Probability & Optimal Transport

Davi Sales Barreira

September 9, 2020

Contents

Abstract

The main goal of these notes is to present an introduction to the transport inequalities, which consist of methods of using Optimal Transport Theory to obtain concentration inequalities in high-dimensional probability. Along the way, as new concepts are presented, some side-lining will be done, with the aim to explore some of these new concepts, before diving back in the proof of the inequalities. The core of these notes are base on the excellent paper "Transport Inequalities. A Survey" by Gozlan and Léonard [2010].

1 Introduction

Given two probability distributions μ, ν, there are many situations where one is interested in defining a way of measuring the distance between them. The Wasserstein distance is a metric that arises from the idea of optimal transport, and which has being gaining attention in Statistics and Machine Learning. One prominent example is the so called Wasserstein Generative Adversarial Network (WGAN), which uses this metric to evaluate how well the model generated distribution approximates the "real" distribution of the data. As will be shown shortly, the Wasserstein metric has several advantages compared to other metrics.

There are several ways of defining distances between two probability measures. Let's assume that ν, μ are defined on $(\Omega, \mathcal{F})$ and that $\nu \ll \lambda$, $\mu \ll \lambda$, for λ representing the Lebesgue measure. Below we present some example of distances:

$$\text{Total Variation :} \quad ||\mu - \nu||_{TV} = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)| = \frac{1}{2}\int_\Omega \left|\frac{d\mu}{d\lambda} - \frac{d\nu}{d\lambda}\right| d\lambda = \sup_{|f|\leq 1}\left|\int f(d\nu - d\mu)\right|$$

$$\text{Hellinger :} \quad \sqrt{\int_\Omega \left(\sqrt{\frac{d\mu}{d\lambda}} - \sqrt{\frac{d\nu}{d\lambda}}\right)^2 d\lambda}$$

$$L_2 : \quad \int_\Omega \left(\frac{d\mu}{d\lambda} - \frac{d\nu}{d\lambda}\right)^2 d\lambda$$

As pointed out by Wasserman [2018] in his lecture notes, although such distances are useful, there are drawbacks:

- If one distribution is discrete and the other is continuous, they cannot be compared. If $X \sim U(0,1)$ and Y is uniform on $\{0, 1/N, 2/N, ..., 1\}$, although this distributions are very similar, their total variation is 1 (which is the maximum value). The Wasserstein distance is $1/N$, which is reasonable.
- These distances ignore the "geometry of the underlying space", while the Wasserstein distance preserves it, as shown in Figure 1.
- When "averaging" different distributions, one might be interested in obtaining a similar distribution, avoiding smoothing. This can be done using the Wasserstein barycenter, as shown in Figure 2.

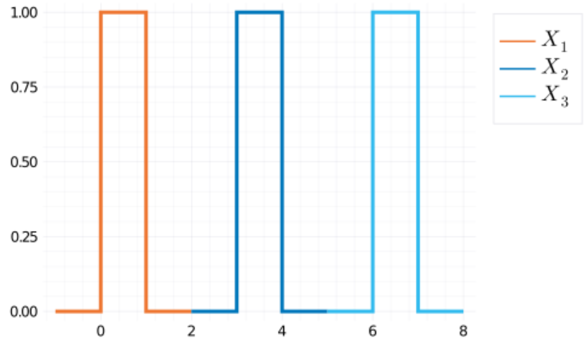

Figure 1: Each pair has the same distance in L_2, Hellinger and TV. But using Wasserstein, X_1 is closer to X_2 than to X_3.

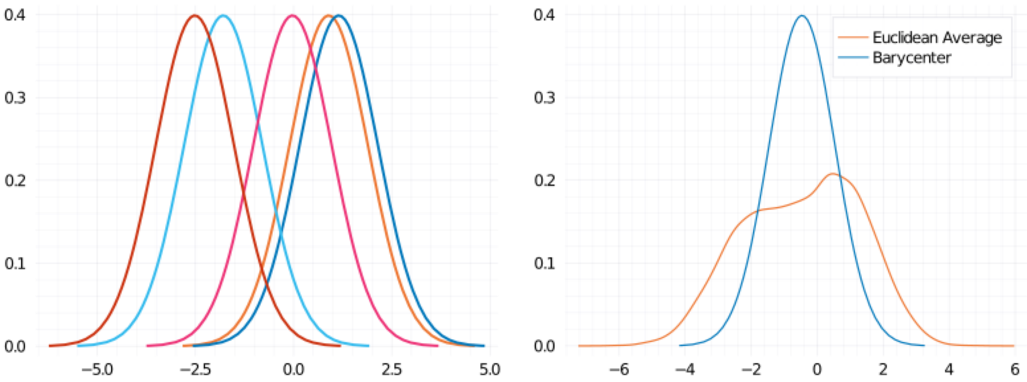

Figure 2: In the left you have different distributions, and in the right, there is a comparison between averaging these distribution versus finding the Wasserstein barycenter.

2 Optimal Transport and the Wasserstein Distance

Let $\mathcal{X}$ be a polish space, and define a function $c : \mathcal{X} \times \mathcal{X} \to [0, \infty)$, where c is a lower semicontinuous function and $\mu, \nu \in P()$ (this means that μ and ν are probability measures on X). Function c is usually called the *cost function*. The Wasserstein distance arises from the Monge-Kantorovich optimal transport problem, which seeks to find a map $\pi : \mathcal{X} \times \mathcal{X} \to \mathcal{X} \times \mathcal{X}$, that transport μ to ν with minimum cost.

Definition 2.1. Coupling A probability measure $\pi \in P(\mathcal{X} \times \mathcal{Y})$ is called a coupling of $\mu \in P(\mathcal{X})$ and $\nu \in P(\mathcal{Y})$ if it's marginal distributions π_1 and π_2 are μ and ν.

$$(MK) \quad \text{Minimize} \pi \in P(\mathcal{X}^2) \mapsto \int_{\mathcal{X}^2} c(x,y) d\pi(x,y) \text{ subject to } \pi \text{coupling of } (\nu, \mu)$$

From solving the above problem, one obtains the optimal transport cost, given by:

$$T_c(\nu, \mu) := \inf \left\{ \int_{\mathcal{X}^2} c(x,y) d\pi(x,y) \ ; \pi_1 = \nu, \pi_2 = \mu \right\} \tag{1}$$

Now, given a metric space $(\mathcal{X}, d)$, we say that $f : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz if

$$\mid f(x) - f(y) \mid \leq L d(x,y) \quad \forall x, y \in \mathcal{X} \tag{2}$$

We will use $||f||_{Lip}$ to denote the smallest L for which this inequality holds. Also, denote $P_1(\mathcal{X})$ the space of probability measures with finite first moment.

Therefore, for $\mu, \nu \in P_1(\mathcal{X})$, the Wasserstein distance is then defined as

$$W_d(\nu, \mu) := \sup_{||f||_{Lip} \leq 1} \left| \int f d\nu - \int f d\mu \right|$$

A classical results in the optimal transport theory, called *Kantorovich-Rubinstein duality* (which will not be proved here), guarantees the equivalence between the Wasserstein distance and the optimal transport cost when the underlying metric space is separable. Hence, by setting $c = d^p$ where , where d is a metric on $\mathcal{X}$ and $p \geq 1$,

$$W_{d^p}(\nu, \mu) := T_{d^p}(\nu, \mu)^{\frac{1}{p}} \tag{3}$$

When the context allows, we will write W_p instead of W_{d^p}. Note that depending on the reference, the Wasserstein distance might be first defined as a transport cost. A simple proof for this duality in the discrete setting is present in the lecture notes by van Handel [2014].

Note that W_p are proper metrics (positive, symmetric and satisfy the triangle inequality) in the space of probability measures. Also, the spaces with p-th moments finite, the metric space $(P_p(\mathcal{X}), W_p)$ is complete and separable if so is $\mathcal{X}$ [Panaretos and Zemel, 2019]. Hence, the Wasserstein distance has the interesting features already presented, such as incorporating the geometry of the "ground space", and allowing the comparison between discrete and continuous distributions. Another interesting aspect, is that convergence of r.v X_n to X in W_p distance is equivalent to convergence in distribution with $E|X_n|^p \to E|X|^p$ [Panaretos and Zemel, 2019].

3 Transport Inequalities and Concentration

In the literature, there are a range of different methods for obtaining inequalities concerning the phenomenon of concentration of measure. The use of transport inequalities is a more recent endeavor and usually not present in more basic texts concerning high-dimensional probability, such as the book by Vershynin [2018]. The relation of optimal transport to concentration inequalities was first pointed out by Marton [1986], and advanced through the nineties, with advances in optimal transport theory. Take a look at Gozlan and Léonard [2010] for a more complete overview on the topic.

Definition 3.1 (Transport Inequalities)**.** Let $J(\cdot \mid \mu) : P(\mathcal{X})$ and $\alpha : [0, \infty) \to [0, \infty)$ an increasing function with $\alpha(0) = 0$. One says that $\mu \in P(\mathcal{X})$ satisfy the transport inequality $\alpha(T_c) \leq J$ if

$$\alpha(T_c(\nu, \mu)) \leq J(\nu \mid \mu), \quad \forall \nu \in P(\mathcal{X})$$

Definition 3.2 (Relative Entropy)**.**

$$H(\nu \mid \mu) = \begin{cases} \int_{\mathcal{X}} \log(\frac{d\nu}{d\mu}) d\nu = \int_{\mathcal{X}} \frac{d\nu}{d\mu} \log(\frac{d\nu}{d\mu}) d\mu, & \text{if } \nu \ll \mu \\ +\infty, & \text{otherwise} \end{cases}$$

When $J = H$, one also calls these transport-entropy inequalities. Among the family of transport inequalities, two classical ones are the $\mathbf{T}_1$ and $\mathbf{T}_2$. We say that $\mu \in P_p(\mathcal{X}) := \{\nu \in P(\mathcal{X}) : \int d(x_o, \cdot)^p d\nu < \infty\}$ satisfies $\mathbf{T}_p(C)$, for $C > 0$ if

$$W_p^2(\nu, \mu) \leq CH(\nu \mid \mu) \tag{4}$$

Note that $\mathbf{T}_1(C)$ is weaker than $\mathbf{T}_2(C)$, because, if μ satisfies $\mathbf{T}_2(C)$, then

$$\begin{aligned} W_2^2(\nu, \mu) \leq CH(\nu \mid \mu) &\implies C^{-1}(T_{d^2}(\nu, \mu)^{1/2})^2 \leq H \\ &\xRightarrow{Jensen} C^{-1}(T_d)^2(\nu, \mu) \leq C^{-1}T_{d^2}(\nu, \mu) \leq H(\nu \mid \mu) \end{aligned}$$

The following inequality is a classical result in information theory.

Theorem 1 (Pinsker-Csiszár-Kullback inequality)**.**

$$||\nu - \mu||_{TV}^2 \leq \frac{1}{2} H(\nu \mid \mu),$$

for all $\mu, \nu \in P(\mathcal{X})$.

Proof. This proof is taken from Gozlan and Léonard [2010] and is originally attributed to Talagrand. Suppose $H(\nu \mid \mu) < \infty$ (otherwise we are done). Let $f = \frac{d\nu}{d\mu}$ and $u = f - 1$, which implies $\int u d\mu = 0$

$$H(\nu \mid \mu) = \int_{\mathcal{X}} f \log f d\mu = \int_{\mathcal{X}} (1 + u) \log(1 + u) - u \; d\mu$$

For $\phi(u) = (1 + u) \log(1 + u) - u$, one can do some manipulations to obtain:

$$\phi(u) = \int_0^y \frac{(u - x)}{1 + u} dx = u^2 \int_0^1 \frac{1 - s}{1 + su} ds, \quad u > -1$$

Then, substituting in the relative entropy equation

$$H(\nu \mid \mu) = \int_{\mathcal{X}} \phi(u) du = \int_{\mathcal{X}} u^2 \int_0^1 \frac{1 - s}{1 + su} \; ds \; d\mu = \int_{\mathcal{X} \times [0,1]} u(x)^2 \frac{1 - s}{1 + su(x)} \; ds \; d\mu(x)$$

We can rewrite some equations and apply Cauchy-Schwarz inequality, to obtain the following

$$\begin{aligned} \left(\int_{\mathcal{X} \times [0,1]} |u(x)|(1 - s) d\mu(x) \; ds \right)^2 \leq &\int_{\mathcal{X} \times [0,1]} u(x)^2 \frac{1 - s}{1 + su(x)} \mu(x) \; ds \; \cdot \\ &\int_{\mathcal{X} \times [0,1]} (1 - s)(1 + su(x)) d\mu(x) \; ds \end{aligned}$$

Noting that:

$$\begin{aligned} \int_{\mathcal{X} \times [0,1]} |u(x)|(1 - s) d\mu(x) \; ds &= \frac{1}{2} \int_{\mathcal{X}} |1 - f| d\mu \\ &= \frac{1}{2} \int_{\mathcal{X}} \left| \frac{d\mu}{d\mu} - \frac{d\nu}{d\mu} \right| d\mu = ||\nu - \mu||_{TV} \end{aligned}$$

And that

$$\int_{\mathcal{X} \times [0,1]} u(x)^2 \frac{1 - s}{1 + su(x)} \mu(x) \; ds \; \cdot \int_{\mathcal{X} \times [0,1]} (1 - s)(1 + su(x)) d\mu(x) \; ds = H(\nu \mid \mu) \cdot \frac{1}{2}$$

We conclude that

$$||\nu - \mu||_{TV}^2 \leq \frac{1}{2} H(\nu \mid \mu),$$

□

The connection with the transport cost is made by defining the Hamming metric

Definition 3.3 (Hamming metric)**.**

$$d_H(x,y) = \mathbb{1}_{x \neq y}, \quad x, y \in \mathcal{X}$$

Now, observe that using d_H, then $|f(x) - f(y)| \leq d_H(x,y) \leq 1$, hence

$$W_{d_h}(\nu, \mu) = \sup_{|f| \leq 1} \left| \int f d\nu - \int f d\mu \right| = ||\mu - \nu||_{TV} \tag{5}$$

Using the Kantorovich-Rubinstein duality, one would then obtain that $T_{d_H}(\mu,\nu) = ||\mu - \nu||_{TV}$. The problem is that $(\mathcal{X}, d_H)$ is not separable, unless $\mathcal{X}$ is discrete. Fortunately, one can still prove that indeed the duality is still valid for this metric. The proof for this assertion is present in Example 4.14 at van Handel [2014]. Briefly, the proof consists of explicitly constructing the optimal coupling, and showing that $||\mu - \nu||_{TV}$ coincides with $T_{d_H}(\nu, \mu)$. Therefore, a probability measure μ defined in the metric space $(\mathcal{X}, d_H)$ satisfies the $\mathbf{T}_1(1/2)$ inequality.

Before showing how this entails a concentration inequality, one needs a little more results.

Definition 3.4 (Concentration of measure)**.** For a metric space $(\mathcal{X}, d)$ and $r \geq 0$. The r-neighborhood of $A \subset \mathcal{X}$ is

$$A^r := \{x \in \mathcal{X} : \ d(x, A) \leq r\}, \quad d(x, A) := \inf_{y \in A} d(x,y)$$

Hence, for $\beta : [0, \infty) \to \mathbb{R}_+$ such that $\beta(r) \to 0$ when $r \to \infty$. One says that the probability measure μ satisfies the concentration inequality with profile β if,

$$\mu(A^r) \geq 1 - \beta(r)$$

for all measurable $A \in \mathcal{X}$ and $\mu(A) \geq 1/2$.

The relation of measure concentration with deviations of Lipschitz functions from their medians is established in the following proposition.

Proposition 1. Let $(\mathcal{X}, d)$ be a metric space with $\mu \in P(\mathcal{X})$ and $\beta : [0, \infty) \to [0, 1]$. The following statements are equivalent:

(i) $\mu(A^r) \geq 1 - \beta(r), \quad r \geq 0$ for all measurable $A \subset \mathcal{X}$ with $\mu(A) \geq 1/2$;

(ii) For $f : \mathcal{X} \to \mathbb{R}$, with f being 1-Lipschitz,

$$\mu(f > m_f + r) \leq \beta(r), \quad r \geq 0$$

where m_f is the median of f.

Proof. $(i) \implies (ii)$. If $x \in A^r$, then $d(x, A) \leq r$. Therefore, for any $y \in A$, $d(x,y) \leq r$. Also, note that since $A = \{y : f(y) \leq m_f\}$, then $f(y) \leq m_f$, for all $y \in A$.

For all f 1-Lipschitz, we then have that

$$f(x) - f(y) \leq d(x,y) \implies f(x) \leq f(y) + r \leq m_f + r$$

Therefore, $x \in \{f \leq m_f + r\}$. We then obtained that $A^r \subset \{f \leq m_f + r\}$. Since $\mu(A) \geq 1/2$, then $\mu(f \leq m_f + r) \geq \mu(A^r) \geq 1 - \beta(r)$.

$(ii) \implies (i)$ Note that for any $A \subset \mathcal{X}$, $f_A(x) = d(x, A) \leq d(x,y)$, hence, it is 1-Lipschitz. Now, consider only the sets A such that $\mu(A) \geq 1/2$.

$$\mu(f_A \leq 0) = \mu(d(x, A) \leq 0) = \mu(x \in A) = \mu(A) = 1/2$$

Therefore, $m_{f_A} = 0$. Note that $A^r = \{f_A \leq r\} = \{x \in \mathcal{X} : d(x, A) \leq r\}$. Finally,

$$\mu(A^r) \geq 1 - \mu(f_A > r) \geq 1 - \beta(r)$$

□

From the proposition above applied to $\pm f$, we have that

$$\mu(\mid f - m_f \mid > r) \leq 2\beta(r), \quad r \geq 0 \tag{6}$$

We finally tie the transport-entropy inequality $\alpha(T_d) \leq H$ with concentration measures by using the so called "Marton's argument" presented in Marton [1986].

Theorem 2 (Marton's Argument)**.** Let $(\mathcal{X}, d)$ be a metric space with $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ be a bijection, and suppose that $\mu \in P(\mathcal{X})$ satisfies the transport-entropy inequality. Then, for all measurable $A \subset \mathcal{X}$ such that $\mu(A) \geq 1/2$, the following can be asserted

$$\mu(A^r) \geq 1 - e^{-\alpha(r-r_o)}, \quad r \geq r_o := \alpha^{-1}(\log 2)$$

In an equivalent manner, for all 1-Lipschitz $f : \mathcal{X} \to \mathbb{R}$,

$$\mu(f - m_f > r + r_o) \leq e^{-\alpha(r)}, \quad r \geq 0$$

Proof. Take a measurable $A \subset \mathcal{X}$ such that $\mu(A) \geq 1/2$ and set $B = \mathcal{X} \setminus A^r$. Next, define the following uniform probability measures in each of these sets:

$$d\mu_A(x) = \frac{d\mu(x)}{\mu(A)} \mathbb{1}_A(x), \quad d\mu_B(x) = \frac{d\mu(x)}{\mu(B)} \mathbb{1}_B(x)$$

For $x \in A$ and $y \in B$, it is easy to see that $d(x, y) \geq r$. Also, if π is a coupling of μ_A, μ_B, then

$$\int_{\mathcal{X}^2} d(x,y) d\pi \geq \int_{\mathcal{X}^2} r d\pi = r \int_{\mathcal{X}^2} d\pi = r$$

Since we used an arbitrary coupling π, we showed that for $\mathcal{C}(\mu_A, \mu_B) := \{\pi \in P(\mathcal{X}^2); \pi_1 = \mu_A, \pi_2 = \mu_B\}$

$$\inf_{\pi \in \mathcal{C}(\mu_A, \mu_B)} \int_{\mathcal{X}^2} d(x,y) d\pi = T_d(\mu_A, \mu_B) \geq r$$

Using the triangle inequality

$$\begin{aligned} r \leq T_d(\mu_A, \mu_B) &\leq T_d(\mu_A, \mu) + T_d(\mu_B, \mu) \\ &\leq \alpha^{-1}(H(\mu_A \mid \mu)) + \alpha^{-1}(H(\mu_B \mid \mu)) \end{aligned}$$

We now calculate $H(\mu_A \mid \mu)$, with $H(\mu_B \mid \mu)$ being analogous.

$$\begin{aligned} H(\mu_A \mid \mu) = \int \frac{d\mu_A}{d\mu} \log(\frac{d\mu_A}{d\mu}) d\mu &= \int \frac{1}{\mu(A)} \mathbb{1}_A \log(\frac{1}{\mu(A)} \mathbb{1}_A) d\mu \\ &= -\frac{\log \mu(A)}{\mu(A)} \int_A d\mu \\ &= -\log \mu(A) \end{aligned}$$

Hence, $H(\mu_A \mid \mu) = -\log \mu(A) \leq \log 2$ and $H(\mu_B \mid \mu) = -\log \mu(B) = -\log(1 - \mu(A^r))$. Then,

$$r \leq \alpha^{-1}(H(\mu_A \mid \mu)) + \alpha^{-1}(H(\mu_B \mid \mu)) \leq \alpha^{-1}(\log 2) + \alpha^{-1}(H(\mu_B \mid \mu))$$

$$r - r_o = r - \alpha^{-1}(\log 2) \leq \alpha^{-1}(-\log(1 - \mu(A^r))$$

Therefore, we conclude that $\mu(A^r) \geq 1 - e^{-\alpha(r-r_o)}$, for all $r \geq r_o = \alpha^{-1}(\log 2)$.

To show the equivalence for 1-Lipschitz functions, just use Proposition 1 and $r' = r + r_o$ in the inequality we just proved. Finally

$$\mu(A^{r'}) \geq 1 - e^{-\alpha(r'-r_o)} = 1 - e^{-\alpha(r)} \therefore \mu(f > m_f + r + r_o) \leq e^{-\alpha(r)}$$

□

Let's use everything we proved so far in an example, so we can better understand how these inequalities relate to concentration.

Example 1. Assume that $X \in [a,b]$ is a random variable and make $L = b-a$. It is clear that using the Hamming metric d_H, that X is L-Lipschitz. Next, define $T_{d_{H_L}}(\mu,\nu) = L \cdot ||\mu-\nu||_{TV}$

Using Pinkser's inequality

$$||\mu-\nu||_{TV}^2 \leq H(\mu \mid \nu)/2 \iff L^2||\mu-\nu||_{TV}^2 \leq L^2 H(\mu \mid \nu)/2 \therefore T_{d_{H_L}}^2(\mu,\nu) \leq \frac{L^2 H(\mu \mid \nu)}{2}$$

Therefore, for $\alpha(x) = \frac{2}{L^2}x^2$, we have $\alpha(T_{d_{H_L}}) \leq H$

We then conclude that

$$\mu(|X-m_X| > r+r_o) \leq 2e^{-2r^2/L^2}$$

Which is very similar to Hoeffding's inequality for bounded random variables.

From this example, it becomes quite clear what we still have to prove. Our results just works for a single random variable. One is usually interested in studying concentration for several random variables, usually independent or with some kind of weak dependence (e.g: Markov).

To extend what we have proved so far to several random variables, we need to obtain estimates no only to μ, but to $\mu^n = \mu \otimes ... \otimes \mu \in \mathcal{X}^n$ (this is just the product measure. Some authors use $\mu_1 \times \mu_2$ instead of $\mu_1 \otimes \mu_2$). This is called the *tensorization of transportation cost.*

Here we will deal with only the product space of two probability measures. But the extension for n cases follows directly from induction.

Let's consider two polish spaces $\mathcal{X}_1, \mathcal{X}_2$, with $\mu_1 \in P(\mathcal{X}_1)$ and $\mu_2 \in P(\mathcal{X}_2)$. Consider $c_1(x_1,y_1)$ and $c_2(x_2,y_2)$ defined on $\mathcal{X}_1 \times \mathcal{X}_1$ and $\mathcal{X}_2 \times \mathcal{X}_2$, respectively. For $\mu_1 \otimes \mu_2$ we then define

$$c_1 \otimes c_2((x_1,y_1),(x_2,y_2)) := c_1(x_1,y_1) + c_2(x_2,y_2) \tag{7}$$

Let ν be a probability measure on $\mathcal{X}_1 \times \mathcal{X}_2$. We write the disintegration of ν as

$$d\nu(x_1,x_2) = d\nu_1(x_1)d\nu_2^{x_1}(x_2) \tag{8}$$

The above definition of disintegration can be thought of a rigorous way of defining the conditional probability measures. For example, if $(X,Y) \sim \nu$, then $X \sim \nu_1$, and $\nu_2^{x_1}$ is $P(Y \in \cdot \mid X = x_1)$.

It is possible to prove that

$$T_{c_1 \otimes c_2}(\nu, \mu_1 \otimes \mu_2) \leq T_{c_1}(\nu_1,\mu_1) + \int_{\mathcal{X}_1} T_{c_2}(\nu_2^{x_1},\mu_2)d\nu_1(x_1) \tag{9}$$

For a proof of the inequality above, look the Appendix 1 of Gozlan and Léonard [2010].

Another result stated without proof is the following:

$$H(\nu \mid \mu_1 \otimes \mu_2) = H(\nu_1 \mid \mu_1) + \int_{\mathcal{X}_1} H(\nu_2^{x_1} \mid \mu_2)d\nu_1(x_1) \tag{10}$$

This is known as the chain rule for relative entropy. A proof can be found on van Handel [2014] Lemma 4.18.

We need one more definition

$$\alpha_1$$

References

Nathael Gozlan and Christian Léonard. Transport inequalities. a survey. *arXiv preprint arXiv:1003.3852*, 2010.

Katalin Marton. A simple proof of the blowing-up lemma (corresp.). *IEEE Transactions on Information Theory*, 32(3):445–446, 1986.

Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.

Ramon van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Larry Wasserman. Statistical methods for machine learning - lecture notes, 2018.