A Suite on the Wasserstein Metric

Ryan Tay

16 September 2024

Contents

Acknowledgements

Funded by the University of Warwick's Undergraduate Research Support Scheme (URSS), this expository piece was written under the supervision of Dr Josephine Evans at the University of Warwick. I thank both her and the University of Warwick for supporting me in this project.

1 Prelude

Suppose we have probability measures μ and ν on a sample space X. How should one quantify how "different" the two probability measures are? If the sample space is $X = \mathbb{R}$ and the measures μ and ν admit probability density functions f and g respectively, so that $\mu(A) = \int_A f(x)\,\mathrm{d}x$ and $\nu(A) = \int_A g(x)\,\mathrm{d}x$, then a quantity which could measure the "distance" between μ and ν is

$$\|f - g\|_{L^1} \coloneqq \int_{\mathbb{R}} |f(x) - g(x)|\,\mathrm{d}x.$$

If $\|f - g\|_{L^1}$ is close to 0 then we would say that the two measures are really close to each other, and if $\|f - g\|_{L^1}$ is close to 2 then we would say that the two measures are really quite different from each other.

There are two problems with this approach. Firstly, some measures may not even admit a probability density function. For instance, the Dirac measure δ_0 on $\mathbb{R}$ given by

$$\delta_0(A) \coloneqq \begin{cases} 1 & \text{if } 0 \in A, \\ 0 & \text{else.} \end{cases}$$

The elephant in the room, though, is that the L^1 distance between f and g may not match up with our intuitive notion of how far measures μ and ν may be from each other. Consider, for instance, the measures μ and ν having densities

$$f(x) \coloneqq \begin{cases} 10^8 & \text{if } 0 \leq x \leq \frac{1}{10^8}, \\ 0 & \text{else,} \end{cases} \qquad \text{and} \qquad g(x) \coloneqq \begin{cases} 10^8 & \text{if } \frac{1}{10^8} \leq x \leq \frac{2}{10^8}, \\ 0 & \text{else,} \end{cases}$$

respectively. Then $\|f - g\|_{L^1} = 2$, which as large as $\|f - g\|_{L^1}$ could possibly be, and so we would say that μ and ν are "really far apart from each other". There is, however, still a sense in which μ and ν are really close to each other: the graph of g is just the graph of f translated to the right by a mere 10^{-8} units. Furthermore, if we consider a third measure ξ with density

$$h(x) \coloneqq \begin{cases} 10^8 & \text{if } 500 \leq x \leq 500 + \frac{1}{10^8}, \\ 0 & \text{else,} \end{cases}$$

then we would also have $\|f - h\|_{L^1} = 2$. The L^1 distance is unable to capture this huge shift of 500 units to the right; it is unable to distinguish μ from ξ any more than it can distinguish μ from ν.

Instead, we turn to a way of quantifying the distance between two probability measures μ and ν based off of this intuitive notion of how much "effort" would it take to "move" from μ to ν. Underpinning this whole theory is the Monge–Kantorovich problem [Gar18, Chapter 20.2] [RR, Chapter 2.1], named after Gaspard Monge and Leonid Kantorovich (Russian: Леонид Канторович):

> **The Monge–Kantorovich problem.**
> Fix Polish spaces X and Y, and a lower semicontinuous cost function $c\colon X \times Y \to \mathbb{R}_{\geq 0}$. Given two Borel probability measures μ and ν on X and Y respectively, the Monge–Kantorovich problem seeks to minimise[1] the following total cost:
>
> $$\int_{X \times Y} c\,\mathrm{d}\pi,$$
>
> with π ranging over the space of all Borel probability measures on $X \times Y$ with marginals μ and ν.

[1] A priori, we seek to find the infimum value of $\int_{X \times Y} c(x, y)\,\mathrm{d}\pi$, rather than the minimum value. It will turn out that the infimum value is actually always achieved (see Proposition 4.1).

An intuitive view of the Monge–Kantorovich problem is as follows. Suppose you had a pile of sand spread about in a space X according to a probability measure μ, and you wished to transport that pile of sand to a space Y and spreading it out according to a probability measure ν. Moving one particle of sand from $x \in X$ to $y \in Y$ costs $\$c(x, y)$. A transport plan π tells you how you should move the pile of sand: a volume of space $B \subseteq Y$ gets $\pi(A \times B)$ of the sand available the volume of space $A \subseteq X$. Of course, there are restrictions for what constitutes a transport plan: after the transportation has finished, the amount of sand in any $B \subseteq Y$ should equal $\pi(X \times B)$. Similarly, the amount of sand leaving any $A \subseteq X$ should equal $\pi(A \times Y)$.

It is these restrictions which are captured by the requirement that π has marginals μ and ν, that is, given the projection maps $p_X \colon X \times Y \to X$ and $p_Y \colon X \times Y \to Y$ defined by

$$p_X(x, y) \coloneqq x \quad \text{and} \quad p_Y(x, y) \coloneqq y,$$

we require that μ is the image measure of π under the map p_X, and ν is the image measure of π under the map p_Y.

A transport plan always exists: the product measure $\pi \coloneqq \mu \times \nu$ works. The Monge–Kantorovich problem, however, seeks to find the optimal transport plan. It is an important result that the Monge–Kantorovich problem *always* has a solution, in the sense that an optimal transport plan always exists (see Proposition 4.1). It is this fact that allows us to develop the Wasserstein metric[2], which we will define in the Allemande section. The subsequent sections of this expository piece explores properties of the Wasserstein metric on the real line. Most of the definitions and results, except for the results in the Courante section, can be generalised to work on $\mathbb{R}^d$ or even on arbitrary Polish spaces. For simplicity, however, we will only be interested in developing the theory on the real line.

Section 2, titled Allemande, defines the Wasserstein metric W_1 on the real line, and demonstrates it with a simple example. Section 3, titled Courante, connects the Wasserstein distance between two probability measures with their cumulative density functions. Section 4, titled Sarabande, fills in all the proofs omitted from the previous three sections. Section 5 and Section 6, titled Bourrée and Gigue respectively, establishes lower and upper bounds on the Wasserstein metric with other objects which also capture some idea of "distance" between measures.

[2] Also known as the Kantorovich–Rubinstein metric or the Earth Mover's distance [Vil09, Chapter 6 Bibliographical Notes].

2 Allemande

From now onwards, for a topological space X, let us use $P(X)$ to denote the space of Borel probability measures on X, though we will mainly be concerned with the spaces $P(\mathbb{R})$ and $P(\mathbb{R}^2)$. For $\mu, \nu \in P(\mathbb{R})$, we define the space

$$\Pi_{\mu,\nu} \coloneqq \{\, \pi \in P(\mathbb{R}^2) : \pi \text{ has marginals } \mu \text{ and } \nu \,\}.$$

We shall use the cost function $c\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ given by $c(x, y) \coloneqq |x - y|$ to define the Wasserstein metric, named after Leonid Vaserstein[3] (Russian: Леонид Васерштейн), on the subspace

$$P_1(\mathbb{R}) \coloneqq \left\{ \mu \in P(\mathbb{R}) : \int_{\mathbb{R}} |x| \, \mathrm{d}\mu(x) < \infty \right\}.$$

Definition 2.1 (The Wasserstein metric W_1 [Gar18, Chapter 21] [Vil09, Definition 6.1, Definition 6.4])
We define the function $W_1\colon P_1(\mathbb{R}) \times P_1(\mathbb{R}) \to \mathbb{R}_{\geq 0}$, called the Wasserstein metric[4], by

$$W_1(\mu, \nu) \coloneqq \inf \left\{ \int_{\mathbb{R}^2} |x - y| \, \mathrm{d}\pi(x, y) : \pi \in \Pi_{\mu,\nu} \right\}.$$

The Wasserstein metric W_1 is indeed a metric on $P_1(\mathbb{R})$ (see Proposition 4.4), so it captures the ideas of "distance" between measures as one would expect, including the triangle inequality.

Calculating $W_1(\mu, \nu)$ for specific probability measures μ and ν can be challenging given only this definition. Often, if there is enough similarity between μ and ν, we can obtain an upper bound for $W_1(\mu, \nu)$ by finding a suitable transport plan from μ to ν. Recall again the measures from the Prelude section:

$$\mu(A) \coloneqq 10^8 \cdot \lambda\left(A \cap \left[0, \frac{1}{10^8}\right]\right) \quad \text{and} \quad \nu(A) \coloneqq 10^8 \cdot \lambda\left(A \cap \left[\frac{1}{10^8}, \frac{2}{10^8}\right]\right).$$

where λ is the Lebesgue measure on $\mathbb{R}$. Consider the transport plan $T\colon \mathbb{R} \to \mathbb{R}$ given by $T(x) \coloneqq x + \frac{1}{10^8}$. Let $G\colon \mathbb{R} \to \mathbb{R}^2$ be defined by $G(x) \coloneqq (x, T(x))$, and define the Borel probability measure π on $\mathbb{R}^2$ to be the image measure of μ under G, that is,

$$\pi(A) \coloneqq \mu(G^{-1}(A)).$$

Then π has marginals μ and ν, and we observe that

$$\int_{\mathbb{R}^2} |x - y| \, \mathrm{d}\pi(x, y) = \frac{1}{10^8},$$

yielding $0 < W_1(\mu, \nu) \leq \frac{1}{10^8}$. Given that the graph of the distribution of ν is just a translation of $\frac{1}{10^8}$ units to the right of the graph of the distribution of μ, would it not be nice if $W_1(\mu, \nu) = \frac{1}{10^8}$? Furthermore, if we recall the other Borel probability measure ξ from the Prelude section, namely

$$\xi(A) \coloneqq 10^8 \cdot \lambda\left(A \cap \left[500,\ 500 + \frac{1}{10^8}\right]\right),$$

[3] Leonid Vaserstein was actually not the first person to come up with the Wasserstein metric. The metric is due to Leonid Kantorovich and Gennadii Rubinstein (Russian: Геннадии Рубинштейн) [Vil09, Chapter 6 Bibliographical Notes] [Vil09, Chapter 3].

[4] The appearance of the subscript "1" in the notations "W_1" and "P_1" is due to the definition of the more general Wasserstein p-metric W_p, defined by

$$W_p(\mu, \nu) \coloneqq \left(\inf \left\{ \int_{\mathbb{R}^2} |x - y|^p \, \mathrm{d}\pi(x, y) : \pi \in \Pi_{\mu,\nu} \right\} \right)^{1/p},$$

where $1 \leq p < \infty$. This metric W_p will be defined on the space $P_p(\mathbb{R})$ consisting of all $\mu \in P(\mathbb{R})$ such that $\int_{\mathbb{R}} |x|^p \, \mathrm{d}\mu(x) < \infty$.

a similar argument to the one above would yield $0 < W_1(\mu, \xi) \leq 500$. Again, it would be wonderful if we indeed had $W_1(\mu, \xi) = 500$.

Theorem 2.2 ([Gar18, Corollary 21.2.3])
If two Borel probability measures μ and ν on $\mathbb{R}$ have respective probability density functions f_μ and f_ν, then

$$W_1(\mu, \nu) \geq \left| \int_{\mathbb{R}} x f_\mu(x) \, \mathrm{d}x - \int_{\mathbb{R}} x f_\nu(x) \, \mathrm{d}x \right|.$$

In other words, the Wasserstein distance between μ and ν is at least the distance between their means.

Proof. This is a corollary of Proposition 4.1 or Proposition 4.3. □

Recall our example Borel probability measures on $\mathbb{R}$:

$$\begin{aligned}
\mu(A) &\coloneqq 10^8 \cdot \lambda\left(A \cap \left[0, \frac{1}{10^8}\right]\right) \\
\nu(A) &\coloneqq 10^8 \cdot \lambda\left(A \cap \left[\frac{1}{10^8}, \frac{2}{10^8}\right]\right), \quad \text{and} \\
\xi(A) &\coloneqq 10^8 \cdot \lambda\left(A \cap \left[500, \ 500 + \frac{1}{10^8}\right]\right).
\end{aligned}$$

Theorem 2.2 together with our earlier discussions yield $W_1(\mu, \nu) = \frac{1}{10^8}$ and $W_1(\mu, \xi) = 500$.

3 Courante

Theorem 3.1 ([San15, Proposition 2.17])
Let $\mu, \nu \in P_1(\mathbb{R})$ have cumulative distribution functions

$$F_\mu(x) \coloneqq \mu((-\infty, x]) \quad \textit{and} \quad F_\nu(x) \coloneqq \nu((-\infty, x])$$

respectively. Then

$$W_1(\mu, \nu) = \int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)| \, \mathrm{d}x.$$

Proof. Let us define the pseudo-inverses $F_\mu^{[-1]}, F_\nu^{[-1]} \colon [0,1] \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{-\infty, \infty\}$, by

$$\begin{aligned} F_\mu^{[-1]}(y) &\coloneqq \inf\{\, x \in \mathbb{R} : F_\mu(x) \geq y \,\}, \quad \text{and} \\ F_\nu^{[-1]}(y) &\coloneqq \inf\{\, x \in \mathbb{R} : F_\nu(x) \geq y \,\}, \end{aligned}$$

where we adopt the convention $\inf \varnothing = \infty$. Denote by $\lambda|_{[0,1]}$ the Lebesgue measure on $[0,1]$. Define $G \colon [0,1] \to \overline{\mathbb{R}}^2$ by $G(y) \coloneqq (F_\mu^{[-1]}(y), F_\nu^{[-1]}(y))$ and define the measure π_{mon} to be the restriction to $\mathbb{R}^2$ of the image measure of $\lambda|_{[0,1]}$ under G. That is,

$$\pi_{\mathrm{mon}}(A) \coloneqq \lambda|_{[0,1]}\big(G^{-1}(A)\big) \quad \text{for all Borel } A \subseteq \mathbb{R}^2.$$

Observe then that $\pi_{\mathrm{mon}} \in \Pi_{\mu,\nu}$, because the image measure of $\lambda|_{[0,1]}$ under $F_\mu^{[-1]}$, when restricted to $\mathbb{R}$, is simply μ itself (and similarly for $F_\nu^{[-1]}$ and ν). We claim that this measure π_{mon} is an optimal solution to the Monge–Kantorovich problem of transporting μ to ν with respect to the cost function $c(x,y) \coloneqq |x - y|$, that is,

$$W_1(\mu, \nu) = \int_{\mathbb{R}^2} |x - y| \, \mathrm{d}\pi_{\mathrm{mon}}(x, y). \tag{3.1}$$

If this is shown, then we would obtain

$$\begin{aligned} W_1(\mu, \nu) = \int_{\mathbb{R}^2} |x - y| \, \mathrm{d}\pi_{\mathrm{mon}}(x, y) &= \int_0^1 |F_\mu^{[-1]}(x) - F_\nu^{[-1]}(x)| \, \mathrm{d}x \\ &= \int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)| \, \mathrm{d}x. \end{aligned}$$

It thus remains to show that we do indeed have $W_1(\mu, \nu) = \int_{\mathbb{R}^2} |x - y| \, \mathrm{d}\pi_{\mathrm{mon}}(x, y)$. We start by approximating the function $t \mapsto |t|$ by strictly convex functions.

Lemma 3.2 ([San15, Lemma 2.10])
For all $\varepsilon > 0$ there exists a continuous strictly convex function $h_\varepsilon \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that for all $t \in \mathbb{R}$,

$$|t| \leq h_\varepsilon(t) \leq (1 + \varepsilon)|t| + \varepsilon.$$

Proof of Lemma 3.2. For any $\varepsilon > 0$, the function

$$h_\varepsilon(t) \coloneqq |t| + \varepsilon \left(\frac{1}{2} \sqrt{4 + t^2} + \frac{1}{2} t \right)$$

works. □

We are interested in these strictly convex functions h_ε due to the following Lemma 3.3, where we will exploit the strict convexity of the functions h_ε.

Lemma 3.3 ([San15, Lemma 2.8, Theorem 2.9])
For $\varepsilon > 0$, define $c_\varepsilon \colon \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ by $c_\varepsilon(x,y) \coloneqq h_\varepsilon(x-y)$. Then π_{mon} is the unique optimal solution to the Monge–Kantorovich problem of transporting μ to ν with respect to the cost function c_ε.

Once Lemma 3.3 has been established, then for any $\pi \in \Pi_{\mu,\nu}$ we have

$$\begin{aligned}\int_{\mathbb{R}^2} |x-y| \,\mathrm{d}\pi_{\mathrm{mon}}(x,y) &\leq \int_{\mathbb{R}^2} h_\varepsilon(x-y) \,\mathrm{d}\pi_{\mathrm{mon}}(x,y) \\ &\leq \int_{\mathbb{R}^2} h_\varepsilon(x-y) \,\mathrm{d}\pi(x,y) \\ &\leq (1+\varepsilon) \int_{\mathbb{R}^2} |x-y| \,\mathrm{d}\pi(x,y) + \varepsilon,\end{aligned}$$

whence taking the limit $\varepsilon \to 0$ yields $\int_{\mathbb{R}^2} |x-y| \,\mathrm{d}\pi_{\mathrm{mon}}(x,y) \leq \int_{\mathbb{R}^2} |x-y| \,\mathrm{d}\pi(x,y)$. Consequently, Equation (3.1) would be established, completing the proof of Theorem 3.1.

The heart of the proof of Theorem 3.1 thus boils down to proving Lemma 3.3. But first, more definitions and lemmas.

Define the support of a measure $\pi \in P(\mathbb{R}^2)$ to be

$$\mathrm{support}(\pi) \coloneqq \{\, (x,y) \in \mathbb{R}^2 \;:\; \pi(B_r((x,y))) > 0 \text{ for all } r > 0 \,\},$$

where $B_r((x,y))$ denotes the open ball of radius $r > 0$ centered around $(x,y) \in \mathbb{R}^2$.

Lemma 3.4 ([San15, Theorem 1.38])
Fix any optimal $\pi_0 \in \Pi_{\mu,\nu}$ solving the Monge–Kantorovich problem of transporting μ to ν with respect to a continuous cost function $\tilde{c} \colon \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$. Then for all $(x_1,y_1), (x_2,y_2) \in \mathrm{support}(\pi_0)$, we have

$$\tilde{c}(x_1,y_1) + \tilde{c}(x_2,y_2) \leq \tilde{c}(x_1,y_2) + \tilde{c}(x_2,y_1).$$

Proof of Lemma 3.4. Suppose, for a contradiction, that

$$\tilde{c}(x_1,y_1) + \tilde{c}(x_2,y_2) > \tilde{c}(x_1,y_2) + \tilde{c}(x_2,y_1).$$

Fix any $\varepsilon > 0$ satisfying

$$0 < \varepsilon < \frac{1}{4}\big(\tilde{c}(x_1,y_1) + \tilde{c}(x_2,y_2) - \tilde{c}(x_1,y_2) - \tilde{c}(x_2,y_1)\big).$$

Since $\tilde{c}$ is continuous, there exists $r > 0$ such that all of the following hold:

- for all $x \in (x_1 - r, x_1 + r)$ and for all $y \in (y_1 - r, y_1 + r)$, we have
$$\tilde{c}(x,y) > \tilde{c}(x_1,y_1) - \varepsilon,$$
- for all $x \in (x_2 - r, x_2 + r)$ and for all $y \in (y_2 - r, y_2 + r)$, we have
$$\tilde{c}(x,y) > \tilde{c}(x_2,y_2) - \varepsilon,$$
- for all $x \in (x_1 - r, x_1 + r)$ and for all $y \in (y_2 - r, y_2 + r)$, we have
$$\tilde{c}(x,y) < \tilde{c}(x_1,y_2) + \varepsilon,$$
- for all $x \in (x_2 - r, x_2 + r)$ and for all $y \in (y_1 - r, y_1 + r)$, we have
$$\tilde{c}(x,y) < \tilde{c}(x_2,y_1) + \varepsilon.$$

Define the open squares $Q_j \coloneqq (x_j - r, x_j + r) \times (y_j - r, y_j + r)$ for $j \in \{1, 2\}$. Since $(x_1, y_1), (x_2, y_2) \in \operatorname{support}(\pi_0)$, we have $\pi_0(Q_1), \pi_0(Q_2) > 0$. Let $\pi_0|_{Q_j}$ denote the restriction of the measure π_0 to the open square V_j. Define the Borel probability measures on Q_j

$$\pi_j \coloneqq \frac{1}{\pi_0(Q_j)} \pi_0|_{Q_j} \quad \text{for each } j \in \{1, 2\}$$

Define the projection maps $p_1, p_2 \colon \mathbb{R}^2 \to \mathbb{R}$ by $p_1(x, y) \coloneqq x$ and $p_2(x, y) \coloneqq y$, and for $j \in \{1, 2\}$:

- define μ_j to be the measure on $\mathbb{R}$ which is the image measure of π_j under the map p_1,
- define ν_j to be the measure on $\mathbb{R}$ which is the image measure of π_j under the map p_2.

Fix any $0 < \varepsilon_0 < \frac{1}{2} \min\{\pi(Q_1), \pi(Q_2)\}$. Define the product measures $\gamma_1 \coloneqq \mu_1 \times \nu_2$ and $\gamma_2 \coloneqq \mu_2 \times \nu_1$. Finally, define a measure γ on $\mathbb{R}^2$ by

$$\gamma \coloneqq \pi_0 - \varepsilon_0(\pi_1 + \pi_2) + \varepsilon_0(\gamma_1 + \gamma_2).$$

It is easy to check that $\gamma \in \Pi_{\mu,\nu}$. We now claim that we have

$$\int_{\mathbb{R}^2} \tilde{c} \, \mathrm{d}\gamma < \int_{\mathbb{R}^2} \tilde{c} \, \mathrm{d}\pi_0,$$

contradicting the assumption that π_0 is optimal. Indeed,

$$\begin{aligned}
\int_{\mathbb{R}^2} \tilde{c} \, \mathrm{d}\pi_0 - \int_{\mathbb{R}^2} \tilde{c} \, \mathrm{d}\gamma &= \varepsilon_0 \left(\int_{\mathbb{R}^2} \tilde{c} \, \mathrm{d}\pi_1 + \int_{\mathbb{R}^2} \tilde{c} \, \mathrm{d}\pi_2 - \int_{\mathbb{R}^2} \tilde{c} \, \mathrm{d}\gamma_1 - \int_{\mathbb{R}^2} \tilde{c} \, \mathrm{d}\gamma_2 \right) \\
&\geq \varepsilon_0 \Big(\big(\tilde{c}(x_1, y_1) - \varepsilon\big) + \big(\tilde{c}(x_2, y_2) - \varepsilon\big) - \big(\tilde{c}(x_1, y_2) + \varepsilon\big) - \big(\tilde{c}(x_2, y_1) - \varepsilon\big) \Big) \\
&> 0.
\end{aligned}$$

□

Lemma 3.5 ([San15, Theorem 2.9])
Let $\tilde{h} \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a continuous strictly convex function. Fix any optimal $\pi_0 \in \Pi_{\mu,\nu}$ solving the Monge–Kantorovich problem of transporting μ to ν with respect to the cost function $\tilde{c} \colon \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ defined by $\tilde{c}(x, y) \coloneqq \tilde{h}(x - y)$. Let $(x_1, y_1), (x_2, y_2) \in \operatorname{support}(\pi_0)$ satisfy $y_1 < y_2$. Then $x_1 \leq x_2$.

Proof of Lemma 3.5. Suppose, for a contradiction, that $x_1 > x_2$. Lemma 3.4 gives us

$$\tilde{h}(x_1 - y_1) + \tilde{h}(x_2 - y_2) \leq \tilde{h}(x_1 - y_2) + \tilde{h}(x_2 - y_1).$$

Then we have, by the strict concavity of $\tilde{h}$, we have

$$\begin{aligned}
\tilde{h}(x_1 - y_1) + \tilde{h}(x_2 - y_2) &\leq \tilde{h}(x_1 - y_2) + \tilde{h}(x_2 - y_1) \\
&= \tilde{h}(t(x_1 - y_1) + (1 - t)(x_2 - y_2)) + \tilde{h}(t(x_2 - y_2) + (1 - t)(x_1 - y_1)) \\
&< t\tilde{h}(x_1 - y_1) + (1 - t)\tilde{h}(x_2 - y_2) + t\tilde{h}(x_2 - y_2) + (1 - t)\tilde{h}(x_1 - y_1) \\
&= \tilde{h}(x_1 - y_1) + \tilde{h}(x_2 - y_2),
\end{aligned}$$

where $t \coloneqq \frac{x_1 - x_2}{(x_1 - x_2) + (y_2 - y_1)} \in (0, 1)$. This strict inequality is a contradiction. □

We are now ready to complete the proof of Lemma 3.3.

Proof of Lemma 3.3. Fix any $\varepsilon > 0$. Certainly, by Proposition 4.1, an optimal solution $\pi_0 \in \Pi_{\mu,\nu}$ exists for the Monge–Kantorovich problem of transporting μ to ν with respect to the cost function $\tilde{c}_\varepsilon$. We aim to show that $\pi_0 = \pi_{\mathrm{mon}}$.

By definition of π_{mon} as the image measure of $\lambda|_{[0,1]}$ under the map $y \mapsto (F_\mu^{[-1]}(y), F_\nu^{[-1]}(y))$, this measure π_{mon} is the unique Borel probability measure on $\mathbb{R}^2$ satisfying

$$\pi_{\text{mon}}((-\infty, a] \times (-\infty, b]) = \min\{F_\mu^{[-1]}(a), F_\nu^{[-1]}(b)\} \qquad \text{for all } a, b \in \mathbb{R}.$$

We will show that π_0 also satisfies the equality above, proving that $\pi_0 = \pi_{\text{mon}}$. Fix any $a, b \in \mathbb{R}$. Then at least one of the sets

$$A \coloneqq (-\infty, a] \times (b, \infty) \quad \text{or} \quad B \coloneqq (a, \infty) \times (-\infty, b]$$

must have measure 0 under π_0. If this were not the case, then there would exist points

$$(x_1, y_1) \in (-\infty, a] \times (b, \infty) \quad \text{and} \quad (x_2, y_2) \in (a, \infty) \times (-\infty, b]$$

violating Lemma 3.5. So we have

$$\begin{aligned}
\pi_0((-\infty, a] \times (-\infty, b]) &= \min\{\pi_0((-\infty, a] \times (-\infty, b] \cup A), \pi_0((-\infty, a] \times (-\infty, b] \cup B)\} \\
&= \min\{\pi_0((-\infty, a] \times \mathbb{R}), \pi_0(\mathbb{R} \times (-\infty, b])\} \\
&= \min\{F_\mu^{[-1]}(a), F_\nu^{[-1]}(b)\}.
\end{aligned}$$

Thus $\pi_0 = \pi_{\text{mon}}$. □

This completes the proof of Lemma 3.3. The proof of Theorem 3.1 is now, at long last, complete. □

4 Sarabande

We have used quite a few results without justifications. This section is dedicated to filling in those gaps.

Recall that for a topological space X, we use $P(X)$ to denote the space of all Borel probability measures on X. Recall that we also defined

$$\Pi_{\mu,\nu} \coloneqq \{\, \pi \in P(X \times X) : \pi \text{ has marginals } \mu \text{ and } \nu \,\}.$$

for $\mu, \nu \in P(X)$. Now, denote by $C(X)$ the space of all continuous real-valued functions on X. We will now show that the Monge–Kantorovich problem has always an optimal solution.

Proposition 4.1 ([Gar18, Theorem 20.2.1, Corollary 20.2.2, Theorem 20.3.1])
Let $c\colon \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ be lower semicontinuous, and fix $\mu, \nu \in P(\mathbb{R})$. Suppose that

$$\inf\left\{ \int_{\mathbb{R}^2} c \,\mathrm{d}\pi : \pi \in \Pi_{\mu,\nu} \right\} < \infty.$$

Define

$$m \coloneqq \inf\left\{ \int_{\mathbb{R}^2} c \,\mathrm{d}\pi : \pi \in \Pi_{\mu,\nu} \right\} \text{ and}$$
$$M \coloneqq \sup\left\{ \int_{\mathbb{R}} f \,\mathrm{d}\mu + \int_{\mathbb{R}} g \,\mathrm{d}\nu : f, g \in C(\mathbb{R}) \text{ and } f(x) + g(y) \leq c(x,y) \text{ for all } x, y \in \mathbb{R} \right\}.$$

Then $m = M$, and there exists $\pi_0 \in \Pi_{\mu,\nu}$ such that

$$\int_{\mathbb{R}^2} c \,\mathrm{d}\pi_0 = m = M.$$

Proof. Let us first solve the Monge–Kantorovich problem on the compact unit square $[0,1]^2$ in the following Lemma 4.2.

Lemma 4.2 ([Gar18, Theorem 20.3.1])
Let $\tilde{c}\colon [0,1] \times [0,1] \to \mathbb{R}_{\geq 0}$ be lower semicontinuous, and fix $\tilde{\mu}, \tilde{\nu} \in P([0,1]^2)$. Suppose that

$$\inf\left\{ \int_{[0,1]^2} \tilde{c} \,\mathrm{d}\tilde{\pi} : \tilde{\pi} \in \Pi_{\tilde{\mu},\tilde{\nu}} \right\} < \infty.$$

Define

$$\tilde{m} \coloneqq \inf\left\{ \int_{[0,1]^2} \tilde{c} \,\mathrm{d}\tilde{\pi} : \tilde{\pi} \in \Pi_{\tilde{\mu},\tilde{\nu}} \right\} \text{ and}$$
$$\tilde{M} \coloneqq \sup\left\{ \int_{[0,1]} \tilde{f} \,\mathrm{d}\tilde{\mu} + \int_{[0,1]} \tilde{g} \,\mathrm{d}\tilde{\nu} : \tilde{f}, \tilde{g} \in C([0,1]) \text{ and } \tilde{f}(x) + \tilde{g}(y) \leq \tilde{c}(x,y) \text{ for all } x, y \in [0,1] \right\}.$$

Then $\tilde{m} = \tilde{M}$, and there exists $\tilde{\pi}_0 \in \Pi_{\tilde{\mu},\tilde{\nu}}$ such that

$$\int_{[0,1]^2} \tilde{c} \,\mathrm{d}\tilde{\pi}_0 = \tilde{m} = \tilde{M}.$$

Proof of Lemma 4.2. Firstly, observe that for all $\tilde{f}, \tilde{g} \in C([0,1])$ with $\tilde{f}(x) + \tilde{g}(y) \leq \tilde{c}(x,y)$ and for all $\tilde{\pi} \in \Pi_{\tilde{\mu},\tilde{\nu}}$, we have

$$\int_{[0,1]} \tilde{f} \,\mathrm{d}\tilde{\mu} + \int_{[0,1]} \tilde{g} \,\mathrm{d}\tilde{\nu} = \int_{[0,1]^2} (\tilde{f}(x) + \tilde{g}(y)) \,\mathrm{d}\tilde{\pi}(x,y) \leq \int_{[0,1]^2} \tilde{c} \,\mathrm{d}\tilde{\pi}.$$

Consequently, $\tilde{M} \leq \tilde{m}$.

We now show that there exists $\tilde{\pi}_0 \in \Pi_{\tilde{\mu},\tilde{\nu}}$ such that $\tilde{m} \leq \int_{[0,1]^2} \tilde{c}\,\mathrm{d}\tilde{\pi}_0 \leq \tilde{M}$. Let L be the space of all functions $\tilde{F}\colon [0,1]^2 \to \mathbb{R}$ of the form $\tilde{F}(x,y) = \tilde{f}(x) + \tilde{g}(y)$ with $\tilde{f}, \tilde{g} \in C([0,1])$. Then L is a linear subspace of $C([0,1]^2)$. Observe that the map $\varphi\colon L \to \mathbb{R}$ defined by

$$\varphi\big(\tilde{f}(x) + \tilde{g}(y)\big) \coloneqq \int_{[0,1]} \tilde{f}\,\mathrm{d}\tilde{\mu} + \int_{[0,1]} \tilde{g}\,\mathrm{d}\tilde{\nu}$$

is a well-defined linear functional on L which is not the zero linear functional on L, since $\varphi(1) = 1$.

Let U be the space of all continuous real-valued functions $\tilde{h}$ on $[0,1]^2$ such that

$$\tilde{h}(x,y) < \tilde{c}(x,y) \quad \text{for all } x, y \in [0,1].$$

Observe that U satisfies all of the following:

- $U \subset C([0,1]^2)$,
- U is non-empty, since $-1 \in U$,
- U is open in $C([0,1]^2)$, where we give $C([0,1]^2)$ the topology induced by the supremum norm $\|\cdot\|_\infty$,
- U is convex,
- $U \cap L$ is non-empty, since $-1 \in U \cap L$,
- φ is bounded above by $\tilde{m}$ on $U \cap L$, because if $\tilde{f}(x) + \tilde{g}(y) \in U \cap L$ and $\tilde{\pi} \in \Pi_{\tilde{\mu},\tilde{\nu}}$ then

$$\varphi\big(\tilde{f}(x) + \tilde{g}(y)\big) = \int_{[0,1]^2} \big(\tilde{f}(x) + \tilde{g}(y)\big)\,\mathrm{d}\tilde{\pi}(x,y) \leq \int_{[0,1]^2} \tilde{c}\,\mathrm{d}\tilde{\pi}.$$

Let $B \coloneqq \sup\{\, \varphi(\tilde{F}) : \tilde{F} \in U \cap L \,\} \leq \tilde{m}$, and let

$$L_B \coloneqq \Big\{\, \tilde{f}(x) + \tilde{g}(y) \in L : \varphi\big(\tilde{f}(x) + \tilde{g}(y)\big) \geq B \,\Big\}.$$

Then L_B satisfies all of the following:

- $L_B \subset C([0,1]^2)$
- L_B is non-empty, since $B \in L_B$,
- L_B is convex,
- L_B is disjoint from U, because if $\tilde{f}(x) + \tilde{g}(y) \in U \cap L$ then there exists $\hat{f} \in C([0,1])$ such that
$$\tilde{f}(x) + \tilde{g}(y) < \hat{f}(x) + \tilde{g}(y) < \tilde{c}(x,y) \quad \text{for all } x, y \in [0,1],$$
and consequently
$$\varphi\big(\tilde{f}(x) + \tilde{g}(y)\big) = \int_{[0,1]} \tilde{f}\,\mathrm{d}\tilde{\mu} + \int_{[0,1]} \tilde{g}\,\mathrm{d}\tilde{\nu} < \int_{[0,1]} \hat{f}\,\mathrm{d}\tilde{\mu} + \int_{[0,1]} \tilde{g}\,\mathrm{d}\tilde{\nu} \leq B.$$

Thus by the Hahn-Banach theorem, there exists a continuous linear functional $\psi\colon C([0,1]^2) \to \mathbb{R}$ such that

$$\text{if } \tilde{h} \in U \text{ then } \psi(\tilde{h}) < K \coloneqq \inf\Big\{\, \psi(\tilde{F}) : \tilde{F} \in L_B \,\Big\}.$$

Observe that ψ is a non-negative linear functional on $C([0,1]^2)$, that is,

$$\text{if } \tilde{h} \geq 0 \text{ then } \psi(\tilde{h}) \geq 0.$$

Indeed, if $\tilde{h} > 0$ then we have $-\alpha\tilde{h} \in U$ for all $\alpha > 0$. So if we had $\psi(\tilde{h}) < 0$ then $\psi(-\alpha\tilde{h}) = -\alpha\psi(\tilde{h}) < K$. But if $\psi(\tilde{h}) < 0$ then this cannot hold for all $\alpha > 0$.

As $\psi \neq 0$, we have $\psi(1) > 0$. Thus if we set $\theta := \frac{\psi}{\psi(1)}$, then θ is a non-negative linear functional on $C([0,1]^2)$ with $\theta(1) = 1$. Riesz's representation theorem thus yields the existence of some $\tilde{\pi}_0 \in P([0,1]^2)$ such that

$$\theta(\tilde{F}) = \int_{[0,1]^2} \tilde{F}(x,y)\,\mathrm{d}\tilde{\pi}_0(x,y) \quad \text{for all } \tilde{F} \in C([0,1]^2).$$

We claim that this $\tilde{\pi}_0$ is our desired transport plan satisfying $\tilde{m} \leq \int_{[0,1]^2} \tilde{c}\,\mathrm{d}\tilde{\pi}_0 \leq \tilde{M}$.

First, let us show that $\tilde{\pi}_0 \in \Pi_{\tilde{\mu},\tilde{\nu}}$. Let

$$\Lambda := \inf\{\, \theta(\tilde{F}) : \tilde{F} \in L_B \,\}.$$

Note that if $\tilde{h} \in U$ then $\theta(\tilde{h}) < \Lambda$. Observe that if $\tilde{F}_0 \in L$ satisfies $\varphi(\tilde{F}_0) = 0$, then

$$\varphi\big(B + \alpha\tilde{F}_0\big) = B \quad \text{for all } \alpha \in \mathbb{R},$$

and so $B + \alpha\tilde{F}_0 \in L_B$ for all $\alpha \in \mathbb{R}$, meaning that

$$B + \alpha\theta(\tilde{F}_0) = \theta\big(B + \alpha\tilde{F}_0\big) \geq \Lambda \quad \text{for all } \alpha \in \mathbb{R},$$

from which it follows that $\theta(\tilde{F}_0) = 0$ and hence

$$B \geq \Lambda. \tag{4.1}$$

Now for any $\tilde{F} \in L$, we can write

$$\tilde{F} = \varphi(\tilde{F}) + \tilde{F}_0$$

for some $\tilde{F}_0 \in L$ with $\varphi(\tilde{F}_0) = 0$. Hence

$$\theta(\tilde{F}) = \theta(\varphi(\tilde{F}) + \tilde{F}_0) = \varphi(\tilde{F}) + \theta(\tilde{F}_0) = \varphi(\tilde{F}).$$

In other words, θ extends φ from L to all of $C([0,1])^2$. Thus both of the following hold:

- if $\tilde{f} \in C([0,1])$ then

$$\int_{[0,1]^2} \tilde{f}(x)\,\mathrm{d}\tilde{\pi}_0(x,y) = \theta(\tilde{f}(x)) = \varphi(\tilde{f}(x)) = \int_{[0,1]} \tilde{f}(x)\,\mathrm{d}\tilde{\mu}(x),$$

- if $\tilde{g} \in C([0,1])$ then

$$\int_{[0,1]^2} \tilde{g}(y)\,\mathrm{d}\tilde{\pi}_0(x,y) = \theta(\tilde{g}(y)) = \varphi(\tilde{g}(y)) = \int_{[0,1]} \tilde{g}(y)\,\mathrm{d}\tilde{\nu}(y).$$

Therefore $\tilde{\pi}_0 \in \Pi_{\tilde{\mu},\tilde{\nu}}$.

Finally, we approximate the cost function $\tilde{c}$ from below by a sequence of non-negative functions $\tilde{h}_1, h_2, h_3, \ldots \in U$ with $h_n \nearrow \tilde{c}$ pointwise as $n \to \infty$ [Gar18, Theorem 4.2.9]. Then

$$\begin{aligned}
\tilde{m} &\leq \int_{[0,1]^2} \tilde{c} \,\mathrm{d}\tilde{\pi}_0, && \text{by the definition of } \tilde{m},\\
&= \lim_{n\to\infty} \int_{[0,1]^2} \tilde{h}_n \,\mathrm{d}\tilde{\pi}_0, && \text{by the monotone convergence theorem},\\
&\leq \sup\left\{ \int_{[0,1]^2} \tilde{h} \,\mathrm{d}\tilde{\pi}_0 : \tilde{h} \in U \right\}, && \text{since } h_1, h_2, h_2, \ldots \in U,\\
&= \sup\left\{ \theta(\tilde{h}) : \tilde{h} \in U \right\}, && \text{since } \tilde{\pi}_0 \text{ represents } \theta,\\
&\leq \Lambda, && \text{by the definition of } \Lambda,\\
&\leq B, && \text{from Equation (4.1)},\\
&= \sup\left\{ \varphi(\tilde{F}) : \tilde{F} \in U \cap L \right\}, && \text{by the definition of } B,\\
&\leq \tilde{M}, && \text{by the definitions of } \tilde{M},\ U,\ \text{and } L.
\end{aligned}$$

Therefore $\int_{[0,1]^2} \tilde{c} \,\mathrm{d}\tilde{\pi}_0 = \tilde{m} = \tilde{M}$, as desired. □

We now turn to proving Proposition 4.1, solving the Monge–Kantorovich problem on $\mathbb{R}$. Recall that

$$\begin{aligned}
m &\coloneqq \inf\left\{ \int_{\mathbb{R}^2} c \,\mathrm{d}\pi : \pi \in \Pi_{\mu,\nu} \right\} \quad \text{and}\\
M &\coloneqq \sup\left\{ \int_{\mathbb{R}} f \,\mathrm{d}\mu + \int_{\mathbb{R}} g \,\mathrm{d}\nu : f, g \in C(\mathbb{R}) \text{ and } f + g \leq c \right\},
\end{aligned}$$

and we aim to show that $m = M$. Arguing similarly as in the start of the proof of Lemma 4.2, we obtain $M \leq m$.

Now embed $\mathbb{R}^2$ into the compact unit square $[0,1]^2$ by identifying $\mathbb{R}^2$ with the open set $(0,1)^2$ via an obvious orientation-preserving contracting homeomorphism $h\colon \mathbb{R}^2 \to (0,1)^2$. Define $\tilde{c}\colon [0,1]^2 \to \mathbb{R}_{\geq 0}$ by

$$\tilde{c}(x,y) \coloneqq \begin{cases} c(h^{-1}(x,y)) & \text{if both } x, y \in (0,1),\\ 0 & \text{else,} \end{cases}$$

and define $\tilde{\mu}, \tilde{\nu} \in P([0,1])$ to be the respective image measures of μ and ν under the map h. Observe that $\tilde{c}$ is lower semicontinuous. Then Lemma 4.2 yields the existence of some $\tilde{\pi}_0 \in \Pi_{\tilde{\mu},\tilde{\nu}}$ satisfying

$$\int_{[0,1]^2} \tilde{c} \,\mathrm{d}\tilde{\pi}_0 = \sup\left\{ \int_{[0,1]} \tilde{f} \,\mathrm{d}\mu + \int_{[0,1]} \tilde{g} \,\mathrm{d}\tilde{\nu} : \tilde{f}, \tilde{g} \in \mathrm{BLip}([0,1]) \text{ and } \tilde{f} + \tilde{g} \leq \tilde{c} \right\},$$

where $\mathrm{BLip}(X)$ denotes the space of bounded Lipschitz functions on a set $X \subseteq \mathbb{R}$. Now note that

$$\tilde{\pi}_0\big([0,1]^2 \setminus (0,1)^2\big) \leq \tilde{\mu}\big(\{0,1\}\big) + \tilde{\nu}\big(\{0,1\}\big) = \mu\big(\{0,1\}\big) + \nu\big(\{0,1\}\big) = 0.$$

Hence, letting $\tilde{\pi}_0|_{(0,1)^2}$ be the restriction of $\tilde{\pi}_0$ to $(0,1)^2$, we may define $\pi_0 \in P(\mathbb{R}^2)$ to be the image measure of $\tilde{\pi}_0|_{(0,1)^2}$ under the map h^{-1}. It is easy to check that we indeed have $\pi_0 \in \Pi_{\mu,\nu}$,

and that

$$\begin{aligned} M &= \sup\left\{ \int_{\mathbb{R}} f \,\mathrm{d}\mu + \int_{\mathbb{R}} g \,\mathrm{d}\nu : f, g \in \mathrm{BLip}(\mathbb{R}) \text{ and } f + g \leq c \right\} \\ &\geq \sup\left\{ \int_{[0,1]} \tilde{f} \,\mathrm{d}\tilde{\mu} + \int_{[0,1]} \tilde{g} \,\mathrm{d}\tilde{\nu} : \tilde{f}, \tilde{g} \in \mathrm{BLip}([0,1]) \text{ and } \tilde{f} + \tilde{g} \leq \tilde{c} \right\} \\ &\geq \int_{[0,1]^2} \tilde{c} \,\mathrm{d}\tilde{\pi}_0 \\ &= \int_{\mathbb{R}^2} c \,\mathrm{d}\pi_0 \\ &\geq m. \end{aligned}$$

This proves that $\int_{\mathbb{R}^2} c \,\mathrm{d}\pi_0 = m = M$, and we are done with the proof of Proposition 4.1. □

In the definition of the Wasserstein metric (Definition 2.1), we defined the space

$$P_1(\mathbb{R}) \coloneqq \left\{ \mu \in P(\mathbb{R}) : \int_{\mathbb{R}} |x| \,\mathrm{d}\mu(x) < \infty \right\}$$

and we defined the Wasserstein metric $W_1 \colon P_1(\mathbb{R}) \times P_1(\mathbb{R}) \to \mathbb{R}_{\geq 0}$ by

$$W_1(\mu, \nu) \coloneqq \inf\left\{ \int_{\mathbb{R}^2} |x - y| \,\mathrm{d}\pi(x, y) : \pi \in \Pi_{\mu,\nu} \right\}$$

Proposition 4.1 shows that we have

$$W_1(\mu, \nu) = \sup\left\{ \int_{\mathbb{R}} f \,\mathrm{d}\mu + \int_{\mathbb{R}} g \,\mathrm{d}\nu : f, g \in C(\mathbb{R}) \text{ and } f(x) + g(y) \leq |x - y| \text{ for all } x, y \in \mathbb{R} \right\}$$

for all $\mu, \nu \in P_1(\mathbb{R})$. In particular, by considering the cases

- $f(x) = x$ and $g(y) = -y$, or
- $f(x) = -x$ and $g(y) = y$,

we obtain

$$W_1(\mu, \nu) \geq \left| \int_{\mathbb{R}} x \,\mathrm{d}\mu(x) - \int_{\mathbb{R}} y \,\mathrm{d}\nu(y) \right|,$$

which proves Theorem 2.2, stating that the Wasserstein distance between μ and ν is at least the distance between their means. In fact, we can do better. Denote by $\mathrm{Lip}(\mathbb{R})$ the space of all Lipschitz continuous functions on $\mathbb{R}$. For $h \in \mathrm{Lip}(\mathbb{R})$, the Lipschitz constant of h is defined to be

$$\|h\|_{\mathrm{Lip}} \coloneqq \sup_{\substack{x,y \in \mathbb{R}, \\ x \neq y}} \frac{|h(x) - h(y)|}{|x - y|}.$$

Proposition 4.3 ([Gar18, Corollary 21.2.3])
For $\mu, \nu \in P_1(\mathbb{R})$ we have

$$W_1(\mu, \nu) \geq \sup\left\{ \left| \int_{\mathbb{R}} h \,\mathrm{d}\mu - \int_{\mathbb{R}} h \,\mathrm{d}\nu \right| : h \in \mathrm{Lip}(\mathbb{R}) \text{ and } \|h\|_{\mathrm{Lip}} \leq 1 \right\}.$$

Proof. This is a corollary of Proposition 4.1. □

The inequality in Proposition 4.3 can actually be proven to be an equality [Gar18, Corollary 21.2.3], but we will not need this fact in this document.

We conclude this section by showing that the Wasserstein metric is indeed a metric on $P_1(\mathbb{R})$, justifying its name.

Proposition 4.4 ([Gar18, Theorem 21.1.2] [Vil09, Chapter 6])
For all $\mu, \nu \in P_1(\mathbb{R})$, the value $W_1(\mu, \nu)$ is finite. Furthermore, $W_1 \colon P_1(\mathbb{R}) \times P_1(\mathbb{R}) \to \mathbb{R}_{\geq 0}$ is a metric, that is,

- *for all $\mu, \nu \in P_1(\mathbb{R})$, we have*
$$W_1(\mu, \nu) = W_1(\nu, \mu),$$
- *for all $\mu, \nu \in P_1(\mathbb{R})$, we have*
$$W_1(\mu, \nu) = 0 \text{ if and only if } \mu = \nu,$$
- *W_1 satisfies the triangle inequality, that is, for all $\mu, \nu, \xi \in P_1(\mathbb{R})$, we have*
$$W_1(\mu, \xi) \leq W_1(\mu, \nu) + W_1(\nu, \xi).$$

Proof. Let us first show that $W_1(\mu, \nu)$ is finite when $\mu, \nu \in P_1(\mathbb{R})$. Indeed, the product measure $\pi \coloneqq \mu \times \nu \in \Pi_{\mu,\nu}$ satisfies

$$\begin{aligned} W_1(\mu, \nu) &\leq \int_{\mathbb{R}^2} |x - y| \, \mathrm{d}\pi(x, y) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} |x - y| \, \mathrm{d}\mu(x) \, \mathrm{d}\nu(y) \\ &\leq \int_{\mathbb{R}} \int_{\mathbb{R}} \big(|x| + |y|\big) \, \mathrm{d}\mu(x) \, \mathrm{d}\nu(y) \\ &= \int_{\mathbb{R}} |x| \, \mathrm{d}\mu(x) + \int_{\mathbb{R}} |y| \, \mathrm{d}\nu(y) \\ &< \infty. \end{aligned}$$

We now turn to proving that W_1 is a metric on $P_1(\mathbb{R})$. The reader might wish to note that this is an immediate corollary of Theorem 3.1. We present, in the remainder of this proof, a different proof which more readily generalises to Wasserstein metrics defined over arbitrary Polish spaces.

Fix $\mu, \nu \in P_1(\mathbb{R})$. It is evident from the definition of W_1 that $W_1(\mu, \nu) = W_1(\nu, \mu)$. Next, if $\mu = \nu$ then clearly the transport plan $\pi_0 \in \Pi_{\mu,\nu}$ which is the image measure of μ under the function $\mathbb{R} \ni x \mapsto (x, x) \in \mathbb{R}^2$ satisfies

$$0 \leq W_1(\mu, \nu) \leq \int_{\mathbb{R}^2} |x - y| \, \mathrm{d}\pi_0(x, y) = 0.$$

On the other hand, if $\mu \neq \nu$, then there exists a compact interval $[a, b] \subset \mathbb{R}$ such that

$$\mu([a, b]) \neq \nu([a, b]).$$

Supposing, without loss of generality, that $\mu([a, b]) > \nu([a, b])$, we let

$$\varepsilon \coloneqq \frac{1}{2}\big(\mu([a, b]) - \nu([a, b])\big) > 0.$$

Now, there exists $\delta > 0$ such that

$$\nu((a - \delta, b + \delta)) < \nu([a, b]) + \varepsilon.$$

Consequently, for any $\pi \in \Pi_{\mu,\nu}$, we have

$$\begin{aligned}
\int_{\mathbb{R}^2} |x-y| \,\mathrm{d}\pi(x,y) &\geq \int_{[a,b]\times\left(\mathbb{R}\setminus(a-\delta,b+\delta)\right)} |x-y| \,\mathrm{d}\pi(x,y) \\
&\geq \int_{[a,b]\times\left(\mathbb{R}\setminus(a-\delta,b+\delta)\right)} \delta \,\mathrm{d}\pi(x,y) \\
&= \delta \cdot \pi\Big([a,b] \times \big(\mathbb{R} \setminus (a-\delta, b+\delta)\big)\Big) \\
&= \delta \cdot \Big(\pi\big([a,b]\times\mathbb{R}\big) - \pi\big([a,b]\times(a-\delta,b+\delta)\big)\Big) \\
&\geq \delta \cdot \Big(\pi\big([a,b]\times\mathbb{R}\big) - \pi\big(\mathbb{R}\times(a-\delta,b+\delta)\big)\Big) \\
&= \delta \cdot \big(\mu([a,b]) - \nu((a-\delta,b+\delta))\big) \\
&\geq \delta\varepsilon.
\end{aligned}$$

Hence $W_1(\mu,\nu) \geq \delta\varepsilon > 0$.

Finally, we show that W_1 satisfies the triangle inequality. Fix $\mu,\nu,\xi \in P_1(\mathbb{R})$. As the Monge–Kantorovich problem always has a solution (see Proposition 4.1), there exist $\pi_1 \in \Pi_{\mu,\nu}$ and $\pi_2 \in \Pi_{\nu,\xi}$ such that

$$W_1(\mu,\nu) = \int_{\mathbb{R}^2} |x-y| \,\mathrm{d}\pi_1(x,y) \quad \text{and} \quad W_1(\nu,\xi) = \int_{\mathbb{R}^2} |x-y| \,\mathrm{d}\pi_2(x,y).$$

We interrupt the proof with the following technical Lemma 4.5 to obtain a measure $\gamma \in P(\mathbb{R}^3)$ which "glues" together the transport plans π_1 and π_2.

Lemma 4.5 (Gluing Lemma [Aki22] [Gar18, Theorem 16.1.1])
Define the projection maps $p_{1,2}, p_{2,3}, p_{1,3} \colon \mathbb{R}^3 \to \mathbb{R}^2$ *by*

$$p_{i,j}(x_1,x_2,x_3) \coloneqq (x_i,x_j).$$

Then there exists $\gamma \in P(\mathbb{R}^3)$ *such that, when we define* $\gamma_{i,j}$ *to be the image measure of* γ *under* $p_{i,j}$*, we obtain*

$$\gamma_{1,2} = \pi_1 \quad \textit{and} \quad \gamma_{2,3} = \pi_2.$$

Proof of Lemma 4.5. By disintegration of measures [Aki22] [Gar18, Theorem 16.10.1], there exists a collection of measures $\{\pi_1(\cdot|y)\}_{y\in\mathbb{R}} \subseteq P(\mathbb{R})$ such that

$$\pi_1(X\times Y) = \int_Y \pi_1(X|y) \,\mathrm{d}\nu(y) \quad \text{for all Borel } X,Y \subseteq \mathbb{R}.$$

Similarly, there exists a collection of measures $\{\pi_2(\cdot|y)\}_{y\in\mathbb{R}} \subseteq P(\mathbb{R})$ such that

$$\pi_2(Y\times Z) = \int_Y \pi_2(Y|y) \,\mathrm{d}\nu(y) \quad \text{for all Borel } Y,Z \subseteq \mathbb{R}.$$

Then the unique measure $\gamma \in P(\mathbb{R}^3)$ satisfying

$$\gamma(X\times Y\times Z) = \int_Y \pi_1(X|y)\pi_2(Z|y) \,\mathrm{d}\nu(y) \quad \text{for all Borel } X,Y,Z \subseteq \mathbb{R}$$

works. □

Returning to the proof of the triangle inequality in Proposition 4.4, we take the measure γ obtained from Lemma 4.5. It is easy to check that $\gamma_{1,3} \in \Pi_{\mu,\xi}$. We thus obtain

$$\begin{aligned}
W_1(\mu,\xi) &\leq \int_{\mathbb{R}^2} |x-z| \, \mathrm{d}\gamma_{1,3}(x,z) \\
&= \int_{\mathbb{R}^3} |x-z| \, \mathrm{d}\gamma(x,y,z) \\
&\leq \int_{\mathbb{R}^3} |x-y| \, \mathrm{d}\gamma(x,y,z) + \int_{\mathbb{R}^3} |y-z| \, \mathrm{d}\gamma(x,y,z) \\
&= \int_{\mathbb{R}^2} |x-y| \, \mathrm{d}\gamma_{1,2}(x,y) + \int_{\mathbb{R}^2} |y-z| \, \mathrm{d}\gamma_{2,3}(y,z) \\
&= \int_{\mathbb{R}^2} |x-y| \, \mathrm{d}\pi_1(x,y) + \int_{\mathbb{R}^2} |y-z| \, \mathrm{d}\pi_2(y,z) \\
&= W_1(\mu,\nu) + W_1(\nu,\xi).
\end{aligned}$$

This completes the proof of Proposition 4.4, showing that W_1 is a metric on $P_1(\mathbb{R})$. □

5 Bourrée

Theorem 5.1 ([Eva24])
Let $\mu, \nu \in P_1(\mathbb{R})$ have probability density functions $f, g \in C^1$ respectively. Then there exists $C > 0$ such that

$$\int_{\mathbb{R}} |f(x) - g(x)| \, \mathrm{d}x \le C \sqrt{\left(\int_{\mathbb{R}} (|f'(x)| + |g'(x)|) \, \mathrm{d}x \right) W_1(\mu, \nu)}.$$

Proof. First, recall that

$$\int_{\mathbb{R}} |f(x) - g(x)| \, \mathrm{d}x \le \sup_{\substack{\varphi \in C_c^\infty \\ \|\varphi\|_\infty \le 1}} \left(\int_{\mathbb{R}} f(x)\varphi(x) \, \mathrm{d}x - \int_{\mathbb{R}} g(x)\varphi(x) \, \mathrm{d}x \right),$$

where C_c^∞ is the space of all smooth functions with compact support, and $\|\cdot\|_\infty$ is the supremum norm.

Fix any smooth positive function $\eta \colon \mathbb{R} \to \mathbb{R}$ with compact support in the open unit interval $(0, 1)$ and $\int_{\mathbb{R}} \eta(x) \, \mathrm{d}x = 1$. For any $\varphi \in C_c^\infty$ with $\|\varphi\|_\infty \le 1$, the convolution

$$(\varphi * \eta)(x) \coloneqq \int_{\mathbb{R}} \varphi(y)\eta(x - y) \, \mathrm{d}y = \int_{\mathbb{R}} \varphi(x - y)\eta(y) \, \mathrm{d}y$$

is differentiable, and $(\varphi * \eta)' = \varphi * \eta'$. Hence the Lipschitz constant $\|\varphi * \eta\|_{\mathrm{Lip}}$ of $\varphi * \eta$ satisfiies

$$\|\varphi * \eta\|_{\mathrm{Lip}} \le \|\varphi * \eta'\|_\infty \le \|\varphi\|_\infty \|\eta'\|_{L^1} \le \|\eta'\|_{L^1}.$$

For $\varepsilon > 0$, define $\eta_\varepsilon \colon \mathbb{R} \to \mathbb{R}$ by $\eta_\varepsilon(x) \coloneqq \eta(\varepsilon x)$. Note that $\|\varphi * \eta_\varepsilon\|_{\mathrm{Lip}} \le \|\eta'\|_{L^1}$.

Now,

$$\int_{\mathbb{R}} f(x)\varphi(x) \, \mathrm{d}x - \int_{\mathbb{R}} g(x)\varphi(x) \, \mathrm{d}x = \mathrm{I} + \mathrm{II} + \mathrm{III},$$

where

$$\begin{aligned}
\mathrm{I} &\coloneqq \int_{\mathbb{R}} f(x)\varphi(x) \, \mathrm{d}x - \varepsilon \int_{\mathbb{R}} f(x)(\varphi * \eta_\varepsilon)(x) \, \mathrm{d}x, \\
\mathrm{II} &\coloneqq \varepsilon \int_{\mathbb{R}} f(x)(\varphi * \eta_\varepsilon)(x) \, \mathrm{d}x - \varepsilon \int_{\mathbb{R}} g(x)(\varphi * \eta_\varepsilon)(x) \, \mathrm{d}x, \text{ and} \\
\mathrm{III} &\coloneqq \varepsilon \int_{\mathbb{R}} g(x)(\varphi * \eta_\varepsilon)(x) \, \mathrm{d}x - \int_{\mathbb{R}} g(x)\varphi(x) \, \mathrm{d}x.
\end{aligned}$$

Note that $\varphi * \eta_\varepsilon$ is Lipschitz, so Proposition 4.3 gives us

$$|\mathrm{II}| \le \varepsilon \|\varphi * \eta_\varepsilon\|_{\mathrm{Lip}} W_1(\mu, \nu) \le \varepsilon \|\eta'\|_{L^1} W_1(\mu, \nu).$$

Next, using the substitution $z \coloneqq x - y$, we have

$$\begin{aligned}
\mathrm{I} &= \int_{\mathbb{R}} f(x)\varphi(x) \, \mathrm{d}x - \varepsilon \int_{\mathbb{R}} f(x)(\varphi * \eta_\varepsilon)(x) \, \mathrm{d}x \\
&= \int_{\mathbb{R}} f(x)\varphi(x) \, \mathrm{d}x - \varepsilon \int_{\mathbb{R}} f(x) \left(\int_{\mathbb{R}} \varphi(x - y)\eta_\varepsilon(y) \, \mathrm{d}y \right) \mathrm{d}x \\
&= \int_{\mathbb{R}} f(x)\varphi(x) \, \mathrm{d}x - \varepsilon \int_{\mathbb{R}} \int_{\mathbb{R}} f(x)\varphi(x - y)\eta_\varepsilon(y) \, \mathrm{d}y \, \mathrm{d}x \\
&= \int_{\mathbb{R}} f(x)\varphi(x) \, \mathrm{d}x - \varepsilon \int_{\mathbb{R}} \int_{\mathbb{R}} f(z + y)\varphi(z)\eta_\varepsilon(y) \, \mathrm{d}y \, \mathrm{d}z.
\end{aligned}$$

Since $\int_{\mathbb{R}} \eta_\varepsilon(y)\,\mathrm{d}y = \frac{1}{\varepsilon}$, we multiply the first term with $\varepsilon \int_{\mathbb{R}} \eta_\varepsilon(y)\,\mathrm{d}y$ to get

$$\begin{aligned}
\mathrm{I} &= \varepsilon \int_{\mathbb{R}} \int_{\mathbb{R}} f(x)\varphi(x)\eta_\varepsilon(y)\,\mathrm{d}y\,\mathrm{d}x - \varepsilon \int_{\mathbb{R}} \int_{\mathbb{R}} f(x+y)\varphi(x)\eta_\varepsilon(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \varepsilon \int_{\mathbb{R}} \int_{\mathbb{R}} (f(x) - f(x+y))\varphi(x)\eta_\varepsilon(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \varepsilon \int_{\mathbb{R}} \int_{\mathbb{R}} \left(\int_{x+y}^{x} f'(z)\,\mathrm{d}z \right) \varphi(x)\eta_\varepsilon(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= -\varepsilon \int_{\mathbb{R}} \int_{\mathbb{R}} \left(\int_0^1 f'(x+ty)y\,\mathrm{d}t \right) \varphi(x)\eta_\varepsilon(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= -\varepsilon \int_0^1 \int_{\mathbb{R}} \int_{\mathbb{R}} f'(x+ty) \cdot \varphi(x) \cdot y \cdot \eta_\varepsilon(y)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t.
\end{aligned}$$

Making the substitution $\xi \coloneqq x + ty$ on the innermost integral, we obtain

$$\mathrm{I} = -\varepsilon \int_0^1 \int_{\mathbb{R}} \int_{\mathbb{R}} f'(\xi) \cdot \varphi(\xi - ty) \cdot y \cdot \eta_\varepsilon(y)\,\mathrm{d}\xi\,\mathrm{d}y\,\mathrm{d}t,$$

hence

$$\begin{aligned}
|\mathrm{I}| &\le \varepsilon \int_0^1 \int_{\mathbb{R}} \int_{\mathbb{R}} |f'(\xi) \cdot \|\varphi\|_\infty \cdot |y| \cdot \eta_\varepsilon(y)\,\mathrm{d}\xi\,\mathrm{d}y\,\mathrm{d}t \\
&\le \varepsilon \left(\int_{\mathbb{R}} |f'(\xi)|\,\mathrm{d}\xi \right) \left(\int_{\mathbb{R}} |y|\eta_\varepsilon(y)\,\mathrm{d}y \right) \\
&= \varepsilon \left(\int_{\mathbb{R}} |f'(\xi)|\,\mathrm{d}\xi \right) \left(\int_{\mathbb{R}} |y|\eta(\varepsilon y)\,\mathrm{d}y \right).
\end{aligned}$$

The substitution $\tilde{y} \coloneqq \varepsilon y$ finally gives us

$$|\mathrm{I}| \le \frac{1}{\varepsilon} \left(\int_{\mathbb{R}} f'(\xi)\,\mathrm{d}\xi \right) \left(\int_{\mathbb{R}} |\tilde{y}|\eta(\tilde{y})\,\mathrm{d}\tilde{y} \right).$$

Similarly,

$$|\mathrm{III}| \le \frac{1}{\varepsilon} \left(\int_{\mathbb{R}} |g'(x)|\,\mathrm{d}x \right) \left(\int_{\mathbb{R}} |y|\eta(y)\,\mathrm{d}y \right).$$

Therefore, from $\int_{\mathbb{R}} f(x)\varphi(x)\,\mathrm{d}x - \int_{\mathbb{R}} g(x)\varphi(x)\,\mathrm{d}x \le |\mathrm{I}| + |\mathrm{II}| + |\mathrm{III}|$, we obtain

$$\int_{\mathbb{R}} |f(x) - g(x)|\,\mathrm{d}x \le \frac{1}{\varepsilon} \left(\int_{\mathbb{R}} |y|\eta(y)\,\mathrm{d}y \right) \left(\int_{\mathbb{R}} |\left(f'(x)| + |g'(x)|\right)\,\mathrm{d}x \right) + \varepsilon \|\eta'\|_{L^1} W_1(\mu,\nu).$$

When we fix constants $A, B > 0$, the value of $\varepsilon > 0$ which minimises the function $\varepsilon \mapsto \frac{1}{\varepsilon}A + \varepsilon B$ is $\varepsilon = \sqrt{\frac{A}{B}}$ where we attain the minimum value of $2\sqrt{AB}$. Therefore

$$\int_{\mathbb{R}} |f(x) - g(x)|\,\mathrm{d}x \le 2\sqrt{\|\eta'\|_{L^1} \left(\int_{\mathbb{R}} |y|\eta(y)\,\mathrm{d}y \right) \left(\int_{\mathbb{R}} \left(|f'(x)| + |g'(x)|\right)\,\mathrm{d}x \right) W_1(\mu,\nu)},$$

as desired, proving Theorem 5.1. □

6 Gigue

We now define Boltzmann's H functional, which represents the relative entropy[5] between two Borel probability measures on $\mathbb{R}$.

Definition 6.1 (Relative entropy, [Vil09, Chapter 22])
Let $\mu, \nu \in P(\mathbb{R})$. If the Radon–Nikodym derivative $\rho \coloneqq \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ exists, we define

$$H(\mu|\nu) \coloneqq \int_{\mathbb{R}} \rho \log(\rho)\, \mathrm{d}\nu.$$

Otherwise we define $H(\mu|\nu) \coloneqq \infty$.

Observe that always have $H(\mu|\nu) \geq 0$. To see this, first note that $\int_{\mathbb{R}} \rho\, \mathrm{d}\nu = \int_{\mathbb{R}} \mathrm{d}\mu = 1$. Consequently,

$$\int_{\mathbb{R}} \rho \log(\rho)\, \mathrm{d}\nu = \int_{\mathbb{R}} (\rho \log \rho - \rho + 1)\, \mathrm{d}\nu \geq 0$$

due to the fact that $x \log(x) - x + 1 \geq 0$ for all $x > 0$. We are thus justified in taking the square root $\sqrt{H(\mu|\nu)}$, and we will do so in Theorem 6.2.

Theorem 6.2 ([Vil09, Definition 22.1, Theorem 22.10])
Let $\nu \in P_1(\mathbb{R})$. Suppose there exists $C > 0$ such that for all Borel sets $A \subseteq \mathbb{R}$,

$$\text{if } \nu(A) \geq \frac{1}{2}, \text{ then for all } r > 0 \text{ we have } \nu(A^r) \geq 1 - e^{-Cr^2},$$

where $A^r \coloneqq \left\{ x \in \mathbb{R} \ : \ \inf\{\, |x - y| : y \in A \,\} \leq r \right\}$. Then there exists a constant $K > 0$ such that for all $\mu \in P_1(\mathbb{R})$ which is absolutely continuous with respect to ν, we have

$$W_1(\mu, \nu) \leq K \sqrt{H(\mu|\nu)}.$$

Proof. For a signed measure η, let $\eta = \eta^+ - \eta^-$ be the Jordan decomposition of η, so η^+ and η^- are non-negative measures. We define the absolute variation of η to be the measure

$$|\eta| \coloneqq \eta^+ + \eta^-.$$

We connect the absolute variation to the Wasserstein distance in the following Lemma 6.3.

Lemma 6.3 ([Vil09, Theorem 6.15])
For all $\mu, \xi \in P_1(\mathbb{R})$ and all $x_0 \in \mathbb{R}$, we have

$$W_1(\mu, \xi) \leq \int_{\mathbb{R}} |x - x_0|\, \mathrm{d}|\mu - \xi|(x).$$

Proof of Lemma 6.3. Define π_1 to be the image measure of $\min\{\mu, \xi\}$ under the map $\mathbb{R} \ni x \mapsto (x, x) \in \mathbb{R}^2$. Define π_2 to be the product measure of $(\mu - \xi)^+$ and $(\mu - \xi)^-$, and define $a \coloneqq (\mu - \xi)^+(\mathbb{R}) = (\mu - \xi)^-(\mathbb{R})$. Then define $\pi \in P(\mathbb{R}^2)$ by

$$\pi \coloneqq \pi_1 + \frac{1}{a}\pi_2.$$

[5] The letter H which appears here is the Greek H (capital η), rather than the Latin H (capital h).

Then $\pi \in \Pi_{\mu,\xi}$, and so

$$\begin{aligned}
W_1(\mu,\xi) &\le \int_{\mathbb{R}^2} |x-y| \,\mathrm{d}\pi(x,y) \\
&= \frac{1}{a} \int_{\mathbb{R}^2} |x-y| \,\mathrm{d}\pi_2(x,y) \\
&= \frac{1}{a} \int_{\mathbb{R}} \int_{\mathbb{R}} |x-y| \,\mathrm{d}(\mu-\xi)^+(x) \,\mathrm{d}(\mu-\xi)^-(y) \\
&= \frac{1}{a} \int_{\mathbb{R}} \int_{\mathbb{R}} |x-x_0| \,\mathrm{d}(\mu-\xi)^+(x) \,\mathrm{d}(\mu-\xi)^-(y) \\
&\quad + \frac{1}{a} \int_{\mathbb{R}} \int_{\mathbb{R}} |x_0-y| \,\mathrm{d}(\mu-\xi)^+(x) \,\mathrm{d}(\mu-\xi)^-(y) \\
&= \int_{\mathbb{R}} |x-x_0| \,\mathrm{d}(\mu-\xi)^+(x) + \int_{\mathbb{R}} |x_0-y| \,\mathrm{d}(\mu-\xi)^-(y) \\
&= \int_{\mathbb{R}} |x-x_0| \,\mathrm{d}((\mu-\xi)^+ + (\mu-\xi)^-)(x) \\
&= \int_{\mathbb{R}} |x-x_0| \,\mathrm{d}|\mu-\xi|(x). \qquad \square
\end{aligned}$$

Let us now turn to proving Theorem 6.2. We first prove the following Lemma 6.4, which establishes the finiteness of an integral we will use in the final proof of Theorem 6.2.

Lemma 6.4 ([Vil09, Theorem 22.10])
There exist $x_0 \in \mathbb{R}$ and $k > 0$ such that the integral

$$\int_{\mathbb{R}} e^{k(x-x_0)^2} \,\mathrm{d}\nu(x)$$

is finite.

Proof of Lemma 6.4. Fix any compact set $A \subset \mathbb{R}$ such that $\nu(A) \ge \frac{1}{2}$. Fix any $x_0 \in A$, and let $R \coloneqq \max\{\, |x-y| : x, y \in A \,\}$ be the diameter of A. Let $d(\cdot, A) \colon \mathbb{R} \to \mathbb{R}_{\ge 0}$ be the distance function from A, defined by

$$d(x,A) \coloneqq \min\{\, |x-y| : y \in A \,\}.$$

Then $d(\cdot, A)$ has median 0, that is,

$$\nu\big(\{\, x \in \mathbb{R} : d(x,A) \ge 0 \,\}\big) \ge \frac{1}{2} \quad \text{and} \quad \nu\big(\{\, x \in \mathbb{R} : d(x,A) \le 0 \,\}\big) \ge \frac{1}{2}.$$

By the hypothesis of Theorem 6.2, for all $r > 0$ and for all $\tilde{r} \in (0, r)$, we have

$$\nu\big(\{\, x \in \mathbb{R} : d(x,A) \le 0 \,\}^{\tilde{r}}\big) \ge 1 - e^{-C\tilde{r}^2}.$$

Consequently,

$$\nu\big(\{\, x \in \mathbb{R} : d(x,A) \ge r \,\}\big) \le e^{-C\tilde{r}^2}.$$

As the above inequality holds for all $\tilde{r} \in (0, r)$, we obtain

$$\nu\big(\{\, x \in \mathbb{R} : |x-x_0| \ge R + r \,\}\big) \le \nu\big(\{\, x \in \mathbb{R} : d(x,A) \ge r \,\}\big) \le e^{-Cr^2}$$

for all $r > 0$.

Now choose any $0 < k < C$. Then we would have

$$\begin{aligned}
\int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x) &= \int_{\mathbb{R}} \int_0^{|x-x_0|} 2kse^{ks^2}\,\mathrm{d}s\,\mathrm{d}\nu(x) + 1 \\
&= \int_{\mathbb{R}} \int_0^{\infty} 2kse^{ks^2} \mathbb{1}_{s \le |x-x_0|}\,\mathrm{d}s\,\mathrm{d}\nu(x) + 1 \\
&= \int_0^{\infty} \int_{\mathbb{R}} 2kse^{ks^2} \mathbb{1}_{s \le |x-x_0|}\,\mathrm{d}\nu(x)\,\mathrm{d}s + 1 \\
&= \int_0^{\infty} 2kse^{ks^2} \nu\big(\{x \in \mathbb{R} : |x-x_0| \ge s\}\big)\,\mathrm{d}s + 1 \\
&= \int_0^{R} 2kse^{ks^2} \nu\big(\{x \in \mathbb{R} : |x-x_0| \ge s\}\big)\,\mathrm{d}s \\
&\quad + \int_R^{\infty} 2kse^{ks^2} \nu\big(\{x \in \mathbb{R} : |x-x_0| \ge s\}\big)\,\mathrm{d}s \\
&\quad + 1 \\
&\le \int_0^{R} 2kse^{ks^2}\,\mathrm{d}s + \int_R^{\infty} 2kse^{ks^2} e^{-C(s-R)^2}\,\mathrm{d}s + 1 \\
&= \int_0^{R} 2kse^{ks^2}\,\mathrm{d}s + 2ke^{-CR^2} \int_R^{\infty} se^{-(C-k)s^2+2CRs}\,\mathrm{d}s + 1 \\
&< \infty.
\end{aligned}$$

□

Returning to the proof of Theorem 6.2, let us take $x_0 \in \mathbb{R}$ and $k > 0$ from the conclusion of Lemma 6.4. By virtue of Lemma 6.3, it suffices to show that

$$\int_{\mathbb{R}} |x - x_0|\,\mathrm{d}|\mu - \nu|(x) \le K\sqrt{H(\mu|\nu)}$$

for some constant $K > 0$.

As μ is absolutely continuous with respect to ν, the Radon–Nikodym derivative $\rho \coloneqq \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ exists. Define $\tilde{\rho}\colon \mathbb{R} \to \mathbb{R}$ by $\tilde{\rho}(x) \coloneqq \rho(x) - 1$, and define $h\colon [-1, \infty) \to \mathbb{R}_{\ge 0}$ by

$$h(s) \coloneqq \begin{cases} (1+s)\log(1+s) - s & \text{if } s > -1, \\ 1 & \text{if } s = -1. \end{cases}$$

Then we can rewrite $H(\mu|\nu)$ as

$$H(\mu|\nu) = \int_{\mathbb{R}} (h \circ \tilde{\rho})\,\mathrm{d}\nu.$$

Now, Taylor expanding h around 0 with integral remainder gives

$$h(s) = \int_0^s \frac{s-t}{1+t}\,\mathrm{d}t = s^2 \int_0^1 \frac{1-t}{1+ts}\,\mathrm{d}t.$$

for $s \ge 0$, since $h(0) = h'(0) = 0$. Hence

$$H(\mu|\nu) = \int_{\mathbb{R}} \int_0^1 \frac{(\tilde{\rho}(x))^2(1-t)}{1+t\rho(x)}\,\mathrm{d}t\,\mathrm{d}\nu(x),$$

and so by the Cauchy–Schwarz inequality,

$$\begin{aligned}
&H(\mu|\nu) \int_{\mathbb{R}} \int_0^1 k(1-t)(x-x_0)^2(1+t\tilde{\rho}(x))\,\mathrm{d}t\,\mathrm{d}\nu(x) \\
&\ge \left(\int_{\mathbb{R}} \int_0^1 (1-t) \cdot \sqrt{k}|x-x_0| \cdot |\tilde{\rho}(x)|\,\mathrm{d}t\,\mathrm{d}\nu(x) \right)^2 \\
&= \frac{1}{4} \left(\int_{\mathbb{R}} \sqrt{k}|x-x_0| \cdot |\tilde{\rho}(x)|\,\mathrm{d}\nu(x) \right)^2.
\end{aligned}$$

Thus we obtain

$$\int_{\mathbb{R}} \sqrt{k}|x-x_0|\,\mathrm{d}|\mu-\nu|(x) = \int_{\mathbb{R}} \sqrt{k}|x-x_0|\cdot|\tilde{\rho}(x)|\,\mathrm{d}\nu(x) \leq K_\mu \sqrt{H(\mu|\nu)}$$

where

$$\begin{aligned}
K_\mu &\coloneqq 2\sqrt{\int_{\mathbb{R}}\int_0^1 k(1-t)(x-x_0)^2(1+t\tilde{\rho}(x))\,\mathrm{d}t\,\mathrm{d}\nu(x)} \\
&= 2\sqrt{\int_{\mathbb{R}}\int_0^1 k(1-t)(x-x_0)^2\,\mathrm{d}t\,\mathrm{d}\nu(x) + \int_{\mathbb{R}}\int_0^1 kt(1-t)(x-x_0)^2\tilde{\rho}(x)\,\mathrm{d}t\,\mathrm{d}\nu(x)} \\
&= 2\sqrt{\int_{\mathbb{R}}\int_0^1 k(1-t)^2(x-x_0)^2\,\mathrm{d}t\,\mathrm{d}\nu(x) + \int_{\mathbb{R}}\int_0^1 kt(1-t)(x-x_0)^2(1+\tilde{\rho}(x))\,\mathrm{d}t\,\mathrm{d}\nu(x)} \\
&= 2\sqrt{\frac{1}{3}\int_{\mathbb{R}} k(x-x_0)^2\,\mathrm{d}\nu(x) + \frac{1}{6}\int_{\mathbb{R}} k(x-x_0)^2\,\mathrm{d}\mu(x)}.
\end{aligned}$$

The only trouble now is to bound the term $\int_{\mathbb{R}}(x-x_0)^2\,\mathrm{d}\mu(x)$ from above in terms of $H(\mu|\nu)$. To do this, we use the following Lemma 6.5.

Lemma 6.5 ([Vil09, Equation 22.7, Equation 22.21] [Led01, Equation 5.13])
We have

$$\int_{\mathbb{R}} k(x-x_0)^2\,\mathrm{d}\mu(x) \leq H(\mu|\nu) + \log\left(\int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x)\right).$$

Proof of Lemma 6.5. We will use the fact that for all $s,t\in\mathbb{R}$ with $s\geq 0$ we have

$$st \leq s\log(s) - s + e^t,$$

which can be obtained from a generalisation of Young's inequality [MN11, Equation 1.3]. Then

$$\begin{aligned}
\int_{\mathbb{R}} k(x-x_0)^2\,\mathrm{d}\mu(x) &= \int_{\mathbb{R}} \rho(x)k(x-x_0)^2\,\mathrm{d}\nu(x) \\
&\leq \int_{\mathbb{R}} \left(\rho(x)\log(\rho(x)) - \rho(x) + e^{k(x-x_0)^2}\right)\mathrm{d}\nu(x) \\
&= H(\mu|\nu) - 1 + \int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x) \\
&\leq H(\mu|\nu) + \int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x).
\end{aligned}$$

□

Returning to the proof of Theorem 6.2, we obtain

$$K_\mu \leq 2\sqrt{\frac{1}{3}\int_{\mathbb{R}} k(x-x_0)^2\,\mathrm{d}\nu(x) + \frac{1}{6}H(\mu|\nu) + \frac{1}{6}\log\left(\int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x)\right)}$$

By Jensen's inequality, as log is concave,

$$\int_{\mathbb{R}} k(x-x_0)^2\,\mathrm{d}\nu(x) = \int_{\mathbb{R}} \log\left(e^{k(x-x_0)^2}\right)\mathrm{d}\nu(x) \geq \log\left(\int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x)\right). \tag{6.1}$$

It follows that

$$K_\mu \leq 2\sqrt{\frac{1}{6}H(\mu|\nu) + \frac{1}{2}\log\left(\int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x)\right)},$$

whence

$$\int_{\mathbb{R}} \sqrt{k}|x-x_0|\,\mathrm{d}|\mu-\nu|(x) \le \sqrt{\frac{2}{3}(H(\mu|\nu))^2 + 2H(\mu|\nu)\log\left(\int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x)\right)}. \tag{6.2}$$

On the other hand, due to the Cauchy-Schwarz inequality and the inequality $|t| \le t+2$ for $t \in [-1,\infty)$, we have

$$\begin{aligned}
&\int_{\mathbb{R}} \sqrt{k}|x-x_0|\,\mathrm{d}|\mu-\nu|(x) \\
&= \int_{\mathbb{R}} \sqrt{k}|x-x_0| \cdot |\tilde{\rho}(x)|\,\mathrm{d}\nu(x) \\
&\le \sqrt{\int_{\mathbb{R}} k(x-x_0)^2|\tilde{\rho}(x)|\,\mathrm{d}\nu(x)}\sqrt{\int_{\mathbb{R}} |\tilde{\rho}(x)|\,\mathrm{d}\nu(x)} \\
&\le \sqrt{\int_{\mathbb{R}} k(x-x_0)^2\rho(x)\,\mathrm{d}\nu(x) + \int_{\mathbb{R}} k(x-x_0)^2\,\mathrm{d}\nu(x)}\sqrt{\int_{\mathbb{R}} \rho(x)\,\mathrm{d}\nu(x) + \int_{\mathbb{R}} \mathrm{d}\nu} \\
&= \sqrt{\int_{\mathbb{R}} k(x-x_0)^2\,\mathrm{d}\mu(x) + \int_{\mathbb{R}} k(x-x_0)^2\,\mathrm{d}\nu(x)}\sqrt{\int_{\mathbb{R}} \mathrm{d}\mu + \int_{\mathbb{R}} \mathrm{d}\nu} \\
&= \sqrt{\int_{\mathbb{R}} k(x-x_0)^2\,\mathrm{d}\mu(x) + \int_{\mathbb{R}} k(x-x_0)^2\,\mathrm{d}\nu(x)} \times \sqrt{2}.
\end{aligned}$$

Applying Lemma 6.5 and Equation (6.1) to the inequality above yields

$$\int_{\mathbb{R}} \sqrt{k}|x-x_0|\,\mathrm{d}|\mu-\nu|(x) \le \sqrt{H(\mu|\nu) + 2\log\left(\int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x)\right)} \times \sqrt{2}. \tag{6.3}$$

Combining Equation (6.2) and Equation (6.3), we obtain

$$\begin{aligned}
&\int_{\mathbb{R}} |x-x_0|\,\mathrm{d}|\mu-\nu|(x) \\
&\le \frac{1}{\sqrt{k}} \min\left\{\sqrt{\frac{2}{3}(H(\mu|\nu))^2 + 2H(\mu|\nu)\log\left(\int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x)\right)},\right. \\
&\qquad \left.\sqrt{2H(\mu|\nu) + 4\log\left(\int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x)\right)}\right\}.
\end{aligned}$$

Recall that $\int_{\mathbb{R}} e^{k(x-x_0)^2}\,\mathrm{d}\nu(x)$ is finite, due to Lemma 6.4.

Finally, for any fixed $a_1, a_2, b_1, b_2 \in \mathbb{R}$ with $a_1, b_1 > 0$, there exists $L > 0$ such that

$$\min\{a_1t^2 + a_2t,\ b_1t + b_2\} \le Lt \quad \text{for all } t \ge 0.$$

Consequently, there exists $K > 0$ such that

$$\int_{\mathbb{R}} |x-x_0|\,\mathrm{d}|\mu-\nu|(x) \le K\sqrt{H(\mu|\nu)}.$$

This, together with Lemma 6.3, completes the proof of Theorem 6.2. □

Bibliography and References

[Aki22] Akira. Gluing lemma in optimal transport. Mathematics Stack Exchange Answer, 2022.
URL: https://math.stackexchange.com/q/4489709.

[Eva24] Josephine Evans. Personal communication, 2024.

[Gar18] David Garling. *Analysis on Polish Spaces and an Introduction to Optimal Transportation*. Cambridge University Press, 2018.

[Led01] Michel Ledoux. *The Concentration of Measure Phenomenon*. Volume 89 of Mathematical Surveys and Monographs. American Mathematical Society, 2001.

[MN11] Flavia-Corina Mitroi and Constantin Niculescu. An extension of Young's inequality. *Abstract and Applied Analysis*, 2011, 2011.

[RR] Svetlozar Rachev and Ludger Rüdenschorf. *Mass Transportation Problems*. Probability and its Applications. Springer New York, NY.

[San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Cham, 2015.

[Vil03] Cédric Villani. *Topics in Optimal Transportation*. Volume 58 of Graduate Studies in Mathematics. American Mathematical Society, 2003.

[Vil09] Cédric Villani. *Optimal Transport: Old and New*. Springer-Verlag Berlin Heidelberg, 2009.