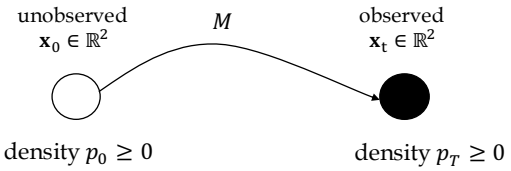
Optimal Transport For Causal Discovery

This is a summary of the paper Optimal Transport For Causal Discovery (Tu et al., 2022).

1 Problem Setup



The observed data points $\mathbf{x}_T = [X, Y]$ at a certain time T are viewed as the product of the dynamic transfer from an unmeasured independent noise $\mathbf{x}_0 = [E_x, E_y]$ at time 0. Let's consider a bi-variate case where X **is a cause of** Y i.e., $X \to Y$ with a Functional Causal Model (FCM): $Y = f(X, E_y)$.

Given $\mathbf{x}_0, \mathbf{x}_T \in \mathbb{R}^2$ and the probability densities $p_0, p_T \geq 0$, the mass transfer scenario realized under the existence of a map $M : \mathbb{R}^2 \mapsto \mathbb{R}^2$ can be described as

$$\mathbf{x}_T = \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} E_x \\ f(X, E_y) \end{bmatrix} = M\left(\begin{bmatrix} E_x \\ E_y \end{bmatrix} \right) = M(\mathbf{x}_0)$$

From the perspective of FCM-based causal discovery approaches, causal influences are represented by FCMs which represent the effect as a function of its direct cause and an unmeasured noise satisfying the FCM constraints:

1. *The map constraint:* the values of X are determined by the values of its corresponding noise, i.e., $X = E_x$. while the values of the effect depend on cause X and noise E_y; for the initial and final values of the cause, the value of its observation X is equal to its noise value E_x.
2. *The independence constraint:* : two random variables (cause and effect) at the initial time have the joint probability density function $p_0(\mathbf{x}_0)$ and independent, i.e., E_x is independent of E_y.

In the remainder, we will see how this dynamical view gives rise to a property unique to the mechanism X is a cause of Y and vice versa, the directionality $X \to Y$ is identifiable if such a condition holds.

2 Characterization of Optimal Map

We first examine the optimality of M under FCM-based causal discovery constraints. The optimal transport M^* is the minimizer of the **dynamical** L^2 **Wasserstein distance** W_2^2.

Benamou & Brenier (2000) formulate the L^2 Monge-Kantorovich problem as a convex space-time minimization problem in a continuum mechanics framework.

Proposition 1. Given a fixed time interval $[0, T]$, the motions of particles are described with the density $\rho_t := \rho(t, \mathbf{x}_t) \geq 0$ and the velocity field $\mathbf{v}_t := \mathbf{v}(t, \mathbf{x}_t)$, the dynamical formulation of W_2^2 is given as

$$W_2^2 = \inf_{\rho, \mathbf{v}} T \int_{\mathbb{R}^2} \int_0^T \rho(t, \mathbf{x}_t) |\mathbf{v}(t, \mathbf{x}_t)|^2 d\mathbf{x}_t dt \tag{1}$$

such that the following conditions are satisfied

$$\begin{cases} \text{initial and final conditions: } \rho(0,\cdot) = p_0, \rho(T,\cdot) = p_T \\ \text{the continuity equation: } \partial_t \rho_t + \nabla \cdot (\rho_t \mathbf{v}_t) = 0 \end{cases}$$

where $\nabla\cdot$ is the divergence in vector calculus. Recall that the divergence of a $3-$dimensional vector field $\mathbf{v}$, for example, is given as

$$\nabla \cdot \mathbf{v} = \frac{\partial \mathbf{v}_x}{\partial x} + \frac{\partial \mathbf{v}_y}{\partial y} + \frac{\partial \mathbf{v}_z}{\partial z}$$

where $\mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z$ are components of $\mathbf{v}$ along the x, y, z axes.

At some point (a, b), the divergence $\nabla \cdot \mathbf{v}(a, b) < 0$ means the fluid flowing along the vector field defined by $\mathbf{x}$ would tend to become **more dense** at the point (a, b). If the divergence $\nabla \cdot \mathbf{v}(a, b) > 0$, the fluid flowing along the vector field becomes **less dense** around (a, b). The zero divergence $\nabla \cdot \mathbf{v}(a, b) = 0$ indicates that even though a fluid flows freely, its density stays **constant**.

Suppose M^* is the solution given by the minimization of W_2^2, the corresponding flows follow the *time evolution equation* with $t \in [0, T]$

$$\mathbf{x}_t = \mathbf{x}_0 + \frac{t}{t}\mathbf{v}(t, \mathbf{x}_t),\ \mathbf{v}(t, \mathbf{x}_t) = \mathbf{v}(0, \mathbf{x}_0) = M^*(\mathbf{x}_0) - \mathbf{x}_0 \tag{2}$$

The time evolution equation shows that $\mathbf{x}_t$ is a convex combination of $\mathbf{x}_0$ and $M^*(\mathbf{x}_0)$ and that the velocity fields do not depend on time. Such optimal flows are *pressureless potential flows* of which the fluid particles are not subject to any pressure or force and the trajectories are determined given their initial positions and velocities or given their initial and final positions. The density ρ and the velocity $\mathbf{v}$ of moving particles can be considered as the probability density and the velocity of changing values of data points.

Proof. We assume ρ_0 and ρ_T to be compactly supported in $\mathbb{R}^d$ and bounded. The framework of the Monge-Kantorovich problem assumes both density functions are bounded with total mass one

$$\int_{\mathbb{R}^d} \rho_T(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^d} \rho_0(\mathbf{x}) d\mathbf{x} = 1$$

We say that a map M from $\mathbb{R}^d$ to $\mathbb{R}^d$ realizes the transfer of ρ_0 to ρ_T if, for all bounded subset $A \subset \mathbb{R}^d$

$$\int_{\mathbf{x} \in A} \rho_T(\mathbf{x}) d\mathbf{x} = \int_{M(\mathbf{x}) \in A} \rho_0(\mathbf{x}) d\mathbf{x} \tag{3}$$

Let us consider (sufficiently smooth) fields ρ and $\mathbf{v}$ satisfying the initial/final conditions and the continuity equation. We use Lagrangian coordinates and define location $\boldsymbol{X}(t, x)$ by

$$\boldsymbol{X}(0, \mathbf{x}) = \mathbf{x}_0, \quad \partial_t \boldsymbol{X}(t, \mathbf{x}) = \mathbf{v}(t, \boldsymbol{X}(t, \mathbf{x})),$$

such that, for all test functions f,

$$\int_{\mathbb{R}^d} f(t, \mathbf{x}) \rho(t, \mathbf{x}) d\mathbf{x} dt = \int_{\mathbb{R}^d} f(t, \boldsymbol{X}(t, \mathbf{x})) \rho_0(\mathbf{x}) d\mathbf{x} dt \tag{4}$$

Note that Eq. (4) implies that Eq. (3) holds i.e., a valid OT problem. For $t \in [0, T]$, we have

$$\begin{aligned}
T \int_{\mathbb{R}^d} \int_0^T \rho(t, \mathbf{x}_t) |\mathbf{v}(t, \mathbf{x}_t)|^2 d\mathbf{x}_t dt &= T \int_{\mathbb{R}^d} \int_0^T \rho_0(\mathbf{x}_0) |\mathbf{v}(t, \boldsymbol{X}(t, \mathbf{x})|^2 d\mathbf{x}_0 dt \\
&= T \int_{\mathbb{R}^d} \int_0^T \rho_0(\mathbf{x}_0) |\partial_t \boldsymbol{X}(t, \mathbf{x})|^2 d\mathbf{x}_0 dt \\
\text{(by Jensen's inequality)} &\geq \int_{\mathbb{R}^d} \rho_0(\mathbf{x}_0) \left(\int_0^T |\partial_t \boldsymbol{X}(t, \mathbf{x})| \right)^2 dt d\mathbf{x}_0 \\
&= \int_{\mathbb{R}^d} \rho_0(\mathbf{x}_0) |\boldsymbol{X}(T, \mathbf{x}) - \boldsymbol{X}(0, \mathbf{x})|^2 d\mathbf{x}_0 \\
&= \int_{\mathbb{R}^d} \rho_0(\mathbf{x}_0) |\boldsymbol{X}(T, \mathbf{x}) - \mathbf{x}_0|^2 d\mathbf{x}_0 \\
&= \int_{\mathbb{R}^d} \rho_0(\mathbf{x}_0) |\mathbf{x}_T - \mathbf{x}_0|^2 d\mathbf{x}_0 \\
(M^* \text{ is optimal}) &\geq \int_{\mathbb{R}^d} \rho_0(\mathbf{x}_0) |M^*(\mathbf{x}_0) - \mathbf{x}_0|^2 d\mathbf{x}_0
\end{aligned}$$

By definition, the L^p Wasserstein distance between p_0 and p_T takes the form of $W_p(p_0, p_T)^p = \inf_M \int |M(\mathbf{x}_0) - \mathbf{x}_0|^p p_0(\mathbf{x}_0) d\mathbf{x}_0$. This indirectly completes the proof for Proposition 1.

Finding the optimal map M in the original OT problem is equivalent to finding the optimal $(\rho_t, \mathbf{v}_t)$ in continuum mechanics. At optimality, the choice of $\boldsymbol{X}(t, \mathbf{x})$ at $t = T$ must satisfy

$$\boldsymbol{X}(T, \mathbf{x}) - \mathbf{x}_0 = M^*(\mathbf{x}_0) - \mathbf{x}_0$$

The appropriately optimal choice for $\boldsymbol{X}(t, \mathbf{x})$ is

$$\boldsymbol{X}(t, \mathbf{x}) - \mathbf{x}_0 = \frac{t}{T} \Big(M^*(\mathbf{x}_0) - \mathbf{x}_0 \Big)$$

This gives rise to

$$\mathbf{x}_t = \mathbf{x}_0 + \frac{t}{T} \big(\mathbf{x}_T - \mathbf{x}_0 \big) = \mathbf{x}_0 + \frac{t}{T} \big(M^*(\mathbf{x}_0) - \mathbf{x}_0 \big)$$

We can also observe that

$$\frac{\mathbf{x}_t - \mathbf{x}_0}{t} = \mathbf{v}(t, \mathbf{x}_t) = \frac{M^*(\mathbf{x}_0) - \mathbf{x}_0}{T}$$

This proves the time evolution equation given in (2). (In the paper, they claim that $\mathbf{v}(t, \mathbf{x}_t) = \mathbf{v}(0, \mathbf{x}_0) = M^*(\mathbf{x}_0) - \mathbf{x}_0$ for $t \in [0, T]$. This is also unclear to me, which is probably a mistake.)

3 Optimal map under FCM constraints

Given the couplings $\mathbf{x}_0, \mathbf{x}_T$, note that the optimal transport M^* is not necessary to be the one generated from the ground-truth FCM. Yet, under FCM constraints, the form of M^* is determined to be

$$M^*(\mathbf{x}_0) = \begin{bmatrix} E_x \\ f(E_x, E_y) \end{bmatrix}$$

Proposition 2. Under FCM constraints, the square of L^2 Wasserstein distance between p_0 and p_T is

$$W_2^2(p_0, p_T) = \mathbb{E}_{E_x} \Big[W_2^2 \big(p(E_y), p(Y|E_x) \big) \Big] \tag{5}$$

Proof. Recall that the L^p Wasserstein distance between p_0 and p_T is defined by $W_p(p_0, p_T)^p = \inf_M \int |M(\mathbf{x}_0) - \mathbf{x}_0|^p p_0(\mathbf{x}_0) d\mathbf{x}_0$. Given optimal map M^*,

$$\begin{aligned} W_2^2(p_0, p_T) &= \int |M^*(\mathbf{x}_0) - \mathbf{x}_0|^2 p_0(\mathbf{x}_0) d\mathbf{x}_0 \\ &= \int_{E_x} \int_{E_y} f\big(E_x, E_y\big) - E_y\big)^2 p(E_y) dE_y p(E_x) dE_x \\ &= \mathbb{E}_{E_x}\big[W_2^2\big(p(E_y), p(Y|E_x)\big)\big] \end{aligned}$$

The derivation at the second line is due the independence of E_x and E_y. Note also that X remains at E_x throughout the process, so the distance along X-axis is zero.

4 Causal Direction Determination

We call an FCM an Additive Noise Model (ANM) if the effect is the sum of noise and a nonlinear function g of cause i.e., the structural assignments are of the form

$$Y := g(X) + E_y$$

Theorem 1. (Zero divergence of the velocity field) Under FCM constraints, the dynamical systems given by the L^2 Wasserstein distance are pressureless flows. Further under ANM constraint, they become volumne-preserving pressureless flows, of which the divergence of the velocity field, $\mathbf{v}(t, \mathbf{x}_t) = \big[v_x(t, x_t), v_y(t, y_t)\big]$, satisfies

$$\nabla \cdot \mathbf{v}(t, \mathbf{x}_t) = \frac{\partial_{v_x}(t, x_t)}{\partial_{x_t}} + \frac{\partial_{v_y}(t, y_t)}{\partial_{y_t}} = 0 \ \forall t \in [0, T],\ x_t, y_y \in \mathbb{R} \tag{6}$$

See Appendix D.3 in the paper for proof.

Proposition 3. (Divergence measure as a causal discovery criterion) Define the divergence measure as

$$D(\mathbf{v}) = \int_{\mathbb{R}^2} |\nabla \cdot \mathbf{v}|^2 p_0(\mathbf{x}) d\mathbf{x}_0 = \mathbb{E}_{\mathbf{x}_0}\big[|\nabla \cdot \mathbf{v}|^2\big] \tag{7}$$

where $\mathbf{v} = M^*(\mathbf{x}_0) - \mathbf{x}_0$. Suppose that the FCM and ANM constraints and identifiability conditions of ANMs are satisfied. The divergence measure of the corresponding dynamical system satisfies $D(\mathbf{v}) = 0$ if and only if X is the direct cause of Y.

See Appendix D.4 in the paper for proof.

5 The Algorithm

Given a set of observed data points of 2 variables X, Y, Proposition 7 suggests computing the expected divergence measure to infer whether $X \to Y$ or $Y \to X$. The method makes standard assumptions of Markov property, faithfulness and no latent confounding.

Step 1. Noise data generation: The first step is to generate data for $\mathbf{x}_0$. Since X is unchanged, we only need to assume the probability distribution of E_y and parameterize it with θ.

Suppose the dataset of $\mathbf{x}_T$ with N samples is given, denoted as $\big\{\big(x^i, y^i\big)\big\}_N$. A dataset of E_y with the sample size N: $\big\{e_y^i\big\}_N$ is generated with the following reparameterization trick

$$e_y^i = f_\theta^{noise}\left(e_y^{source}\right) = \theta \times e_y^{source} \sim \mathcal{N}(0,1)/\mathcal{U}(0,1)$$

where f_θ^{noise} a monotonic function parameterized by a neural network and e_y^{source} is sampled from a standard normal or uniform distribution. We randomly match the noise data $\big\{e_y^i\big\}_N$ with the data of X: $\big\{x^i\big\}_N$, which gives $\mathbf{x}_0 := \big\{\big(x^i, e_y^i\big)\big\}_N$.

Step 2. Couplings matching: To compute the divergence measure, we need to compute the velocity field $\mathbf{v} = M^*(\mathbf{x}_0) - \mathbf{x}_0$, which requires the appropriate couplings $\mathbf{x}_0$ and $\mathbf{x}_T$. Therefore, we now solve the classic matching problem by minimizing W_2^2 given in Eq. (5) which can be computing with Monte Carlo estimator.

Under the FCM constraints, only the y-coordinates need matching. Thus, the optimal transport problem is reduced to one-dimensional, which can be implemented with a sorting operation (Kolouri et al., 2019).

Step 2 yields the optimal map M^* and matching tuples $\big(\mathbf{x}_0; \mathbf{x}_T\big) := \big\{\big(x^i, e_y^i\big); \big(x^i, y^i\big)\big\}_N$.

Step 3. Variance-based Divergence Measure We now compute the divergence measure. The velocity field is given as

$$\mathbf{v} = M^*(\mathbf{x}_0) - \mathbf{x}_0 = \mathbf{x}_T - \mathbf{x}_0 = \begin{bmatrix} X \\ Y \end{bmatrix} - \begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} 0 \\ Y - E_y \end{bmatrix}$$

Thus, $|\nabla \mathbf{v}|^2 = \big(\frac{dv_y}{dy}\big)^2$. A straightforward way to approximate the derivative is using its nearest neighbour pair $\big(e_{xb}^i, e_{yb}^i\big)$ and $\big(xb^i, yb^i\big)$ and approximate it with

$$\frac{d}{dy}\big(v_y\big)|_{y=y^i} = \frac{(y^i - e_y^i) - (yb^i - e_{yb}^i)}{y^i - yb^i}$$

However in practice, the authors find that such approximation introduces several numerical issues:

- the denominator is in general a small number, and the distance to the nearest neighbour can be large in the few-sample case, which makes the computation unstable and inaccurate
- the deviation on the X-axis makes the approximation a biased estimate especially when the gradient of $g(X) + E_y$ at $X = x$ is large.

Therefore, they proposed the *variance-based divergence measure*, which is defined as

$$D_{var}(\mathbf{v}) = \mathbb{E}_X\big[\mathbb{V}[V_{y|X}|X]\big] \tag{8}$$

where $\mathbb{V}[V_{y|X}|X]$ represents a conditional variance of the velocity field V_y at position X at the initial time.

Since the velocity along the X-axis is zero, under all above constraints, the value of the divergence measure of an ANM is zero if and only if the value of the variance-based divergence measure is zero. The velocity field V_y is a random variable and computing the conditional variance simply involves computing the variance over velocities v_y^i among pairs with the same value of x^i.

Formally, let $\big\{v_{y|x}^i\big\}_{N_x}$ denote all the velocities v_y^i at position $X = x$ at the initial time, where the sample size is N_x and the mean value is $\bar{v}_{y|x}$.

$$\mathbb{V}[V_{y|X}|X = x] = \sum_{i=1}^{N_x}\big(v_{y|x}^i - \bar{v}_{y|x}\big)^2/\big(N_x - 1\big)$$

The variance-based divergence measure is computed as

$$D_{var}(\mathbf{v}) \approx \frac{1}{N}\sum_{x \in \{x^i\}_N} \frac{\|\mathrm{sort}(\overrightarrow{y_x}) - \mathrm{sort}(\overrightarrow{e_y}) - \mathrm{ave}(\overrightarrow{y_x} - \overrightarrow{e_y})\|_2^2}{N_x - 1}$$

where $\overrightarrow{y_x}$ is the vector of the Y samples where $X = x$; $\overrightarrow{e_y}$ is the vector of the E_y samples where $E_x = x$; $\mathrm{sort}\big(\,\cdot\,\big)$ sorts a vector; $\mathrm{ave}\big(\,\cdot\,\big)$ computes the vector mean; and $\|\cdot\|_2^2$ is the square of a l_2 norm.

Step 4. Fitting the noise distribution: We only initialize θ with some random value, and $p(E_y;\theta)$ is not necessary to be the true distribution or even significantly different from the true one, which can lead to the wrong result of the divergence measure. We therefore minimize the divergence measure w.r.t θ. Meanwhile, the divergence measure in the causal direction is zero if and only if $p(E_y;\theta^*)$ with the optimal parameters θ^* being the true noise distribution, implied by the identifiability of ANMs.

6 Extension to Multivariate case

In the case of multiple variables, one can use a constraint-based method to find the causal skeleton (the undirected causal graph) and then use the extension of the method for the edge orientation.

Suppose the general ANM is $X_i = g_i(\mathrm{Pa}_i) + E_i$ where $i = 1, ..., d$, E_i is the noise term of X_i and Pa_i denotes the parent variables of X_i. The square of L^2 Wasserstein distance is

$$W_2^2\big(p_0, p_T\big) = \sum_{i=1}^{m} \mathbb{E}_{\mathrm{Pa}_i}\big[W_2^2\big(p(E_i), p(X_i|\mathrm{Pa}_i)\big)\big]$$

Given the couplings of E_i and X_i, the corresponding dynamical system has zero divergence of its velocity field; the corresponding dynamical system which moves the samples of $\mathbf{x}_0$ to the samples of $\mathbf{x}_T$ under ANM constraints has zero divergence on each dimension.

$$D_{var}(\mathbf{v}) \approx \sum_{i=1}^{m} \frac{1}{N} \sum_{k \in \{\text{samples of } \mathrm{Pa}_i\}_N} \frac{\|\mathrm{sort}(\overrightarrow{x_{i|Pa_i=k}}) - \mathrm{sort}(\overrightarrow{e_i}) - \mathrm{ave}(\overrightarrow{x_{i|Pa_i=k}} - \overrightarrow{i})\|_2^2}{N_k - 1}$$

One could enumerate all possible DAGs of the causal skeleton and compute their measure values, of which the minimum value is corresponding to the causal graph.

Because the causal skeleton is given, it must be the case where one of the two variables of an edge is the cause and the other one is the effect. So the enumerated graphs have two situations:

1. all the edges are correctly oriented;
2. the causal direction of at least one edge is wrong such that the measure value of at least one causal module is significantly larger than the correct one (note that considering a child as the direct cause leads to increasing the measure value, while omitting a cause is not necessary to increase the measure value of the causal module).

Thus, we can simply choose the graph with the minimum measure value as the causal one.

References

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

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Ruibo Tu, Kun Zhang, Hedvig Kjellström, and Cheng Zhang. Optimal transport for causal discovery. *arXiv preprint arXiv:2201.09366*, 2022.