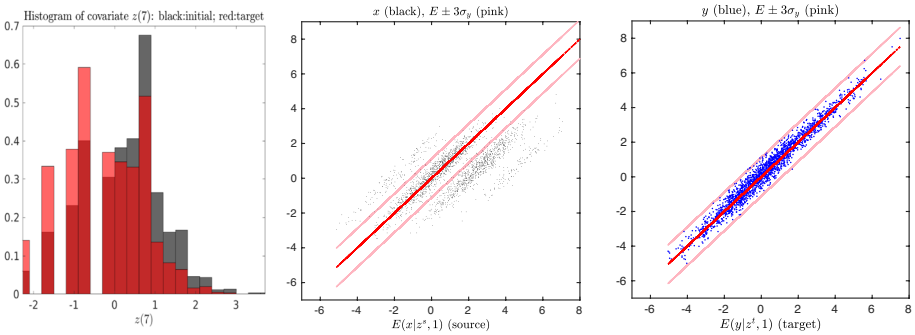


Fig. 3 Left panel: Unbalance in the distribution of z_7 between the source and the target data set. Center: Response variable x for patients before the treatment plotted as a function of the theoretical expected value that the same patients would have if they would undergone the treatment. Right: Response variable y of the treated patients as a function of the theoretical expected value of the same patients

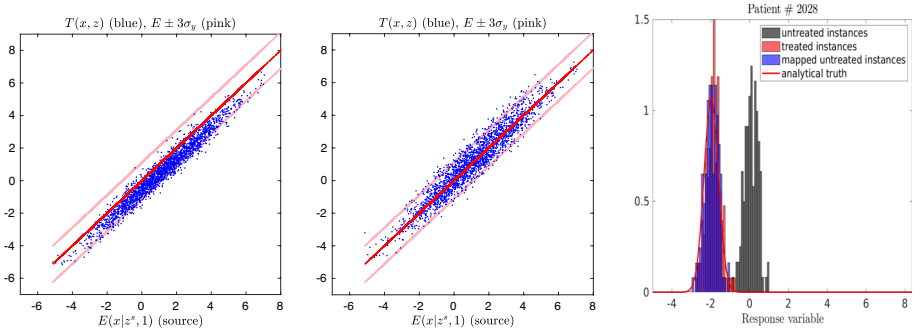
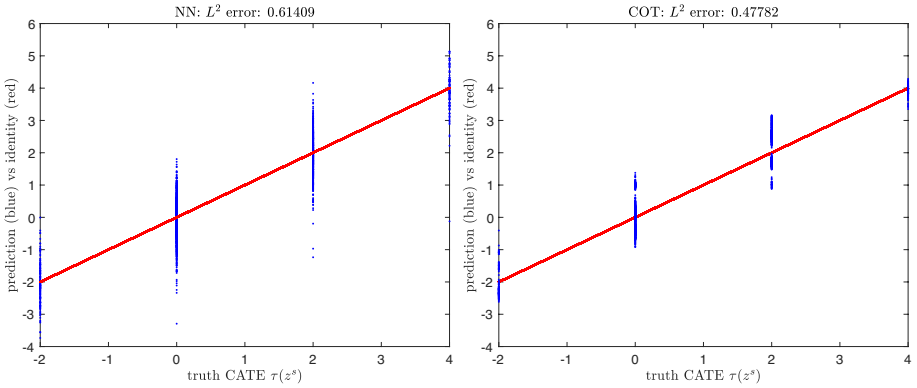


Fig. 4 Left: numerical value of the map $T(x_i, z_i)$ obtained using only the discrete covariates, which are the only ones that the true T depends on. The result is biased due to the unbalance between γ^x and γ^y for $\gamma(z_7)$). Middle: numerical value of the map $T(x_i, z_i)$ obtained using all the covariates. Right: comparison between the application of the map $T(x, z)$ to all untreated instances in the full 250 batches to the histogram of the response y for all treated instances of the patient

The middle panel of Fig. 3 displays the untreated values x^i as a function of the expected value that they would have under treatment given the values z_s^i of their covariates in the source distribution corresponding to the untreated patients:

$$E(x|z_s, 1) = \int \big(x + \tau(z_s)\big)\rho(x|z_s)\, dx,$$

while the right panel displays similarly the treated values y^i. The 1 above refers to the situation under treatment, while a zero would denote the absence of treatment. An exact quantification of the effect of the treatment would recover the map $T(x, z) = x + \tau(z)$. The left panel of Fig. 4 displays the map $T(x^i, z^i)$ obtained using only the discrete covariates, which are the ones that the true T of the underlying model depends on. However, because of the unbalance between γ^x and γ^y (see the left panel of Fig. 3 for $\gamma(z_7)$), the results are biased, much as in the synthetic example in the introduction. The middle panel shows that, when all covariates are considered, this biased is resolved. The right panel compares the



(a) Benchmark: nearest neighbor in z space

(b) Conditional Optimal Transport

Fig. 5 Predicted CATE using K-nearest neighbor in the latent space with $K = 1$ (left) and conditional optimal transport (right)

Table 1 Error in CATE and number of unique predictions using KNN benchmarks with different values of $K \geq 1$ and conditional optimal transport, performed on additive dataset with maps restricted to rigid translations

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
CATE: L_∞ error	3.9054	2.2877	2.0965	2.2592	2.3504
CATE: L_2 error	0.6079	0.5274	0.5128	0.5053	0.5006
Number of unique predictions	888	1103	1238	1319	1382
	$K = 7$	$K = 10$	$K = 20$	$K = 48 \approx \sqrt{N}$	COT
CATE: L_∞ error	2.2042	2.2027	4.6307	4.7494	1.5933
CATE: L_2 error	0.5071	0.5227	0.5646	0.6988	0.4778
Number of unique predictions	1436	1474	1522	1554	2283

application of the map $T(x, z)$ to all untreated instances of one specific patient in the full 250 batches, to the histogram of the response y for all treated instances of the same patient. The prediction agrees very accurately with the underlying model, even though the patient appeared only once in the batch used for training.

Figure 5 benchmarks the predicted CATE using conditional optimal transport (right panel) versus nearest-neighbor estimation (left panel), which, given x and the corresponding value of z^*, estimates $\tau(z)$ by the difference between the y with closest z to z^* and x. As can be observed, the estimate of τ obtained via conditional optimal transport has a smaller variance than the one obtained using the KNN (k-nearest neighbors algorithm) with $K = 1$ in the latent (z) space. Error in CATE with $K > 1$ and the total number of unique predictions can be found in Table 1. The conditional optimal transport approach achieves the smallest error in L_2 and $L\infty$ norms compared with all benchmarks, and is able to produce a unique prediction for each individual.

The methodology of conditional optimal transport naturally extends to non-additive cases by controlling the complexity of maps and test functions. The following results are produced using the non-additive dataset in Hahn et al. (2017), where the causal effect is determined by a nonlinear map that involves both x and z that can not be expressed exactly by polynomials. In our approach, we include one extra degree of freedom by proposing an affine map $T(x, z) = x + \tau(z) + \tilde{\tau}(z)x$ and corresponding quadratic test function. We compare the results with the same KNN benchmarks by searching for the nearest neighbors in the latent (z) space, and the results are reported in Table 2. As can be seen, a modest increase in the complexity of our elementary map result in errors that are comparable with KNN after selecting the best value of K.

Yet we do not claim that our methodology offers the most accurate estimate of CATE in this particular application. In fact, most of the recent winners of the ACIC competition adopted methodologies based on Bayesian Additive Regression Trees (BART, Chipman et al., 2010) or variations thereof, which often display higher accuracy than our method. We argue in the section below that this success is due not only to virtues of BART, but also to the design principles underlying the generation of data for ACIC. We also argue that CATE is not optimal as a quantifier for the quality of a prediction. We chose to include this example non-the-less in order to show that our map-based methodology can compute CATE with an accuracy comparable with out-of-the shelf methods such as KNN. In the next section we specify in which sense the use of optimal map goes beyond the estimate of the CATE and discuss in more detail a comparison with BART.

4.2 Beyond the conditional averaged treatment effect

The estimate of CATE is essentially a regression problem where, given samples (x_i, z_i) of an outcome x (say blood pressure) and cofactor(s) (z) from untreated patients and samples (y_j, z_j) from treated patients, we estimate $E[Y|Z] - E[X|Z]$. One can go one step further and estimate from the data the two full conditional densities $\rho(x|z)$ and $\mu(y|z)$ and compare them with tests beyond their conditional means. *Yet this is still not equivalent to finding a conditional map T(x, z) between these two densities.*

One thing is to have an idea of how a given treatment affects the probability distribution of the blood pressure of a patient with given characteristics z (this is equivalent to

Table 2 Error in CATE and number of unique predictions using KNN benchmarks with different values of $K \geq 1$ and conditional optimal transport (COT), performed on nonadditive dataset with affine maps

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
CATE: L_∞ error	4.7776	4.1451	4.4079	4.8251	5.4309
CATE: L_2 error	1.2527	1.1131	1.0803	1.0636	1.0475
Number of unique predictions	888	1103	1238	1319	1382
	$K = 7$	$K = 10$	$K = 20$	$K = 48 \approx \sqrt{N}$	COT
CATE: L_∞ error	2.2042	2.2027	4.6307	4.7494	1.5933
CATE: L_2 error	0.5071	0.5227	0.5646	0.6988	0.4778
Number of unique predictions	1436	1474	1522	1554	2283

comparing $\rho(x|z)$ and $\mu(y|z)$) and quite another is to predict the effect of a treatment on a patient with characteristics z *and* blood pressure x. To stress that these two points are not the same consider that:

1. The value of x may encode other, idiosyncratic traits of the patient that the observed factors z do not capture. This is the case when, in addition to a set of observed cofactors z_o there is a set of hidden cofactors z_h affecting the outcome of the treatment.
2. The effect of a treatment may indeed depend on the blood pressure before treatment, providing much more specific information than just $\rho(x|z)$.

Even though the second of these considerations could be addressed through a regression procedure for y that includes x as an additional factor z, this is not doable when the data does not provide pairs (x, y), but only samples from the treated and untreated populations separately. This is the case of long term treatments, of quantifyng the effect of life-long habits, and also of the dataset from ACIC.

Summarizing, we seek the effect of the treatment on that specific patient given that the treatment on a population of patients with same characteristics z has the effect of changing $\rho(x|z)$ to $\mu(y|z)$.

We will now consider two toy models in which the data are generated to reproduce the two scenarios introduced above.

4.2.1 Model 1

Consider a model with two covariates, one observed z_o and another hidden z_h. Assume that z_h is a Gaussian random variable with mean zero and variance σ_h^2:

$$z_h \sim \mathcal{N}(0, \sigma_h^2).$$

The data relative to the outcome of untreated and treated patients (x and y respectively) are Gaussian:

$$x \sim \mathcal{N}(z_o + z_h, \sigma^2), \quad y \sim \mathcal{N}(az_o + bz_h, \sigma^2).$$

In this case the reference value of the CATE, in which we ideally assume that both z_o and z_h are known is given by

$$\text{CATE} = (a-1)z_o + (b-1)z_h. \tag{6}$$

We want to compare this value with the value obtained by COT and BART when only z_o is known. Let's first estimate the treatment effect via COT by first noticing that the conditional distributions $\rho(x|z_o)$ and $\mu(y|z_o)$ relative to patients before and after the treatment are $\mathcal{N}(z_o, \sigma_h^2 + \sigma^2)$ and $\mathcal{N}(az_o, b^2\sigma_h^2 + \sigma^2)$ respectively. The optimal map is then:

$$y = T(x, z_o) = \left(a - \sqrt{\frac{b^2\sigma_h^2 + \sigma^2}{\sigma_h^2 + \sigma^2}}\right) z_o + \sqrt{\frac{b^2\sigma_h^2 + \sigma^2}{\sigma_h^2 + \sigma^2}}\, x$$

and the treatment effect is estimated as $T(x, z_o) - x$. As anticipated, the treatment effect, computed via COT, includes the dependence on z_h through the dependence on x. By taking the expectation over the true distribution of x, we recover the CATE:

$$\mathbb{E}[T(x, z_o) - x | z_o, z_h] = (a-1)z_o + \left(\sqrt{\frac{b^2\sigma_h^2 + \sigma^2}{\sigma_h^2 + \sigma^2}} - 1 \right) z_h, \tag{7}$$

which is close to the truth if $b^2\sigma_h^2$ is much larger than σ^2, namely if the dependency of x on the hidden covariates z_h is still detectable in the post-treatment population.

The value in (7) should be compared with the value obtained by computing the regression of x and y as a function of z_o. Such value will only take the difference between the mean of y and x given z_o, which is $(a-1)z_o$ containing no information relative to z_h.

For a numerical comparison we generated the data according to the model defined above with parameters $a = 1/2$, $b = 2$, $\sigma_h = 3$ and $\sigma = 0.1$ and estimated CATE via COT and BART. We then computed the L_2 error between the result obtained via COT and BART with reference value of the CATE in (6). The error relative to COT is 0.61 and to BART is 3.09, confirming the analytical result derived above.

4.2.2 Model 2

The model used to generate the data in the ACIC competition has the form

$$\begin{cases} y = f(z, \epsilon_y) \\ x = \mu(z) + \epsilon_x \end{cases}$$

where ϵ_x and ϵ_y are normally distributed, independent random variables *and different samples of z are used to generate independent samples for x and y*. Hence, in this model, x and y are conditionally independent given z. As mentioned at the end of the previous section, this is exactly the model that BART is designed for, and hence it is not surprising that the winners of the last few years of the ACIC competitions have used BART.

The strength of our method is not in the precision with which it can estimate the CATE, but rather in its capacity to access information contained in the data that cannot be captured by a procedure based purely on regression. In order to illustrate how computing the map between ρ and μ goes beyond the estimate of the CATE, consider data generated, more realistically, according to the second scenario described in Sect. 4.2:

$$\begin{cases} y = f(z, x, \epsilon_y) \\ x = \mu(z) + \epsilon_x, \end{cases} \tag{8}$$

where now y and x are not conditionally independent given z. In this case, computing the CATE is of limited use, because fixing z does not result in a response to a given treatment that in average is independent of the value of x. The response to the treatment in this case depends not only on the set of observed factors z but also on the pre-treatment value of blood pressure x of that specific patient. Therefore computing $E[Y|X = x, Z = z]$ in this case is much more meaningful than $E[Y|Z = z]$.

Computing $E[Y|X, Z]$ from data generated using (8) where, as before, different samples of z are used to generate samples for x and y is, to the best of our knowledge, not doable with plain regression procedures, since the data set does not come in triplets (x_i, y_i, z_i) but in pairs (x_i, z_i), (y_j, z_j), where for a given value of z_i we have either x_i or y_i but not both. By contrast, the information contained in the optimal map $T(x, z)$ allows for such an estimate as given z_i and x_i we can recover $y_i = T(x_i, z_i)$.

4.3 Lightness transfer

Next we apply conditional optimal transport to lightness transfer. Consider the first column of Fig. 6, corresponding to two flowers photographed under different light conditions. We seek to transform the first photograph so as to present it under the light conditions of the second. This goes beyond merely changing lightness uniformly, since for instance at sunset certain colors are perceived as having become darker than others.

An image can be represented in the three dimensional CIELAB (L*a*b) space whose coordinates are the lightness L, the red/green contrast A and blue/yellow contrast B. The right column of Fig. 6 shows the images of the flowers in this L*a*b space, where each point corresponds to a superpixel, defined through a clustering procedure to introduce information about the geometry of the image (Rabin et al., 2014). We follow Tai et al. (2005) to define a similarity metric by means of a Gaussian kernel, map the obtained superpixels through our procedure, and use a TMR filter after the map to recover sharp details (Rabin et al., 2011).

Figure 7 shows the results obtained by changing lightness through three different procedures. First (left column) we use one-dimensional optimal transport (with quadratic cost) to map the L coordinate, ignoring the values of A and B. The L*a*b diagram shows that this results in a nearly uniform shift of L towards smaller values. The third column shows the effect of mapping the starting image to the target image through 3d optimal transport in the full L*a*b space. In this case the point clouds overlap to a much better degree, yet we observe that the color of the lotus has been changed too much towards the color of the poinsettia in the target image. The second column is obtained performing optimal transport of L conditioned on A and B. Contrasting to

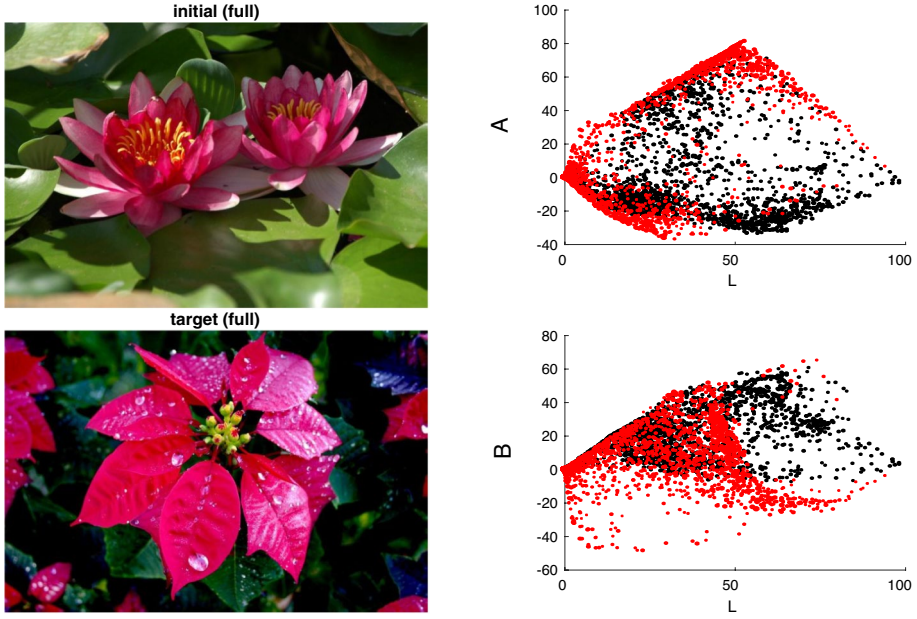


Fig. 6 Left column: initial (top) and target (bottom) image. Right column: L*a*b coordinates for the initial (in red) and the target (in black) image (Color figure online)

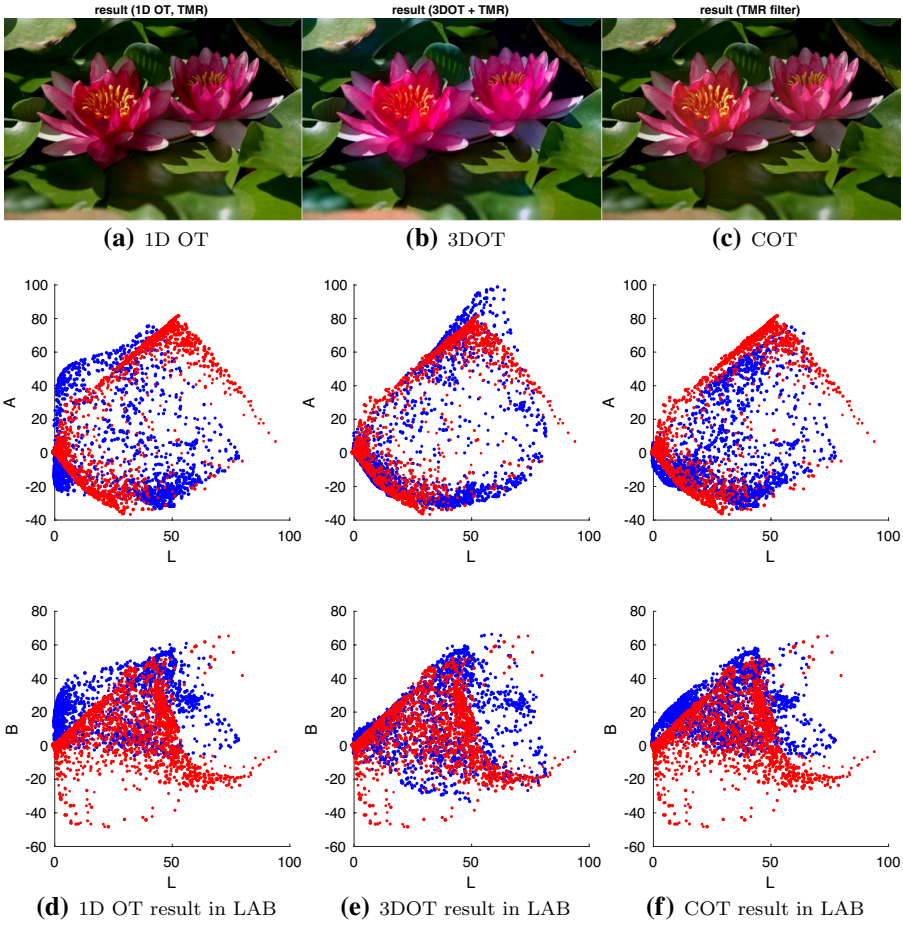


Fig. 7 Left Column: image obtained performing one dimensional optimal transport for the Luminosity (L) coordinate ignoring the A and B coordinates. Second column: plain three dimensional optimal transport in L*a*b space. Third column: image obtained by performing optimal transport on luminosity conditioned on color (Color figure online)

the other two results, here the lotus has kept its original color, and the lightness has changed to a different degree for the lotus than for the background leaves.

This is a general advantage of conditional optimal transport: unlike its unconditional cousin, it is not constrained to preserve total mass (in this case, transferring fully one color palette to the other), but only the mass for each value of z. This points to an additional application of conditional optimal transport: its capacity to address possible unbalances between source and target by parameterizing the transfer map by means of convenient labels z. In work in progress, we expand on this notion, finding those latent covariates z that help resolve unbalances optimally.

4.4 Color contrast and Lightness transfer

In this section we show an example in which performing color transfer, in addition to lightness transfer, can be used to simulate the effect of a restoration under different conditions of lightness (L) and color contrast in the CIELAB space. We present this example to display the broad set of options provided to the user by condition optimal transport.

The source image represents Michelangelo's Jesse spandrel in the Sistine Chapel before the restoration that took place in the period 1984–1994. We chose this particular image because of the controversy that followed its restoration (Beck & Daley, 1993). Comparing the first and the last panels of the second row of Fig. 8, representing the image before and after the restoration respectively, one can notice the disappearance of the eyes of Jesse and the loss of depth in his vest. In order to simulate a series of possible effects of the restoration process, we perform lightness and color transfer between the source image and two target images corresponding to two frescos dating back to roughly the same time period as the source image. These two frescos have been chosen because they underwent a successful restoration process. In the first row of Fig. 8, we chose a fresco by Luca Signorelli (San

Fig. 8 Three color transfers obtained with same source image and different target images, one for each row. 2D OT: two dimensional optimal transport in the a*b space, independently from the value of L. 2D COT: two dimensional optimal transport conditional on the value of L. 3D OT: 3 dimensional optimal transport in the L*a*b space. 1+2 OT: One dimensional optimal transport performed on L alone followed by two dimensional optimal transport in a*b space. OT + COT: one dimensional optimal transport in the L space followed by two dimensional optimal transport conditional on the value of L. Since the lightness and the color are correlated, the ability of conditioning over lightness (third and sixth column) provides an additional tool to mitigate the presence of an unbalanced color palette between the source and the target image (Color figure online)

Brizio Chapel - Orvieto) and in the third row, a fresco by Michelangelo himself, the conversion of Saint Paul in the Pauline chapel. The results obtained using these two frescos as targets should be compared to the second row of Fig. 8, where we use for target the actual restored version of the source image, which corresponds to the colors actually applied by Michelangelo, after eliminating those surface layers that have deteriorated the most.

It is worth noticing that, when conditioning on lightness (third and sixth column of Fig. 8), we obtain a figure that is less affected by the difference in color density between the target and the source image. For instance, the target image of the first row is characterized by a larger amount of vivid red that is not present in the source image. This results in most of the transported images (second to sixth columns) to be characterized by a shift in the red color. Similarly, the target image of the third row is characterized by much more yellow/orange palette than the source image.

Hence using lightness as a variable to condition over, has the effect of mitigating the presence of an unbalanced color palette between the source and the target image. A possible reason is that lightness can often be used as as surrogate variable for the color label, as different colors are often attached to different values of lightness. Therefore, conditioning on lightness has the effect of establishing a correspondence between areas in the target and source image characterized by the same colors, even if these areas do not have the same size.

The purpose of this experiment is not to show that conditional optimal transport performs better than simple optimal transport when used for color transfer. Instead, it intends to show that the ability of conditioning over lightness provides additional tools that the practitioner of color transfer can use with different target images to create alternative if-scenarios when restoring a fresco.

5 Conclusion

This work develops conditional optimal transport (COT), a variant of the classical optimal transport problem where the distributions to match are conditioned to cofactors. In particular, the data-driven case is considered, where the two conditional probabilities $\rho(x|z_1, \dots, z_L)$ and $\mu(y|z_1, \dots, z_L)$ are known only through samples. A formulation is developed that integrates all conditional maps $T(x, z)$ into a single minimax problem, providing an adaptive, adversarial game theoretical framework for the satisfaction of the push forward conditions.

Ignoring the dependence on cofactors can lead to wrong estimates for the map for two reasons: the map may truly depend on these ignored covariates, as when the effect of as treatment depends on the age of the patient, and/or the distributions of the covariates may differ in the source and target distributions, as when comparing hospitals which serve populations with different ratio of ethnicities. These two effects appear prominently in our application of COT to the ACIC Data Analysis Challenge 2017 data-set, where the effect of a medical treatment depends on a set of discrete covariates, and the distributions of the diagnostic variable x in the treated and untreated populations differ not only due to the effect of the treatment, but also to the unbalance in a different set of continuous covariates.

COT provides a flexible tool for data analysis. For instance, in cases where there are no explicit covariates, one can choose some of the variables x as covariates z. This choice may be driven by field knowledge and, when ambiguous, experiments with COT may shed light on the effect of each particular choice. This is illustrated through simple synthetic

examples and through applications to lightness and color transfer. The various choices can be used, for instance, to change the lightness condition of an image, and to simulate the effect of the restoration of frescos under different assumptions on the effect of the passing of time on color contrast and lightness.

Still another use of COT, currently under development, would seek those hidden cofactors z under which the conditional transfer is optimal under a user-determined criterion.

Acknowledgements The work of E. G. Tabak and W. Zhao was partially supported by NSF Grant DMS-1715753 and ONR Grant N00014-15-1-2355.

References

Beck, J., & Daley, M. (1993). *Art restoration: The culture, the business and the scandal*. John Murray.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics, 4*(1), 266–298.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. Wiley.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2016). Density estimation using real nvp. arXiv preprint arXiv:1605.08803.

Donsker, M. D., & Varadhan, S. R. S. (1975). Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics, 28*(1), 1–47.

Essid, M., Tabak, E. G., & Trigila, G. (2019). An implicit gradient-descent procedure for minimax problems. *Submitted to Machine Learning*.

Finlay, C., Jacobsen, J.-H., Nurbekyan, L., & Oberman, A. (2020). How to train your neural ode: The world of Jacobian and kinetic regularization. In *International conference on machine learning* (pp. 3154–3164). PMLR.

Grover, A., Dhar, M., & Ermon, S. (2017). Flow-gan: Combining maximum likelihood and adversarial learning in generative models. arXiv preprint arXiv:1705.08868.

Hahn, P. R., Dorie, V., & Murray, J. S. (2019). Atlantic causal inference conference (acic) data analysis challenge 2017. arXiv preprint arXiv:1905.09515.

Nowozin, S., Botond C., & Ryota T (2016). f-gan: Training generative neural samplers using variational divergence minimization. In: *Proceedings of the 30th international conference on neural information processing systems* (pp. 271–279).

Oliva, J. B., Dubey, A., Zaheer, M., Poczos, B., Salakhutdinov, R., Xing, E. P., & Schneider, J. (2018). Transformation autoregressive networks. arXiv preprint arXiv:1801.09819.

Rabin, J., Delon, J., & Gousseau, Y. (2011). Removing artefacts from color and contrast modifications. *IEEE Transactions on Image Processing, 20*(11), 3073–3085.

Rabin, J., Ferradans, S., & Papadakis, N. (2014). Adaptive color transfer with relaxed optimal transport. In *2014 IEEE international conference on image processing (ICIP)* (pp. 4852–4856). IEEE.

Rezende, D. J., & Mohamed, S. (2015). Variational inference with normalizing flows. arXiv preprint arXiv:1505.05770.

Tabak, E. G., & Turner, C. V. (2013). *A family of non-parametric density estimation algorithms*. In CPAM, LXVI.

Tabak, E. G., & Vanden-Eijnden, E. (2010). Density estimation by dual ascent of the log-likelihood. *Communications of the Mathematical Science*, 8.

Tai, Y.-W., Jia, J., & Tang, C.-K. (2005). Local color transfer via probabilistic segmentation by expectation-maximization. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 747–754). IEEE.

Trigila, G., & Tabak, E. G. (2016). Data-driven optimal transport. *Communications on Pure and Applied Mathematics, 69*(4), 613–648.

Villani, C. (2003). *Topics in optimal transportation*. Number 58. American Mathematical Soc.

Yang, H., & Tabak, E. G. (2019). Conditional density estimation, latent variable discovery and optimal transport. arXiv preprint arXiv:1910.14090.