

Data Driven Conditional Optimal Transport

Esteban G. Tabak · Giulio Trigila ·
Wenjun Zhao

Received: date / Accepted: date

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 [9]. 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

Esteban G. Tabak
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA
E-mail: tabak@cims.nyu.edu

Giulio Trigila
Baruch College, CUNY, 55 Lexington avenue, New York, NY 10035, USA
E-mail: giulio.trigila@baruch.cuny.edu

Wenjun Zhao
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA
E-mail: wenjun@cims.nyu.edu

quantities that characterize the health 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_i . 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_i . The procedure is data driven, as it uses only samples $\{x^i, z_1^i, \dots, z_L^i\}$ and $\{y^j, z_1^j, \dots, z_L^j\}$ from ρ and μ . Notice that we do not seek a pairwise matching between $\{x^i, z_1^i, \dots, z_L^i\}$ and $\{y^j, z_1^j, \dots, z_L^j\}$: 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 section 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 section 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 Figure 1 for a sketch relative 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 [9]. 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.

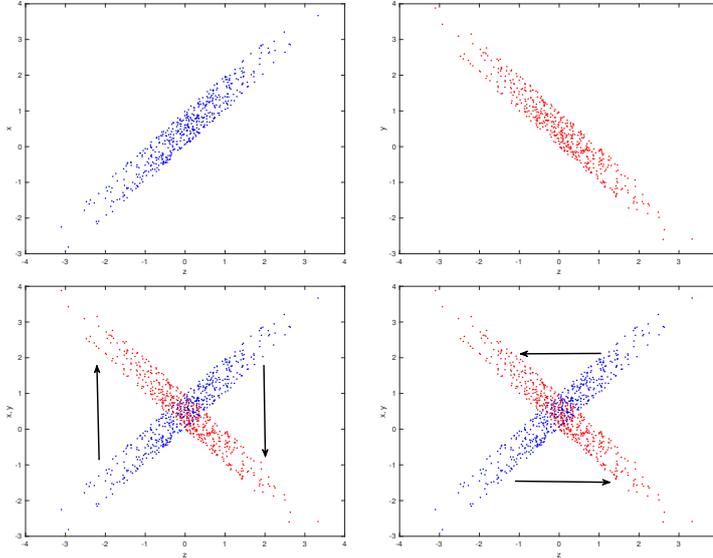


Fig. 1: Upper row: source (left) and target (right) distributions. Lower left: optimal transport of x conditioned on z . The arrows indicate that the lower left branch and the upper right branch of the source distribution are mapped respectively to the upper left branch and the lower right branch of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

The plan of this article is as follows: after this introduction, section 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 section 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 :

$$\begin{cases} \min_{T(\cdot, z)} \int c(T(x, z), x) \rho(x|z) dx \\ T\#\rho(\cdot|z) = \mu(\cdot|z), \end{cases} \quad (1)$$

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 $\mu(y|z) = T\#\rho(\cdot|z)$. Since $T(\cdot, z)$ 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

$$\begin{cases} \min_{T(\cdot, z)} \int c(T(x, z), x) \rho(x, z) dx dz \\ \forall z T\#\rho(\cdot|z) = \mu(\cdot|z), \end{cases} \quad (2)$$

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 $\{x^i, z_x^i\}$ and $\{y^j, z_y^j\}$. 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

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

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 :

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

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:

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

the conditional Kullback-Leibler divergence between ρ_1 and ρ_2 ([2]). Since this is non-negative, we can rewrite the problem in (2) as

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

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 $\lambda \rightarrow \infty$, 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

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

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 [2]:

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

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

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

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 [3]:

$$D_{KL}(\rho_1||\rho_2) = \max_g \left[\int g(x, z) \rho_1(x, z) dx dz - \log\left(\int e^{g(x, z)} \rho_2(x, z) dx dz\right) \right], \quad (3)$$

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

$$\begin{aligned} & \min_T \max_g \left[\int c(T(x, z), x) \rho(x, z) dx dz + \lambda \left[\int g(y, z) \mu(y, z) dy dz \right. \right. \\ & \left. \left. - \log\left(\frac{1}{2} \int e^{g(y, z)} \mu(y, z) dy dz + \frac{1}{2} \int e^{g(T(x, z), z)} \rho(x, z) dx dz\right) \right] \right] \end{aligned} \quad (4)$$

or, in terms of samples,

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

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:

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

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.

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 point-wise, 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.
4. Be apt to robust and effective optimization.

These conditions leave space for many proposals. For instance, we could define both T and g through neural networks. 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 .

3.1 Evolving Gaussian mixtures

We adopt as elementary map the gradient of a convex potential function: $E(x, z) = \nabla_x \Phi(x, z)$, chosen from a family that includes the identity map. Convexity of the potential guarantees that the resulting elementary map is one-to-one. The potential Φ is built from a quadratic form in x with coefficients that depend on z , plus a combination of Gaussians in (x, z) space, and similarly for the test function g . By having the centers and amplitudes of these Gaussians evolve, we can approximate quite general functions Φ and g .

In order to guarantee the convexity of Φ , notice that the gradient 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 $\frac{1}{2d^2} \frac{\|\mathbf{x}\|_2^2}{2} \pm G_d(\mathbf{z}, \mathbf{m}_{z_i}) G_d(\mathbf{x}, \mathbf{m}_i)$ is convex, so we propose

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

with $\mathbf{C}_{2,0}, \mathbf{C}_{2,1}$ lower triangular. 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

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

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 $d = \text{quantile}(\text{pdist}([y; z]), 1/K)$, where pdist is the pairwise distance function. With this choice there are approximately $1/K$ points in the effective support of each Gaussian.

For the test function, we propose

$$g(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^K \alpha_i G_d(\mathbf{z}, \mathbf{m}_{z_i}) G_d(\mathbf{x}, \mathbf{m}_i) + (\beta_0^T + \mathbf{z}^T \beta_1) \mathbf{x} + \mathbf{x}^T (\beta_2 + \sum_i \beta_{3,i} z_i) \mathbf{x},$$

with each iteration starting at the parameter values from the previous step. The Gaussian centers are treated differently in the test function g , where they are extra parameters to ascend, and in the potential Φ , 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.

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 . 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 in z space, through the composition of function of the form

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

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

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

and the test function via

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

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 min-max formulation.

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 (5) we use the general procedure described in [4].

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 [5] (<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 [5]. 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.

The middle panel of Figure 2 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:

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

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 Figure 3 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 Figure 2 for $\gamma(z\tau)$), the results are biased, much as in the synthetic example in the introduction. The middle panel shows that, when

all covariates are considered, this bias is resolved. The right panel compares the 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.

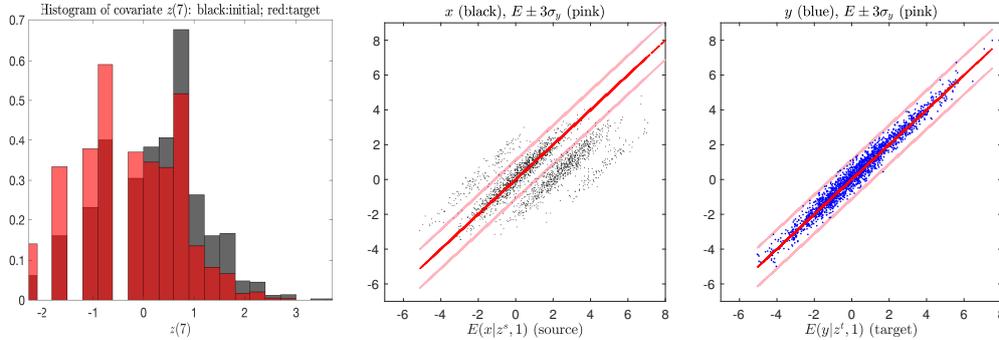


Fig. 2: Left panel: Unbalance in the distribution of z_7 between the source and the target data set. Center: Response variable x for patients before the treatment plotted as a function of the theoretical expected value that the same patients would have if they would undergone the treatment. Right: Response variable y of the treated patients as a function of the theoretical expected value of the same patients.

Figure 4 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 nearest neighbors in z space.

4.2 Lightness transfer

Next we apply conditional optimal transport to lightness transfer. Consider the first column of Figure 5, 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 Figure 5 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

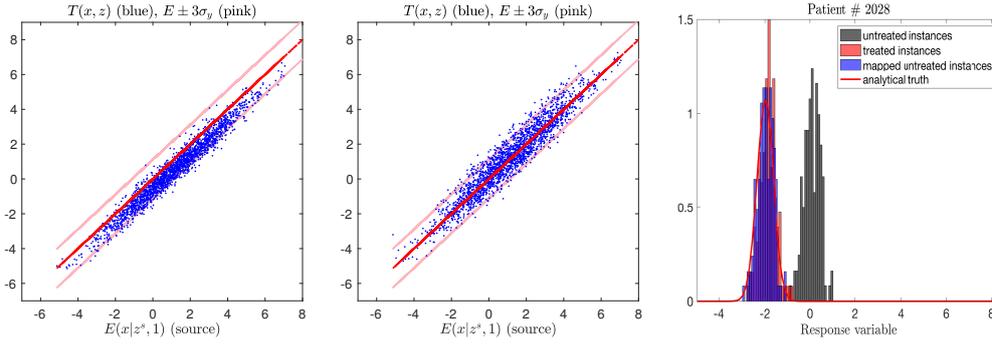


Fig. 3: Left: numerical value of the map $T(x_i, z_i)$ obtained using only the discrete covariates, which are the only ones that the true T depends on. The result is biased due to the unbalance between γ^x and γ^y for $\gamma(z_\tau)$. Middle: numerical value of the map $T(x_i, z_i)$ obtained using all 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.

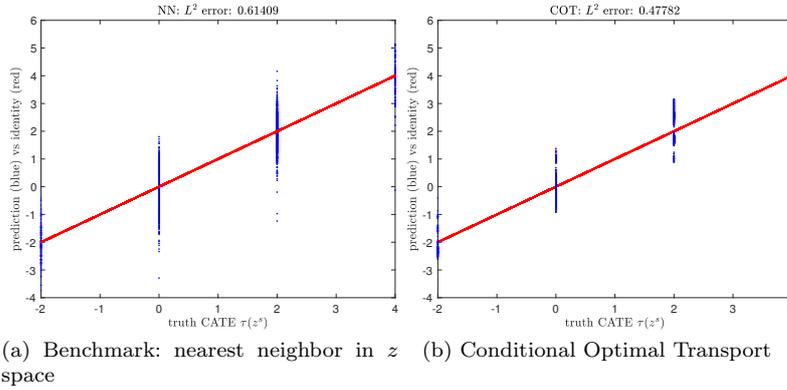


Fig. 4: Predicted CATE using benchmark (left) and conditional optimal transport (right).

[7]. We follow [8] 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 [6].

Figure 6 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 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.

4.3 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 [1]. Comparing the first and the last panels of the second row of Figure 7, 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 Figure 7, we chose a fresco by Luca Signorelli (San 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 Figure 7, 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 Figure 7), 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.

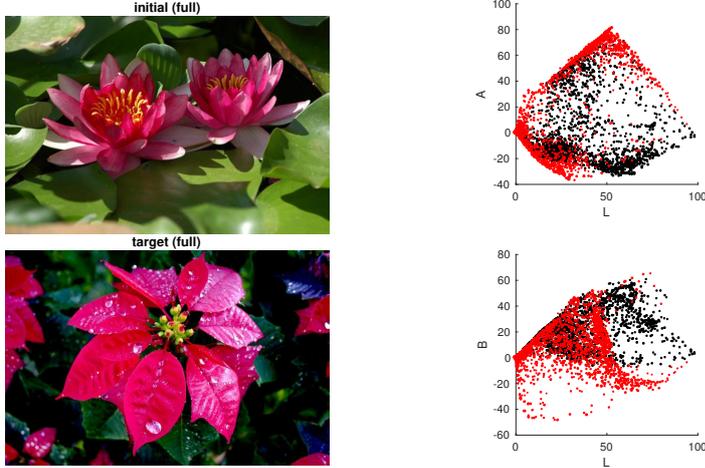


Fig. 5: Left column: initial (top) and target (bottom) image. Right column: $L^*a^*b^*$ coordinates for the initial (in red) and the target (in black) image

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 a 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 .

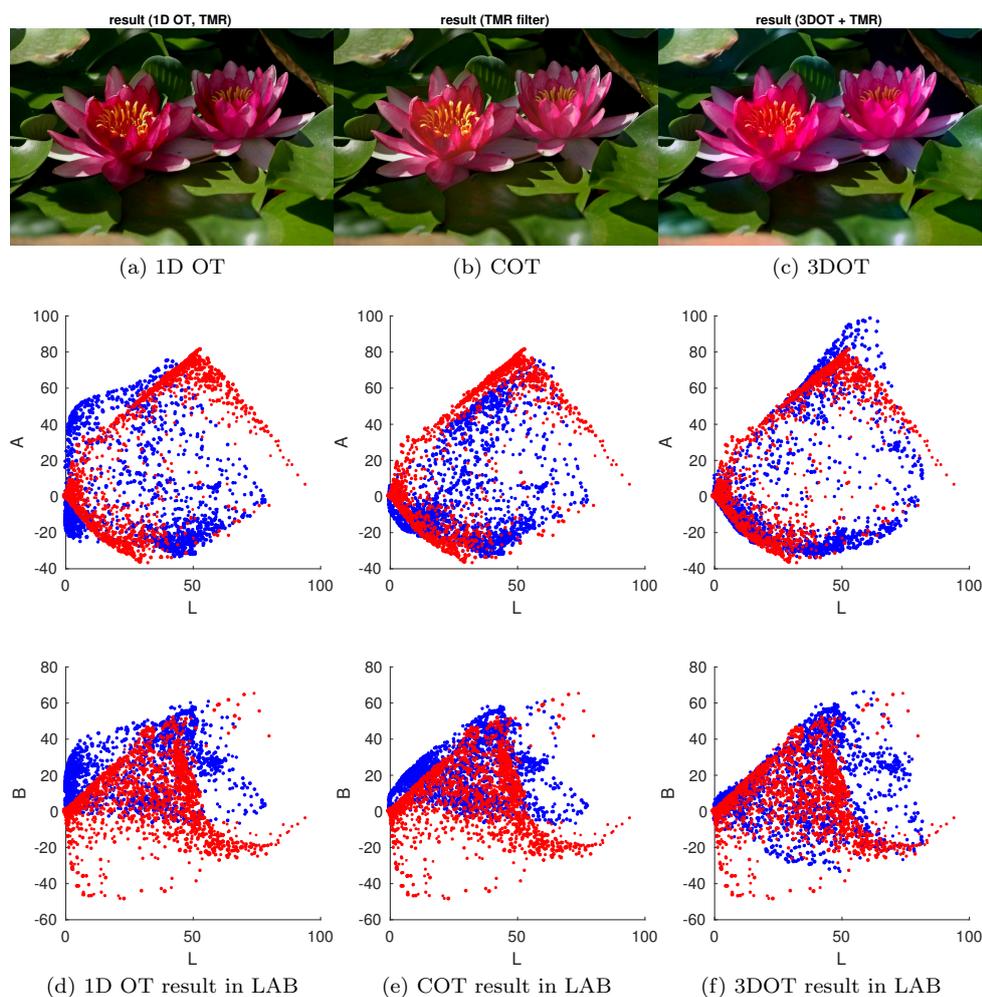


Fig. 6: Left Column: image obtained performing one dimensional optimal transport for the Luminosity (L) coordinate ignoring the A and B coordinates. Second column: image obtained by performing optimal transport on luminosity conditioned on color. Third column: plain three dimensional optimal transport in L^*a^*b space.

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.

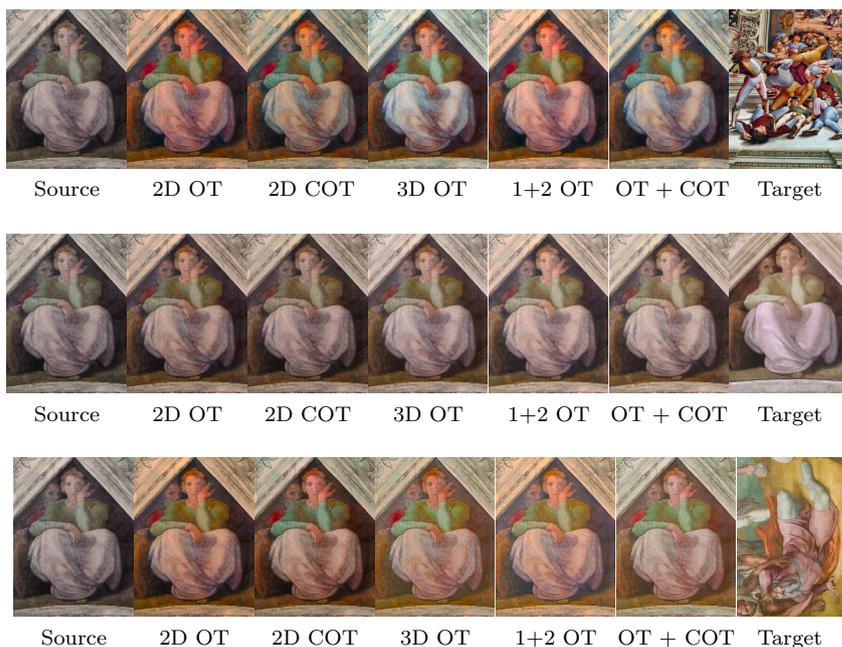


Fig. 7: Three color transfers obtained with same source image and different target images, one for each row. 2D OT: two dimensional optimal transport in the $a*b$ space, independently from the value of L . 2D COT: two dimensional optimal transport conditional on the value of L . 3D OT: 3 dimensional 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 .

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

1. James Beck and Michael Daley. *Art restoration: the culture, the business and the scandal*. John Murray London, 1993.
2. Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
3. Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.
4. Montacer Essid, Esteban G. Tabak, and Giulio Trigila. An implicit gradient-descent procedure for minimax problems. *Submitted to Machine Learning*, 2019.
5. P Richard Hahn, Vincent Dorie, and Jared S Murray. Atlantic causal inference conference (acic) data analysis challenge 2017. *arXiv preprint arXiv:1905.09515*, 2019.
6. Julien Rabin, Julie Delon, and Yann Gousseau. Removing artefacts from color and contrast modifications. *IEEE Transactions on Image Processing*, 20(11):3073–3085, 2011.
7. Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4852–4856. IEEE, 2014.
8. Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang. Local color transfer via probabilistic segmentation by expectation-maximization. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 747–754. IEEE, 2005.
9. Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.