# Some brief notes on inverse optimal transport

Clarice Poon

September 26, 2025

**Abstract**

These notes are intended to supplement my lecture at the 2025 Chemnitz summer school on applied mathematics. They provide a brief overview of inverse optimal transport and provide limited pointers to relevant literature.

## 1 Optimal transport

We begin with a few key equations for optimal transport. Refer to [14] for details on optimal transport.

For simplicity, we assume throughout that $\mathcal{X}, \mathcal{Y}$ are compact spaces.

**Monge's formulation (1781)** Given $\alpha \in \mathcal{P}(\mathcal{X})$ and a measurable map $T : \mathcal{X} \to \mathcal{Y}$, the *push-forward* of $\alpha$ by $T$ is that measure $T_\sharp\alpha$ such that given any measurable set $A$, $(T_\sharp\alpha)(A) = \alpha(T^{-1}(A))$ and given any measurable function $\varphi$, $\int \varphi(y) dT_\sharp\alpha(y) = \int \varphi(T(x)) d\alpha(x)$.

Given $\alpha \in \mathcal{P}(\mathcal{X})$, $\beta \in \mathcal{P}(\mathcal{Y})$, and some cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the goal of Monge is to find $T : \mathcal{X} \to \mathcal{Y}$ with $T_\sharp\alpha = \beta$ such that it minimizes the following problem:

$$\mathrm{OT}(\alpha, \beta) \stackrel{\text{def.}}{=} \inf \left\{ \int c(x, T(x)) d\alpha(x) \setminus T_\sharp\alpha = \beta \right\}.$$

This is highly nonlinear in $T$ and was difficult to analyse directly.

**Kantorovich formulation (1942)** Kantorovich introduced a relaxation of Monge's problem as

$$\mathrm{OT}(\alpha, \beta) = \inf \left\{ \int c(x, y) d\pi(x, y) \setminus \pi \in \mathcal{U}(\alpha, \beta) \right\}$$

where $\mathcal{U}(\alpha, \beta) \stackrel{\text{def.}}{=} \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \setminus \pi_1 = \alpha,\ \pi_2 = \beta\}$ is the set of probability couplings of $\alpha$,$\beta$. We write $\pi_1 = (P_\mathcal{X})_\sharp\pi$ and $\pi_2 = (P_\mathcal{Y})_\sharp\pi$ as the first marginal and second marginal of $\pi$.

*Remark* 1 (Kantorovich formulation is a relaxation). If $T$ is such that $T_\sharp\alpha = \beta$, then $(\mathrm{Id}, T)_\sharp\alpha \in \mathcal{U}(\alpha, \beta)$. Indeed,

$$\int \varphi(x, y)(\mathrm{Id}, T)_\sharp\alpha = \int \varphi(x, T(x)) d\alpha(x).$$

*Remark* 2 (Kantorovich allows the splitting of mass). The constraint of the Monge formulation can be empty. For example, if $\alpha = \delta_0$ and $\beta = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, then there does not exist $T$ such that $T_\sharp\alpha = \beta$. On the other hand, $\mathcal{U}(\alpha, \beta)$ is always nonempty, since it contains the product measure $\alpha \otimes \beta$.

The Kantorovich problem is a *convex* optimization problem, and if you think in finite dimensions, this is a linear programme. The **convex dual** is

$$\sup_{f,g} \{\langle f,\, \alpha\rangle + \langle g,\, \beta\rangle \setminus c \geqslant f \oplus g\}\,. \tag{1}$$

where the supremum is over $f \in \mathcal{C}(\mathcal{X})$ and $g \in \mathcal{C}(\mathcal{Y})$. We write $f \oplus g$ for the function $(x, y) \mapsto f(x) + g(y)$.

**Entropic optimal transport** The entropy regularized OT problem is: for $\varepsilon > 0$,

$$\mathrm{eOT}(\alpha, \beta) = \inf \left\{ \int c(x, y) d\pi(x, y) + \varepsilon \mathrm{KL}(\pi | \alpha \otimes \beta) \setminus \pi \in \mathcal{U}(\alpha, \beta) \right\}.$$

For $\varepsilon = 0$, this is the standard OT problem. As $\varepsilon \to \infty$, the minimizing coupling becomes $\pi = \alpha \otimes \beta$.

This is again a convex problem with dual formulation

$$\mathrm{eOT}(\alpha, \beta) = \sup_{f,g} \langle f,\, \alpha\rangle + \langle g,\, \beta\rangle - \varepsilon \int \exp\left(\frac{f(x) + g(y) - c(x, y)}{\varepsilon}\right) d\alpha(x) d\beta(y). \tag{2}$$

The last decades have seen a sharp growth in applications of eOT for machine learning, some of the attractions of using entropy regularization is the availability of fast algorithms such as Sinkhorn [5] and also the favorable statistical properties of eOT [11].

In the following, we will consider the inverse problem that arises from entropic optimal transport. The use of entropy is often considered a natural modelling assumption, describing the uncertainty in the coupling $\pi$ between $\alpha$ and $\beta$ [10]. In fact, this formulation was naturally derived in many economic application such as the so-called gravity model [7], independently from the development of OT.

## 2 Inverse optimal transport (iOT)

**Forward problem** Let $\varepsilon > 0$ be a fixed entropic regularization parameter. Given two probability measures $\alpha \in \mathcal{P}(\mathcal{X})$ and $\beta \in \mathcal{P}(\mathcal{Y})$, a cost function $c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$, find $\pi(c) \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ that solves

$$\pi(c) = \operatorname*{argmin}_{\pi \in \mathcal{U}(\alpha, \beta)} \int c(x, y) d\pi(x, y) + \varepsilon \mathrm{KL}(\pi | \alpha \otimes \beta).$$

**Inverse problem** The *inverse problem* is to recover the cost function $c$ given $n$ samples $(x_i, y_i) \overset{iid}{\sim} \pi(c)$. We can think of the data as the empirical measure $\hat{\pi}^n = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i, y_i)}$. Note that these samples also give access to the empirical marginals $\hat{\alpha}^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\beta}^n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$.

The problem of inverse optimal transport was initially proposed for studying matching problems in economics (e.g. understanding labour or marriage markets) [8, 9, 6, 4], but has since garnered attention in the machine learning community [16, 12].

## 3 A loss function for iOT

There are two natural and equivalent ways to derive a loss function between a cost $c$ and $\hat{\pi}^n$.

### 3.1 Maximum likelihood estimation

The first approach for iOT is via MLE [6]. Given cost function $c$, the optimal coupling has density

$$\frac{d\pi(c)}{d\alpha \otimes \beta} = \exp\left(\frac{f_c(x) + g_c(y) - c(x,y)}{\varepsilon}\right)$$

where $f_c, g_c$ are the Kantorovich potentials. One can therefore parameterize the cost $c_\theta$ by some $\theta \in \mathbb{R}^p$, and perform MLE

$$\theta = \operatorname*{argmax}_\theta \mathbb{E}_{\hat{\pi}}[\log(d\pi(c_\theta)/d\alpha \otimes \beta)]$$

One can think of this as a bilevel problem, since

$$\begin{aligned}\mathbb{E}_{\hat{\pi}}[\log(\pi(c)/\alpha \otimes \beta)] &= \mathbb{E}_{\hat{\pi}}[-\log(d\hat{\pi}/d\pi(c)) + \log(d\hat{\pi}/d\alpha \otimes \beta)] \\ &= -\mathrm{KL}(\hat{\pi}|\pi(c)) + \mathbb{E}_{\hat{\pi}}[\log(d\hat{\pi}/d\alpha \otimes \beta)],\end{aligned}$$

so, computing the MLE is equivalent to solving

$$\min_\theta \mathrm{KL}(\hat{\pi}|\pi_\theta) \quad \text{where} \quad \pi_\theta \in \operatorname{argmin}_{\pi \in \mathcal{U}(\alpha,\beta)} \langle c_\theta, \pi\rangle + \varepsilon \mathrm{KL}(\pi|\alpha \otimes \beta)$$

### 3.2 Fenchel-Young loss and inverse optimization

The second (equivalent) approach is to view iOT as an *inverse optimization* problem. Consider the following inverse optimization problem: Recover the parameter $c$ from noisy/sampled observations of an optimization solution $\pi(c) = \operatorname{argmin}_c \langle c, \pi\rangle + \Omega(\pi)$. Given observation $\hat{\pi}$, the Fenchel-Young loss [3] is

$$\mathcal{L}(c; \hat{\pi}, \Omega) := \langle c, \hat{\pi}\rangle + \Omega(\hat{\pi}) - \inf_\pi \left\{\langle c, \pi\rangle + \Omega(\pi)\right\}. \tag{3}$$

As a function of $c$, this loss satisfies the following three properties:

1. For all $c$, $\mathcal{L}(c; \hat{\pi}, \Omega) \geqslant 0$ and $L(c, \hat{\pi}) = 0$ if $\hat{\pi} = \pi(c)$;
2. It is differentiable with respect to $c$ if the inner problem over $\pi$ has a unique solution;
3. It is convex in $c$ since the infimum over affine functions is concave.

*Remark* 3. Note that $\Omega^*(c) \stackrel{\text{def.}}{=} \sup_\pi \left\{\langle c, \pi\rangle + \Omega(\pi)\right\}$ is the convex conjugate of $\Omega$. So,

$$\mathcal{L}(c; \hat{\pi}, \Omega) = \langle c, \hat{\pi}\rangle + \Omega(\hat{\pi}) + \Omega^*(-c).$$

The fact that this is non-negative and zero when $-c \in \partial\Omega(\pi)$ is due to the *Fenchel-Young inequality*.

In practice, we parameterize $c$ in a linear manner[1] $c_\theta = \theta^\top \varphi := \sum_{j=1}^S \theta_j \varphi_j$ for some basis $\{\varphi_j\}_j$, $\hat{\pi}$ corresponds to an empirical measure (from sampled data), and $\Omega$ is only given approximately as $\hat{\Omega}$ since it often incorporates empirical data. In the following, since we are interested in minimizing over $c$, we drop the $\Omega(\hat{\pi})$ term when writing the loss.

[1] other nonlinear parameterizations are possible, but linear or affine parametrizations will preserve convexity

**F-Y loss for iOT** Let us now specialize to the case of iOT [1]: In the context of iOT, $\Omega(\pi) := \mathrm{KL}(\pi|\alpha \otimes \beta) + \iota_{\mathcal{U}(\alpha,\beta)}(\pi)$ where $\iota_{\mathcal{U}(\alpha,\beta)}$ denotes the indicator function on $\mathcal{U}(\alpha, \beta)$. The sampled loss, given data $\hat{\pi}^n$ with marginals $\hat{\alpha}^n$ and $\hat{\beta}^n$, is

$$F_n(\theta) = \langle c_\theta, \hat{\pi}^n \rangle - \inf_{\pi \in \mathcal{U}(\hat{\alpha}^n, \hat{\beta}^n)} \left\{ \langle c_\theta, \pi \rangle + \mathrm{KL}(\pi|\hat{\alpha}^n \otimes \hat{\beta}^n) \right\}.$$

By convex duality on the inner problem, one can write

$$F_n(\theta) = \inf_{f,g} \langle c_\theta - (f \oplus g), \hat{\pi}^n \rangle + \varepsilon \int \exp\left( \frac{f(x) + g(y) - c_\theta(x,y)}{\varepsilon} \right) d\hat{\alpha}^n(x) d\hat{\beta}^n(y). \tag{4}$$

Finally, due to the noisy data, we consider the regularized problem

$$\min_{\theta \in \mathbb{R}^S} \lambda R(\theta) + F_n(\theta), \tag{5}$$

for some (convex lower semi-continuous) regularizer $R$ with parameter $\lambda > 0$, which is often taken as the $\ell_1$ norm (to enforce sparsity) or nuclear norm (to enforce low-rankness).

**MLE and FY losses are equivalent** In general, the FY viewpoint is perhaps more versatile as it allows to develop losses for problems where there is no clear MLE formulation. However, in the case of iOT, the two approaches are *equivalent*:

$$\begin{aligned} -\mathbb{E}_{\hat{\pi}}[\log(d\pi(c_\theta)/d\alpha \otimes \beta)] &= \langle c_\theta - f_\theta - g_\theta,\, \hat{\pi} \rangle \\ &= \langle c_\theta,\, \hat{\pi} \rangle - \langle f_\theta,\, \alpha \rangle - \langle g_\theta,\, \beta \rangle. \end{aligned}$$

At optimality, $\int \exp((f_\theta \oplus g_\theta - c_\theta)/\varepsilon) d\hat{\alpha} \otimes \hat{\beta} = \pi_{c_\theta}(\mathcal{X} \times \mathcal{Y}) = 1$. So,

$$\langle f_\theta,\, \hat{\alpha} \rangle + \langle g_\theta,\, \hat{\beta} \rangle - \varepsilon = \sup_{f,g} \langle f,\, \alpha \rangle + \langle g,\, \beta \rangle - \varepsilon \int \exp\left( \frac{f(x) + g(y) - c_\theta(x,y)}{\varepsilon} \right) d\hat{\alpha}(x) d\hat{\beta}(y).$$

We can therefore replace $\langle f_\theta,\, \alpha \rangle + \langle g_\theta,\, \beta \rangle$ with the $\sup_{f,g}$ problem, and this gives precisely the FY loss.

# 4 Solvers

**Proximal gradient descent** The F-Y loss $F_n$ is differentiable in $\theta$, the most direct approach is to apply proximal gradient descent [6]. The gradient of the loss is

$$\nabla F_n(\theta) = \left( \int \varphi_j(x,y) d(\hat{\pi} - \pi_\theta)(x,y) \right)_j$$

where $\pi_\theta$ is the solution to the inner problem. So, one can apply pGD with stepsize $\tau_k$, $k = 1, 2, \ldots,$

$$\theta_{k+1} = \mathrm{Prox}_{\tau_k \lambda R}\left(\theta_k - \tau_k \nabla F_n(\theta)\right)$$

where $\mathrm{Prox}_{\tau R}(\theta') = \mathrm{argmin}_\theta \, \frac{1}{2}\|\theta - \theta'\|_2^2 + \tau R(\theta)$. This requires solving the inner problem at each iteration (e.g. using Sinkhorn).

**Sinkhorn-ISTA** This algorithm was introduced in [4] where $R(\theta) = \|\theta\|_1$, although their method applies easily to any convex regularizer with an easy-to-compute proximal map. Their idea was to avoid solving fully the inner problem at every iteration by alternating between Sinkhorn iterations (i.e. doing block coordinate descent) and proximal steps.

From the Kantorovich formulation (4), we have $\inf_\theta \lambda R(\theta) + F_n(\theta) = \inf_{\theta,f,g} \lambda R(\theta) + K(\theta, f, g)$ where

$$K(\theta, f, g) \stackrel{\text{def.}}{=} \langle c_\theta - (f \oplus g), \hat{\pi}^n \rangle + \varepsilon \int \exp\left(\frac{f(x) + g(y) - c_\theta(x,y)}{\varepsilon}\right) d\hat{\alpha}^n(x) d\hat{\beta}^n(y) \qquad (6)$$

One therefore computes for $k = 1, 2, \ldots,$

1. $f_{k+1} = \operatorname{argmin}_f K(\theta_k, f, g_k)$.
2. $g_{k+1} = \operatorname{argmin}_g K(\theta_k, f_{k+1}, g)$.
3. $\theta_{k+1} = \operatorname{Prox}_{\tau_k \lambda R}\left(\theta_k - \tau_k \partial_\theta K(\theta, f_{k+1}, g_{k+1})\right)$ .

The two minimization problems in $f$ and $g$ have closed form solutions are correspond exactly to Sinkhorn steps. One can prove that for $\tau_k = \tau$ sufficiently small, there is linear convergence in $\|\theta_{k+1} - \theta^\star\| = \mathcal{O}(\rho^k)$ to a minimizer $\theta^\star$.

**Applying quasi-Newton solvers** In the case where $R$ admits a quadratic variational form, one can reparameterize $\theta$ to obtain a smooth problem. We focus on the case of $R(\theta) = \|\theta\|_1$, although this can also be done for nuclear norm and group l1. The Hadamard product over-parameterization of the $\ell^1$ norm

$$\|\theta\|_1 = \min_{u \odot v} \frac{\|u\|_2^2}{2} + \frac{\|v\|_2^2}{2}.$$

where the Hadamard product is $u \odot v \stackrel{\text{def.}}{=} (u_i v_i)_i$. We can therefore replace $\theta$ by $u \odot v$ to obtain a smooth optimization problem: $\frac{\lambda}{2}(\|u\|^2 + \|v\|^2) + K(u \odot v, f, g)$.

To obtain a better-conditioned optimization problem, we consider the *semi-dual* problem, which is derived by leveraging the closed-form expression for the optimal $g$, given $f, c$. In particular, with $f, c$ fixed, the optimal $g$ in (6) satisfies for $\hat{\beta}^n$ a.e.

$$e^{g(y)/\varepsilon} = \left(\int \exp\left(\frac{f(x) - c(x,y)}{\varepsilon}\right) d\hat{\alpha}(x)\right)^{-1}.$$

and so,

$$g(y) = -\varepsilon \log\left(\int \exp\left(\frac{f(x) - c(x,y)}{\varepsilon}\right) d\hat{\alpha}(x)\right)$$

Plugging this into (6), one can write $\inf_g K(\theta, f, g)$ as

$$\langle c_\theta, \hat{\pi}^n \rangle - \langle f, \hat{\alpha}^n \rangle + \varepsilon \int \log\left(\int \exp\left(\frac{f(x) - c_\theta(x,y)}{\varepsilon}\right) d\hat{\alpha}^n(x)\right) d\hat{\beta}^n(y) + \varepsilon.$$

So, combining this with the Hadamard product form for $\theta$, $\min_\theta F_n(\theta)$ is equivalent to solving

$$\inf_{u,v,f} \langle c_{uv}, \hat{\pi}^n \rangle - \langle f, \hat{\alpha}^n \rangle + \varepsilon \int \log\left(\int \exp\left(\frac{f(x) - c_{uv}(x,y)}{\varepsilon}\right) d\hat{\alpha}^n(x)\right) d\hat{\beta}^n(y) + \frac{\lambda}{2}\|u\|^2 + \frac{\lambda}{2}\|v\|^2.$$

The minimizer $u, v$ is related to the optimal $\theta$ by $\theta = u \odot v$. This is now a smooth optimization problem, for which we employ a quasi-Newton solver (L-BFGS) [1].

# 5 Extension to inverse gradient flows

*Inverse gradient flow (iJKO).* Suppose one observes samples iid samples of probability distributions $\rho_k$ for $k = 1, 2, \ldots$, where

$$\rho_{k+1} = \text{argmin}_{\rho \in \mathcal{P}(\mathcal{X})} \mathcal{F}(\rho) + \frac{1}{2\tau} W_2^2(\rho, \rho_k)$$

where $W_2^2$ is the (entropy regularized) Wasserstein distance with Euclidean metric. The *inverse problem* is to recover the functional $\mathcal{F} : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$. This is (on a very formal level) the so-called Jordan Kinderlehre Otto discretization of the PDE $\text{div}(\mu_t \nabla \delta F(\mu_t)) + \partial_t \mu_t = 0$ with $\mu_{k\tau} \approx \rho_k$ when $\tau$ is small. We will call this the iJKO problem and for simplicity, consider the case where we have observations of two snapshots $\rho_0$ and $\rho_1$ One particular example of interest is where $\mathcal{F}(\rho) = \int V(x) d\rho(x)$ and in this context, we are tasked with recovering the potential function $V$ from iid samples of $\rho_k$. Such problems are of particular interest for understanding cell population dynamics in single-cell genomics [15].

**Application of the FY loss** In the notation of the inverse optimization problem from before, for iJKO, we have $\Omega(\rho) = W_2^2(\rho, \rho_0)$. Although one could write down the FY loss (3) in this case, one can consider an extension where we insert the data into the inner problem: Given observation $\hat{\pi}$ and a discrepancy $D : \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0, \infty]$ with $D(\rho, \rho) = 0$, the sharpened Fenchel-Young loss is

$$L(c, \hat{\pi}, \Omega, D) := \langle c, \hat{\pi} \rangle + \Omega(\hat{\pi}) - \inf_{\pi} \left\{ \langle c, \pi \rangle + \Omega(\pi) + D(\pi, \hat{\pi}) \right\}.$$

One can contrast this with (3) where $D \equiv 0$. If $D$ is the Bregman distance induced by $\Omega$, then this is the so-called Fitzpatrick function.

For iJKO, if one seeks to recover a functional of the form $\mathcal{F}(\rho) = \int V_\theta(x) d\rho(x)$ where $V_\theta$ is the potential of interest, parameterized by $\theta$, the sampled loss given empirical data $\hat{\rho}_0, \hat{\rho}_1$ is

$$F_n(\theta) = \langle V_\theta, \hat{\rho}_1 \rangle - \inf_{\alpha \in \mathcal{P}(\mathcal{X})} \left\{ \langle V_\theta, \alpha \rangle + W_{2,\varepsilon}^2(\alpha, \hat{\rho}_0) + r\text{KL}(\alpha | \hat{\rho}_1) \right\},$$

where we have taken $D(\alpha, \rho) = r\text{KL}(\alpha | \hat{\rho}_1)$ for some $r > 0$.

**Equivalence to inverse unbalanced optimal transport** In this case, one can write the inner minimization problem as

$$\inf_{\alpha \in \mathcal{P}(\mathcal{X}), \pi \in \mathcal{U}(\alpha, \hat{\rho}_0)} \int \left( V_\theta(x) + \|x - y\|^2 \right) d\pi(x, y) + \varepsilon \text{KL}(\pi | \alpha \otimes \hat{\rho}_0) + r\text{KL}(\alpha | \hat{\rho}_1)$$

Note that since $\alpha \ll \hat{\rho}_1$, we have

$$\int \log \left( \frac{d\pi}{d(\alpha \otimes \hat{\rho}_0)} \right) d\pi = \int \log \left( \frac{d\pi}{d(\hat{\rho}_1 \otimes \hat{\rho}_0)} \right) d\pi - \int \log \left( \frac{d\alpha}{d\hat{\rho}_1} \right) d\alpha.$$

Plugging this in, and using $\pi_1 = \alpha$, we arrive at

$$\inf_{\pi_2 = \hat{\rho}_0} \int \left( V_\theta(x) + \|x - y\|^2 \right) d\pi(x, y) + \varepsilon \text{KL}(\pi | \alpha \otimes \hat{\rho}_0) + (r - \varepsilon) \text{KL}(\pi_1 | \hat{\rho}_1)$$

The inner problem corresponds to an unbalanced OT problem between $\hat{\rho}_1$ and $\hat{\rho}_0$ where we replace the hard constraint of $\pi_1 = \hat{\rho}_1$ with the KL term. The 'cost' that one seeks to find is then $V_\theta(x) + \|x - y\|^2$. One can therefore study this setting under the general framework of inverse unbalanced optimal transport [2].

# 6 Recovery guarantees

## 6.1 Uniqueness

We first mention that if $\hat{\pi} = \pi(c) = \pi(c')$, i.e. $c$ and $c'$ both give rise to the same data, then there exists functions $f$ and $g$ such that for a.e. $x, y$,

$$(c - c')(x, y) = f(x) + g(y).$$

In particular, we have uniqueness only up to functions $f \oplus g$.

*Example* 1. Let $c_\theta(x, y) = x^\top \theta y$. Suppose that $c_\theta(x, y) = x^\top \theta y = f(x) + g(y)$, we will show that this implies that $\theta \equiv 0$:

$$\begin{aligned}(x - \alpha(\mathcal{X}))^\top \theta (y - \beta(\mathcal{Y})) &= c_\theta - \int c_\theta(x, y) d\alpha(x) - \int c_\theta(x, y) d\beta(y) + \int c_\theta(x, y) d\alpha(x) d\beta(y) \\ &= c_\theta(x, y) - \langle f, \alpha \rangle - g(y) - \langle g, \beta \rangle - f(x) + \langle f, \alpha \rangle + \langle g, \beta \rangle \\ &= c_\theta - f \oplus g = 0.\end{aligned}$$

This is true for all $x$ and $y$, so $\theta = 0$.

In general, to cater for this invariance of $f \oplus g$, we make the following assumption

**Assumption 1.** *Given parameterization $c_\theta(x, y) = \theta^\top \varphi(x, y)$, the centred functions $\bar{\varphi}$ are linearly independent:*

$$\bar{\varphi}(x, y) = \varphi - \int \varphi(x, y) d\alpha(x) - \int \varphi(x, y) d\beta(y) + \int \varphi(x, y) d\alpha(x) d\beta(y)$$

*are such that $\mathbb{E}_{\alpha \otimes \beta}[\varphi \varphi^\top] \in \mathbb{R}^{p \times p}$ is invertible.*

By construction, $\int \bar{\varphi}(x, y) d\alpha(x) = 0$ and $\int \bar{\varphi}(x, y) d\beta(y) = 0$ and one can see by emulating the argument with the quadratic parameterization that there is at most one solution to $\min_\theta \mathcal{L}(c_\theta; \hat{\pi})$.

## 6.2 Stability

We consider the case where $R$ is the $\ell_1$ regularizer, so

$$\inf_\theta \lambda \|\theta\|_1 + F_n(\theta) = \inf_{\theta, f, g} \lambda \|\theta\|_1 + K(\theta, f, g)$$

To obtain stability guarantees, it is not sufficient that $\theta \mapsto F_n(\theta)$ is a convex function. We would require some local curvature properties, such as local strong convexity and local Lipchitz smoothness.

**An abstract stability result for inverse optimization** Let

$$P_\Omega(c) \stackrel{\text{def.}}{=} \operatorname{argmin}_\pi \langle c, \pi \rangle + \Omega(\pi).$$

Let $\hat{F}(\theta) = \mathcal{L}_{\hat{\Omega}}(c_\theta, \hat{\pi})$ be the Fenchel-Young loss constructed from noisy $\hat{\Omega} \approx \Omega$ and noisy data $\hat{\pi}$. For

$$\hat{\theta} \in \operatorname{argmin}_\theta \hat{F}(\theta) + \lambda \|\theta\|,$$

we have the following abstract result.

**Theorem 1.** *[2] Suppose that $\pi^\star = P_\Omega(c_{\theta^\star})$. Let $\hat{\pi}$ be an approximation to $\pi^\star$ and $\hat{\Omega}$ be an approximation to $\Omega$. Suppose that*

1. *Measurement stability $\|\langle \varphi, \hat{\pi} - \pi^\star \rangle\| \leqslant \varepsilon$.*

2. *Forward stability:* $\|\langle \varphi,\, P_{\Omega}(c_{\theta^\star}) - P_{\hat{\Omega}}(c_{\theta^\star})\rangle\| \leqslant \varepsilon$.
3. *Local curvature of the loss:* $\hat{F}(\theta) = \mathcal{L}(c_\theta; \hat{\Omega}, \hat{\pi})$ *is locally Lipschitz smooth and locally strongly convex.*

*Then, for all* $\varepsilon$ *and* $\lambda$ *sufficiently small, there exists a unique minimizer to* (3) *with*

$$\|\theta - \theta^\star\| = \mathcal{O}(\varepsilon + \lambda).$$

By checking these three conditions for iOT, one can derive the following stability result.

**Theorem 2.** *[2, 1] Fix the entropy regularization parameter* $\varepsilon > 0$. *Let* $\pi^\star$ *be the entropic OT plan associated with cost* $c^\star = (\theta^\star)^\top \varphi$, *and let* $\alpha^\star, \beta^\star$ *be its marginals. Assume that*

- $\alpha^\star, \beta^\star$ *are compactly supported,*
- *the cost parameterization* $\varphi$ *is such that its centered version is nondegenerate: define* $\bar{\varphi}(x,y) = \varphi(x,y) - \int \varphi(x,y) d\alpha^\star(x) - \int \varphi(x,y) d\beta^\star(y) + \int \varphi(x,y) d\alpha^\star(x) d\beta^\star(y)$, *and assume that*
$$\left(\mathbb{E}_{\alpha^\star \otimes \beta^\star}[\bar{\varphi}_i(x,y)\bar{\varphi}_j(x,y)]\right)_{i,j}$$
*is invertible.*

*Then, the iOT loss* $F$ *defined with the full data* $\pi^\star$ *is locally strongly convex, locally Lipschitz smooth and is twice differentiable. Moreover, for all* $t > 0$, *with probability at least* $1 - e^{-t}$, *the minimizer* $\hat{\theta}_n^\lambda$ *to the sampled problem* (5) *is unique and satisfies*

$$\|\hat{\theta}_n^\lambda - \theta^\star\|_2 = \mathcal{O}\left(\sqrt{\frac{m_\alpha m_\beta (\log(n) + t)}{n}}\right) + \mathcal{O}(\lambda). \tag{7}$$

Let us comment on the three conditions of Theorem 1 the first condition on measurement stability is generally straightforward. In the iOT setting, $\hat{\pi} = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$ where $(x_i, y_i)$ are iid samples drawn from $\pi^\star$. So, using concentration inequalities such as Hoeffding's inequality, one has $\|\langle \varphi,\, \hat{\pi} - \pi^\star\rangle\| = \mathcal{O}(1/\sqrt{n})$ with high probability.

**Forward stability** The second on forward stability is non-trivial to prove, but is well-studied in the context of iOT. It is essentially asking that the transport plan computed from empirical data $\hat{\alpha}^n, \hat{\beta}^n$

$$\tilde{\pi} \overset{\text{def.}}{=} \operatorname{argmin}_{\pi \in \mathcal{U}(\hat{\alpha}^n, \hat{\beta}^n)} \langle c_{\theta^\star},\, \pi\rangle + \varepsilon \mathrm{KL}(\pi | \hat{\alpha}^n \otimes \hat{\beta}^n)$$

is close to the one computed from true data $\pi^\star$. In fact, one has

$$\|\langle \varphi,\, \tilde{\pi} - \pi^\star\rangle\| = \mathcal{O}(1/\sqrt{n})$$

with high probability [13]. The only slight caveat here is that in the iOT setting, $\hat{\alpha}^n, \hat{\beta}^n$ are the marginals of $\hat{\pi}^n$ are therefore not independent as in the setting of [11, 13], so a slightly more careful analysis needs to be carried out [2].

**Local curvature** The main work is checking the local curvature properties of $F$. To do so, one shows that the loss function constructed full data $F = F_\infty$ has local curvature properties and these are retained with high probability when considering the empirical loss. Recall that $F(\theta) = \inf_{f,g} K(\theta, f, g)$, where

$$K(\theta, f, g) \overset{\text{def.}}{=} \langle c_\theta - (f \oplus g), \pi\rangle + \varepsilon \int \exp\left(\frac{f(x) + g(y) - c_\theta(x,y)}{\varepsilon}\right) d\alpha(x) d\beta(y).$$

At the heart of the local curvature results on $F$, is that fact that the exponential function is locally lipschitz smooth and locally strongly convex.

Below, I list a few results to give some idea of the main mechanisms behind the proof for stability of the recovered cost: the first lemma shows that the functiion $K$ is coercive. As a result, given the cost $c$, we can restrict the optimization of $K$ over $f$ and $g$ to bounded sets. The third lemma shows that $K$ is locally strongly convex thanks to the exponential term.

**Lemma 1** (Coercivity)**.**

$$K(f,g,c) \geqslant \int |f \oplus g - c| d\hat{\pi} + \text{constant}.$$

**Lemma 2** (Restriction to bounded sets)**.** *Fix $c$ and assume that $c$ is Lipschitz. Then, there exists $M > 0$ such that*

$$\inf_{f,g} K(f,g,c) = \inf_{(f,g)\in\mathcal{S}_M} K(f,g,c)$$

*where* $\mathcal{S}_M \overset{\text{def.}}{=} \big\{(f,g) \setminus \|f\| \leqslant M, \|g\|_\infty \leqslant M, \int f d\alpha = 0\big\}$.

The following lemma shows $K$ is locally strongly convex.

**Lemma 3** (local strong convexity)**.** *If $\|f \oplus g - c\|_\infty, \|f' \oplus g' - c'\|_\infty \leqslant M$,*

$$K(f,g,c) \geqslant K(f',g',c') + \langle \nabla K(f',g',c'),\, (f,g,c) - (f',g',c')\rangle + C\|(f-f')\oplus(g-g')-(c-c')\|^2_{L^2(\alpha\otimes\beta)}$$

*If $\int f(x) d\alpha(x) = 0$ and $\int c(x,y) d\alpha(x) = \int c(x,y) d\beta(y) = 0$, then*

$$\|(f-f')\oplus(g-g')-(c-c')\|^2_{L^2(\alpha\otimes\beta)} = \|f-f'\|_{L^2(\alpha)} + \|g-g'\|_{L^2(\beta)} + \|c-c'\|_{L^2(\alpha\otimes\beta)}$$

## 6.3 Preservation of sparsity patterns

One often takes $R(\theta) = \|\theta\|_1$ to enforce sparsity of the solution. One can therefore ask:

*Let the ground truth $\theta^\star$ be $s$-sparse. Under $\ell_1$-norm regularization, for appropriately chosen regularization parameter $\lambda$ and sufficiently many samples $n$, is the recovered solution $\hat{\theta}$ also $s$-sparse?*

One can analyse such questions using dual certificates which govern the structural stability of our reconstruction method. In particular, under certain properties of the so-called dual certificate, one can guarantee that the recovered solution $\hat{\theta}^n_\lambda$ has the same support as the underlying ground truth $\theta^\star$.

**The dual certificate** The iOT loss $F$ with full data $\pi^\star$ can be shown to be twice differentiable. Let $H := \nabla^2 F(\theta^\star)$ and define the certificate

$$z^\star := \operatorname{argmin} \left\{ \langle z, (H^\star)^{-1} z\rangle :\ z \in \partial\|\theta^\star\|_1 \right\}.$$

It is said to be *nondegenerate* if $z^\star$ is in the relative interior of $\partial\|\theta^\star\|_1$. That is,

$$\forall i \in \operatorname{Supp}(\theta^\star), \quad (z^\star)_i = \operatorname{sign}(\theta^\star_i) \quad \text{and} \quad \forall i \notin \operatorname{Supp}(\theta^\star), \quad |(z^\star)_i| < 1.$$

*Remark* 4. In general, the optimality condition to $\theta^\lambda \in \operatorname{argmin}_\theta F(\theta) + \lambda\|\theta\|_1$ reads as $z^\lambda = -\frac{1}{\lambda}\nabla F(\theta^\lambda) \in \partial\|\theta^\lambda\|_1$. The points where $|z^\lambda_i| = 1$ contain the support of $\theta^\lambda$. One can show that $z^\star$ is the limit of $z^\lambda$, so it gives one certificate to analyse the behaviour of $\theta^\lambda$ when $\lambda$ is small.

We have the following result:

**Theorem 3.** *Consider the setting of Theorem 2. Suppose that the certificate $z^*$ is nondegenerate. Let $\hat{\theta}$ minimize* (5) *with $R(\theta) = \lambda\|\theta\|_1$. Let $\delta > 0$. Then, for all sufficiently small regularization parameters $\lambda$ and sufficiently many number of samples $n$ with $\lambda \geqslant C\sqrt{n^{-1}\log(n/\delta)}$, with probability at least $1-\delta$, the minimizer $\hat{\theta}$ has the same support as $\theta^\star$.*

### 6.3.1 The Gaussian setting

For simple settings (such as sampling from Gaussians), the non-degeneracy condition can be checked numerically and we carry out such a numerical investigation in [1]. Similar investigations are carried out for the iJKO setting in [2].

In the case where $\alpha = \mathcal{N}(m_\alpha, \Sigma_\alpha)$ and $\beta = \mathcal{N}(m_\beta, \Sigma_\beta)$, and the cost function is $c_\theta(x, y) = x^\top \theta y$, the eOT plan is a Gaussian, has a closed form expression $\pi_\varepsilon(\theta) = \mathcal{N}(\binom{m_\alpha}{m_\beta}, \Sigma)$ where

$$\Sigma = \begin{pmatrix} \Sigma_\alpha & \hat{\Sigma} \\ \hat{\Sigma}^\top & \Sigma_\beta \end{pmatrix}.$$

where the cross covariance has a closed form expression, but asymptotically, for small $\varepsilon$, it has the expansion $\Sigma = \Sigma_\alpha^{\frac{1}{2}} U V^\top \Sigma_\beta^{\frac{1}{2}} + \varepsilon A^{\top,\dagger} + \mathcal{O}(\varepsilon^2)$. Since everything is explicit here, one can compute the Hessian

$$\partial^2 F(\theta^\star) = 2\varepsilon \Big[ 4\varepsilon^2 (\Sigma_\beta - \Sigma^T \Sigma_\alpha \Sigma)^{-1} \otimes (\Sigma_\alpha - \Sigma {\Sigma_\beta}^{-1} \Sigma^\top)^{-1} + ({\theta^\star}^\top \otimes \theta^\star) \Big]^{-1}$$

and hence the certificate can be computed in closed form.

In general, it is still difficult to analyse theoretically, but one can consider its limits as $\varepsilon \to \infty$ and $\varepsilon \to 0$. As $\varepsilon \to \infty$, the problem converges $\min_\theta \lambda/\varepsilon \|\theta\|_1 + J(\theta; \pi_\varepsilon(\theta^\star))$ to a Lasso problem

$$\min_\theta \frac{1}{2} \|(\Sigma_\beta^{\frac{1}{2}} \otimes \Sigma_\alpha^{\frac{1}{2}})(\theta - \theta^\star)\|_F^2 + \lambda \|\theta\|_1.$$

The certificate corresponds to a Lasso certificate and is typically non-degenerate.

On the other hand, in the special case where $\Sigma_\alpha = \Sigma_\beta = \mathrm{Id}$, as $\varepsilon \to 0$, the problem $\min_\theta \lambda\varepsilon \|\theta\|_1 + J(\theta; \pi_\varepsilon(\theta^\star))$ converges to a *graphical lasso* problem

$$\min_{\theta \succ 0} \lambda \|\theta\|_1 + \langle \theta, (\theta^\star)^\dagger \rangle + \log\det(\theta).$$

The certificate corresponds to a graphical lasso certificate and is typically degenerate.

# References


[1] F. Andrade, G. Peyré, and C. Poon. Sparsistency for inverse optimal transport. In *The Twelfth International Conference on Learning Representations*.

[2] F. Andrade, G. Peyré, and C. Poon. Learning from samples: Inverse problems over measures via sharpened fenchel-young losses. *arXiv preprint arXiv:2505.07124*, 2025.

[3] M. Blondel, A. F. Martins, and V. Niculae. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

[4] G. Carlier, A. Dupuy, A. Galichon, and Y. Sun. Sista: learning optimal transport costs under sparsity constraints. *Communications on Pure and Applied Mathematics*, 76(9):1659–1677, 2023.

[5] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Adv. in Neural Information Processing Systems*, pages 2292–2300, 2013.

[6] A. Dupuy, A. Galichon, and Y. Sun. Estimating matching affinity matrices under low-rank constraints. *Information and Inference: A Journal of the IMA*, 8(4):677–689, 2019.

[7] S. Erlander and N. F. Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. Vsp, 1990.

[8] A. Galichon. *Optimal transport methods in economics*. Princeton University Press, 2016.

[9] A. Galichon and B. Salanié. Matching with trade-offs: Revealed preferences over competing characteristics. *Preprint hal-00473173*, 2010.

[10] A. Galichon and B. Salanié. Cupid's invisible hand: Social surplus and identification in matching models. *The Review of Economic Studies*, 89(5):2600–2629, 2022.

[11] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR, 2019.

[12] R. Li, X. Ye, H. Zhou, and H. Zha. Learning to match via inverse optimal transport. *Journal of machine learning research*, 20(80):1–37, 2019.

[13] P. Rigollet and A. J. Stromme. On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*, 2022.

[14] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[15] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[16] L. Shi, G. Zhang, H. Zhen, J. Fan, and J. Yan. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. 2023.