From Here to There: Transport-Entropy Inequalities

a thesis presented by

Emma Lucia Byrnes Finn

to

The Department of Mathematics

in partial fulfillment of the requirements

for the degree of

Bachelor of Arts with Honors

in the subject of

Mathematics and Classics

Advised by
Mark Sellke and Kevin Yang

Harvard College
Cambridge, Massachusetts
March 23, 2026

Contents

0.1 Abstract

This thesis explores classical transport-entropy inequalities, their connection to Log-Sobolev inequalities, and the concentration of measure phenomenon. In particular, it seeks to introduce a reader familiar with analysis and measure-theoretic probability, but without a strong information or optimal transport background to these beautiful and understudied tools. We begin with a historical perspective, addressing the development of the optimal transport problem and the history of entropy. We then develop the theory of transport-entropy inequalities, with a view towards modern applications of these tools to diffusion models and machine learning techniques. Chapter 2 provides a brief introduction to key tools from measure-theoretic probability, defines relative entropy, and briefly introduces the optimal transport problem, focusing on concrete examples in cases where direct computation is possible. Chapter 3 develops the classical theory of transport-entropy inequalities through the lens of T_p inequalities, including concrete examples, showcasing some of the most important results like Pinsker's Inequality and Talagrand's T_2 inequality for Gaussian measures. Chapter 4 develops the machinery needed to state Log-Sobolev inequalities precisely and introduces Log-Sobolev inequalities with three concrete examples. It then connects transport-entropy inequalities to log-Sobolev inequalities through the Otto-Villani theorem and presents a detailed outline of the proof of this major theorem. Finally Chapter 5 concludes the thesis with a short survey of the application of these tools to modern machine learning methods, emphasizing how Log-Sobolev inequalities have helped to develop the theory of diffusion models and explaining how transport-entropy inequalities are relevant to score-based generative modeling as a whole.

0.2 Acknowledgements

First, I owe an enormous debt of gratitude to Mark Sellke and Kevin Yang, both of whom were incredibly generous advisors. This thesis would not have been possible without their support in every stage of the process, from brainstorming topics to offering invaluable edits on my drafts. This thesis would not have been possible without their help and encouragement throughout this process.

I'm also enormously grateful to the many professors in the mathematics and statistics department who have encouraged me to pursue my interests and find beautiful problems. I would never have studied math or statistics if it weren't for Professor Blitzstein's classes. I would never have fallen in love with probability and stochastic processes if it weren't for Professor Sen's class. Professor Harris and Professor Chen taught me to write proofs that were not just correct but elegant and clear. I'm especially grateful to Professor Demba Ba, who encouraged me to begin research and gave me the confidence to pursue the problems that mattered to me. I am also enormously grateful to Andy Keller and Binxu Wang, who have been so generous with their time and help across many research projects and late nights. Their mentorship, encouragement, and help have made all of this possible. I would not be the person or mathematician I am today without the help of these professors and mentors.

Finally, I'm grateful to my friends and family for so much more than I can explain here. To my friends, who were there with me through many late-night work sessions and early-morning bagel runs, I could not have done this without you. To Nico, whose infinite patience, support, and encouragement made this possible, I am enormously thankful for you. To my family, Mom, Dad, Papa, Nani, June, and Hopper, thank you for teaching me to love the world around me, to study it, and to try to do beautiful things.

Chapter 1

Introduction and Motivation

> Why study inequalities? Let us turn this problem on its head and ask instead why we study equalities. After all, pick two random numbers; now, aren't they almost surely unequal?
>
> Joe Blitzstein [7]

If someone handed you two distributions and asked you to compare them, you might first be puzzled not only by the task, but also by the fact that you now appear to be in physical possession of something as intangible as a probability measure. Even setting that aside, without further context, it is not clear what it should mean to say that two distributions are "close" or "different." There are many natural approaches: pick your favorite event and compare how these two measures assign mass to that event, graph the densities of these distributions, compare their moments, or compare the expectations of suitable test functions. Two less obvious but especially useful points of view are the transport viewpoint and the information-theoretic viewpoint. Roughly speaking, the optimal transport approach asks how much effort is required to move one distribution to the other, while the information-theoretic approach measures the discrepancy between them in terms of relative entropy.

This thesis is devoted to the study of inequalities that relate transport distances and relative entropy. I hope that by the time readers reach the end of this thesis, they will feel fully confident in how to respond if handed two distributions.

1.1 Transport: History of the Optimal Transport Problem

It is rare for a mathematical problem to arise from an immediate, life-or-death practical concern. Yet, Gaspard Monge, serving as a scientific advisor to Napoleon, found himself in precisely that position. According to historical accounts, when tasked with designing the most efficient layout for defensive fortifications, Monge confronted what would

become known as the optimal transport problem [30]. The challenge was deceptively simple: when excavating a moat and using the displaced earth to construct a wall, what is the most efficient way to transport the soil from the recently dug moat to the wall [39]?

This question opens a series of profound mathematical inquiries: what mathematical structures best represent the distributions of earth in the moat and wall? By what criterion should one transport plan be judged superior to another? Under what conditions is the optimal plan unique? And can the geometric or probabilistic properties of these distributions be used to bound the total effort required for transport?

This formulation was eventually made precise. A much fuller discussion of Optimal Transport is reserved for Section 2.3.

Definition 1.1.1. For a cost function $c : \mathscr{X} \times \mathscr{X} \to [0,\infty)$ and two distributions $\mu, \nu \in \mathscr{P}(\mathscr{X})$, the Kantorovich optimal transport problem is to find the coupling $(X, Y) \sim \pi$ with marginals $X \sim \mu$ and $Y \sim \nu$ that minimizes

$$\mathbb{E}_{(X,Y)\sim\pi}\big[c(X,Y)\big].$$

Let $\Pi(\mu,\nu)$ denote the set of all such couplings. This is a convex minimization problem, and its optimal value is denoted

$$\mathscr{T}_c(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\big[c(X,Y)\big].$$

1.2 Entropy: History of Entropy

Nearly two hundred years later, Claude Shannon developed his statistical theory of entropy, thankfully under less dire circumstances. Shannon was working at Bell Labs, trying to understand how much information was lost in the course of a phone call. Designed to measure the reduction of the amplitude of a signal through a telephone line, Shannon proposed this new quantity as a measure of surprise.

Relative entropy (also called KL-divergence) is the quantity most of interest for comparison to transport distance.

Definition 1.2.1. For two probability measures $\mu, \nu \in \mathscr{P}(\mathscr{X})$ with μ absolutely continuous with respect to ν (written $\mu \ll \nu$), the relative entropy, also called Kullback–Leibler divergence of μ from ν is

$$D_{\mathrm{KL}}(\mu\|\nu) = \mathbb{E}_{X\sim\mu}\left[\log\frac{d\mu}{d\nu}(X)\right],$$

where $\frac{d\mu}{d\nu}$ is the Radon–Nikodym derivative of μ with respect to ν. If μ is not absolutely continuous with respect to ν, we set $D_{\mathrm{KL}}(\mu\|\nu) = +\infty$.

1.3 Transport-Entropy Inequalities

This thesis will explore the relationship between these two ideas. Heuristically, a transport-entropy inequality says that distributions that are "information-wise" close (ie those which have low relative entropy) are also "geometrically" close (low transport cost). If one distribution has low relative entropy with respect to another, then it must also be close in transport cost.

Definition 1.3.1. Let $\mu \in \mathscr{P}(\mathscr{X})$ be a reference distribution and let $c : \mathscr{X} \times \mathscr{X} \to [0,\infty)$ be a cost function. We say that μ satisfies a transport-entropy inequality if there exists a non-decreasing function $\alpha : [0,\infty) \to [0,\infty]$ with $\alpha(0) = 0$ such that for all $\nu \in \mathscr{P}(\mathscr{X})$,

$$\alpha\big(\mathscr{T}_c(\mu,\nu)\big) \leq D_{\mathrm{KL}}(\nu\|\mu).$$

1.4 Motivation and Outline

This thesis will aim to develop the theory of transport-entropy inequalities from basic principles. Chapter 2 provides a brief introduction to measure-theoretic probability, relative entropy, and optimal transport. Chapter 3 develops the classical theory of transport-entropy inequalities through the lens of T_p inequalities, including concrete examples, showcasing some of the most important results like Pinsker's Inequality and Talagrand's T_2 inequality for Gaussian measures, closing with a brief discussion of the relationship between concentration inequalities and transport-entropy. Chapter 4 develops the machinery needed to state Log-Sobolev inequalities precisely and introduces Log-Sobolev inequalities with concrete examples. It then connects transport-entropy inequalities to log-Sobolev inequalities through the Otto-Villani theorem and presents a sketch of the proof of this major theorem. Finally Chapter 5 concludes the thesis with a short survey of the application of these tools to modern machine learning methods, emphasizing how log-Sobolev inequalities have helped to develop the theory of diffusion models and explaining how transport-entropy inequalities are relevant to score-based generative modeling as a whole.

Chapter 2

Entropy and Optimal Transport Basics

This thesis aims to be a relatively self-contained introduction to the application of transport-entropy inequalities to diffusion models. In that spirit, this chapter introduces the key tools from information theory, probability, and optimal transport that will be used to develop results and intuition. First, in Section 2.1, we'll review some definitions and theorems in probability that will be used repeatedly in this thesis. Next, in Section 2.2, we'll provide an introduction to discrete and differential entropy, eventually building to a definition of relative entropy, which will be used constantly in Chapter 3 and generalized entropy and its tensorization properties, which will be used in Chapter 4. Finally, in Section 2.3, we'll introduce readers to the basic theory of optimal transport focusing on concrete examples, which will be very relevant for the material developed in Chapter 3.

2.1 Probability Background

We'll begin by reviewing the probability theory which developed in the years between Gaspard Monge and Claude Shannon, which made a formal statement of Monge's ideas possible. In particular, Leonid Kantorovich, a Soviet mathematician and economist roughly contemporaneous with Claude Shannon made progress in understanding the optimal transport problem in terms of so-called coupled measures [39]. But before we get there, we'll formulate the problem as it is stated in most modern treatments and cover a few key definitions.

Definition 2.1.1 (Complete)**.** A metric space (X, d) is defined to be complete if every Cauchy sequence is convergent.

Definition 2.1.2 (Separable)**.** A metric space (X, d) is defined to be separable if there exists a countable set $S \subseteq X$ which is dense in (X, d).

Definition 2.1.3 (Polish)**.** A metric space $\mathscr{X}$ is called a Polish space if it is complete and separable.

An example of a Polish space is $\mathbb{R}$ or any L^p space for $p < \infty$. A non-example of a Polish space would be $\mathbb{Q}$ since it is not complete.

Remark 1. We go to the trouble to define and explain some basic properties of Polish spaces here, because Polish spaces can be thought of as the minimum viable setting for such transport problems to be well-posed. In Chapters 2 and 3, we primarily work over general Polish spaces. In Chapter 4, we restrict ourselves to $\mathbb{R}^n$ because some of the results stated do not hold for a fully general Polish space or require substantially different approach to show.

In the context of transport-entropy inequalities, we work on a Polish space $\mathscr{X}$, which we equip with $B(\mathscr{X})$, the σ-algebra generated by the open sets in $\mathscr{X}$, which is called the Borel σ-algebra. We write $P(\mathscr{X})$ for the set of Borel probability measures on $\mathscr{X}$ [39].

We also briefly review these definitions and a few others here.

Definition 2.1.4 (σ Algebra)**.** A collection of subsets of Ω, $\mathscr{F}$, is called a σ algebra if it contains the empty set, is closed under complements, and is closed under countable unions.

Definition 2.1.5 (Borel$-\sigma$ Algebra)**.** For a topological space (X,τ), the Borel σ algebra over X is the σ-algebra generated by all the open sets in X.

Definition 2.1.6 (Measure)**.** A map $\mu : \mathscr{F} \to [0,\infty]$ is called a measure if it satisfies:

(a) $\mu(\varnothing) = 0$,

(b) $\mu(A) \geq 0$ for all $A \in \mathscr{F}$,

(c) (Countable Additivity) for any pairwise disjoint sequence $(A_n)_{n=1}^{\infty} \subset \mathscr{F}$,

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

We might be interested in how to approximate the measure of a set $A \subseteq X$ under μ. The notion of 'inner regularity' gives us a tool to do this.

Definition 2.1.7 (Inner Regular)**.** [1] Let $(\mathscr{X}, T)$ be a topological space and $\mathscr{B}_{\mathscr{X}}$ be the Borel σ-algebra on X and μ be a measure on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$. Then we say that μ is inner regular if for all measurable subsets $A \subset \mathscr{X}$,

$$\mu(A) = \sup\{\mu(F) : F \subseteq A \text{ where F is compact and measurable}\}.$$

Before we arrive at the final piece of machinery presented in this introduction, we need one more definition.

Definition 2.1.8 (Absolute Continuity)**.** [5] Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. Let ν be another measure on that space. We say that ν is absolutely continuous with respect to μ if for all $A \in \mathscr{F}$, we have that $\mu(A) = 0 \implies \nu(A) = 0$.

This gives rise to another useful idea: the equivalence of two measures.

Definition 2.1.9 (Equivalence of Measures)**.** [5] We say that two measures are equivalent if $\mu(A) = 0 \iff \nu(A) = 0$, that is if $\nu \ll \mu$ and $\mu \gg \nu$.

Intuitively, this corresponds to the same notion of 'impossibility.' Thus, if an event happens $\nu-$almost surely then it also happens $\mu-$almost surely.

Now, we arrive at a very useful theorem which ensures the existence of a crucial quantity: the Radon-Nikodym derivative. Intuitively, we know that we can build new measures ν on the metric space $(\Omega, \mathscr{F}, \mu)$ by playing the following game: choose a non-negative function f and define a measure ν as follows. For all $A \in \mathscr{F}$, let $\nu(A) := \int_A f \, d\mu$. Then, conveniently, for any A such that $\mu(A) = 0$, we must also have $\nu(A) = \int_A f \, d\mu = 0$.

The natural question, then, is can we go the other way? If I have a measure ν on the metric space $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$, can I always find a non-negative function f such that $\nu(A) = \int_A f \, d\mu$?

The Radon Nikodym theorem says that under extremely mild conditions, this is possible.

Theorem 2.1.10. *In particular, if μ and ν are $\sigma-$finite and ν is absolutely continuous with respect to μ, then there exists a $\mu-$unique non-negative function f such that $\nu(A) = \int_A f \, d\mu$ for all $A \in \mathscr{B}(\mathscr{X})$.*

Definition 2.1.11 (Radon-Nikodym Derivative)**.** [5] The function f such that $\nu(A) = \int_A f \, d\mu$ is called the Radon-Nikodym derivative of ν with respect to μ and is usually written $\frac{d\nu}{d\mu}$.

Theorem 2.1.12 (Portmanteau Theorem)**.** *[6] Let $\mathscr{X}$ be a metric space with Borel-σ algebra $\mathscr{F}$. A sequence of probability measures $\{\mu_n\}_{n\in\mathbb{N}}$ converges in distribution to μ, if any of the following equivalent conditions hold*

(a) $\mathbb{E}_{\mu_n}[f] \to \mathbb{E}_{\mu}[f]$ *for all bounded, continuous f*

(b) $\limsup_{n\to\infty} \mu_n(F) \leq \mu(F)$ *for all measurable closed sets $F \in \mathscr{F}$*

(c) $\liminf_{n\to\infty} \mu_n(U) \geq \mu(U)$ *for all measurable open subsets $U \in \mathscr{F}$.*

(d) $\limsup_{n\to\infty} \mathbb{E}_{\mu_n}[f] \leq \mathbb{E}_{\mu}[f]$ *for all upper semi-continuous functions bounded above*

(e) $\liminf_{n\to\infty} \mathbb{E}_{\mu_n}[f] \geq \mathbb{E}_{\mu}[f]$ *for all lower semi-continuous functions bounded below.*

Remark 2. There are, of course, more equivalent conditions, but in this presentation we've chosen to only include the conditions which are relevant to later chapters.

Because this is a standard theorem and proved in [6], we will omit the proof here.

2.2 Entropy Background

2.2.1 History of Entropy

Entropy is a term with origins in fields as disparate as thermodynamics and information theory. It originally appeared in Boltzmann's work on statistical mechanics but rose to prominence when Claude Shannon rediscovered it in the context of information theory [25]. Intuitively, if you learn that a low probability event has occurred, you should be surprised. This is the idea that "surprisal" and "entropy" formalize.

As it turns out, in the discrete case, entropy behaves well and basically accords with our intuition. For that reason, this section is split into three parts: first, we will introduce discrete entropy, then we'll explain how these ideas extend (sometimes fail to) to the case of continuous random variables, and finally we'll offer a fully precise measure-theoretic definition which encompasses both perspectives.

2.2.2 Discrete Entropy

Definition 2.2.1 (Surprise)**.** [34] Let A be an event with probability p. The surprise of A $:= \log_2(\frac{1}{p})$ when measured in bits or $\ln(\frac{1}{p})$ when measured in nats.

Remark 3. In this thesis we'll always work in nats, and as is common in probability, denote ln as log.

For a discrete random variable, the entropy is simply the expected surprise over all possible values.

Example 1. This should make sense in the context of a six-sided die, for example. If the die is loaded and always rolls a two with probability 1, we should not be surprised at all when, inevitably, a two is rolled. The surprise is 0 and the entropy is also 0. However, on a standard six-sided die with equal probability of rolling the values one through six, it is relatively surprising to roll a two, but it is also surprising to roll a three or a one. Entropy should capture that.

Definition 2.2.2 (Discrete Entropy [12])**.** Let X be a random variable taking states $x \in \mathscr{X}$ with probability mass function $p : \mathscr{X} \to [0,1]$. Then we call the entropy of X $H(X) := \sum_{x\in\mathscr{X}} p(x)\log(\frac{1}{p(x)})$

Example 2. Returning to the previous example, consider three random variables D_1 representing the value rolled by the deterministic always two die, D_2 representing the value rolled by a standard, fair six sided die, and D_3 a weighted die that rolls a one with probability 1/2 and a six with probability 1/2. Let's compute the entropy of each die.

$$H(D_1) = \sum_{i=1}^{6} p(x_i)\log(\frac{1}{p(x_i)}) = 1\cdot\log(\frac{1}{1}) = 1\cdot 0 = 0$$

$$H(D_2) = \sum_{i=1}^{6} p(x_i)\log(\frac{1}{p(x_i)}) = \sum_{i=1}^{6} \frac{1}{6}\log(\frac{1}{(1/6)}) = \log(6)$$

$$H(D_3) = \sum_{i=1}^{6} p(x_i)\log(\frac{1}{p(x_i)}) = \frac{1}{2}\log(\frac{1}{1/2}) + \frac{1}{2}\log(\frac{1}{1/2}) = \log(2)$$

Intuitively, this makes sense. The deterministic D_1 has zero entropy, since we're never surprised to see that it rolls a two. On the other hand, any outcome we see on die D_2 is relatively surprising, since there are six equi-probable options. Finally, the "goldilocks" die D_3 has a little more variability than the deterministic die, but less than the fair die. This is accurately represented by the ordering of their entropies

$$H(D_1) = 0 \leq H(D_3) = \log(2) \leq H(D_2) = \log(6)$$

2.2.3 Differential Entropy

This notion is extended to continuous distributions in the natural way, by computing the expectation with respect to the density of the random variable.

Definition 2.2.3 (Differential Entropy [12])**.** For a random variable Y with density $f(y)$, we define the entropy of Y as $h(Y) = \mathbb{E}[\log(\frac{1}{f(Y)})] = \int_{-\infty}^{\infty} f(y)\log(\frac{1}{f(y)})dy$

However, there are a few subtleties to keep in mind. Differential entropy has some relatively counter-intuitive quirks as compared to discrete entropy. For one, differential entropy can be negative! Consider the following simple example.

Example 3. Let $X \sim \text{Unif}[0,\theta]$. Let's compute the (differential) entropy of X.

$$h(X) = \int_{\text{supp}(X)} \frac{1}{\theta}\log(\frac{1}{(1/\theta)})dx = \int_0^{\theta} \frac{1}{\theta}\log(\theta)dx = \log(\theta)$$

Notice that for $\theta \in (0,1)$ $h(X) < 0$ while for $\theta > 1$ $h(X) > 0$. This suggests some important intuition: sufficiently "concentrated" continuous distributions have negative entropy. However, it demonstrates that differential entropy is no longer measuring simple "surprisal," but instead something more complicated.

To see what our intuition for differential entropy should be, consider the following comparison of the entropy of a scaled discrete random variable to the entropy of a scaled continuous random variable.

Example 4. [12] Let X be a discrete random variable with entropy $H(X)$. Suppose we want to compute the entropy of $a \cdot X$ where $a \in \mathbb{R} \setminus 0$ is some non-zero real scalar. Intuitively, changing the units of X (converting from centimeters to meters, for example) should not change how surprised we are, intuitively. This is borne out by reality in the discrete case:

$$H(aX) = \sum_{\text{supp}(aX)} p(ax_i)\log(\frac{1}{p(ax_i)}) = \sum_{\text{supp}(X)} P(aX = ax_i)\log(\frac{1}{P(aX = ax_i)})$$

$$= \sum_{\text{supp}(X)} p(X = x_i) \log(\frac{1}{P(X = x_i)}) = H(X)$$

This is because rescaling is a bijection, so it is obvious that for $a \neq 0$, $P(aX = ax_i) = P(X = x_i)$, which is the key fact which allows us to move from the first equality to the second.

Now consider the continuous case. Let Y be a continuous random variable with differential entropy $h(Y)$. Now consider the entropy of $Z = aY$:

$$h(Z) = -\int_{\mathbb{R}} f_Z(z) \log(f_Z(z)) dz$$

Then, by the change of variables formula, the fact that the Jacobian of $g(z) = az$ is $\frac{dy}{dz} = |\frac{1}{|a|}|$, and u-substituting we find

$$\begin{aligned} h(Z) = h(aY) &= -\int_{\mathbb{R}} f_{aY}(x) \log f_{aY}(x)\, dx \\ &= -\int_{\mathbb{R}} \frac{1}{|a|} f_Y\left(\frac{x}{a}\right) \log\left(\frac{1}{|a|} f_Y\left(\frac{x}{a}\right)\right) dx \\ &\overset{u=x/a}{=} -\int_{\mathbb{R}} f_Y(u) \log\left(\frac{1}{|a|} f_Y(u)\right) du \\ &= -\int_{\mathbb{R}} f_Y(u) \log f_Y(u)\, du + \log|a| \int_{\mathbb{R}} f_Y(u)\, du \\ &= h(Y) + \log|a|. \end{aligned}$$

Thus, in the differential entropy context, changing units also changes entropy. This is surprising unless we think of entropy as a measure of surprise relative to some reference distribution or measure. This motivates the more general notion of entropy, which we will employ from here on out.

2.2.4 Relative and Generalized Entropy

As we've seen above, differential entropy and discrete entropy are useful. However, they are unified under a more general framework: relative entropy, also called $KL-$divergence. Intuitively, KL-divergence measures some notion of how far apart a distribution p is from a distribution q.

Definition 2.2.4 (Kullback-Leibler Divergence, also called Relative Entropy)**.** [1]
If μ is absolutely continuous with respect to ν, the KL-Divergence is

$$D_{KL}(\mu \,\|\, \nu) = \int_{\mathscr{X}} \log\left(\frac{d\mu}{d\nu}(x)\right) d\mu(x),$$

where $\frac{d\mu}{d\nu}$ denotes the Radon–Nikodym derivative of μ with respect to ν.

[1] In this thesis, we'll use the terms KL-Divergence and relative entropy interchangeably.

Here, we use the convention mentioned in [12] that $0\log(\frac{0}{0}) = 0$, $0\log(\frac{0}{q}) = 0$, and $q\log(\frac{q}{0}) = \infty$.

Note that KL-Divergence is not a true metric because it is not symmetric and does not satisfy the triangle inequality [7]. It is however always non-negative, which can be shown relatively easily with Jensen's inequality.

In practice, this quantity is often difficult to compute. Except for densities which follow relatively nice forms (in particular, densities which follow a known exponential family), the KL-Divergence is not usually available in closed form.

Example 5. In the case where p, q are both Gaussian measures, however, we can compute this quantity. Let p be the measure corresponding to the distribution $\mathcal{N}(\mu_p, \sigma_p^2)$ and q be the measure corresponding to the distribution $\mathcal{N}(\mu_q, \sigma_q^2)$. Then,

$$D_{KL}(p||q) = \mathbb{E}_{X\sim p}[\log\left(\frac{p(X)}{q(X)}\right)] = \mathbb{E}_{X\sim p}[\log p(X) - \log q(X)] = \mathbb{E}_{X\sim p}[\log p(X)] - \mathbb{E}_{X\sim p}[\log q(X)]$$

Then, substituting in the normal densities, we find

$$D_{KL}(p||q) = -\frac{1}{2}\log(\sigma_p^2) - \frac{1}{2}\log(\sigma_q^2) - \mathbb{E}_{X\sim p}\left[\frac{(X-\mu_p)^2}{2\sigma_p^2}\right] + \mathbb{E}_{X\sim p}\left[\frac{(X-\mu_q)^2}{2\sigma_q^2}\right]$$

Simplifying and using the definition of variance, we find

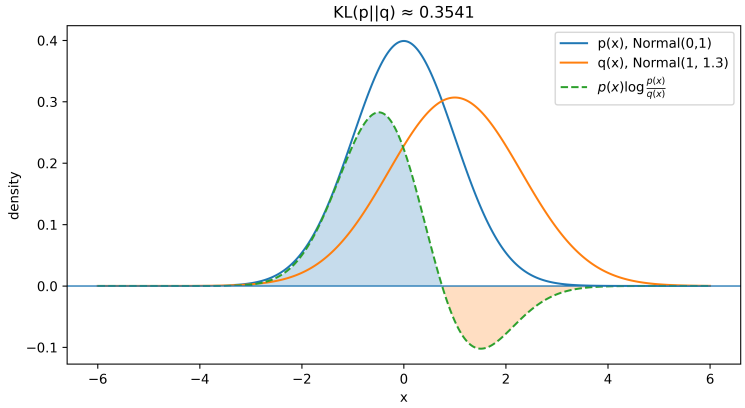
$$D_{KL}(p||q) = -\frac{1}{2}\log\left(\frac{\sigma_p^2}{\sigma_q^2}\right) - \mathbb{E}_{X\sim p}\left[\frac{(X-\mu_p)^2}{2\sigma_p^2}\right] + \mathbb{E}_{X\sim p}\left[\frac{(X-\mu_q)^2}{2\sigma_q^2}\right]$$

$$D_{KL}(p||q) = \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2} + \frac{1}{2\sigma_q^2}\mathbb{E}_{X\sim p}\left[(X-\mu_p)^2 + 2(\mu_p-\mu_q)(X-\mu_p) + (\mu_p-\mu_q)^2\right]$$

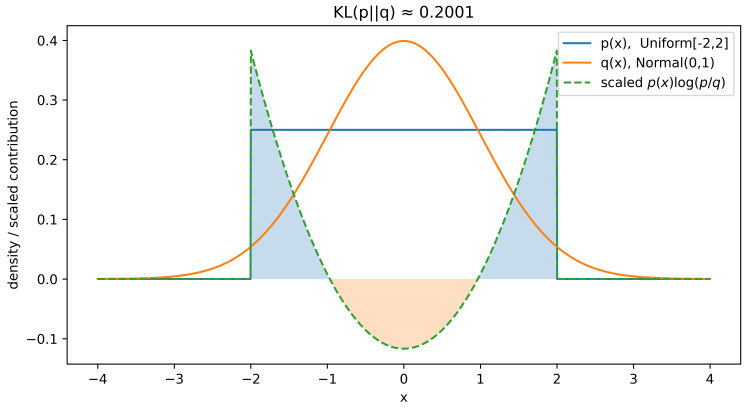
$$D_{KL}(p||q) = \log\left(\frac{\sigma_q}{\sigma_p}\right) + \frac{\sigma_p^2 + (\mu_p-\mu_q)^2}{2\sigma_q^2} - \frac{1}{2}$$

2.2.5 Generalized Entropy

Now, we only need one more conception of entropy to complete our discussion. Though we've already defined entropy above, we need to update our definition slightly to accommodate the a context we will encounter in Chapter 4. In particular, as we've described above in Example 4, differential entropy is not scale-invariant. Generalized-entropy is also not scale-invariant, but has been re-normalized such that $\text{Ent}_\rho(c\cdot f) = c\text{Ent}_\rho(f) \quad \forall c \geq 0$.



(a) Gaussian vs Gaussian



(b) Gaussian vs Uniform

Figure 2.1: Examples of KL-divergence behavior under different distribution pairs.

Definition 2.2.5 (Generalized Entropy [18])**.** Let f be any non-negative function on $\mathscr{X}$ and μ be some measure on $\mathscr{X}$. We define

$$\text{Ent}_\mu(f) := \int_{\mathscr{X}} f\log(f)d\mu - \left(\int_{\mathscr{X}} f\,d\mu\right)\cdot\log\left(\int_{\mathscr{X}} f\,d\mu\right)$$

Remark 4. Note as in the rest of Section 2.2, we use the convention that $0\log 0 = 0$.

Remark 5 (Connection Between Generalized Entropy and Relative Entropy). Notice that if f is already normalized such that $\int_{\mathscr{X}} f\,d\mu = 1$, then it is a probability density. In that case, we have that

$$\text{Ent}_\mu(f) = \int f\log f\,d\mu = H(f\mu|\mu)$$

Finally, we will state and prove two lemmas that will be extremely useful later. First, we'll offer a second way to express this generalized entropy as a supremum over a set of functions, and second we'll describe how this entropy behaves on product measures.

Lemma 2.2.6 (Generalized Entropy is a Supremum, [22])**.**

$$Ent_\mu(f) = \sup_{g:\mathbb{E}[e^g]\le 1} \mathbb{E}[fg]$$

where expectations are with respect to μ*.*

The following proof expands the presentation given in [22], where the proof of this lemma is compressed into the proof of Proposition 2.2.

Proof. As usual, in order to show equality, we'll show both directions of the inequality. First, recall that

$$\text{Ent}_\mu(f) = \int_{\mathscr{X}} f\log(f)d\mu - \left(\int_{\mathscr{X}} f\,d\mu\right)\cdot\log\left(\int_{\mathscr{X}} f\,d\mu\right)$$

Then, consider the normalized version of f, which we'll define as

$$h = \frac{f}{\int f\,d\mu}$$

Then, we have

$$\text{Ent}_\mu(f) = \int f\log\frac{f}{\int f\,d\mu}d\mu = \left(\int f\,d\mu\right)\int h\log h\,d\mu$$

Then, by Young's Inequality (see [22], page 25), we have that

$$a\cdot b \le a\log a - a + e^b \quad \text{for } a\ge 0, b\in\mathbb{R}$$

Then,

$$h\cdot g \le h\log h - h + e^g \quad \text{for } g \text{ any function } \mathbb{R}\to\mathbb{R}$$

Then, integrating both sides, we have

$$\int hg\,d\mu \leq \int h\log h - h + e^g\,d\mu = \int h\log h\,d\mu - 1 + \int e^g\,d\mu$$

Re-normalizing to recover f, we multiply by $\int f\,d\mu$ on both sides to find

$$\int fg\,d\mu \leq \left(\int f\,d\mu\right)\int h\log h\,d\mu$$

Then, we have

$$\mathbb{E}[fg] \leq \mathrm{Ent}_\mu(f)$$

Then, noticing that this inequality holds for arbitrary g as defined above, we may take the supremum on both sides to conclude

$$\sup_{g:\mathbb{E}[e^g]\leq 1} \mathbb{E}[fg] \leq \mathrm{Ent}_\mu(f)$$

Then, for the other direction, we observe that we can take

$$g^* = \log\frac{f}{\int f\,d\mu}$$

since

$$\int e^{g^*}\,d\mu = 1$$

Then, consider

$$\int fg^*\,d\mu = \int f\log\frac{f}{\int f\,d\mu}\,d\mu = \mathrm{Ent}_\mu(f)$$

which finishes the other direction of the inequality and therefore completes the proof. □

The next lemma is very useful in Chapter 4.

Lemma 2.2.7 (Tensorization of Entropy, Proposition 2.2 in [22])**.** *Let $(\mathcal{X}_i,\mathcal{F}_i,\mu_i)$, $i = 1,\ldots,n$, be probability spaces, define $\mathcal{X} := \mathcal{X}_1\times\cdots\times\mathcal{X}_n$ and $P := \mu_1\otimes\cdots\otimes\mu_n$. Let $f : \mathcal{X}_1\times\cdots\times\mathcal{X}_n \to \mathbb{R}$ be measurable and non-negative. Define*

$$f_i(x_i) = f(x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)$$

where we treat $(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$ as fixed. Then, we have that for any such f, the following inequality holds:

$$Ent_P(f) \leq \sum_{i=1}^n \mathbb{E}[Ent_{\mu_i}(f_i)].$$

As above, the proof of this lemma expands the treatment given by [22]in Proposition 2.2.

Proof of Lemma 2.2.7. First, consider a function $g : \mathscr{X} \to \mathbb{R}$ which satisfies $\int e^g dP \leq 1$. We want to understand what the ith coordinate contributes when we decompose g. In particular, consider

$$G_i(x_i, \dots, x_n) := \int_{\mathscr{X}} e^{g(x_1,\dots,x_n)} d\mu_1(x_1) \dots d\mu_{i-1}(x_{i-1})$$

and

$$G_{i+1}(x_{i+1}, \dots, x_n) := \int_{\mathscr{X}} e^{g(x_1,\dots,x_n)} d\mu_1(x_1) \dots d\mu_i(x_i)$$

Then, define

$$g^i(x_1, \dots, x_n) = \log(G_i / G_{i+1})$$

Then, g^i can be thought of the additional log-weight gained by observing x_i once you've already seen $x_{i+1}, x_{i+2}, \dots x_n$. Concretely, we can think of

$$g^i(X_i, \dots, X_n) = \log(\mathbb{E}[e^g | X_i, \dots X_n]) - \log(\mathbb{E}[e^g | X_{i+1}, \dots, X_n])$$

which can be thought of as the 'reverse martingale increment.'
Now, notice that g^i satisfies our requirement from Lemma 2.2.6,

$$\int_{\mathscr{X}} e^{g^i(x_i,\dots,x_n)} d\mu_i \leq 1$$

since, for each fixed $(x_{i+1}, \dots, x_n)$,

$$\int_{\mathscr{X}_i} e^{g^i(x_i,\dots,x_n)} \, d\mu_i(x_i) = \frac{1}{G_{i+1}(x_{i+1}, \dots, x_n)} \int_{\mathscr{X}_i} G_i(x_i, \dots, x_n) \, d\mu_i(x_i) = 1.$$

Then, we can also notice that

$$g(x_1, \dots x_n) \leq \sum_{i=1}^n g^i(x_1, \dots, x_n) \quad \text{for all } (x_1, \dots, x_n) \in \mathscr{X}_1 \times \mathscr{X}_2 \times \dots \times \mathscr{X}_n$$

since the sum telescopes as follows:

$$\sum_{i=1}^n g^i = \sum_{i=1}^n (\log(G_i) - \log(G_{i+1})) = \log(G_1) - \log(G_{n+1}) = g - \log(\int e^g dP) \geq g.$$

The final inequality holds because we have by hypothesis that $\int e^g dP \leq 1 \implies \log(\int e^g dP) \leq 0$.
Now, notice that the above gives immediately that

$$\mathbb{E}_P[fg] \leq \sum_{i=1}^n \mathbb{E}_P[fg^i]$$

Now, we'll look at the LHS of this inequality. Consider one term in the sum. Then, rewriting the expectation as an integral and interchanging the inner and outer integral by Fubini's theorem, we find

$$E_P[fg^i] = \mathbb{E}_{\mu_{-i}}\left[\int_{X_i} f_i(t)(g^i)_i(t)d\mu_i(t)\right]$$

Where μ_{-i} means we integrate over all coordinates which are not i.
Therefore,

$$\mathbb{E}[fg] \leq \sum_{i=1}^{n} \mathbb{E}_P[fg^i] = \sum_{i=1}^{n} \mathbb{E}_{\mu_{-i}}\left[\int_{X_i} f_i(t)(g^i)_i(t)d\mu_i(t)\right] \leq \sum_{i=1}^{n} \mathbb{E}_P[\mathrm{Ent}_{\mu_i}(f_i)]$$

Thus, we can take the supremum on both sides over all admissible test functions g and finally apply the result of Lemma 2.2.6 Therefore, we have

$$\mathrm{Ent}_P(f) = \sup_{g:\mathbb{E}[e^g]\leq 1} \mathbb{E}[fg] \leq \sum_{i=1}^{n} \mathbb{E}_P[\mathrm{Ent}_{\mu_i}(f_i)]$$

which completes the proof. □

2.3 Optimal Transport Background

Just like Gaspard Monge, our first task is to make precise what we mean by an "efficient" transport plan. There are two ways to frame the optimal transport problem: one is due to Gaspard Monge and the other is due to Kantorovich. The latter is usually seen as a generalization of Monge's approach [39].

2.3.1 Transport Plans

This begins with the definition of a transport map, which requires a review of the notion of a "push forward measure."

Definition 2.3.1 (Pushforward)**.** Let $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$ and $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$ be two measurable spaces. Let μ be a measure on $\mathscr{X}$. Further, let $T : \mathscr{X} \to \mathscr{Y}$ be a measurable function. We say that $T_{\#}\mu$ the push forward measure of μ by T on $\mathscr{Y}$ is defined by

$$(T_{\#}\mu)(B) := \mu(T^{-1}(B)) \quad \forall B \in \mathscr{B}(\mathscr{Y})$$

Then, we can define the transport map.

Definition 2.3.2 (Transport Map)**.** Let $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$ and $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$ be two measurable spaces. Let μ be a measure on $\mathscr{X}$ and ν be a measure on $\mathscr{Y}$. We say that $T : X \to Y$ is a transport map if $\nu = T_{\#}\mu$.

This idea is closely related to how we think about couplings of measures μ and ν.

Definition 2.3.3 (Coupling)**.** [41] Let $(\mathscr{X},\mu)$ and $(\mathscr{Y},\nu)$ be two probability spaces with μ and ν two probability measures. We say that $\pi \in \mathscr{P}(\mathscr{X}\times\mathscr{Y})$ is a coupling of μ and ν if it has the property that $\pi(A\times\mathscr{Y}) = \mu(A)$ and $\pi(\mathscr{X}\times B) = \nu(B)$. The set of all such couplings of μ and ν is denoted $\Pi(\mu,\nu)$.

Example 6. The simplest example of a coupling is the 'product' measure, $\mu\otimes\nu$. Let $(\mathscr{X},\mu)$ and $(\mathscr{Y},\nu)$ be two probability spaces with μ and ν two probability measures as in the definition. Let $\pi := \mu\otimes\nu$ be a measure on the space $\mathscr{X}\times\mathscr{Y}, \mathscr{B}(\mathscr{X})\otimes\mathscr{B}(\mathscr{Y})$ defined by $\pi(A\times B) = \mu(A)\nu(B)$. We can easily verify that this satisfies our definition. Namely $\pi(A\times\mathscr{Y}) = \mu(A)\nu(\mathscr{Y}) = \mu(A)$ as desired, since the measure of the whole space is always 1. Similarly $\pi(\mathscr{X}\times B) = \mu(\mathscr{X})\nu(B) = \nu(B)$. Conveniently, this formulation assures us that for any two measures (μ,ν), the set of all their couplings $\Pi(\mu,\nu)$ is never empty.

Example 7. For concreteness, consider the set of couplings of $\mu = \nu = \text{Bernoulli}(1/2)$. We know at least that the independent coupling exists from the example above. Since $\mu = \nu$ there exists a diagonal coupling, ie $\pi(\{(x,y)\in\mathscr{X}\times\mathscr{X} : x = y\}) = 1$. We also have the perfect correlation case: $X = Y \implies P(X = Y) = 1$. A natural question is what are all such couplings. Because Bernoulli random variables are simple, we can give a complete description. First notice that the coupling is fully determined by the set of joint probabilities $p_{ij} := P(X = i, Y = j)$ for $i,j\in\{0,1\}$ where $p_{ij}\geq 0$, $\sum_{i,j} p_{ij} = 1$ and we have the following marginal constraints to ensure that $\mu(A) = \text{Bern}(1/2)$ and $\nu(A) = \text{Bern}(1/2)$, namely that the row and column sums are all $1/2$ in the following table.

	$Y=0$	$Y=1$	Row sum
$X=0$	p_{00}	p_{01}	$\frac{1}{2}$
$X=1$	p_{10}	p_{11}	$\frac{1}{2}$
Col sum	$\frac{1}{2}$	$\frac{1}{2}$	

We can solve this system to express the set of all couplings of μ,ν. These marginal constraints imply the system is uniquely determined by a single parameter $\theta = p_{00}$. Solving in terms of θ, we find $p_{10} = \frac{1}{2}-\theta$ by the column constraint, $p_{01} = \frac{1}{2}-\theta$ by the row constraint, and using the fact that $\sum_{i,j} p_{ij} = 1$, we find $p_{11} = \theta$. Thus, any coupling of these two measures μ and ν can be expressed (for $0\leq\theta\leq\frac{1}{2}$, to ensure we have non-negative probabilities) as

$$P_\theta = \begin{bmatrix} \theta & \frac{1}{2}-\theta \\ \frac{1}{2}-\theta & \theta \end{bmatrix}$$

In particular this means the set of couplings $\Pi(\mu,\nu) = \{P_\theta : \theta\in[0,\frac{1}{2}]\}$ is a 1-dimensional convex set, which we can think of as a line segment in this context.

This example also builds the useful intuition that for discrete marginal distributions μ (which assigns positive mass to m values, that is, has support size m) and ν (which assigns positive mass to n values, that is, has support size n) the set of all possible couplings, $\Pi(\mu,\nu)$, is a convex set with dimension $(m-1)\cdot(n-1)$ [23].

The idea of a "push forward measure" also gives rise to a natural coupling. In particular, if $\nu = T_{\#}(\mu)$ is a push forward, then the coupling $\pi := (Id, T)_{\#}\mu$ is a coupling of μ and ν. Unfortunately, not every coupling arises from such a map T. Consider the following counterexample.

Example 8 (Sometimes, there exist couplings between measures, but no transport map!). Consider two discrete measures, μ and ν where $\mu = \delta_x$ is a Dirac mass at x and $\nu = \frac{2}{3}\delta_{y_1} + \frac{1}{6}\delta_{y_2} + \frac{1}{6}\delta_{y_3}$, where $y_1 \neq y_2 \neq y_3$. Then $\nu(y_1) = \frac{2}{3}$, $\nu(y_2) = \frac{1}{6}$, and $\nu(y_3) = \frac{1}{6}$, but $\mu(T^{-1}(y_i)) \in 0, 1 \forall i \in \{1,2,3\}$, since μ must be zero unless $x \in T^{-1}(y_1) \cup T^{-1}(y_2) \cup T^{-1}(y_3)$. The key structural problem here is that transport maps can't split mass from a single atom using a deterministic map, so any push forward measure $T_{\#}\mu = \delta_{T(x)}$ must also be a Dirac mass.

This asymmetry will be relevant when comparing the Monge and Kantorovich formulations of the optimal transport problem later.

2.3.2 Cost Functions

The notion of efficiency is made formal by a "cost" function which we require to be lower semi-continuous and bounded below at 0.

Definition 2.3.4. Recall that a function $f : D \to \mathbb{R}$ is lower semi continuous at $\bar{x}$ if

$$\forall \varepsilon > 0 \; \exists \delta > 0 \text{ s.t. } f(\bar{x}) - \varepsilon < f(x) \quad \forall x \in N_\delta(\bar{x}) \cap D.$$

There are many natural cost functions, which are of varying degrees of utility depending on the context.

(a) Euclidean Distances

(i) $c(x,y) = |x-y|$ is a very natural notion of distance which simply penalizes the absolute distance between the initial point x and the end point y. Suppose we're in Monge's situation of moving dirt from one set of piles to another set of locations. Perhaps the cost corresponds to how much fuel is needed to move the dirt and our cars have constant gas mileage over distance.

(ii) $c(x,y) = (x-y)^2$ is the most commonly used cost. This might correspond to a context in which humans move dirt and get more and more tired as the distance increases.

(iii) $c(x,y) = |x-y|^p$ is the natural generalization of cost functions based on Euclidean notions of distance

(b) Discrete Cost $\mathbb{I}(x \neq y)$ penalizes disagreement with a constant factor of 1 and says that staying in place costs nothing. This cost offers a convenient connection to the total variation distance between two measures, as we'll see later.

(c) Stranger Costs: so far, we've seen relatively nice functions, but cost functions need not correspond to metrics

(i) Asymmetric Costs: $c(x,y) \neq c(y,x)$ Perhaps transporting dirt from x to y requires moving uphill and is thus more difficult than moving dirt downhill where gravity helps. In that case a sensible cost function would be necessarily asymmetric.

(ii) Possibility / Impossibility costs: suppose we require that our transport plan is supported on some area Γ. A cost function like $c(x,y) = \begin{cases} \tilde{c}(x,y) & \text{for } (x,y) \in \Gamma \\ +\infty & \text{for } (x,y) \notin \Gamma \end{cases}$ would account for that constraint.

2.3.3 Monge and Kantorovich Formulations

Now, we finally have all the necessary machinery for stating Monge's Optimal Transport problem.

Definition 2.3.5 (Monge's Optimal Transport Problem)**.** [39] Given two probability spaces $(\mathscr{X}, \mu)$ and $(\mathscr{Y}, \nu)$ with μ and ν two probability measures, find the $T_{\min}$ such that

$$\int_{\mathscr{X}} c(x, T_{\min}(x)) d\mu(x) = \inf_{\mu-\text{measurable maps } T:\mathscr{X}\to Y : \nu = T_{\#}\mu} \int_{\mathscr{X}} c(x, T(x)) d\mu(x)$$

What this definition says is that we're looking for the measurable map T that sends μ to ν and that makes the total cost of moving mass from x to $T(x)$ as small as possible with some initial allocation of mass specified by μ. This problem treats each x as indivisible. If we have three piles of dirt x_1, x_2, x_3, then we're obliged in this formulation to send all of the dirt in pile 1 to $T(x_1)$. Here we treat our initial divisions of the mass as fixed, unsplittable objects. But what if it is more convenient to move half of the mass of dirt in pile 1 to one location and half the mass to another? This possibility is allowed under a more general formulation of the OT problem.

Definition 2.3.6 (Kantorovich's OT Problem)**.** [39] Given two probability spaces $(\mathscr{X}, \mu)$ and $(\mathscr{Y}, \nu)$ with μ and ν two probability measures, find the coupling that minimizes

$$\int_{\mathscr{X}\times\mathscr{Y}} c(x,y) d\pi(x,y)$$

over all couplings $\pi \in \Pi(\mu, \nu)$.

We know that there exists at least one coupling, since the independent coupling exists (as described above).

Let's consider a simple example to see what we're actually comparing in the Kantorovich formulation.

Example 9. Consider $\mathcal{X} = \mathbb{R}$ and suppose ν and μ are the laws of familiar random variables. Let $X \sim \nu = \text{Unif}[0,1]$ and $Y \sim \mu = \text{Unif}[1,2]$. Consider two potential couplings, either $Y = X + 1$ or X and Y independent. Consider the cost function given by $c(x,y) = (x-y)^2$. Now, we'll compute the cost of each coupling.
For the independent random variables, we find

$$\int_{\mathbb{R}^2} (x-y)^2 d\pi(x,y) = \int_1^2 \int_0^1 x^2 - 2xy + y^2 dx dy = \int_1^2 1/3 - y + y^2 dy = 7/6$$

Now consider the coupling $Y = X + 1$.[2] We find that the cost is given by

$$\int_{\mathbb{R}^2} (x-y)^2 d\pi(x,y) = \mathbb{E}_\pi[(X-(X+1))^2] = 1$$

Thus, with the quadratic cost, we observe that the independent coupling is strictly more expensive, which should accord with our intuition, since quadratic cost penalizes spreading mass widely, and the deterministic coupling bounds cost between any two x and y at 1, preventing scenarios possible under the independent coupling, like $(x = 0.001, y = 2)$.

Remark 6. It is also important to observe that the Monge and Kantorovich problems need not coincide when no transport maps exist. In particular, the Monge problem is not guaranteed to have a solution, while the Kantorovich problem is.

It is worth noting at this juncture that for continuous distributions, this problem is only analytically tractable in a few special cases. Monge's formulation, for example, with L^1 cost $|x-y|$ was only solved correctly in 1999, as mentioned in [39]. This is a major motivation for the transport-entropy inequalities presented in Chapter 3. In particular, when we cannot compute the transport distance directly, it is often sufficient to upper bound it with some tractable quantity, like entropy.

However, in discrete cases, there is a clear algorithmic approach. The following example is inspired by the wonderful blog post, [42].

Example 10. Suppose you are Gaspard Monge practicing moat building in a nice grassy field. Consider five piles of dirt and five holes scattered around this field. Piles of dirt

[2] Interestingly, as we'll be able to prove later $Y = X + 1$ is the optimal coupling. To understand why, we need to dig into the structure of the problem. The key fact here is that our cost function $c(x,y)$ is particularly nice in that it looks like the Euclidean distance in $\mathbb{R}^2$. It turns out that this analogy is incredibly important.

correspond to our initial distribution μ and holes correspond to our desired final configuration ν. Suppose that the dirt is initially distributed as follows:

$$\mu = \frac{1}{10}\delta_{(x_1,y_1)} + \frac{3}{10}\delta_{(x_2,y_2)} + \frac{4}{10}\delta_{(x_3,y_3)} + \frac{1}{10}\delta_{(x_4,y_4)} + \frac{1}{10}\delta_{(x_5,y_5)}$$

where $\delta_{(x_i,y_i)}$ is a Dirac mass at location (x_i, y_i). We suppose our field is divided up into 25 squares, each of which is 1 meter by 1 meter. Each pile of dirt and each hole lives entirely in one of those squares.
Suppose our holes are distributed according to ν as follows

$$\nu = \frac{1}{5}\delta_{(u_1,w_1)} + \frac{1}{5}\delta_{(u_2,w_2)} + \frac{1}{5}\delta_{(u_3,w_3)} + \frac{1}{5}\delta_{(u_4,w_4)} + \frac{1}{5}\delta_{(u_5,w_5)}$$

Suppose our cost function is is

$$c(\vec{a},\vec{b}) = \sqrt{(a_1-b_1)^2 + \lambda(a_2-b_2)^2}, \qquad \lambda > 1$$

because moving in the left-right direction in our field is much easier than moving in the up-down direction. We can compute our cost matrix C as follows

$$\mathbf{C} = \left[\sqrt{(x_i-u_j)^2 + \lambda(y_i-w_j)^2}\right]_{i,j\in\{1,2,\ldots,5\}} \in \mathbb{R}^{5\times 5}$$

Here, the (i,j)th entry of C represents the cost to move from pile i to hole j.
Then, the transport plan T^* is determined as follows.

$$T^* = \operatorname{argmin}_{T\in\mathbb{R}^{5\times5}_{\geq 0}} \langle T, C\rangle = \operatorname{argmin}_{T\in\mathbb{R}^{5\times5}_{\geq 0}} \sum_{i=1}^{5}\sum_{j=1}^{5} T_{ij}C_{ij} \quad \text{subject to } T1 = \mu \text{ and } T^T 1 = \nu$$

where the constraint enforces that we have the correct marginals. Then, T^*_{ij} is the amount of dirt transported from pile i to hole j. Notice that this is just a simple constrained linear optimization problem, so linear programming techniques can solve this easily.

Lemma 2.3.7 (Existence of Kantorovich Transport Plans)**.** *[39] Given two Polish spaces $(\mathscr{X},\mu)$ and $(\mathscr{Y},\nu)$ with μ and ν two probability measures, and a lower-semi continuous cost function $c : \mathscr{X}\times\mathscr{Y} \to [0,\infty)$, there exists a $\pi_{\min} \in \Pi(\mu,\nu)$ that minimizes $\mathbb{K}(\pi) = \int_{\mathscr{X}\times\mathscr{Y}} c(x,y)d\pi(x,y)$ over all couplings $\pi \in \Pi(\mu,\nu)$.*

Proof. [39] We'll approach this directly. First, we want to show that the set we're minimizing over is compact, so our first task is to verify that $\Pi(\mu,\nu)$ is compact in the topology of convergence in distribution[3]. Then, we'll consider a minimizing sequence of measures π_n such that $\mathbb{K}(\pi_n) \to \inf_{\pi\in\Pi(\mu,\nu)}\mathbb{K}(\pi)$ and apply lower-semi-continuity to conclude that $\pi_{\min}$ is a true minimizer.

[3] Often called weak convergence by probabilists and weak* convergence by folks from functional analysis

Step 1: Compactness We'll show that the space of couplings $\Pi(\mu,\nu)$ is compact first by verifying that it is tight and then applying Prokhorov's theorem. Recall that the family of measures $\Pi(\mu,\nu)$ is tight if $\forall \epsilon > 0$, there exists a compact set $A \times B$ such that $\pi(A \times B) \geq 1-\epsilon \quad \forall \pi \in \Pi(\mu,\nu)$.

To see that this is the case here, recall that since μ and ν are Borel measures on Polish spaces, we know they are inner regular. Thus for any $\delta > 0$ we can find compact sets $K \subset X$ and $L \subset Y$ such that $\mu(X \setminus K) \leq \delta$ and similarly for ν such that $\nu(Y \setminus L) \leq \delta$. Notice that this bounds how much mass can sit outside $K \times L$ under any $\pi \in \Pi(\mu,\nu)$. In particular, any point $(x,y) \notin (K \times L)$ means that either $x \notin K$ or $y \notin L$. This means we can use sub additivity to find the bound

$$\pi((X \times Y) \setminus (K \times L) \leq \pi(X \times (Y \setminus L)) + \pi((X \setminus K) \times Y)$$

Now, we can apply the definition of a coupling which preserves marginals. Since above we're taking the measure of the whole sample space in one coordinate and a subset in the other, we find

$$\pi((X \times Y) \setminus (K \times L)) \leq \nu(Y \setminus L) + \mu(X \setminus K)$$

Then, applying the bounds from inner regularity above, this directly implies

$$\pi((X \times Y) \setminus (K \times L)) \leq \delta + \delta \implies \pi(K \times L) \geq 1 - 2\delta$$

Choosing $\delta = \frac{\epsilon}{2}$ gives the exact statement of tightness.
Now, we need to invoke Prokhorov's Theorem [6] which says that a collection of probability measures $\Pi(\mu,\nu)$ is tight if and only if the closure of $\Pi(\mu,\nu)$ is sequentially compact under the topology of convergence in distribution. We've just shown that $\Pi(\mu,\nu)$ is tight, thus its closure must be sequentially compact. If we can show that $\Pi(\mu,\nu)$ is its own closure (that is, it is closed), then we have demonstrated that $\Pi(\mu,\nu)$ is sequentially compact.

Now, recall that a set is closed if it contains all its limit points. Consider a sequence $\pi_n \in \Pi(\mu,\nu)$ that converges in distribution to π, which means precisely that for any continuous bounded test function f, we have

$$\int f(x,y) d\pi_n(x,y) \to \int f(x,y) d\pi(x,y)$$

We need to show that $\pi \in \Pi(\mu,\nu)$. The trick here is to choose a clever test function $\tilde{f}$ which allows us to conclude that $\mu = P^X_{\#}\pi$ and $\nu = P^Y_{\#}\pi$. Pick $f(x,y) = \tilde{f}(x)$ where $\tilde{f}$ is continuous and bounded (which amounts to a nice projection onto the x coordinate). First, note that since for every n, π_n has μ as its marginal in the x-coordinate, we know

$$\int \tilde{f}(x) d\mu(x) = \int f(x,y) d\pi_n(x,y)$$

Then, by hypothesis

$$\int f(x,y)d\pi_n(x,y) \to \int f(x,y)d\pi(x,y)$$

Then, by definition of the pushforward measure, this means

$$\begin{aligned}\int \tilde{f}(x)d\mu(x) = \int f(x,y)d\pi_n(x,y) &\to \int f(x,y)d\pi(x,y)\\ = \int f(x,y)dP^X_{\#}\pi(x)\end{aligned}$$

Then, since we're working in Polish spaces, if two measures agree when integrated against every bounded continuous test function, they must be equal. Thus $\mu = P^X_{\#}\pi$. We can play exactly the same game with $f(x,y) = \tilde{g}(y)$ to conclude that $\nu = P^Y_{\#}\pi$. Therefore, $\lim_{n\to\infty}\pi_n = \pi \in \Pi(\mu,\nu)$ so we have that $\Pi(\mu,\nu)$ is weakly closed.

Now, recall that on Polish spaces, sequential compactness is equivalent to compactness, so we've finished step 1, showing $\Pi(\mu,\nu)$ is compact.

Step 2: π^* is a Minimizer If we knew that our cost function $c(x,y)$ was bounded and continuous, we would be done, but in this context there are two reasons why we can't apply convergence directly. First, note that $c(x,y)$ is lower-semi-continuous on the product space $\mathcal{X}\times\mathcal{Y}$ but what we actually want to say is that $\mathbb{K}(\pi)$ is lower semi continuous on the space of measures with the topology of convergence in distribution. The Portmanteau Theorem allows us to move between these ideas, since it gives equivalent definitions for convergence in distribution [6].

Let $\pi_n \in \Pi(\mu,\nu)$ be a minimizing sequence, that is $K(\pi_n) \to \inf_{\pi\in\Pi(\mu,\nu)} K(\pi)$. Then, since $\Pi(\mu,\nu)$ is sequentially compact as established above, we have that there exists a subsequence π_{n_k} and some $\pi^* \in \Pi(\mu,\nu)$ such that $\pi_{n_k} \to \pi_*$. Now we just need to verify that

$$\inf_{\pi\in\Pi(\mu,\nu)} \mathbb{K}(\pi) \geq \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi^*(x,y)$$

Note first that the Portmanteau theorem gives us that

$$\begin{aligned}\mathbb{K}(\pi^*) &= \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi^*(x,y)\\ &\leq \liminf_{n\to\infty}\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi_n(x,y) = \lim_{n\to\infty}\mathbb{K}(\pi_n) = \inf_{\pi\in\Pi(\mu,\nu)}\mathbb{K}(\pi)\end{aligned}$$

Then, by the definition of inf, we have $\mathbb{K}(\pi^*) \geq \inf_{\pi\in\Pi(\mu,\nu)}\mathbb{K}(\pi)$. Therefore, we have

$$\mathbb{K}(\pi^*) = \inf_{\pi\in\Pi(\mu,\nu)}\mathbb{K}(\pi)$$

which says exactly that $\pi^* = \pi_{\min}$ is a minimizer. □

Now that we know that such minimizer exists for the Kantorovich formulation, it still remains to understand why this formulation which we've gone to a lot of trouble to establish might be useful. To see this involves stepping back for a moment and considering three ways to compute the distance between two measures that we've seen thus far. First, we've discussed the KL Divergence between two measures μ and ν, second we've (implicitly) seen the notion of an L^p distance between two functions, which we can extend to the L^p distance between two cumulative distribution functions corresponding to measures, and third we've seen the notion of optimal transport distance, namely the smallest total cost to move some mass from some configuration μ to some other configuration ν.

Example 11. This example is loosely inspired by a discussion in [39].
Consider the following simple example: let μ be the measure corresponding to $\mathbb{I}(X \in [0,1])$ and ν be the measure corresponding to $\mathbb{I}(X \in [\delta, 1+\delta])$ for some $\delta > 0$. Let's compute all three notions of distance between these two measures.

(a) KL-Divergence

$$D_{KL}(\mu||\nu) = \infty$$

since for any $\delta > 0$, there exists the region $[0,\delta)$ of length δ where μ assigns some positive probability and ν assigns none. Therefore μ is not absolutely continuous with respect to ν so the KL-divergence is infinite. Note also that this is not an issue of ordering either. While the other examples are symmetric and D_{KL} is not in general, in this case $D_{KL}(\nu||\mu) = \infty$ because of the support mis-match at the upper end, this time, on the region $[1, 1+\delta)$. This clearly does not result in a sensible notion of distance.

(b) L^p Distance
A reasonable thing to do in this case might be to consider the L^p distance between the two density functions corresponding to μ and ν. Recall that $f_\mu = \mathbb{I}(X \in [0,1])$ and $f_\nu = \mathbb{I}(X \in [\delta, 1+\delta])$. Then

$$d_{L^p}(f_\mu, f_\nu) = ||f_\mu - f_\nu||_p^p$$

Then, note that

$$f_\mu(x) - f_\nu(x) = \begin{cases} 1 & x \in [0,\delta) \\ 0 & x \in [\delta, 1] \\ -1 & x \in (1, 1+\delta] \end{cases}$$

Thus

$$d_{L^p}(f_\mu, f_\nu) = \begin{cases} (2\delta)^{1/p} & \text{for } |\delta| < 1 \\ 2^{1/p} & \text{for } |\delta| \geq 1 \end{cases}$$

This approach is more reasonable in the sense that it as $\delta \to 0$, f_μ and f_ν get closer together. However it says unreasonably that the distance between f_μ and f_ν is the same for all $\delta > 1$. In particular, it says that f_μ and f_ν are exactly $2^{1/p}$ apart whether

$\delta = 1$ or $\delta = 1000000000$, which seems not to respect the underlying geometry of the problem. If, for example, you were responsible for moving dirt from μ to ν, you would certainly notice the difference between $\delta = 1$ and $\delta = 1000000000$.

(c) OT Distance with $p-$norm Cost
The optimal transport distance is exactly given as

$$\inf_{\pi \in \Pi(\mu,\nu)} \int \int (x-y)^p \, d\pi(x,y)$$

Consider the deterministic coupling induced by $T(x) = x + \delta$, which resembles the coupling we considered in Example 9. This transport map corresponds exactly to $\pi^* = (Id \times T)_{\#}\mu$ and has the correct marginals, which says $x \sim$ Unif[0,1] and $x + \delta \sim$ Unif$[\delta, 1+\delta]$. Then $\pi * \in \Pi(\mu,\nu)$ and it has cost

$$\int \int |x-y|^p \, d\pi^*(x,y) = \int_0^1 |x - (x+\delta)|^p \, dx = \delta^p$$

We know that this is the optimal coupling by the Monotone Rearrangement Theorem.

Theorem 2.3.8 (Monotone Rearrangement Theorem [33]). *Let ν and μ be two Borel-measurable probability measures on $\mathbb{R}$ with CDFs F_ν and F_μ respectively. Let $F_\mu^{-1}(u) := \inf\{y \in \mathbb{R} : F_\mu(y) \geq u\}$ for $u \in [0,1]$ be the quantile function. Then, let $c(x,y) = d(x-y)$ be a convex, continuous cost function such that*

$$\int_0^1 d\left(F_\nu^{-1}(t) - F_\mu^{-1}(t)\right) dt < \infty.$$

Define the non-decreasing map

$$T(x) := F_\mu^{-1}\left(F_\nu(x)\right), \qquad x \in \mathbb{R}.$$

Then the coupling $\pi := (\mathrm{id}, T)_{\#}\nu$ is an optimal solution of Kantorovich's optimal transport problem between ν and μ with cost c. If, in addition, if F_μ is continuous and d is strictly convex, then T is the unique optimal transport map from ν to μ.

The theorem is proved in [33], another nice exposition of it is given in Theorem 2.1 in [39]. We do not prove this theorem here because it is not crucial to the point of this thesis, which argues in effect that rather than direct computation of transport distances, an upper bound provided by an easy and elegant entropy computation usually suffices. We will provide a sketch of the main ideas of the proof below for the interested reader.

Proof. The idea of the proof is to first notice that any optimal coupling γ has a crucial monotone property which holds for all $(x, y) \in \sup \gamma$.
In particular the set $\sup \gamma$ is monotone in a particular way which means that for any pair of points (x, y) and $(\tilde{x}, \tilde{y})$, we have that

$$d(x-y) + d(\tilde{x}-\tilde{y}) \leq d(x-\tilde{y}) + d(\tilde{x}-y).$$

The key insight which is usually shown via contradiction is that this monotone property implies monotone behavior of the transport map, that is

$$x \leq \tilde{x} \implies y \leq \tilde{y}.$$

This result finishes the strictly convex case. For the non-strict case, the idea is to use a sequence of approximating functions, each of which is strictly convex, and simply take the limit. □

This result tells us that the minimum distance is given exactly by

$$\int_0^1 (F_\mu^{-1}(t) - F_\nu^{-1}(t))^p \, dt = \int_0^1 (t - (t+\delta))^p \, dt = \delta^p$$

since the inverse CDFs here are precisely $F_\mu^{-1}(t) = t$ and $F_\nu^{-1}(t) = t + \delta$ for $t \in [0, 1]$. Thus our coupling is the optimal coupling and therefore the optimal transport distance between μ and ν is exactly δ^p. This has very nice, intuitive properties. In particular, it is strictly increasing in δ and goes to 0 as $\delta \to 0$. In particular, the optimal transport distance with $p-$norm cost respects that difference between $\delta = 1$ and $\delta = 1000000000$ that both KL Divergence and L^p distance ignore.

Notice though that this is as much a property of our choice of cost function $c(x, y)$ as it is of the optimal transport problem formulation. If we had instead chosen a cost function like $\tilde{c}(x, y) = \min(1, |x-y|^2)$, we could re-create the problems with the L^2 formulation in the problem. In particular, the optimal transport distance with this new cost function $\tilde{c}(x, y) = \min(1, |x-y|^2)$ would be the same for $\delta < 1$, ie the transport distance would be δ^2, but for $\delta > 1$ the transport distance would get "stuck" at 1. In particular for $\delta > 1$, the OT distance is 1 regardless of whether $\delta = 1$ and $\delta = 1000000000$.

This gives us the crucial intuition for defining Wasserstein distances: we want something that respects the underlying geometry of the problem.

Whenever we discuss Wasserstein distances, we'll restrict ourselves to the following setting. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^d$, and let's work only with distributions whose p^{th} moment is bounded. In particular, let

$$\mu, \nu \in \mathcal{P}_p(\mathcal{X}) = \{\gamma \in \mathcal{P}(\mathcal{X}) \text{ such that } \int_{\mathcal{X}} |x|^p \, d\mu(x) < \infty\}$$

Definition 2.3.9 (Wasserstein Distance)**.** [39] [33] Let $\mu, \nu \in \mathscr{P}_p(\mathscr{X})$. Then, the Wasserstein distance is defined as

$$W_p(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathscr{X}\times\mathscr{X}} |x-y|^p d\pi(x,y) \right)^{\frac{1}{p}}$$

Remark 7. This should be very familiar! The Wasserstein distance is exactly the p^{th} root of the minimum of the Kantorovich optimal transport problem with cost function $c(x,y) = |x-y|^p$. From here on out, we'll denote the solution to the Kantorovich optimal transport problem with cost function c as $T_c(\nu,\mu)$. This means we can write $W_p(\mu,\nu) = \mathscr{T}_{d^p}(\mu,\nu)^{\frac{1}{p}}$.

Lemma 2.3.10. *Notice that the Wasserstein distance is a valid metric on the space of the space of distributions with bounded p^{th} moment.*

Proof. The metric inherits its properties from the $c(x,y) = |x-y|^p$. Symmetry and non-negativity are immediate from the structure of the cost function. The triangle inequality comes from 'gluing' measures together and Minkowski's inequality, though this takes some effort to prove. See [33] Lemma 5.2 and Lemma 5.3 for full details. □

Chapter 3

Classical Transport-Entropy Inequalities

The key idea behind transport-entropy inequalities is that there should be some way to connect how far distribution ν is from distribution μ (the optimal transport distance) to how surprising ν looks if you expected μ (the relative entropy). Intuitively these ideas are closely connected as the many examples given in Chapter 2 seem to suggest.

This chapter will first introduce a general formulation for transport-entropy inequalities in Section 3.1. In Section 3.2, we introduce the most common class of transport-entropy inequalities, T_p inequalities which compare Wasserstein distance to relative entropy and prove that Gaussian measures satisfy a T_2 inequality. Next, in Section 3.3, we present a classical inequality, Pinsker's inequality as a kind of 'T_0' transport-entropy inequality. Finally, in Section 3.4, we demonstrate one application of transport-entropy inequalities: establishing concentration inequalities.

3.1 Introduction to Transport-Entropy Inequalities

Definition 3.1.1 (Transport-Entropy Inequality [17])**.** Let $c : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be a lower semi-continuous cost function and let

$$H(\cdot \mid \mu) : \mathscr{P}(X) \to [0, \infty)$$

be the relative entropy of some measure with respect to μ. Let $\alpha : [0,\infty) \to [0,\infty)$ be an increasing function satisfying $\alpha(0) = 0$.
We say that $\mu \in \mathscr{P}(X)$ satisfies a transport-entropy inequality if

$$\alpha\big(T_c(\nu,\mu)\big) \leq 2CH(\nu \mid \mu) \qquad \text{for all } \nu \in \mathscr{P}(X),$$

where $T_c(\nu,\mu)$ denotes the optimal transport cost with respect to the lower semi-continuous cost function and C is a fixed, non-negative constant.

Remark 8. This requires that $\alpha\big(T_c(\mu,\mu)\big) = 0$, since we know that $H(\mu \mid \mu) = 0$. To ensure that $\alpha\big(T_c(\mu,\mu)\big) = 0$, it is enough to require $c(x,x) = 0$ for all $x \in \mathscr{X}$ and $\alpha(0) = 0$, which we

assume from here on out. Notice that if we don't require that $\alpha\big(T_c(\mu,\mu)\big) = 0$, no measure will ever satisfy this inequality, since the relative entropy of μ with respect to itself is always 0, as made clear by Definition 2.2.4.

3.2 T_p Inequalities

Arguably the most important class of such inequalities offer a bound on Wasserstein distance by relative entropy. These inequalities are the so-called T_p inequalities which provide bounds on the Wasserstein$-p$ distance.

Definition 3.2.1 (T_p Inequality)**.** Let $(\mathscr{X}, d)$ be a metric space and μ a measure of $\mathscr{X}$ such that $\int_{\mathscr{X}} d(x_0,x)^p d\mu(x) < \infty$ for some $x_0 \in \mathscr{X}$. We say that μ satisfies a T_p inequality with constant $C > 0$ if for all probability measures ν on $\mathscr{X}$ such that $\int_{\mathscr{X}} d(x_0,x)^p d\nu(x) < \infty$ for some $x_0 \in \mathscr{X}$, we have

$$W_p(\mu,\nu)^2 \leq 2CH(\nu|\mu)$$

or equivalently

$$W_p(\mu,\nu) \leq \sqrt{2CH(\nu|\mu)}$$

where $H(\cdot|\mu)$ is the relative entropy with respect to μ and W_p is the Wasserstein-p distance.

Remark 9. In general, we restrict to measures $\mu,\nu \in \mathscr{P}_p(\mathscr{X}) = \{\gamma \in \mathscr{P}_p(\mathscr{X}) : \int_{\mathscr{X}} |x|^p d\gamma(x) < \infty\}$, that is measures that have finite p^{th} moment because we need the Wasserstein$-p$ distance to be finite in order for this inequality to be meaningful.

Remark 10. Different authors have different preferences and conventions about whether to absorb the 2 into the constant C. We choose not to and will be consistent with this throughout Chapters 3 and 4.

The two most common transport-entropy inequalities are the T_1 and T_2 inequalities.

Definition 3.2.2 (T_1 Inequality: $T_1(C) : W_1(\nu,\mu) \leq \sqrt{2C\,H(\nu\,|\,\mu)}$)**.** [18]
Define

$$\mathscr{P}_1(X) := \Big\{\nu \in \mathscr{P}(X) : \int d(x_0,x)\,d\nu(x) < \infty \text{ for some } x_0 \in X\Big\}.$$

For $\mu \in \mathscr{P}_1(X)$, we say that μ satisfies the transport inequality $T_1(C)$ with constant $C > 0$ if for every $\nu \in \mathscr{P}_1(X)$,

$$W_1(\nu,\mu) \leq \sqrt{2\,C\,D_{KL}(\nu\,|\,\mu)},$$

where W_1 denotes the 1-Wasserstein distance and $D_{KL}(\nu\,|\,\mu)$ is the Kullback–Leibler divergence, which is also called the relative entropy and denoted $H(\nu|\mu)$.

To build intuition, we consider two simple reference measures μ on $\mathbb{R}$ and compute the quantities in the T_1 inequality along a concrete parametric subfamily $\{\nu_\theta\}$ of absolutely continuous measures.

Proving the $T_1(C)$ inequality from its definition is often difficult, because it requires verifying that

$$W_1(\nu,\mu) \le \sqrt{2C\,H(\nu \mid \mu)}$$

holds for *all* $\nu \ll \mu$ with $H(\nu \mid \mu) < \infty$ and finite first moment. In the short examples below we instead verify the inequality only for a named parametric subfamily $\{\nu_\theta\}$. This direct calculation is useful for intuition, but it does *not* by itself establish that μ satisfies $T_1(C)$ universally.

Example 12 (Beta Distribution). Consider μ as the measure corresponding to a Beta$(1,1)$ distribution and define the test family as $\nu_t = \{\text{Beta}(t+1,1) : t > -1\}$. First, let's compute the relative entropy.

First, note both densities have the same support, so the relative entropy is finite and well defined.

$$H(\nu_t|\mu) = \int_0^1 \frac{d\nu_t}{d\mu}(x)\log(\frac{d\nu_t}{d\mu}(x))\,dx = \int_0^1 (1+t)x^t(\log(1+t)+t\log(x))\,dx$$
$$= (1+t)\int_0^1 x^t\log(1+t)\,dx + (1+t)\int_0^1 x^t\,t\log(x)\,dx = \log(1+t) - \frac{t}{1+t}$$

which results from applying nice results about β integrals [7].

Now, we'll compute the Wasserstein-1 Cost $W_1(\nu_t,\mu)$. Here, we'll apply the Monotone Rearrangement Theorem (Theorem 2.3.8) which gives us a straightforward way to compute the Wasserstein-1 distance. First, note that the distribution functions of μ and ν_t on $[0,1]$ are

$$F_\mu(x) = x, \qquad F_{\nu_t}(x) = \int_0^x (1+t)\,u^t\,du = x^{1+t}.$$

By Theorem 2.3.8, the optimal transport from μ to ν_t for the cost $c(x,y) = |x-y|$ is given by the monotone map

$$T(x) = F_{\nu_t}^{-1}\big(F_\mu(x)\big) = F_{\nu_t}^{-1}(x) = x^{\frac{1}{1+t}},$$

and therefore

$$W_1(\nu_t,\mu) = \int_0^1 |x - T(x)|\,d\mu(x) = \int_0^1 \left|x - x^{\frac{1}{1+t}}\right| dx.$$

Set $p = \frac{1}{1+t} > 0$. Then

$$\int_0^1 x\,dx = \frac{1}{2}, \qquad \int_0^1 x^p\,dx = \frac{1}{p+1}.$$

Note, for $t > 0$ (i.e. $p < 1$) we have $x^p \ge x$ on $(0,1)$, while for $-1 < t < 0$ (i.e. $p > 1$) we have $x^p \le x$. Thus in all cases

$$W_1(\nu_t,\mu) = \left|\frac{1}{2} - \frac{1}{p+1}\right| = \left|\frac{1}{2} - \frac{1}{1+\frac{1}{1+t}}\right| = \left|\frac{1}{2} - \frac{1+t}{2+t}\right|.$$

Finally, simplifying gives the closed form solution

$$W_1(\nu_t, \mu) = \frac{|t|}{2(2+t)}, \qquad t > -1$$

Now, note that we have two functions which we want to compare over our entire parametric family ν_t.

$$H(\nu_t|\mu) = \log(1+t) - \frac{t}{t+1} \quad \text{and } W_1(\nu_t, \mu) = \frac{|t|}{2(2+t)}$$

To see if μ satisfies the $T_1(C)$ inequality *for the family* ν_t, we want to compare

$$\frac{t^2}{4(2+t)^2} \text{vs} \quad 2 \cdot C \cdot (\log(1+t) - \frac{t}{t+1})$$

Now, finally, we need to show that there is one scalar C which ensures this holds for all $t > -1$. Consider the ratio

$$R(t) = \frac{W_1(\nu_t, \mu)^2}{2H(\nu_t|\mu)} = \frac{\frac{t^2}{4(2+t)^2}}{2(\log(1+t) - \frac{t}{t+1})} = \frac{t^2}{8(t+2)^2(\log(t+1) - \frac{t}{t+1})}$$

Now, note that on the domain $(-1, \infty)$ this function which is a composition of continuous functions is continuous. At the boundaries $\lim_{t\to -1} R(t) = 0$, which can be verified with L'Hopitals rule and note that $\lim_{t\to\infty} R(t) = 0$. Since it is continuous and bounded at endpoints, it attains a finite supremum. This is effectively by the extreme value theorem, since, although the interval $(-1, \infty)$ is not itself compact, for any $\epsilon > 0$, the set $\{t : R(t) \geq \epsilon\}$ is contained in a compact interval to which we can safely apply the extreme value theorem, which ensures that the maximum is itself contained in $(-1, \infty)$. Then we can safely set

$$C := \sup_{t \in (-1,\infty)} R(t)$$

and be assured that this C is large enough to ensure that the family ν_t does satisfy the $T_1(C)$ inequality, since for some constant C, we have for all $t > -1$

$$\frac{t^2}{4(2+t)^2} \leq 2 \cdot C(\log(1+t) - \frac{t}{t+1})$$

See Figure 3.1 to see the comparison between the optimal transport distance (in blue) and the KL-divergence (in red) for all $t > -1$. Intuitively, it should make sense that the KL-divergence grows faster than the Wasserstein distance in this example for the following reason.

Notice that the parameter t "tilts" the Beta distribution from the uniform distribution on the unit interval (given by a Beta(1,1)) towards one of the two end points: $t > 0$ piles mass near 1 and $-1 < t < 0$ piles mass near 0. As $t \to -1$, we know $\text{Beta}(1+t, 1) \xrightarrow{d} \delta_0$ while also

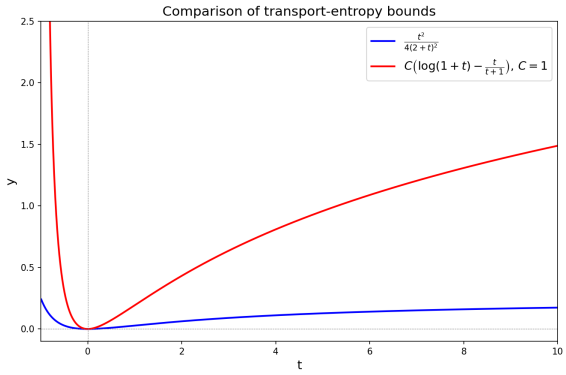


Figure 3.1: Relative entropy and KL-Divergence plotted against all possible values of $t > -1$

$t \to \infty$, we have that $\text{Beta}(1+t,1) \xrightarrow{d} \delta_1$. This means in some sense which we will make precise later that this family of distributions is 'concentrated.' In one dimension, the optimal W_1-transport is monotone and is explicitly given by the monotone rearrangement theorem (2.3.8) so W_1 is the roughly the average amount of dirt displaced resulting from this transformation of the unit interval. Crucially, transport on a bounded interval has a natural saturation point. No coupling can move mass more than distance 1, and as we've seen,

$$W_1(\nu_t, \mu) \leq \frac{1}{2}.$$

However, the relative entropy does not saturate: as $t \to \infty$ (or $t \downarrow -1$) the tilted density becomes highly non-uniform (and converges to one of two Dirac delta measures. Therefore, the Kullback–Leibler divergence blows up. Thus the ratio $W_1(\nu_t, \mu)^2 / H(\nu_t \mid \mu)$ is small in the extremes of t, and the "worst case" must occur at an intermediate tilt where transport is noticeable but entropy has not yet exploded to ∞.

Now we'll move on to the arguably more common T_2 transport inequality. In this case it is possible to verify directly a very useful result which we'll state after providing the definition of the T_2 inequality.

Definition 3.2.3 (T_2 Inequality: $T_2(C) : W_2(\nu, \mu) \leq \sqrt{2C\, H(\nu \mid \mu)}$)**.** [18] Define

$$\mathscr{P}_2(X) := \left\{ \nu \in \mathscr{P}(X) : \int d(x_0, x)^2 \, d\nu(x) < \infty \text{ for some } x_0 \in X \right\}.$$

For $\mu \in \mathscr{P}_2(X)$, we say that μ satisfies the transport inequality $T_2(C)$ with constant $C > 0$ if for every $\nu \in \mathscr{P}_2(X)$,

$$W_2(\nu, \mu) \leq \sqrt{2\, C\, D_{KL}(\nu \mid \mu)},$$

where W_2 denotes the 2-Wasserstein distance.

It turns out that Gaussian measures satisfy this inequality. This fact is crucial to the study of these inequalities and will come up over and over again in the following sections. This

was a major result proved by Talagrand in 1996 [37]. A very nice presentation of this proof is given in [18], and for the most part, the subsequent proof follows [18].

Proposition 3.2.4 (Gaussian Measures on $\mathbb{R}$ satisfy T_2)**.** *The Gaussian measure γ corresponding to a standard normal distribution on $\mathbb{R}$ satisfies the following T_2 inequality with $C = 1$ for all $\nu \in \mathscr{P}(\mathbb{R})$:*

$$W_2^2(\nu, \gamma) \leq 2H(\nu|\gamma)$$

Proof. The key idea of this proof is to take advantage of the similarity in form between the standard normal density, which gives rise to a nice relative entropy formula, and the quadratic cost function used in the OT computation of the Wasserstein-2 distance.

We'll begin by rewriting the normal density in potential form to emphasize this quadratic form.

$$d\mu(x) = e^{-V(x)}dx$$

where $V(x) = \frac{x^2}{2} + \frac{\log(2\pi)}{2}$.

Let ν be any probability measure on $\mathbb{R}$ with finite second moment. Now, recall the Monotone Rearrangement Theorem 2.3.8. This handy theorem tells us what the optimal map T on $\mathbb{R}$ under any convex, continuous cost function, including this one, $c(x, y) = |x - y|^2$. In this case, by Theorem 2.3.8 the optimal transport plan is deterministic, given by $T(x) = F_\nu^{-1}(F_\mu(x))$ where $F_\nu^{-1}(\cdot)$ is the quantile function of ν and F_μ is the CDF of μ. We can write this more explicitly as $T(x) = F_\nu^{-1}(\Phi(x))$ where Φ is the normal CDF.

Now, we'd like to compute the entropy and show that it upper bounds the Wasserstein-2 distance.

In the case where ν is not absolutely continuous with respect to μ, we're done since that means that $H(\nu|\mu)$ is infinite which is trivially an upper bound on $W_2^2(\nu, \mu)$. Therefore, we can assume without loss of generality that $\nu \ll \mu$ and that the Radon-Nikodym derivative is well defined.

To do this, first write out the definition of the entropy in terms of $f = \frac{d\nu}{d\mu}$ the Radon Nikodym derivative.

$$H(\nu|\mu) = \int \log(f)d\nu = \int \log(f(T(x))d\mu$$

where the last equality follows from the fact that ν is the push-forward of μ by T, ie $\nu = T_\# \mu$ where T is the map given by the Monotone Rearrangement Theorem.

Now, it remains to find what exactly $f(T(x))$ is in terms of our density. We can find this by noticing that $T(x)$ is monotone increasing and therefore differentiable almost everywhere.

Notice that $T(x)$ is monotone increasing in x by construction, since $\Phi(x)$ is a CDF and therefore increasing in x and the quantile function is similarly increasing in its argument, by its definition. In particular, for any $x_1 \leq x_2$, we know $\Phi(x_1) \leq \Phi(x_2)$ so it suffices to check that for any $p_1 \leq p_2$, we have $F_\nu^{-1}(p_1) \leq F_\nu^{-1}(p_2)$.

To see this, recall that $F_\nu^{-1}(p) := \inf\{x \in \mathbb{R} : F_\nu(x) \geq p\}$, so for any $p_1 \leq p_2$, we must have $\{y : F_\nu(y) \geq p_2\} \subseteq \{y : F_\nu(y) \geq p_1\}$, which implies exactly that F_ν^{-1} is monotone increasing.

Now, we'll observe that the map T is almost everywhere differentiable by recalling that a monotone map on $\mathbb{R}$ is almost everywhere Lebesgue differentiable (for a proof of this fact see [38]).

Now, note that we can re-write our MRT coupling to solve for this quantity. Recall $T(x) = F_\nu^{-1}(F_\mu(x)) \iff F_\nu(T(x)) = F_\mu(x)$. Then, using the definition of the CDF, we find

$$\int_{-\infty}^{T(x)} f(z)e^{-V(z)}dz = \int_{-\infty}^{x} e^{-V(z)}dz$$

Now, we can differentiate both sides with respect to x. Clearly

$$\frac{d}{dx}(\int_{-\infty}^{x} e^{-V(z)}dz) = e^{-V(x)}$$

and by the chain rule, we find exactly

$$\frac{d}{dx}(\int_{-\infty}^{T(x)} f(z)e^{-V(z)}dx) = T'(x)f(T(x))e^{-V(T(x))}.$$

Therefore, we have

$$e^{-V(x)} = T'(x)f(T(x))e^{-V(T(x))} \implies f(T(x)) = \frac{e^{-V(x)}}{T'(x)e^{-V(T(x))}} = \frac{1}{T'(x)}e^{-V(x)+V(T(x))}$$

Now, we can compute our entropy by substituting in this value

$$H(\nu|\mu) = \int_{\mathbb{R}} \log(f(T(x)))d\mu = \int_{\mathbb{R}} (-V(x) + V(T(x)) - \log(T'(x)))e^{-V(x)}dx$$

Now, this is beginning to look promising, because the quadratic form of $V(x)$ makes the $-V(x) + V(T(x))$ look promising if we're hoping to extract a squared difference term from our integral to make it match the Wasserstein-2 distance. The problem is precisely to remove the $-\log(T'(x))$ term and at the same time use it to argue that

$H(\nu|\mu) = W_2^2(\nu,\mu) +$ positive constant. There are three clever algebraic tricks which will allow us to do just that.

First, we'll add zero! Notice that $-\log(T'(x)) = T'(x) - 1 - \log(T'(x)) - (T'(x) - 1)$. Therefore, we have:

$$H(\nu|\mu) = \int_{\mathbb{R}} (-V(x) + V(T(x)) + T'(x) - 1 - \log(T'(x)) - (T'(x) - 1)e^{-V(x)}\,dx$$

Now, we can split the integral:

$$H(\nu|\mu) = \underbrace{\int_{\mathbb{R}} (-V(x) + V(T(x)) - (T'(x) - 1))e^{-V(x)}\,dx}_{I_1} + \underbrace{\int_{\mathbb{R}} (T'(x) - 1 - \log(T'(x)))e^{-V(x)}\,dx}_{I_2}$$

Considering I_2 we note that the integrand is positive over the entire domain of integration $\mathbb{R}$, by recalling that for any $a \geq 0$, we know that $a - 1 - \log(a) \geq 0$. This applies since we know that $T'(x) \geq 0$ almost everywhere. Similarly, e^{-x} is always positive. Therefore, we know that this integral $I_2 > 0$.

Now, turning to integral I_1, we want to interpret $T'(x) - 1$ as the result of integration by parts in order to convert $-V(x) + V(T(x)) + T'(x) - 1$ into something that looks like the $c(x,y) = |x - T(x)|^2 = |x^2 - 2T(x) + T(x)^2|$.

In particular, consider

$$\int_{\mathbb{R}} (T'(x) - 1)e^{-V(x)}\,dx$$

as the remaining term in integration by parts. Suppose instead of the current problem, we were interested in solving

$$\int_{\mathbb{R}} (T(x) - x)V'(x)e^{-V(x)}\,dx$$

Now, let's integrate this by parts. Let $f = (T(x) - x)$ and $g' = V'(x)e^{-V(x)}$. Then, we find that

$$\begin{aligned}\int_{\mathbb{R}} (T(x) - x)V'(x)e^{-V(x)}\,dx &= -[(T(x) - x)e^{-V(x)}]_{-\infty}^{\infty} + \int_{\mathbb{R}} (T'(x) - 1)e^{-V(x)}\,dx \\ &= \int_{\mathbb{R}} (T'(x) - 1)e^{-V(x)}\,dx\end{aligned}$$

Where the first term vanishes at the boundary.

Now, we notice that

$$\underbrace{\int_{\mathbb{R}} (-V(x) + V(T(x)) + T'(x) - 1)e^{-V(x)}\,dx}_{I_1} = \int_{\mathbb{R}} (-V(x) + V(T(x)))e^{-V(x)}\,dx$$

$$+\int_{\mathbb{R}}(T(x)-x)V'(x)e^{-V(x)}dx$$

Then substituting our integration by parts result, we find

$$\underbrace{\int_{\mathbb{R}}(-V(x)+V(T(x))+T'(x)-1)e^{-V(x)}dx}_{I_1}=\int_{\mathbb{R}}(-V(x)+V(T(x)))e^{-V(x)}dx$$
$$+\int_{\mathbb{R}}(T(x)-x)V'(x)e^{-V(x)}dx$$

Factoring our integrals, we find

$$\underbrace{\int_{\mathbb{R}}(-V(x)+V(T(x))+T'(x)-1)e^{-V(x)}dx}_{I_1}=\int_{\mathbb{R}}(-V(x)+V(T(x))-V'(x)(T(x)-x))e^{-V(x)}dx$$

Then, we have

$$H(\nu|\mu)=\underbrace{\int_{\mathbb{R}}(-V(x)+V(T(x))-V'(x)(T(x)-x))e^{-V(x)}dx}_{I_1}$$
$$+\underbrace{\int_{\mathbb{R}}(T'(x)-1-\log(T'(x)))e^{-V(x)}dx}_{I_2}.$$

We know that I_2 is non-negative, so we have that

$$H(\nu|\mu)\geq\int_{\mathbb{R}}(-V(x)+V(T(x))-V'(x)(T(x)-x))e^{-V(x)}dx$$

Now we'll finally leverage the form of $V(x)$. It is really the connection between the quadratic form of the cost and the quadratic form of the Gaussian potential that makes this proof work. Notice that

$$-V(x)+V(T(x))-V'(x)(T(x)-x)=-(\frac{x^2}{2}+\frac{\log(2\pi)}{2})+\frac{T(x)^2}{2}+\frac{\log(2\pi)}{2}-x(T(x)-x).$$
$$T(x)^2-xT(x)+\frac{x^2}{2}=\frac{(T(x)-x)^2}{2}$$

Therefore, we can finally conclude

$$H(\nu|\mu)\geq\int_{\mathbb{R}}\frac{(T(x)-x)^2}{2}e^{-V(x)}dx=\frac{W_2^2(\nu,\gamma)}{2}.$$

Therefore, we've proved that for every $\nu\in\mathscr{P}(\mathbb{R})$, $W_2^2(\nu,\gamma)\leq 2H(\nu|\gamma)$ as desired.
Intuitively, what this property means is that you can't 'move' a Gaussian density very far in W_2 distance without having to re-weight its density by quite a lot (ie having a very large KL-Divergence). The magic of the Gaussian case is that almost every portion of this inequality is computable and available in a nice closed form. □

Example 13. A natural question after going to all this trouble to prove this inequality is whether we can do better with the constant. In particular we have proved this for $C = 1$, but is it possible to make C smaller? No, as it turns out. To see why, consider a very simple $\nu = \mathcal{N}(m, 1)$, a translation of the normal distribution $\gamma = \mathcal{N}(0, 1)$.

We'll first compute the entropy of $H(\nu|\gamma)$. To do this, recall Example 5, where we found that

$$D_{KL}(p||q) = \log(\frac{\sigma_q}{\sigma_p}) + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}$$

Therefore,

$$H(\nu|\gamma) = \log(1) + \frac{1+m^2}{2} - \frac{1}{2} = \frac{m^2}{2}$$

Again, applying the Monotone Rearrangement Theorem 2.3.8, we find that the optimal coupling between ν and γ is given by $T(x) = F_\nu^{-1}(\Phi(x))$. Here, since $\nu \sim \mathcal{N}(m, 1)$, the CDF of ν is given by $F_\nu(x) = \Phi(x - m) \implies F_\nu^{-1}(p) = m + \Phi^{-1}(p)$, so the composition

$$T(x) = F_\nu^{-1}(\Phi(x)) = x + m$$

Thus, the Wasserstein-2 Distance is given by

$$W_2(\nu, \gamma) = \left(\int_{\mathbb{R}} (x - T(x))^2 d\gamma(x)\right)^{1/2} = \left(\int_{\mathbb{R}} (x - (x+m))^2 d\gamma(x)\right)^{1/2} = \left(m^2 \int_{\mathbb{R}} d\gamma(x)\right)^{1/2} = |m|$$

Therefore, we have exact equality, that is $W_2^2(\nu, \gamma) = m^2 = m^2 = 2H(\nu|\gamma)$, so the inequality is tight for Normal translations and $C = 1$, so we can do no better than in Proposition 3.2.4

Another very natural question is how do the T_1 and T_2 inequalities relate. In particular if we know that $\gamma \sim \mathcal{N}(0, 1)$ satisfies $T_2(1)$, we might wonder if it also satisfies $T_1(2)$? As it turns out, it does. To see this, we'll prove a general fact: namely that as p increases, the 'strength' of the transport-entropy inequality increases.

Theorem 3.2.5 (For $p \geq q$, $T_p(C) \implies T_q(C)$)**.** *Let $T_p(C)$ be the inequality defined above, namely that for $(\mathcal{X}, d)$ a metric space and a μ a measure with finite p^{th} moment, there exists $C > 0$ such that for all measures ν on $\mathcal{X}$ with finite p^{th} moment, we have $W_p(\mu, \nu) \leq \sqrt{2CH(\nu|\mu)}$. Then for $p \geq q$, we have $T_p(C) \implies T_q(C)$.*

To prove this, we need a convenient way to compare $p-$norms and $q-$norms of functions.

Lemma 3.2.6 (Hölder's Inequality)**.** *[31] For some metric space $(\mathcal{X}, d)$ and $r, s \in [1, \infty]$ such that $\frac{1}{s} + \frac{1}{r} = 1$. Then, for any measurable functions g, h on $\mathcal{X}$, we have*

$$||gh||_1 \leq ||g||_r ||h||_s$$

Now, we'll prove Theorem 3.2.5.

Proof. First, note that by Hölder's Inequality (see Lemma 3.2.6), we have that

$$\left(\int_{\mathscr{X}} d(x,y)^q d\pi(x,y)\right)^{\frac{1}{q}} \leq \left(\int_{\mathscr{X}} d(x,y)^p d\pi(x,y)\right)^{\frac{1}{p}}$$

We can see this by applying Lemma 3.2.6 with $g = 1$, $f(x) = d(x,y)^q$, $r = \frac{p}{q}$ and $s = \frac{p}{p-q}$.

This gives exactly

$$\int_{\mathscr{X}\times\mathscr{X}} d(x,y)^q d\pi = \int_{\mathscr{X}\times\mathscr{X}} d(x,y)^q \cdot 1 d\pi(x,y) \leq \left(\int_{\mathscr{X}\times\mathscr{X}} d(x,y)^p d\pi(x,y)\right)^{q/p}$$

Then, raising both sides to the power of $\frac{1}{q}$ gives the desired result.

Therefore, we can take the infimum over all couplings of μ, ν. This preserves the inequality, yielding

$$W_q(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \left(\int_{\mathscr{X}\times\mathscr{X}} d(x,y)^q d\pi(x,y)\right)^{\frac{1}{q}} \leq \inf_{\pi\in\Pi(\mu,\nu)} \left(\int_{\mathscr{X}\times\mathscr{X}} d(x,y)^p d\pi(x,y)\right)^{\frac{1}{p}} = W_p(\mu,\nu)$$

Therefore, if we have $p \geq q$, we have $W_p(\mu,\nu) \geq W_q(\mu,\nu)$. Thus if μ satisfies a transport-entropy inequality $T_p(C)$ with constant C, we know that $W_q(\mu,\nu) \leq W_p(\mu,\nu) \leq \sqrt{CD_{KL}(\nu|\mu)}$. □

This leads to a nice corollary of the T_2 result.

Corollary 3.2.7. *If $\gamma \sim \mathcal{N}(0,1)$ then γ satisfies a $T_1(1)$ inequality.*

Proof. This follows directly from Theorem 3.2.5. □

A natural response to this very nice result would be to try to prove the $T_p(C)$ inequality for some measure μ for some p as large as possible and simply use the Theorem 3.2.5 to derive all of the smaller p examples. This would lead quite naturally to considering the case $T_\infty(C)$ where we take the limit as $p \to \infty$ of the Wasserstein-p distance. This turns out not to be a fruitful approach for some very interesting reasons. To learn a little more about transport-entropy inequalities, we'll spend a little time on this non-example.
Before we do anything else, however, we need to define the W_∞ distance. Following [2], we define the 'worst case' transport cost as follows.

Definition 3.2.8. For a given transport plan $\pi \in \Pi(\mu,\nu)$ we define the T_∞ cost (which is intuitively the furthest mass has to move according to this transport plan) as

$$T_\infty(\pi) := \operatorname{ess\,sup}_{x,y\in X\times X:\pi_{x,y}>0} |x-y|$$

where ess sup denotes the essential supremum

Then we can interpret the Wasserstein-infinity distance as the mini-max cost.

Definition 3.2.9.

$$W_\infty(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} T_\infty(\pi) = \inf_{\pi\in\Pi(\mu,\nu)} \operatorname{ess\,sup}_{(x,y)\sim\pi} |x-y|$$

It is not immediate from this definition that this is the limit as $p\to\infty$ of the Wasserstein-p distance we had above. We verify this fact briefly, though it is given as an exercise in [33].

Lemma 3.2.10. *Let μ and ν be compactly supported measures on $\mathscr{X}$.*

$$W_\infty(\mu,\nu) = \lim_{p\to\infty} W_p(\mu,\nu)$$

Proof. We'll prove equality in the usual way, first by showing

$$\lim_{p\to\infty} W_p(\mu,\nu) \le W_\infty(\mu,\nu)$$

and then by showing

$$\lim_{p\to\infty} W_p(\mu,\nu) \ge W_\infty(\mu,\nu).$$

First, note that the map $p\mapsto W_p(\mu,\nu)$ is non-decreasing by Theorem 3.2.5, so we know that $L := \lim_{p\to\infty} W_p(\mu,\nu) \in [0,\infty]$ and exists.

Let $\pi^*\in\Pi(\mu,\nu)$ be optimal for the essential supremum loss, that is let π^* be the coupling that yields the smallest essential supremum and thus is used in the the $W_\infty(\mu,\nu)$ distance. Then, note $W_p(\mu,\nu) \le (\int_{\mathscr{X}} |x-y|^p d\pi^*)^{1/p}$ since π^* is not necessarily optimal for the cost $|x-y|^p$.

Then, taking the limit of both sides as $p\to\infty$ (which is permitted and well defined in both cases since, first π^* is a probability measure, and second $c(x,y)=|x-y|^p$ is bounded, since we assume μ and ν to be compactly supported), we have that the inequality is preserved in the limit:

$$\lim_{p\to\infty} W_p(\mu,\nu) \le \lim_{p\to\infty} (\int_{\mathscr{X}} |x-y|^p d\pi^*)^{1/p} = W_\infty(\mu,\nu)$$

Now, we'll show the other direction, which takes a bit more work $\lim_{p\to\infty} W_p(\mu,\nu) \ge W_\infty(\mu,\nu)$.

The key idea here is to use the fact that by construction, we know that $W_\infty \le \operatorname{ess\,sup}_\pi |x-y|$. Notice that since this is the case, if we can show that $\operatorname{ess\,sup}_\pi |x-y| \le \lim_{p\to\infty} W_p(\mu,\nu) := L$, then we're done.

To see why $\operatorname{ess\,sup}_\pi |x-y| \le \lim_{p\to\infty} W_p(\mu,\nu) := L$, recall the definition of the essential supremum. Notice that

$$\operatorname{ess\,sup}_\pi |x-y| \le L \iff \quad \forall \epsilon>0, \quad \pi(|x-y| > L+\epsilon) = 0$$

Thus, we should consider sets of the form

$$G_\epsilon := \{(x,y) : |x-y| > L+\epsilon\}$$

and try to show that as n gets large the probability of the event G_ϵ occurring gets small.

In particular, let p_n be a sequence such that $\lim_{n\to\infty} p_n \to \infty$, and for each n, let π_n be the optimal coupling for the cost function $|x-y|^{p_n}$ which gives rise to the $W_{p_n}(\mu,\nu)$ distance. Then, since the π_n are compactly supported (which follows since μ and ν are compactly supported), we have that $\pi_n \to \pi^* \in \Pi(\mu,\nu)$ up to a subsequence.

Now, fix $\epsilon > 0$ and consider the set $G_\epsilon := \{(x,y) : |x-y| > L+\epsilon\}$. Clearly on the set G_ϵ, we have

$$|x-y|^{p_n} \geq (L+\epsilon)^{p_n}$$

since we're just exponentiating the criterion for being in G_ϵ. Then, integrating both sides over G_ϵ, we find

$$(L+\epsilon)^{p_n}\pi_n(G_\epsilon) = \int_{G_\epsilon} (L+\epsilon)^{p_n} d\pi_n \leq \int_{G_\epsilon} |x-y|^{p_n} d\pi_n$$

Now, we need some way to compare this to W_{p_n}. Notice that since $W_{p_n} \to L$, we know that for n sufficiently large, we have that

$$W_{p_n}(\mu,\nu) \leq L + \frac{\epsilon}{2}$$

This is exactly the inequality we need, because we can now compare this quantity to $\int_{G_\epsilon} |x-y|^{p_n} d\pi_n$. Notice that other than the domain of integration, $(W_{p_n}(\mu,\nu))^{p_n} = \int_{\mathscr{X}} |x-y|^{p_n} d\pi_n$. Thus, since the function is non-negative on its entire domain we have

$$(L+\epsilon)^{p_n}\pi_n(G_\epsilon) = \int_{G_\epsilon} (L+\epsilon)^{p_n} d\pi_n \leq \int_{G_\epsilon} |x-y|^{p_n} d\pi_n \leq \int_{\mathscr{X}} |x-y|^{p_n} d\pi_n \leq (L+\frac{\epsilon}{2})^{p_n}$$

More simply, this gives

$$(L+\epsilon)^{p_n}\pi_n(G_\epsilon) \leq (L+\frac{\epsilon}{2})^{p_n} \implies \pi_n(G_\epsilon) \leq (\frac{L+\epsilon/2}{L+\epsilon})^{p_n}$$

Then, since $\epsilon > 0$, we have $\frac{L+\epsilon/2}{L+\epsilon} < 1$, so $\lim_{n\to\infty}(\frac{L+\epsilon/2}{L+\epsilon})^{p_n} = 0$. Therefore,

$$0 \leq \lim_{n\to\infty} \pi_n(G_\epsilon) \leq \lim_{n\to\infty} (\frac{L+\epsilon/2}{L+\epsilon})^{p_n} = 0$$

Then, as desired we have for all $\epsilon > 0$, that $\pi_n(G_\epsilon) = 0$. Now, we can apply the Portmanteau theorem to extend this to a claim about π^*. Recall that the Portmanteau Theorem gives equivalent definitions for convergence in distribution [6]. Here recall that it says (in part)

that convergence in distribution is equivalent to the statement that for all measurable open subsets U,

$$\pi^*(U) \leq \liminf_{n\to\infty} \pi_n(U)$$

Now, taking G_ϵ to be our set (which is open because the map $(x,y) \mapsto |x-y|$ is continuous and the preimage of an open set under a continuous function is open), we find that

$$\pi^*(G_\epsilon) \leq \liminf_{n\to\infty} \pi_n(U) \leq 0$$

which says directly that $\pi^*(G_\epsilon) = 0$, so, as described above, we know $\operatorname{ess\,sup}_\pi |x-y| \leq \lim_{p\to\infty} W_p(\mu,\nu) := L$, which says directly that $\lim_{p\to\infty} W_p(\mu,\nu) \geq W_\infty(\mu,\nu)$. Therefore,

$$W_\infty(\mu,\nu) = \lim_{p\to\infty} W_p(\mu,\nu)$$

□

While this is certainly a nice and natural extension of the $W_p(\mu,\nu)$ metric, it has some problems which make it ill-suited to the transport-entropy inequality context.

For one, there is still no known analytic way to find $W_\infty(\mu,\nu)$ where μ and ν are two arbitrary measures on a Polish space [2]. This is because for $p=\infty$ instead of minimizing some integral quantity, you are minimizing an essential supremum, which is in general non-convex in the couplings $\pi \in \Pi(\mu,\nu)$.

The other problem is that this distance penalty is incredibly strict. In effect, the W_∞ distance gives the 'minimax' distance, because the cost function is the largest point-wise distance between the initial configuration of dirt piles$x$ and the final arrangement y. It is not natural that the KL divergence of ν with respect to γ would have to respect that 'worst-case scenario' transport cost, because KL Divergence is a global quantity, so asking it to upper bound a point-wise 'worst case' distance is unnatural. It is perhaps for this reason that the T_∞ case is rarely studied in the transport-entropy literature.

To see this concretely, consider the following simple example.

Example 14. As usual for these examples, we'll work on $\mathcal{X} = \mathbb{R}$ and consider the standard Gaussian measure γ. We'd like to compare the W_∞ distance between $\gamma = \mathcal{N}(0,1)$ and the family of measures $\nu_t = \{\mathcal{N}(0,\sigma^2) : \sigma^2 > 0\}$. We'll also compare this to $D_{KL}(\nu||\gamma)$ to see exactly how strict an inequality of the form $W_\infty(\gamma,\nu) \leq \sqrt{2CD_{KL}(\nu||\gamma)}$ for the family ν_t would be.

In this simple example in 1-D, we can compute the optimal Wasserstein distance between γ and $\nu \in \nu_t$ just by using the Monotone Rearrangement Theorem 2.3.8. In particular, we have

$$W_\infty(\nu,\gamma) = \sup_{t\in[0,1]} |F_\nu^{-1}(t) - F_\gamma^{-1}(t)|$$

Then, recalling that any measures $\nu \in \nu_t = \{\mathcal{N}(0,\sigma^2) : \sigma^2 > 0\}$ is just a scaled Gaussian, we know its inverse CDF is of the form $F_\nu^{-1}(t) = \sigma\Phi^{-1}(t)$. Applying this fact, we know

$$W_\infty(\nu,\gamma) = \sup_{t\in[0,1]} |\sigma\Phi^{-1}(t) - \Phi^{-1}(t)| = |\sigma - 1| \sup_{t\in[0,1]} \Phi^{-1}(t) = \begin{cases} +\infty & \text{for } \sigma \neq 1 \\ 0 & \text{for } \sigma = 1 \end{cases}$$

Now, recall our computation of the KL Divergence between two Gaussian measures in Example 5. Here, clearly we have

$$D_{KL}(\nu||\gamma) = \frac{1}{2}(\sigma^2 - 1 - \log(\sigma^2)) < \infty$$

Thus, in this case, it is clearly not possible to upper bound the infinite $W_\infty(\gamma,\nu)$ distance by the finite KL-Divergence. Intuitively, this corresponds to the fact that in the worst-case, when moving mass between two Gaussians with different variances, we have to move some mass infinitely far, but if we are expecting to see a Gaussian with unit variance, we are not that surprised to see a Gaussian with non-unit variance. This is the problem with comparing 'worst-case scenario' point wise costs with global measurements.

3.3 T_0 and Pinsker's Inequality

The natural thing to consider now is whether the limit as $p \to 0$ is more useful. In fact, it is and gives rise to a very famous and useful inequality, called Pinsker's Inequality, which compares the total variation distance between two measures and the KL-Divergence between them. In order to interpret this as a transport-entropy inequality, we first need to give an interpretation of TV distance as an optimal transport cost which is intimately related to the limit as $p \to 0$ of the Wasserstein-p distance.

In particular, we'll show that Total Variation Distance is exactly $||\mu-\nu||_{TV} = \lim_{p\to 0} W_p(\mu,\nu)^p$ distance. Before we do, however, we'll define total variation distance, show that it is an optimal transport cost, and then, finally, connect it to $\lim_{p\to 0} W_p(\mu,\nu)^p$.

Definition 3.3.1 (Total Variation Distance)**.** The total variation distance between two probability measures μ and ν is given by

$$\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A\in\mathscr{F}} |\mu(A) - \nu(A)|.$$

Before we prove Pinsker's inequality, we need two helpful lemmas which help us connect total variation distance to entropy.

Lemma 3.3.2 (Integral Expression of TV Distance)**.** *For μ, ν two measures on a Polish space $\mathscr{X}$ where $\nu \ll \mu$, we have that*

$$||\nu - \mu||_{TV} = \frac{1}{2}\int_{\mathscr{X}} |f - 1| d\mu$$

where $f = \frac{d\nu}{d\mu}$ the Radon Nikodym derivative.

Proof. To see this, begin with the ordinary definition of TV distance:

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{A \in \mathscr{F}} |\mu(A) - \nu(A)|$$

Now, consider the RHS.

$$|\mu(A) - \nu(A)| = \int_A \frac{d\nu}{d\mu} d\mu - \mu(A) = \int_A (f-1) d\mu$$

Now, take the sup of both sides:

$$\sup_{A \in \mathscr{F}} |\nu(A) - \mu(A)| = \sup_{A \in \mathscr{F}} \int_A (f-1) d\mu$$

Now, notice that in order to maximize the RHS, we should pick $A := \{x : f(x) > 1\}$, that is all the points where ν puts more mass than μ. The, we have

$$\sup_{A \in \mathscr{F}} |\nu(A) - \mu(A)| = \sup_{A \in \mathscr{F}} \int_{\{f>1\}} (f-1)^+ d\mu$$

Then, notice that the positive part and the negative part have equal area, thus we have

$$\int_{\{f>1\} \cup \{f<1\}} f - 1 d\mu = 0 \implies \int_{\{f>1\}} (f-1)^+ d\mu = \int_{\{f<1\}} (f-1)^- d\mu$$

Therefore, we have

$$\sup_{A \in \mathscr{F}} |\nu(A) - \mu(A)| = \frac{1}{2} \int_{\mathscr{X}} |f-1| d\mu$$

□

Lemma 3.3.3 (Total Variation Distance is an Optimal Transport Cost)**.** *Let μ, ν be probability measures on the same measurable space $(\mathscr{X}, \mathscr{F})$. Then*

$$\|\mu - \nu\|_{TV} = \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathscr{X} \times \mathscr{X}} \mathbf{1}_{\{x \neq y\}} \, d\pi(x,y).$$

Proof. This proof is given in [24] and has been lightly adapted to our notation.
We show the equality by proving matching upper and lower bounds. First, we show that

$$\|\mu - \nu\|_{TV} \leq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathscr{X} \times \mathscr{X}} \mathbf{1}_{\{x \neq y\}} \, d\pi(x,y).$$

Let $\pi \in \Pi(\mu,\nu)$ be arbitrary, and let $(X,Y) \sim \pi$. Then

$$\int_{\mathscr{X} \times \mathscr{X}} \mathbf{1}_{\{x \neq y\}} \, d\pi(x,y) = \mathbb{E}_\pi \left[\mathbf{1}_{\{X \neq Y\}} \right] = \pi(X \neq Y).$$

Now let $A \in \mathscr{F}$. Since π has marginals μ and ν,

$$\mu(A) - \nu(A) = \mathbb{E}_\pi [\mathbf{1}_A(X) - \mathbf{1}_A(Y)] .$$

Moreover, on the event $\{X = Y\}$ we have $\mathbf{1}_A(X) - \mathbf{1}_A(Y) = 0$, so

$$\mu(A) - \nu(A) = \mathbb{E}_\pi\left[(\mathbf{1}_A(X) - \mathbf{1}_A(Y))\mathbf{1}_{\{X \neq Y\}}\right].$$

Hence

$$\begin{aligned}
|\mu(A) - \nu(A)| &= \left|\mathbb{E}_\pi\left[(\mathbf{1}_A(X) - \mathbf{1}_A(Y))\mathbf{1}_{\{X \neq Y\}}\right]\right| \\
&\leq \mathbb{E}_\pi\left[|\mathbf{1}_A(X) - \mathbf{1}_A(Y)|\,\mathbf{1}_{\{X \neq Y\}}\right] \\
&\leq \mathbb{E}_\pi\left[\mathbf{1}_{\{X \neq Y\}}\right] \\
&= \pi(X \neq Y).
\end{aligned}$$

Taking the supremum over $A \in \mathscr{F}$, we obtain

$$\|\mu - \nu\|_{TV} \leq \pi(X \neq Y).$$

Since $\pi \in \Pi(\mu, \nu)$ was arbitrary, it follows that

$$\|\mu - \nu\|_{TV} \leq \inf_{\pi \in \Pi(\mu,\nu)} \pi(X \neq Y) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathscr{X} \times \mathscr{X}} \mathbf{1}_{\{x \neq y\}}\, d\pi(x, y).$$

Now, in the other direction, we want to show that $||\mu - \nu||_{TV} \geq \int_{\mathscr{X} \times \mathscr{X}} \mathbf{1}_{\{x \neq y\}}\, d\pi(x, y) = \mathbb{E}_\pi\left[\mathbf{1}_{\{X \neq Y\}}\right] = \pi(X \neq Y)$.

To see this we begin by decomposing $||\mu - \nu||_{TV}$ analogously to Lemma 3.3.2. In particular, we note that if μ and ν are dominated by a common measure λ (if we're working on $\mathbb{R}^n$ that would be the Lebesgue measure). Let g and h be the densities corresponding to μ and ν respectively.

$$\begin{aligned}
L = ||\mu - \nu||_{TV} &= \int_{\mathscr{X}} (g - h)^+ d\lambda \\
&= \int_{\mathscr{X}} (\min(g(x), h(x)) + (g - h)^+ d\lambda = 1 - \int_{\mathscr{X}} \min(g(x), h(x)) d\lambda(x)
\end{aligned}$$

The idea here is to decompose each density into a common part and an excess part unique to μ or ν respectively. Then, we couple the common part of μ and ν along the 'diagonal' where $x = y$ which costs nothing and then couple the un-matched mass, which has total mass exactly L. In particular, we define the coupling

$$\begin{aligned}
\pi(dx, dy) &= L\pi_1(dx, dy) + (1 - L)\,\pi_2(dx, dy) \\
&= L\left(\frac{(g - h)^+(x)}{L}\,\lambda(dx)\,\frac{(h - g)^+(y)}{L}\,\lambda(dy)\right) + (1 - L)\left(\frac{\min\{g(x), h(x)\}}{1 - L}\,\lambda(dx)\,\delta_x(dy)\right).
\end{aligned}$$

This is clearly a measure on $\mathscr{X} \times \mathscr{X}$ since it is a convex combination of probability measures and L corresponds to the mass of the un-matched part, while $1 - L$ corresponds to the mass of the matched part. It is easy to verify that the marginals are μ and ν respectively, so it is clearly a valid coupling.

Now, we need to compute the optimal transport distance with the discrete cost $\mathbb{I}(x \neq y)$. We've already shown

$$\int_{\mathscr{X}\times\mathscr{X}} \mathbf{1}_{\{x\neq y\}}\, d\pi(x,y) = \mathbb{E}_\pi\left[\mathbf{1}_{\{X\neq Y\}}\right] = \pi(X \neq Y) = L\pi_1(X \neq Y) \leq L$$

Since π_2 has mass concentrated on the diagonal $X = Y$.

Therefore, we have exactly that

$$\inf_{\tilde{\pi}\in\Pi(\mu,\nu)} \tilde{\pi}(X \neq Y) \leq L = ||\mu - \nu||_{TV}$$

Combining this with the first bound, we get the desired equality

$$\|\mu - \nu\|_{TV} = \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathscr{X}\times\mathscr{X}} \mathbf{1}_{\{x\neq y\}}\, d\pi(x,y)$$

□

Now, we have all the materials to relate this metric to $\lim_{p\to 0} W_p(\mu,\nu)^p$. Crucially, here, we're relating this to the transport cost, not the standard Wasserstein distance. This lemma was suggested by a comment in [15].

Lemma 3.3.4. *Let $(\mathscr{X}, d)$ be a Polish space with metric $d(x,y) = |x-y|$. Then,*

$$\lim_{p\to 0} W_p(\mu,\nu)^p = \lim_{p\to 0} T_p(\mu,\nu) = \lim_{p\to 0} \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathscr{X}\times\mathscr{X}} d(x,y)^p d\pi(x,y) = ||\mu - \nu||_{TV}$$

for μ, ν with at least finite first moments.

Proof. The proof of this lemma follows very closely the proof of Lemma 3.2.10. They are both examples of standard technique.

Let p_n be a non-negative sequence converging to 0 where $0 \leq p_n < 1$.

First, we'll show the upper bound. Recall by Lemma 3.3.3 that $||\mu-\nu||_{TV} = \inf_{\pi\in\Pi(\nu,\mu)} \pi(X \neq Y)$. Therefore, for any $\epsilon > 0$, we can pick $\pi^\epsilon \in \Pi(\mu,\nu)$ such that

$$\pi^\epsilon(X \neq Y) \leq ||\mu - \nu||_{TV} + \epsilon$$

Now, as mentioned in Section 6 of [15], we know that $d(x,y)^{p_n} \to \mathbb{I}(X \neq Y)$ point wise. Then, dominated convergence under the fixed coupling π_ϵ gives

$$\limsup_{n\to\infty} T_{p_n}(\mu,\nu) \leq \int_{\mathscr{X}\times\mathscr{X}} \mathbb{I}(x \neq y) d\pi^\epsilon(x,y) \leq ||\mu - \nu||_{TV} + \epsilon$$

Then, taking the limit as $\epsilon \to 0$, we find

$$\limsup_{n\to\infty} T_{p_n}(\mu,\nu) \leq ||\mu - \nu||_{TV}$$

which gives the desired upper bound.

Now, we'll show the lower bound. Choose a sequence of 'almost-minimizing' couplings π_n such that for each π_n, we have

$$\int_{\mathscr{X}} d(x,y)^{p_n} d\pi_n(x,y) \leq T_{p_n}(\mu,\nu) + \frac{1}{n}$$

Now, notice that since $\mathscr{X}$ is a Polish space we have that a subsequence $\pi_{n_k} \to \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{K}(\pi)$.

Next, we'll fix some $\delta > 0$. Notice that we have

$$d(x,y)^{p_n} \geq \delta^{p_n} \mathbb{I}(d(x,y) \geq \delta)$$

Then, integrating both sides, we have

$$\int_{\mathscr{X}\times\mathscr{X}} d(x,y)^{p_n} d\pi_n(x,y) \geq \int_{\mathscr{X}\times\mathscr{X}} \delta^{p_n} \mathbb{I}(d(x,y) \geq \delta) d\pi_n(x,y) = \delta^{p_n} \pi_n(d(x,y) \geq \delta)$$

Now, consider the limit as $p_n \to 0$. Clearly $\lim_{n\to\infty} \delta^{p_n} = 1$. We now want to apply the Portmanteau Theorem 2.1.12. In this form, it says that weak convergence (ie $\pi_n \xrightarrow{d} \pi$) is equivalent to the fact that for all measurable closed subsets C, $\limsup_{n\to\infty} \pi_n(C) \leq \pi(C)$. Observe here that the set of interest is

$$\{(x,y) : d(x,y) \geq \delta\}$$

is the preimage of a closed set under a continuous function and so is itself closed. Therefore we know that

$$\liminf_{n\to\infty} \int_{\mathscr{X}\times\mathscr{X}} d(x,y)^{p_n} d\pi_n(x,y) \geq \pi(\{(x,y) : d(x,y) \geq \delta\})$$

Then, taking the limit as $\delta \to 0$, we have

$$\liminf_{n\to\infty} \int_{\mathscr{X}\times\mathscr{X}} d(x,y)^{p_n} d\pi_n(x,y) \geq \pi(\{(x \neq y\})$$

since $\lim_{\delta\to 0}\{(x,y) : d(x,y) \geq \delta\} = \{(x,y) : x \neq y\}$.

Now, notice that the mass π assigns to the so-called 'diagonal' part can't exceed the mass that is actually shared by the two marginals μ and ν. Therefore, we have exactly that

$$\pi(\{x = y\}) \leq 1 - ||\mu - \nu||_{TV} \implies \pi(\{x \neq y\}) \geq ||\mu - \nu||_{TV}$$

Thus, we have that

$$\liminf_{n\to\infty} \int_{\mathscr{X}\times\mathscr{X}} d(x,y)^{p_n} d\pi_n(x,y) \geq \pi(\{(x \neq y\}) \geq ||\mu - \nu||_{TV}$$

as desired. Thus we have the upper and lower bounds which completes the proof. □

A nice corollary of this result is that Pinsker's Inequality can be seen (roughly speaking) as T_0 inequality, in the sense that

$$W_{d_0}(\nu,\mu)^2 = ||\mu-\nu||_{TV}^2 \leq \frac{1}{2}H(\nu|\mu)$$

This T_0 notation is not standard in the field, but is beginning to be more common among papers like [4] which interprets TV, as we do, as a Wasserstein distance. More formally, we would say Pinsker's inequality is the discrete-metric analogue of a Talagrand transport inequality.

Theorem 3.3.5 (Pinsker's Inequality)**.** *The inequality*

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

or equivalently

$$\|\nu-\mu\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2}H(\nu|\mu)}$$

holds for all probability measures μ on $\mathscr{X}$.

Proof. Here, we'll use the two lemmas above to prove Pinsker's Inequality 3.3.5. This proof is adapted from the version given in [18], which is in turn taken from Remark 22.12 and Theorem 22.10 in [41]. There are many nice proofs of this fact, which are summarized in Section 1 of [10]. I've presented the one below because it is relatively self-contained, but there is a very nice proof in [43] which makes use of the so-called "data-processing" inequality.

The key idea of this proof is to use Lemma 3.3.2 to rewrite the LHS. Let $f = \frac{d\nu}{d\mu}$ be the Radon-Nikodym derivative. In particular, we have

$$||\mu-\nu||_{TV} = \frac{1}{2}\int_{\mathscr{X}} |f-1|d\mu$$

Then, we can notice that the entropy term, which is the term we want to use to bound this total-variation distance, has a similar algebraic form.

$$H(\nu|\mu) = \int_{\mathscr{X}} f\log(f)d\mu$$

We'll exploit this algebraic similarity to actually write the proof using Taylor expansion and the Cauchy-Schwarz inequality.

Let's begin first by supposing that $H(\nu|\mu) < \infty$, since otherwise, this is trivially true. We're going to take the upper bound $H(\nu|\mu)$ and try to algebraically extract the quantity $||\mu-\nu||_{TV} = \frac{1}{2}\int_{\mathscr{X}}|f-1|d\mu$ from it. We'll begin to do this by u-substituting $u = f-1$. Therefore

$$H(\nu|\mu) = \int_{\mathscr{X}} f\log(f)\,d\mu = \int (1+u)\log(1+u) - u\,d\mu = \int \phi(u(x))\,d\mu$$

Now, consider the function $\phi(t) = (1+t)\log(1+t) - t$. We'll compute the derivatives

$$\phi'(t) = \log(1+t)$$

$$\phi''(t) = \frac{1}{t+1} \quad \text{for } t > -1$$

Now, we'll use Taylor's Theorem with remainder at $x = 0$ to algebraically relate this quantity to $\int_{\mathscr{X}} |f-1|\,d\mu$.

$$\phi(t) = \phi(t)|_{t=0} + \phi'(t)|_{t=0}(0-t) + \int_0^t (t-x)\phi''(x)\,dx$$

Now, notice that $\phi(0) = (1+0)\log(1+0) - 0 = 0$ and $\phi'(0) = \log(1+0) = 0$. Therefore, we have

$$\phi(t) = (1+t)\log(1+t) - t = \int_0^t (t-x)\phi''(x)\,dx = \int_0^t \frac{t-x}{1+x}\,dx$$

We need to do one more u-substitution before we'll put this $\phi(t)$ back into the entropy expression $H(\nu|\mu)$. Let $s = \frac{x}{t} \iff x = st$ where $s \in [0,1]$. Then, we have that

$$\phi(t) = \int_0^1 \frac{t-st}{1+st}\,t\,ds = t^2 \int_0^1 \frac{1-s}{1+st}\,ds$$

Now that we've made $\phi(t)$ easier to relate algebraically, we'll substitute back in. In particular, we have

$$H(\nu|\mu) = \int_{\mathscr{X}} \phi(u(x))\,d\mu = \int_{\mathscr{X}} \int_{[0,1]} u(x)^2 \frac{1-s}{1+su(x)}\,ds\,d\mu(x)$$

Now, we'll apply the Cauchy-Schwarz inequality (which is exactly Hölder's Inequality 3.2.6 with $p = q = 2$). Recall that (with very informal notation) says that $(\int gh)^2 \le \int g^2 \cdot \int h^2$. Now, we need to take a nice product form that is going to decompose into two components, one of which, when squared is exactly $u(x)^2\frac{1-s}{1+su(x)}$. This motivates the choice of our g and h.

$$g(x,s) = \frac{|u(x)|\sqrt{1-s}}{\sqrt{1+su(x)}} \implies g^2(x,s) = \frac{u(x)^2(1-s)}{1+su(x)}$$

$$h(x,s) = \sqrt{(1-s)(1+su(x))}$$

Then,

$$g(x,s)\cdot h(x,s) = |u(x)|(1-s)$$

Now, we'll apply the Cauchy-Schwarz inequality

$$\left(\int_{\mathscr{X}} \int_{[0,1]} |u(x)|(1-s)\,ds\,d\mu(x)\right)^2 \le \left(\int_{\mathscr{X}} \int_{[0,1]} \frac{u^2(x)(1-s)}{1+su(x)}\,ds\,d\mu(x)\right)$$

$$\cdot \left(\int_{\mathscr{X}} \int_{[0,1]} (1-s)(1+su(x))\, ds\, d\mu(x) \right).$$

Then, note that this first term is exactly

$$\left(\int_{\mathscr{X}} \int_{[0,1]} \frac{u^2(x)(1-s)}{1+su(x)} ds d\mu(x) \right) = \int_{\mathscr{X}} f \log f \, d\mu = H(\nu|\mu)$$

by the above.

Now, observe that

$$\begin{aligned}
&\int_{\mathscr{X}} \int_0^1 (1-s)(1+s\,u(x))\, ds\, d\mu(x) \\
&= \int_{\mathscr{X}} \left(\int_0^1 (1-s)\, ds + u(x) \int_0^1 s(1-s)\, ds \right) d\mu(x) \\
&= \int_{\mathscr{X}} \left(\frac{1}{2} + \frac{1}{6} u(x) \right) d\mu(x) \\
&= \frac{1}{2}.
\end{aligned}$$

Therefore, we have that

$$\left(\int_{\mathscr{X}} \int_{[0,1]} |u(x)|(1-s) ds d\mu(x) \right)^2 \leq \frac{1}{2} H(\nu|\mu)$$

Now, it remains to relate the LHS to the total variation distance. Note that substituting back $u(x) = f(x) - 1$, we find

$$\begin{aligned}
&\left(\int_{\mathscr{X}} \int_0^1 |u(x)|(1-s)\, ds\, d\mu(x) \right)^2 \\
&= \left(\int_{\mathscr{X}} \int_0^1 |f(x)-1|(1-s)\, ds\, d\mu(x) \right)^2 \\
&= \left(\int_{\mathscr{X}} |f(x)-1| \left(\int_0^1 (1-s)\, ds \right) d\mu(x) \right)^2 \\
&= \left(\frac{1}{2} \int_{\mathscr{X}} |f(x)-1|\, d\mu(x) \right)^2 \qquad \text{by Lemma 3.3.2} \\
&= ||\mu - \nu||_{TV}^2
\end{aligned}$$

Thus we have exactly as desired that

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

□

There is another very nice, but less well known inequality of this flavor which is introduced in [8] and described very well in [10]. This inequality improves on Pinsker's Inequality 3.3.5 in the sense that the total variation distance is bounded above by 1, by constriction, ie for any two measures μ and ν we must have that

$$||\mu - \nu||_{TV} \leq 1 \implies ||\mu - \nu||_{TV}^2 \leq 1.$$

However, as we've seen often with KL-divergence 2.2.4, it is unbounded. In particular, for any $H(\mu|\nu) > 1$, Pinsker's Inequality 3.3.5 is vacuously true. This suggests we can do better, but, as it turns out, not simply by changing the constant in Theorem 3.3.5. Instead, we need to venture beyond the extended $T_p(C)$ inequality universe, and employ a more general transport-entropy inequality of the more general form described in Definition 3.3.5.

This theorem is originally stated in [8], Lemma 2.1, but I found it in [10] and follow that presentation.

Theorem 3.3.6 (Bretagnolle-Huber Inequality)**.** *Let $\mathscr{X}$ be a Polish space and μ, ν two measures on $\mathscr{X}$. Then*

$$||\mu - \nu||_{TV} \leq \sqrt{1 - e^{-H(\nu|\mu)}}$$

Remark 11 (BH Bound is a transport-entropy Inequality). Note that we can rewrite Theorem 3.3.6 as a transport-entropy inequality in the form of Definition 3.1.1. In particular, after a bit of algebra,it is clear that

$$||\mu - \nu||_{TV} \leq \sqrt{1 - e^{-H(\nu|\mu)}} \iff -\log(1 - ||\mu - \nu||_{TV}^2) \leq H(\nu|\mu)$$

which is clearly of the form

$$\alpha(T_c(\nu, \mu)) \leq H(\nu|\mu)$$

where we take $\alpha(t) := -\log(1 - t^2)$, which verifies $\alpha(0) = 0$, $c(x, y) = \mathbb{I}(x \neq y)$ as discussed in Lemma 3.3.3, and $C = 1$.

It just remains to prove this statement.

Proof. As with Pinsker's Inequality, there are many ways to prove this statement, but I think that the proof given in [10] which adapts a proof given of Lemma 2.6 in [40]. This proof mirrors the style of our proof of Theorem 3.3.5, which uses a nice expression of the TV distance and applies Cauchy-Schwarz to reach the desired conclusion.

We begin by noting that for two measures μ, ν, with corresponding densities g and h, that are dominated by a common measure λ we have (as demonstrated in the proof of Lemma 3.3.3)

$$||\mu - \nu||_{TV} = 1 - \int \min(g(x), h(x)) d\lambda(x) = \int \max(g(x), h(x)) d\lambda(x) - 1$$

Then we can nicely express

$$\begin{aligned}
&\|\mu-\nu\|_{\mathrm{TV}}^2 \\
&= 1-(1+\|\mu-\nu\|_{\mathrm{TV}})(1-\|\mu-\nu\|_{\mathrm{TV}}) \\
&= 1-\left(\int_{\mathscr{X}} \min(g(x),h(x))\,d\lambda(x)\right)\left(\int_{\mathscr{X}} \max(g(x),h(x))\,d\lambda(x)\right).
\end{aligned}$$

If we can bound $\left(\int_{\mathscr{X}} \min(g(x),h(x))d\lambda(x)\right)\cdot\left(\int_{\mathscr{X}} \max(g(x),h(x))d\lambda(x)\right)$ above by $e^{-H(\nu|\mu)}$ then we're done.
To do this, recall the informal statement the Cauchy Schwarz inequality: $(\int vw)^2 \le \int v^2 \cdot \int w^2$. If we take

$$w = \sqrt{\max(g(x),h(x))}$$

and

$$v = \sqrt{\min(g(x),h(x))} \implies v\cdot w = \sqrt{\max(g(x),h(x))\cdot\min(g(x),h(x))}$$

We have

$$\begin{aligned}
\left(\int_{\mathscr{X}} \sqrt{\max(g(x),h(x))\,\min(g(x),h(x))}\,d\lambda(x)\right)^2 \le &\int_{\mathscr{X}} \max(g(x),h(x))\,d\lambda(x) \\
&\cdot\int_{\mathscr{X}} \min(g(x),h(x))\,d\lambda(x).
\end{aligned}$$

Crucially, we can recognize that $\sqrt{\max(g(x),h(x))\cdot\min(g(x),h(x))} = \sqrt{g(x)h(x)}$ since multiplication is commutative. Therefore, we have

$$\left(\int_{\mathscr{X}} \sqrt{g(x)h(x)}\quad d\lambda(x)\right)^2 \le \int_{\mathscr{X}} \max(g(x),h(x))d\lambda(x)\cdot\int_{\mathscr{X}} \min(g(x),h(x))d\lambda(x)$$

Now we can cleverly upper bound $\left(\int_{\mathscr{X}} \sqrt{g(x)h(x)}\quad d\lambda(x)\right)^2$ to finish the proof. We begin by rewriting the expression as an exponential, since in the end we want to bound the quantity by $e^{-H(\nu|\mu)}$.

$$\begin{aligned}
&\left(\int_{\mathscr{X}} \sqrt{g(x)h(x)}\,d\lambda(x)\right)^2 \\
&= e^{2\log\mathbb{E}_\lambda\left[\sqrt{h(x)g(x)}\right]} \\
&= e^{\log\left(\mathbb{E}_\mu\left[\sqrt{\frac{h(x)}{g(x)}}\right]\right)^2} \\
&\text{by Jensen's Inequality} \\
&\ge e^{\mathbb{E}_\mu\log\left[\frac{h(x)}{g(x)}\right]} = e^{-H(\nu|\mu)}
\end{aligned}$$

Therefore, we have that

$$\left(\int_{\mathscr{X}} \min(g(x),h(x))\,d\lambda(x)\right)\left(\int_{\mathscr{X}} \max(g(x),h(x))\,d\lambda(x)\right) \ge e^{-H(\nu|\mu)}$$

Thus, as desired, we have

$$\begin{aligned}&\|\mu-\nu\|_{\mathrm{TV}}^2\\&\leq 1-e^{-H(\nu|\mu)} \implies \|\mu-\nu\|_{\mathrm{TV}} \leq \sqrt{1-e^{-H(\nu|\mu)}}\end{aligned}$$

□

Note that this bound is never vacuous, since

$$\lim_{H(\nu|\mu)\to\infty} \sqrt{1-e^{-H(\nu|\mu)}} = 1$$

3.4 Concentration Inequalities and Marton's Argument

For readers familiar with probability theory, the previous sections may seem to be missing something crucial. In particular, the intuition for transport-entropy inequalities (namely that the optimal way to rearrange mass from one configuration μ to another ν should be related to how surprised we are to observe ν if we expected μ seems to be demand a connection to concentration inequalities.

3.4.1 Concentration Inequalities

Definition 3.4.1 (Concentration Inequality [18])**.** Let $\mathscr{X}$ be a Polish space and d some metric on $\mathscr{X}$. Further, let $\beta : [0,\infty) \to \mathbb{R}^+$ be a 'profile' such that $\lim_{r\to\infty}\beta(r) = 0$. We say that a measure μ verifies a concentration equality with profile β if

$$\mu(\{x \in \mathscr{X} : d(x,A) \leq r\}) \geq 1-\beta(r) \quad \text{for } r \geq 0$$

for all measurable $A \subseteq \mathscr{X}$ such that $\mu(A) \geq \frac{1}{2}$

Theorem 3.4.2 (Marton's Argument)**.** *Consider a transport-entropy inequality for measure μ on some measure space $\mathscr{X}$ of the form*

$$\mathscr{T}_d(\mu,\nu) \leq \alpha^{-1}(H(\nu|\mu)) \quad \textit{for all } \nu \in \mathscr{P}(\mathscr{X})$$

where $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a bijection such that $\alpha(0) = 0$. In particular $\mathscr{T}_d(\mu,\nu)$ is the total optimal transport cost for moving μ to ν where the cost function $d(x,y)$ is a metric d on $\mathscr{X}$. Then, for all measurable subsets $A \subseteq \mathscr{X}$ such that $\mu(A) \geq \frac{1}{2}$, the following concentration inequality holds

$$\mu(A^r) \geq 1-e^{-\alpha(r-\alpha^{-1}(\log(2)))}$$

where $r \geq \alpha^{-1}\log(2)$ and $A^r := \{x \in \mathscr{X} : d(x,A) \leq r\}$.

Proof of Marton's Argument. We'll prove this statement by considering two measures, one which corresponds to the mass in A and one which corresponds to the mass not contained in the ball A^r. The idea is to apply the fact that d is a metric and therefore has obeys the triangle inequality. This behavior allows us to bound the mass not contained in A^r.

Formally, fix some arbitrary measurable set $A \subset \mathscr{X}$ which has at least half the total mass (ie. $\mu(A) \geq \frac{1}{2}$). Consider the complement of the enlargement of A defined as $B := (A^r)^c = \mathscr{X} \setminus A^r$. Define the following two measures

$$d\mu_A(x) = \frac{1}{\mu(A)} \mathbb{I}(x \in A) d\mu(x)$$

$$d\mu_B(x) = \frac{1}{\mu(B)} \mathbb{I}(x \in B) d\mu(x)$$

Now notice for any coupling π of μ_A and μ_B we have

$$\int d(x,y) d\pi(x,y) \geq r$$

since if $x \in A$ and $y \in B$, $d(x,y) \geq d(y,A) \geq r$. Therefore, the optimal coupling must have cost at least r if for every possible coupling the cost is at least r.

$$\mathscr{T}_d(\mu_A, \mu_B) = \inf_{\pi \in \Pi(\mu_A, \mu_B)} \left(\int_{\mathscr{X}} d(x,y) d\pi(x,y) \right) \geq r$$

Then, we'll first take advantage of the fact that d is a metric. Thus, by the triangle inequality we know for any set of three points $(x,y,z) \in \mathscr{X}^3$ we have $d(x,y) \leq d(x,z) + d(z,y)$. Then consider the measure γ on $\mathscr{X}^3$ which has (x,z)-marginal $\pi_1 \in \Pi(\mu_A, \mu)$ and $(z,y)-$marginal $\pi_2 \in \Pi(\mu, \mu_B)$. We know such a measure exists by the gluing lemma. Then, we integrate both sides of the point-wise triangle inequality, which gives

$$\int d(x,y) d\gamma \leq \int d(x,z) d\gamma + \int d(z,y) d\gamma$$

Then, using the marginals, we get

$$\int d(x,y) d\tilde{\pi}(x,y) \leq \int d(x,z) d\pi_1 + \int d(z,y) d\pi_2$$

where $\tilde{\pi}$ is a coupling of μ_A and μ_B. Then taking the infimum over all such π_1 and π_2 we have

$$\mathscr{T}_d(\mu_A, \mu_B) \leq \mathscr{T}_d(\mu_A, \mu) + \mathscr{T}_d(\mu, \mu_B)$$

Putting it all together, we have

$$r \leq \mathscr{T}_d(\mu_A, \mu_B) \leq \mathscr{T}_d(\mu_A, \mu) + \mathscr{T}_d(\mu, \mu_B) \leq \alpha^{-1}(H(\mu_A|\mu)) + \alpha^{-1}(H(\mu_B|\mu))$$

where the final inequality follows by the assumption that μ satisfies a transport-entropy inequality with constant C and so, taking $\nu = \mu_A$ in the first case and $\nu = \mu_B$ in the second gives the above.
Then, finally, we'll use the definition of relative entropy given in Definition 2.2.4. In particular

$$H(\mu_A|\mu) = \int \log\left(\frac{d\mu_A}{d\mu}\right) d\mu_A = \int_A \log\left(\frac{1}{\mu(A)}\right) d\mu_A = \log\left(\frac{1}{\mu(A)}\right)\mu_A(A) = \log\left(\frac{1}{\mu(A)}\right)$$

Then, by hypothesis since $\mu(A) \geq \frac{1}{2}$, therefore

$$H(\mu_A|\mu) = -\log(\mu(A)) \leq \log(2)$$

An identical computation applies to $\mu(B)$. Therefore

$$H(\mu_B|\mu) = -\log(\mu(B)) = -\log(1-\mu(A^r))$$

Therefore

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

Applying α to both sides, multiplying by -1 and exponentiating gives

$$e^{-\alpha(r)} \geq e^{\log(2)+\log(1-\mu(A^r))}$$

Then, re-arranging, we get the desired concentration inequality

$$\mu(A^r) \geq 1 - e^{-\alpha(r-\alpha^{-1}(\log(2)))}$$

□

Chapter 4

Log-Sobolev Inequalities and Transport-Entropy Inequalities

4.1 Motivation

As we've shown in Chapter 3, Transport-Entropy inequalities are an incredibly useful tool for understanding how distributions compare. However, we've also seen that these inequalities are difficult to prove, especially in high-dimensional settings. In Chapter 3, We explored one potential (but unsuccessful) way to prove a large class of transport-entropy inequalities: namely, observing in Theorem 3.2.5 that for $p \geq q, T_p(C) \implies T_q(C)$ and trying to prove a transport-entropy inequality for the limit $p \to \infty$. This, as we've demonstrated, does not work. It is far too strict in the point-wise sense to be useful. However there is something to learn from this failure.

In particular, we do want to use a stronger, more local statement to control global deviations, and we want this stricter condition to imply the transport-entropy inequalities of interest. As it turns out, the correct tool in this setting is a log-Sobolev inequality.
This chapter will first introduce the tools and structures needed to define a log-Sobolev inequality precisely, namely Markov Processes and Dirichlet Forms. Then, in Section 4.2, we offer three examples of measures which satisfy log-Sobolev inequalities. Finally, in Section 4.3, we introduce and sketch a proof of the Otto-Villani theorem, which shows how Log-Sobolev inequalities can be used to derive a certain class of transport-entropy inequalities.

4.1.1 Log-Sobolev Inequalities

Definition 4.1.1 (General Log-Sobolev Inequality [18])**.** We say that a probability measure μ on $\mathscr{X}$ satisfies a log Sobolev inequality for a constant $C > 0$ if

$$H(f\mu|\mu) \leq 2CI(f\mu|\mu)$$

for all sufficiently smooth, non-negative functions $f : \mathcal{X} \to \mathbb{R}^+$ such that $\int f d\mu = 1$.

Remark 12 (Note on Constant Conventions). Some authors use $\tilde{C} = 2C$ as their constant. For the rest of this chapter I'll use the $2C$ convention because it means that Gaussian measures satisfy a $LS(1)$ inequality, which is simple to remember and it is consistent with the convention to preserve the 2 in our Transport-Entropy inequalities in Chapter 3. Similarly, some authors prefer a different normalization convention for $I(f\mu|\mu)$ which we will discuss in greater detail in Section 4.1.3.

Intuitively what this inequality says is that the relative entropy of ν with respect to μ is bounded above by the some function of the local-roughness of ν. This should make some sense: a density should not be able to distribute mass in a very non-uniform way without having local roughness.

This differs from transport-entropy inequalities in a few key ways. The first and most obvious is that in $T_p(C)$ inequalities (recall Definition 3.1.1), we upper bound the Wasserstein distance with the entropy, rather than upper bounding the entropy by some function of the local roughness of the density. Second, and more subtly, as we've discussed, T_p inequalities compare two measures by studying two different notions of global difference, while here we're comparing a global quantity (relative entropy) to a local one (information).

The natural question, then, is why such a local statement should imply a global transport-entropy inequality. The heuristic answer is dynamical. In particular, as you learn in calculus, the way to go from local statements to global ones is to integrate. We can think (heuristically) of the time-dependent process which causes some measure ν (such that $\nu \ll \mu$) to evolve over time with limiting distribution μ. This produces a path $(\nu_t)_{t\geq 0}$ from ν to equilibrium μ. Along this flow, the information governs the instantaneous decay of entropy:

$$-\frac{d}{dt} H(\nu_t \mid \mu) = I(\nu_t \mid \mu)$$

A Log-Sobolev inequality says that $H(\nu_t|\mu) \leq CI(\nu_t|\mu)$, which means exactly that whenever entropy is large, the rate of change is large. This suggests that there is a strong push to be close to the stationary distribution μ. If we imagine integrating both sides, we can think of the stochastic process as describing a 'path' in Wasserstein space. In particular, the total transport cost from ν to μ can be bounded by integrating this local quantity along the flow. Since log-Sobolev controls entropy by information, it forces the flow to reach equilibrium efficiently enough that the total transport cost displacement is bounded by the initial entropy.

This is a rough sketch of the intuition which we will formalize in the following chapter. The first step is to clarify the new ideas and terms we've introduced in Definition 4.1.1.

In order to make the above heuristics formal, there are quite a few new constructions involved in this definition which we'll define carefully before moving on to the application of these inequalities.

First, we need to account for the fact that our heuristic description introduced a stochastic process which evolves over time, while in the definition, we only consider a fixed measure μ and a set of functions. This is because embedded in the definition of a Log-Sobolev inequality is an implied stochastic process, which we use to define I. In particular, we need to formalize the class of stochastic processes that we will work with and explain a few of their relevant properties.

4.1.2 Markov Processes

This section is intended to introduce readers familiar with the basic theory of Markov chains to a somewhat more abstract view of these processes through the lens of semigroups and generators.

Since we are working with a process which sends $\nu \to \mu$, we need to formalize and clarify the properties of such a process. One desirable property is that the future of ν_t should only depend on its present state, rather than the entirety of the past.

Definition 4.1.2 (Markov Process)**.** A stochastic process $\{X_t\}_{t\geq 0}$ is called a Markov Process if for every n and $t_1 < t_2 < \cdots < t_n$, we have that

$$X_{t_n}|X_{t_{n-1}},...,X_{t_1} \overset{d}{=} X_{t_n}|X_{t_{n-1}}.$$

Similarly, we require that this process has μ as an equilibrium (also called stationary) distribution and this Markov process is $\mu-$reversible.

Definition 4.1.3 (Stationary Distribution)**.** [13] A measure μ on $\mathscr{X}$ is said to be a stationary measure if

$$\mu(A) = \int_{\mathscr{X}} P_t(x, A)\mu(dx) \quad \text{for every measurable } A \subseteq \mathscr{X}.$$

Equivalently, and more intuitively, if $X_0 \sim \mu$ then for all $t \geq 0$, $X_t \sim \mu$.

Though it may be more familiar to describe a Markov process by its transition kernel (or, in discrete time, its transition matrix), a more general formulation is in terms of semi-groups and generators. For an algebraist, a semi-group is just a set with a binary associative operation, but in the context of Markov processes, it means something a bit more specific. First, we consider the set over which we operate. The elements of the set are so-called Markov operators.

Definition 4.1.4 (Markov Operator [3])**.** Consider a space $(\mathscr{X},\mathscr{F})$ and a set of real, measurable functions $W = \{f : (\mathscr{X},\mathscr{F}) \to (\mathbb{R},\mathscr{B}(\mathbb{R}))\}$. Then, an linear operator P on W (i.e. an operator P such that $P(af+\tilde{f}) = aP(f)+P(\tilde{f})$). Then, we say that P on W is a Markov operator if for all $f \in W$ such that f is bounded and measurable, $P(f)$ is also bounded and measurable. We also require $P(1) = 1$ where 1 is the constant function $x \mapsto 1$. Then, finally we require that for all $f \geq 0$, we have $P(f) \geq 0$.

Now, as introduced in [3], consider the family of Markov operators indexed by $t \geq 0$: $\mathscr{P} = \{P_t\}_{t\geq 0}$ which describe the family of distributions of a Markov process $\{X_t\}_{t\geq 0}$ on some measurable space $(\mathscr{X},\mathscr{F})$. In particular we have

$$P_t(f(x)) = \mathbb{E}[f(X_t)|X_0 = x]$$

for $t \geq 0$ and $x \in \mathscr{X}$ and a bounded measurable function $f : \mathscr{X} \to \mathbb{R}$.

Definition 4.1.5 (Markov Semi-Group)**.** Take $\mathscr{P}$ as above. We say that $\mathscr{P}$ is a Markov semigroup if it has an identity $P_0 = Id$, it has an associative binary operation $P_t \circ P_s = P_{t+s}$ for any $s, t \geq 0$ and finally, there exists a $\sigma-$finite measure μ which is a stationary distribution for any P_t with $t \geq 0$.

Now, we want to make formal the notion of $\frac{d}{dt}P_t$, which will lead us to the definition of the 'generator' of a Markov-semi-group. Omitting some of the nitty-gritty details because they are primarily technical and not conceptual (we refer interested readers to Appendix 1 of [3]), we note that there exists a dense linear subspace $\mathscr{D}$ of $L^2(\mu)$ on which our family of bounded linear operators $\mathscr{P} = \{P_t\}_{t\geq 0}$ act, such that $dP_t|_{t=0}$ exists and is $\mu-$square integrable.

Definition 4.1.6 (Infinitesimal Generator)**.** The operator $L : \mathscr{D} \to L^2(\mu)$ such that

$$L(f) := \lim_{t\downarrow 0} \frac{P_t f - f}{t}$$

is called the Markov generator of the semi-group $\mathscr{P} = \{P_t\}_{t\geq 0}$ and we say it has domain

$$\mathscr{D}(L) := \{f \in L^2(\mu) : Lf \text{ exists and } Lf \in L^2(\mu)\}.$$

To see how these definitions work in practice, we'll build up the theory from the simplest Markov process: a discrete time random walk on the integers $\mathbb{Z}$. Here we'll introduce the convention that discrete time is indexed by $n \in \mathbb{N}$ and continuous time by $t \in \mathbb{R}^+$.

Example 15 (Discrete Time Simple Symmetric Random Walk). For concreteness, let $\{X_n\}_{n\in\mathbb{N}}$ be a simple symmetric random walk on $\mathbb{Z}$ beginning at the origin. Thus, let $X_0 = 0$. The Markov process evolves as follows:

$$X_{n+1} = X_n + \xi_{n+1} \quad \text{where } \xi_{n+1} = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

The transition matrix of this chain is infinite but is of the form

$$[p_{ij}]_{(i,j)\in\mathbb{Z}^2} \quad \text{where } p_{ij} = P(X_{n+1} = j | X_n = i) = \begin{cases} \frac{1}{2} & \text{when } j = i \pm 1 \\ 0 & \text{otherwise} \end{cases}$$

Now, let's turn to the semi-group and generator perspective. What are the Markov operators in this context? Well, they are precisely the linear operators on $W := \{f : \mathbb{Z} \to \mathbb{R} | f \text{ is bounded}\}$ that satisfy Definition 4.1.4. We leave it to the reader to verify that these operators satisfy Definition 4.1.4. They are of the form

$$P_1 f(x) = \mathbb{E}[f(X_1)|X_0 = x].$$

Notice that we can compute this exactly. We simply notice that there are two cases, either $X_1 = x+1$ or $X_1 = x-1$, each of which occurs with probability $\frac{1}{2}$, since we've decided to work with a symmetric random walk. Therefore,

$$P_1 f(x) = \mathbb{E}[f(X_1)|X_0 = x] = \frac{1}{2} f(x+1) + \frac{1}{2} f(x-1)$$

What about the n step operator? That's easy! We apply the same logic:

$$P_n f(x) = \mathbb{E}[f(X_n)|X_0 = x] = P^n(f(x))$$

which can be unwound recursively, conditioning on successive X_is, where P^n indicates the composition of P_1 with itself n times. However, to verify these operators satisfy Definition 4.1.5, that is form a semi-group, we do not even need to compute the explicit form of this operator. Instead, simply notice that $P_0 f(x) = \mathbb{E}[f(X_0)|X_0 = x] = f(x)$ for all $f \in W$ and note that the composition property also follows from the rules of conditional expectation. In particular, we note that

$$P_{n+m} f(x) = \mathbb{E}[f(X_{n+m})|X_0 = x]$$

while

$$P_n(P_m(f(x))) = \mathbb{E}[P_m(f(X_n))|X_0 = x] = \mathbb{E}[\mathbb{E}[f(X_m)|X_0 = X_n]|X_0 = x] = \mathbb{E}[f(X_{n+m})|X_0 = x]$$

where the last equality follows from the tower law.

We'll leave the verification of the stationary measure property to the reader.

Now, we need to consider the generator L. Notice since we are working in discrete time, clearly the infinitesimal generator $L(f) := \lim_{t\downarrow 0} \frac{P_t f - f}{t}$ of Definition 4.1.6 is undefined. The useful analog in the discrete case is the 1-step transition probability. Since this chain is time-homogeneous, we have that the discrete generator L is given by

$$Lf := \frac{(P-I)f}{1} = \frac{Pf - f}{1} = Pf(x) - f(x) = \frac{1}{2} f(x+1) + \frac{1}{2} f(x-1) - f(x)$$

Even in the relatively simple case of a discrete state space in discrete time the virtues of this semi-group formulation are beginning to be clear – they are certainly more wieldy and informative than an infinite dimensional matrix.

As suggested by the fact that we had to go a bit out of our way to define the generator in the discrete time setting of Example 15, the continuous time setting is where the real value of the semi-group/generator perspective begin to shine. For concreteness and simplicity, consider the following example which adapts the discrete time random walk into a continuous time random walk. See page 87 of [32] for an introduction to compound Poisson processes, which are a fairly common technique for embedding discrete time stochastic processes into continuous time.

Example 16. Define the continuous-time-random walk as $\{X_t\}_{t\geq 0}$ by

$$X_t = \sum_{i=1}^{N_t} \xi_i = S_{N_t}$$

where N_t is a Poisson process with rate $\lambda = 1$ and

$$\xi_i = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

and S_n is the simple, symmetric random walk described in Example 15 called S here instead of X to avoid confusion. This process is exactly the continuous time analog of the simple symmetric random walk. This verification is also left as an exercise to the reader. Now we're ready to compute the semi-group corresponding to this process as well as find its infinitesimal generator.

As above, we know that the elements of the semi-group are of the form

$$P_t(f(x)) = \mathbb{E}[f(X_t)|X_0 = x] = \mathbb{E}[f(S_{N_t})|X_0 = x] = \mathbb{E}[f\left(\sum_{i=1}^{N_t} \xi_i\right)|X_0 = x] = \sum_{i=1}^{\infty} P_n(f(x))P(N_t = n)$$

where P_n is the Markov operator defined in Example 15. Now, substituting in the density of N_t, we find

$$P_t(f(x)) = \left(\sum_{n=0}^{\infty} e^{-t}\frac{(t)^n}{n!}P^n\right)(f(x)) = \left(e^{-t}\sum_{n=0}^{\infty}\frac{(tP)^n}{n!}\right)(f(x)) = \left(e^{-t}e^{tP}\right)(f) = e^{t(P-I)f}$$

Now, we're in a position to compute the infinitesimal generator. Recall that

$$L(f) = \lim_{t\downarrow 0}\frac{P_t f - f}{t} = \lim_{t\downarrow 0}\frac{e^{t(P-I)f} - f}{t}$$

Now, to simplify, we Taylor expand the numerator in t about $t = 0$

$$e^{\lambda t(P-I)f(x)} = f + (P-I)f + \frac{t^2}{2}(P-I)^2 f + \dots$$

Then, substituting this in, we find

$$L(f) = \lim_{t \downarrow 0} \frac{P_t f - f}{t} = \lim_{t \downarrow 0} \frac{f + (P-I)f + \frac{t^2}{2}(P-I)^2 f + \dots - f}{t}$$

All terms with exponent greater than or equal to 2 are sent to zero, so we're left with

$$L(f) = \lim_{t \downarrow 0} \frac{P_t f - f}{t} = P_n f(x) - f(x) = \frac{1}{2} f(x+1) + \frac{1}{2} f(x-1) - f(x)$$

As we had hoped, this is exactly the discrete Laplacian from Example 15, which makes sense because the instantaneous behavior of this process should match the one-step behavior of the discrete time chain.

Now, finally, we can move on to a continuous time, continuous state-space example. The most natural example to take here is the scaling limit of a simple symmetric random walk, which is Brownian motion. Recall (for full proof, see Theorem 1.9 in [29]). In particular, we have by Donsker's theorem that

$$\lim_{n \to \infty} (Z_t^{(n)})_{t \in [0,1]} = \lim_{n \to \infty} (\frac{S_{\lfloor nt \rfloor}}{\sqrt{n}})_{t \in [0,1]} \stackrel{d}{=} (B_t)_{t \in [0,1]}$$

where $(B_t)_{t \in [0,1]}$ is standard Brownian motion on the unit interval and convergence is in distribution.

Example 17 (Brownian Motion). The utility of the semi-group perspective is even clearer in the context of Brownian motion. Consider standard Brownian motion on the unit interval $(B_t)_{t \in [0,1]}$ as our Markov process. It is easy to verify that B_t is a Markov process. As before, we'll first consider a Markov operator

$$P_t f(x) = \mathbb{E}[f(B_t) | B_0 = x] = \mathbb{E}[f(B_t + x)] = \mathbb{E}[f(x + \sqrt{t} Z]$$

where the last equality follows from the fact that $B_t \sim \mathcal{N}(0, t)$ and we take Z to be an independent standard normal random variable.

Now, supposing f is sufficiently smooth, we can Taylor expand in $\sqrt{t}$. We're allowed to do this because we assume $f \in C_b^2(\mathbb{R})$ since that's the natural domain on which the infinitesimal generator is defined, so it is appropriate to use it for the semi-group computations too.

Therefore, we have

$$f(x + \sqrt{t} Z) = f(x) + \sqrt{t} Z f'(x) + \frac{t}{2} Z^2 f''(x) + o(t)$$

Taking expectations on both sides and recalling that the first moment of a normal random variable is zero, we have

$$P_t f(x) = \mathbb{E}[f(x+\sqrt{t}Z] = f(x) + \frac{t}{2}f''(x) + o(t)$$

Then, computing the generator is comparatively easy:

$$L(f) = \lim_{t \downarrow 0} \frac{P_t f - f}{t} = \lim_{t \downarrow 0} \frac{f(x) + \frac{t}{2}f''(x) + o(t) - f(x)}{t} = \frac{1}{2}f''(x)$$

This derivation is somewhat heuristic, but should communicate clearly that exactly as we had hoped, the continuous time, continuous state space generator is the continuous Laplacian!

Now, we can define one last quantity which will be helpful in the rest of the chapter.

Definition 4.1.7 (Carré du Champ (Square of a Field) Operator, see Section 1.4.2 in [3])**.** Consider the Markov semi-group $(P_t)_{t\geq 0}$ with an infinitesimal generator $\mathscr{L}$ and a domain $\mathscr{D}(\mathscr{L})$ which has a vector subspace of nice functions $\mathscr{A}$ such that for any pair $(f,g) \in \mathscr{A}$ such that $fg \in \mathscr{D}(\mathscr{L})$. The Carré du Champ operator is a bilinear map $\Gamma : \mathscr{A} \times \mathscr{A} \to \mathbb{R}$ such that

$$\Gamma(f,g) = \frac{1}{2}[\mathscr{L}(fg) - f\mathscr{L}(g) - g\mathscr{L}(f)]$$

Remark 13. Notice that when $\mathscr{L} = \Delta$ is the Laplacian on $\mathbb{R}$ as in Example 17, we have $\Gamma(f,g) = \nabla f \cdot \nabla g = f' \cdot g'$.

4.1.3 Dirichlet Forms, Information, and Entropy

Now that we know what a Markov process is,it is possible to define some familiar quantities in a new way. In particular, many readers will be familiar with the theory of Maximum Likelihood Estimation. In that context, there is some parameter of interest θ which we'd like to estimate a parametric family of densities from which our data is drawn which are indexed by θ. Suppose for simplicity that this family $f(\theta)$ is twice differentiable in θ. Suppose we observe data X and we compute our likelihood $f(\theta, X)$. Then, the Fisher information is given by

$$\mathscr{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\left(\log(f(X,\theta))\right)\right)^2\right]$$

where the expectation is taken over the data.

This setting may seem initially to be far removed from our own, here, but crucially, some intuition connects them. We should think of Fisher information as describing the curvature of the log-likelihood function. In particular, large Fisher information corresponds to a sharp peak, which means that only a small neighborhood of values of θ provide a

high log likelihood. Conversely, small Fisher information indicates we shouldn't be very confident in our maximum likelihood estimate of θ since there is a large neighborhood of possible θ values which correspond to similar values of the log-likelihood.

The message from the Fisher information case is that curvature and relative roughness are crucial quantities in understanding the behavior of a density that information captures. We'll now turn to the more general setting:

Definition 4.1.8 (Donsker-Varadhan information)**.** Consider two measures μ and ν on some measure space $\mathscr{X}$. Then, the Donsker-Varadhan information of

$$I(\nu|\mu) = \begin{cases} \mathscr{E}(\sqrt{f},\sqrt{f}) & \text{when } \nu \ll \mu \text{ with density } f \text{ that is } \nu = f\mu \quad \text{for} \sqrt{f} \in \mathscr{D}(\mathscr{E}) \\ +\infty & \text{otherwise} \end{cases}$$

we write the Donsker-Varadhan information of a function as

$$I_\mu(f) = \mathscr{E}(\sqrt{f},\sqrt{f})$$

for all non-negative functions $f \in \mathscr{D}(\mathscr{E})$.

There are still a few pieces of machinery we need to define here. First, we need to understand what $\mathscr{E}(\sqrt{f},\sqrt{f})$ means. Intuitively, $\mathscr{E}(g,g)$ measures the variability or 'energy' of g. Now, we can make use of the definitions in the previous section to formalize this

Definition 4.1.9 (Dirichlet Form)**.** [18] [1] Let $(X_t)_{t\geq 0}$ be a Markov Process with state space $\mathscr{X}$, semi-group $(P_t)_{t\geq 0}$ with generator $\mathscr{L}$, and stationary measure μ. Additionally, let $(X_t)_{t\geq 0}$ be reversible with respect to μ. For any $g \in \mathscr{D}(\mathscr{L}))$, we define the Dirichlet form

$$\mathscr{E}(g,g) := \langle -\mathscr{L}g, g\rangle_\mu = \int_{\mathscr{X}} (-\mathscr{L}g)(x)g(x)\mu(dx)$$

Remark 14. Notice, that when $\mathscr{E}(\sqrt{f},\sqrt{f}) = \int_{\mathscr{X}} |\nabla\sqrt{f}|^2 d\mu$, we recover four times the standard Fisher information. See page 40 of [18] for a more detailed description.

Example 18. For concreteness, consider the measure $\mu = e^{-V(x)}dx$ on $\mathscr{X} = \mathbb{R}$. Let μ be the law of a standard normal distribution (so $V(x) = \frac{x^2}{2} + \frac{1}{2}\log(2\pi)$ and consider the generator given by

$$\mathscr{L} = \Delta - \nabla V \cdot \nabla = \frac{d^2}{dx} - x\frac{d}{dx}$$

Then, consider diffusion $(X_t)_{t\geq 0}$ which is generated by the infinitesimal generator $\mathscr{L}$. Then, Donsker-Varadhan information for a measure $\nu = f\mu$ is given exactly by

$$I(\nu|\mu) = \mathscr{E}(\sqrt{f},\sqrt{f})$$

[1] There is a small measure-theoretic subtlety here, which is explained on page 39 of [18]. In particular, we define the Dirichlet form on the closure $(\mathscr{E},\mathscr{D}(\mathscr{E}))$.

since we've assumed that $\nu \ll \mu$. For notational convenience, we'll let $g := \sqrt{f}$

$$\begin{aligned} I(\nu \mid \mu) &= \int_{\mathbb{R}} (-\mathscr{L} g)(x)\, g(x)\, \mu(dx) \\ &= \int_{\mathbb{R}} \left(-g''(x) + x g'(x)\right) g(x)\, \phi(x)\, dx \\ &= \int_{\mathbb{R}} (-g''(x)) g(x) \phi(x) dx + \int_{\mathbb{R}} x g'(x) g(x) \phi(x) dx \end{aligned}$$

where $\phi(x)$ is the pdf of a standard normal distribution.

Then, we'll integrate by parts with the additional hypothesis that the boundary term vanishes. Taking $u'(x) = -g''(x)$ and $v(x) = g(x)\phi(x)$ for our first term gives

$$\begin{aligned} -\int_{\mathbb{R}} g''(x)\, g(x)\, \phi(x)\, dx &= \int_{\mathbb{R}} g'(x) \left(g(x)\phi(x)\right)' dx \\ &= \int_{\mathbb{R}} (g'(x))^2 \phi(x)\, dx + \int_{\mathbb{R}} g'(x) g(x)\, \phi'(x)\, dx. \end{aligned}$$

Therefore we have

$$\begin{aligned} I(\nu \mid \mu) &= \int_{\mathbb{R}} \left(-g''(x) + x g'(x)\right) g(x)\, \phi(x)\, dx \\ &= \int_{\mathbb{R}} (g'(x))^2 \phi(x)\, dx. \end{aligned}$$

So, finally, substituting back $g = \sqrt{f}$, we conclude that

$$I(\nu \mid \mu) = \int_{\mathbb{R}} \left(\frac{d}{dx} \sqrt{f(x)} \right)^2 \mu(dx).$$

Generalized Entropy

Now, we only need one more piece to understand Definition 4.1.1 formally. Recall Definition 2.2.5. In effect, then, this generalized entropy construction simply provides the correct re-scaling to allow us to discuss relative entropy between a non-negative function and a reference measure.

To conclude this section, we'll review a very useful identity which concisely relates generalized entropy and information.

Proposition 4.1.10 (de Bruijn's Identity on $\mathbb{R}$, a simplified version of proposition 5.2.2 in [3])**.** *Let μ (with $d\mu = e^{-V(x)} dx$) be a measure on $\mathscr{X} = \mathbb{R}^n$ such that μ is a stationary measure for some diffusion $(X_t)_{t \geq 0}$ with corresponding semi-group $(P_t)_{t \geq 0}$ and generator*

$\mathscr{L}$, which gives a Dirichlet form $\mathscr{E}$ and domain $\mathscr{D}(\mathscr{E})$. Then, for all positive $f \in \mathscr{D}(\mathscr{E})$, we have

$$\frac{d}{dt} Ent_\mu(P_t f) = -2 I_\mu(P_t f) = -\frac{1}{2} \mathscr{I}_{Fisher}(P_t f)$$

Proof. This proof follows just from working with the definitions and applying properties of integrals. First, choose a sufficiently nice $f \in \mathscr{D}(\mathscr{E})$. We can extend to $f \in \mathscr{D}(\mathscr{E})$ by approximation, which we leave to the reader to complete, since the key ideas of the proof are demonstrated by the nice f case.

$$\frac{d}{dt} \mathrm{Ent}_\mu(P_t f) = \frac{d}{dt} \left(\int_{\mathscr{X}} P_t f \log(P_t f) d\mu - \left(\int_{\mathscr{X}} P_t f \, d\mu \right) \cdot \log \left(\int_{\mathscr{X}} P_t f \, d\mu \right) \right)$$

Now, differentiating under the integral sign (which is allowed since we're restricting our interest to nice f), we get

$$\frac{d}{dt} \mathrm{Ent}_\mu(P_t f) = \int_{\mathscr{X}} (1 + \log(P_t f)) \frac{d}{dt}(P_t f) d\mu - \left(1 + \log(\int_{\mathscr{X}} P_t f \, d\mu) \right) \left(\frac{d}{dt} \int_{\mathscr{X}} P_t f \, d\mu \right)$$

Then, noticing that μ is a stationary measure for the Markov process, we have

$$\frac{d}{dt} \int_{\mathscr{X}} P_t f \, d\mu = \frac{d}{dt} \int_{\mathscr{X}} f \, d\mu = 0$$

since $\forall t \geq 0$, we have $\int P_t f \, d\mu = \int f \, d\mu$.

Therefore, we have

$$\frac{d}{dt} \mathrm{Ent}_\mu(P_t f) = \int_{\mathscr{X}} (1 + \log(P_t f)) \frac{d}{dt}(P_t f) d\mu$$

Now, recall that in the nice, symmetric diffusion case (see Example 18)

$$\frac{d}{dt}(P_t f) = L(P_t f) = (\Delta - \nabla V \cdot \nabla)(P_t f) = \Delta P_t f - \nabla V \cdot \nabla P_t f$$

Therefore, we have

$$\begin{aligned} \frac{d}{dt} \mathrm{Ent}_\mu(P_t f) &= \int_{\mathscr{X}} \left(1 + \log(P_t f)\right) \left(\Delta P_t f - \nabla V \cdot \nabla P_t f\right) d\mu \\ &= \int_{\mathscr{X}} \left(1 + \log(P_t f)\right) \Delta P_t f \, e^{-V} \, dx - \int_{\mathscr{X}} \left(1 + \log(P_t f)\right) (\nabla V \cdot \nabla P_t f) \, e^{-V} \, dx \end{aligned}$$

Then, we can integrate the first term by parts, so taking $u'(x) = \Delta P_t f e^{-V}$ and $v(x) = 1 + \log(P_t f)$. Thus, we have

$$\int_{\mathscr{X}} \left(1+\log(P_t f)\right) \Delta P_t f e^{-V} dx = \int_{\mathscr{X}} \nabla(1+\log(P_t f)) \nabla P_t f e^{-V} dx - \int (1+\log(P_t f)) \nabla V \cdot \nabla P_t f e^{-V} dx$$

Now, substituting back, we get

$$\begin{aligned}\frac{d}{dt}\operatorname{Ent}_\mu(P_t f) &= \int_{\mathscr{X}} \big(1+\log(P_t f)\big)\,\Delta P_t f\, e^{-V}\,dx - \int_{\mathscr{X}} \big(1+\log(P_t f)\big)\,(\nabla V\cdot\nabla P_t f)\, e^{-V}\,dx \\ &= \int_{\mathscr{X}} \nabla(1+\log(P_t f))\nabla P_t f e^{-V}\,dx - \int (1+\log(P_t f))\nabla V\cdot\nabla P_t f e^{-V}\,dx \\ &\quad - \int_{\mathscr{X}} \big(1+\log(P_t f)\big)\,(\nabla V\cdot\nabla P_t f)\, e^{-V}\,dx\end{aligned}$$

Then, noticing that terms cancel nicely, we get

$$\frac{d}{dt}\operatorname{Ent}_\mu(P_t f) = -\int_{\mathscr{X}} \nabla(1+\log(P_t f))\nabla(P_t f)e^{-V}\,dx = -\int_{\mathscr{X}} \frac{|\nabla P_t f|^2}{P_t f} e^{-V}\,dx = -2I(P_t f\mu|\mu)$$

as desired, which we find by applying Definition 4.1.8. □

4.2 Examples of Log-Sobolev Inequalities

Now that we have all the ingredients, we can actually work with Log-Sobolev inequalities. As in Chapter 3, we'll investigate some examples of measures which satisfy Log-Sobolev Inequalities. As usual, we'll begin with the well behaved Gaussian case on $\mathbb{R}$.

4.2.1 Gaussian Measures

Standard Gaussian on $\mathbb{R}$

The following statement is originally due to Gross [19], but I'll follow the notation in [3].

Theorem 4.2.1 (Gaussian measure satisfies an $LS(1)$ inequality)**.** *Let γ be the standard Gaussian measure on $\mathbb{R}$,*

$$d\gamma(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx.$$

Consider the Ornstein–Uhlenbeck generator

$$Lf = f'' - xf',$$

the carré du champ operator

$$\Gamma(f,g) = f'g',$$

and the associated Dirichlet form

$$\mathscr{E}(f,g) = \int_{\mathbb{R}} f'(x)g'(x)\,d\gamma(x).$$

Its domain is

$$\mathscr{D}(\mathscr{E}) = \left\{f \in L^2(\gamma) : f' \in L^2(\gamma)\right\},$$

Then, for every $g \in \mathscr{D}(\mathscr{E})$,

$$Ent_\gamma(f^2) \leq 2\mathscr{E}(f,f) = 2\int_{\mathbb{R}} |f'(x)|^2 \, d\gamma(x).$$

Equivalently, γ satisfies LS(1).

As is natural for such a useful theorem, there are many different proofs of this statement. We'll follow the original proof presented by [19], but restricted to the one-dimensional case to begin with. We'll begin by proving the one-dimensional case, also referencing the presentation by [22].

Proof. The structure of this proof is very clever. We'll first show that Bernoulli random variables satisfy such a Log-Sobolev inequality and then apply the central limit theorem to deduce that Gaussian measures also satisfy this inequality. The nice thing about this idea is that we can compute all of the quantities of interest directly in the Bernoulli case.

First, consider the measure μ on $\{0,1\}$ corresponding to the Bernoulli($\frac{1}{2}$) random variable. The analog of Theorem 4.2.1 is exactly

$$\mathrm{Ent}_\mu(f^2) \leq \frac{1}{2}\mathbb{E}[|Df|^2]$$

where D is the discrete analog of the infinitesimal generator $\mathscr{L}$ as introduced in Example 15. In particular,

$$D(f) = f(1) - f(0)$$

Then,

$$\begin{aligned}\mathrm{Ent}_\mu(f^2) &= \int_{\mathscr{X}} f^2 \log(f^2)\, d\mu - \left(\int_{\mathscr{X}} f^2\, d\mu\right)\log\left(\int_{\mathscr{X}} f^2\, d\mu\right) \\ &= \frac{1}{2}f(0)^2\log\big(f(0)^2\big) + \frac{1}{2}f(1)^2\log\big(f(1)^2\big) - \frac{f(0)^2+f(1)^2}{2}\log\left(\frac{f(0)^2+f(1)^2}{2}\right).\end{aligned}$$

Then, we have

$$\mathbb{E}[|Df|^2] = \mathbb{E}[|f(1)-f(0)|^2] = \big(f(0)-f(1)\big)^2$$

Therefore, all we need to show is that

$$\begin{aligned}\mathrm{Ent}_\mu(f^2) &= \frac{1}{2}f(0)^2\log\big(f(0)^2\big) + \frac{1}{2}f(1)^2\log\big(f(1)^2\big) - \frac{f(0)^2+f(1)^2}{2}\log\left(\frac{f(0)^2+f(1)^2}{2}\right) \\ &\leq \frac{1}{2}\big(f(0)-f(1)\big)^2\end{aligned}$$

To see this, consider the function

$$d(f) = \frac{1}{2}\big(f(0)-f(1)\big)^2 - \frac{1}{2}f(0)^2\log\big(f(0)^2\big) - \frac{1}{2}f(1)^2\log\big(f(1)^2\big) - \frac{f(0)^2+f(1)^2}{2}\log\left(\frac{f(0)^2+f(1)^2}{2}\right)$$

and it suffices to show that $d(f) \geq 0$. Notice this is minimized as $f(0) = f(1)$, that $d(f)$ is convex and therefore minimized at 0 (to convince yourself, substitute $u = \frac{1}{4}(f(0) - f(1))^2$, rewrite $d(f)$ in terms of u and compute the first and second derivatives).

Thus, we have $\text{Ent}_\mu(f^2) \leq \frac{1}{2}\mathbb{E}[|Df|^2]$.

Now, we need to understand how this inequality behaves on the n-dimensional cube because we need to consider a sum of Bernoulli variables in order to apply the central limit theorem. Now, consider the product measure $\mu \otimes \mu \cdots \otimes \mu = \mu^n$ on the $n-$cube $\{0,1\}^n$. Then, we claim the following lemma holds:

Lemma 4.2.2. *[22]*

$$Ent_{\mu^n}(f^2) \leq \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}[|D_i f|^2]$$

where $D_i f(x_1, \ldots, x_n) = Df_i(x_i) = f_i(x^{(i,1)}) - f_i(x^{(i,0)})$, *and* $x^{(i,z)}$ *means replace the* i*th coordinate with* $z \in \{0,1\}$.

We can deduce Lemma 4.2.2 from the so-called Tensorization property of entropy, which we proved in Lemma 2.2.7, in Section 2.2.
In particular, applying the two-point inequality

$$\text{Ent}_\mu(f^2) \leq \frac{1}{2}\mathbb{E}[|Df|^2]$$

in each coordinate to

$$\text{Ent}_{\mu_1 \otimes \mu_2 \cdots \otimes \mu_n}(f) \leq \sum_{i=1}^{n} \mathbb{E}[\text{Ent}_{\mu_i}(f_i)]$$

gives exactly

$$\text{Ent}_{\mu^n}(f^2) = \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}[|D_i f|^2].$$

Now, we will construct a suitable quantity to apply the Central Limit Theorem to. In particular, consider $X_1, \ldots, X_n \overset{iid}{\sim} \mu$. We want the centralized sample mean. Define

$$S_n = \frac{2}{\sqrt{n}}\sum_{i=1}^{n}(X_j - \frac{1}{2})$$

where we subtract off the sample mean of each Bernoulli random variable and rescale appropriately.

Consider some test function $\psi \in C_b^2$, where as usual we'd like ψ to be bounded and at least twice continuously differentiable. We will extend this to any function $g \in \mathscr{D}(\mathscr{E}) = \{f \in L^2(\gamma) : f' \in L^2(\gamma)\}$. Consider

$$F_n(x_1, x_2, \ldots, x_n) = \psi(S_n).$$

Then, we can apply $\mathrm{Ent}_{\mu^n}(f^2) = \frac{1}{2}\sum_{i=1}^n \mathbb{E}[|D_i f|^2]$ to F.
Thus

$$\mathrm{Ent}_{\mu^n}(\psi(S_n)^2) \leq \frac{1}{2}\sum_{i=1}^n \mathbb{E}[|D_i F_n|^2]$$

All that's left then, is to carefully take the limit as $n \to \infty$ of both sides applying the central limit theorem appropriately.

First, recall the statement of the central limit theorem in the simplest independent and identically distributed (IID) case, which is all we need here.

Theorem 4.2.3 (Central Limit Theorem IID Case [7])**.** *Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mu$ with finite variance σ^2 and let $S_n = \sum_{i=1}^n X_i$. Then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0,1).$$

Now, we'll begin with the LHS. We want to understand

$$\lim_{n\to\infty} \mathrm{Ent}_{\mu^n}(\phi(S_n)^2).$$

First, notice that the central limit theorem tells us that

$$S_n \xrightarrow{d} Z \sim \gamma.$$

Then, recall that the Portmanteau Theorem 2.1.12 says that for any bounded, continuous function, h, if $Y_n \xrightarrow{d} Y$, then $\mathbb{E}[h(Y_n)] \to \mathbb{E}[h(Y)]$.

Now, recall the definition of generalized entropy:

$$\mathrm{Ent}_{\mu^n}[\psi(S_n)^2] = \mathbb{E}[\psi(S_n)^2 \log(\psi(S_n)^2)] - \mathbb{E}[\psi(S_n)^2]\log(\mathbb{E}\psi(S_n)^2)).$$

Conveniently, $h_1(s_n) = \psi(s_n)^2$ and $h_2(s_n) = \psi(s_n)^2\log(\psi(s_n)^2)$ are both bounded and continuous by our assumptions on ψ, since for h_2 we maintain the convention that $0\log 0 = 0$. Then Theorem 2.1.12 gives exactly that

$$\mathrm{Ent}_{\mu^n}[\psi(S_n)^2] \to \mathrm{Ent}_{\gamma}(\psi^2).$$

Now, turning our attention to the RHS, we first notice by the fact that $X_i \overset{iid}{\sim} \mu$

$$\frac{1}{2}\sum_{i=1}^n \mathbb{E}_{\mu^n}\left[(D_i F_n)^2\right] = \frac{n}{2}\mathbb{E}_{\mu^n}\left[(D_1 F_n)^2\right]$$

Now, notice that the discrete gradients $D_i F_n \to \frac{1}{\sqrt{n}} f'$. In particular, we have

$$D_1F_n = \psi(\frac{2}{\sqrt{n}}\sum_{i=1}^{n}(X_j - \frac{1}{2}) + \frac{1}{\sqrt{n}}) - \psi(\frac{2}{\sqrt{n}}\sum_{i=1}^{n}(X_j - \frac{1}{2}) - \frac{1}{\sqrt{n}})$$

Thus, we have

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}_{\mu^n}\left[(D_iF_n)^2\right] &= \frac{n}{2}\mathbb{E}_{\mu^n}\left[(D_1F_n)^2\right] \\ &= \frac{n}{2}\mathbb{E}_{\mu^n}\left[\left(\psi\left(\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\left(X_i - \frac{1}{2}\right) + \frac{1}{\sqrt{n}}\right) - \psi\left(\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\left(X_i - \frac{1}{2}\right) - \frac{1}{\sqrt{n}}\right)\right)^2\right].\end{aligned}$$

Now note that by the mean value theorem there exists some random $\theta_n \in [-1,1]$ such that

$$\psi\left(\frac{2}{\sqrt{n}}\sum_{i=2}^{n}\left(X_i - \frac{1}{2}\right) + \frac{1}{\sqrt{n}}\right) - \psi\left(\frac{2}{\sqrt{n}}\sum_{i=2}^{n}\left(X_i - \frac{1}{2}\right) - \frac{1}{\sqrt{n}}\right) = \frac{2}{\sqrt{n}}\psi'\left(\frac{2}{\sqrt{n}}\sum_{i=2}^{n}\left(X_i - \frac{1}{2}\right) + \frac{\theta_n}{\sqrt{n}}\right)$$

which is permissible since we've assumed that $\psi \in C_b^2$ so ψ' is uniformly continuous.

Then, we have that

$$(D_iF_n)^2 = \frac{4}{n}\left(\psi'\left(\frac{2}{\sqrt{n}}\sum_{i=2}^{n}\left(X_i - \frac{1}{2}\right) + \frac{\theta_n}{\sqrt{n}}\right)\right)^2$$

Now notice that

$$\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}_{\mu^n}\left[(D_iF_n)^2\right] = 2\mathbb{E}[\left(\psi'\left(\frac{2}{\sqrt{n}}\sum_{i=2}^{n}\left(X_i - \frac{1}{2}\right) + \frac{\theta_n}{\sqrt{n}}\right)\right)^2]$$

Now, since $(\psi')^2$ has bounded continuous derivatives, it is Lipschitz, and therefore, we have as $n \to \infty$

$$\lim_{n\to\infty}\mathbb{E}[(\psi'(\frac{2}{\sqrt{n}}\sum_{i=2}^{n}\left(X_i - \frac{1}{2}\right) + \frac{\theta_n}{\sqrt{n}}))^2] - \mathbb{E}[(\psi'(\frac{2}{\sqrt{n}}\sum_{i=2}^{n}\left(X_i - \frac{1}{2}\right))^2] = 0$$

Now, recall that by the Central Limit Theorem (and a little bit of algebra),

$$\frac{2}{\sqrt{n}}\sum_{i=2}^{n}\left(X_i - \frac{1}{2}\right) \xrightarrow{d} Z \sim \gamma$$

Therefore, we have that by the Portmanteau Theorem again

$$\mathbb{E}[\left(\psi'(\frac{2}{\sqrt{n}}\sum_{i=2}^{n}\left(X_i - \frac{1}{2}\right)\right)] = \mathbb{E}\left[\left(\psi'(Z)\right)^2\right]$$

Then, putting everything together we have

$$\mathbb{E}[(\psi'(\frac{2}{\sqrt{n}}\sum_{i=2}^{n}\left(X_i-\frac{1}{2}\right)+\frac{\theta_n}{\sqrt{n}}))^2]\to\mathbb{E}\left[\left(\psi'(Z)\right)^2\right]$$

Now multiplying both sides by two, we have

$$\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}_{\mu^n}\left[(D_iF_n)^2\right]\to 2\mathbb{E}\left[\left(\psi'(Z)\right)^2\right]$$

Finally, since taking limits preserves the inequality we have for all $\psi\in C_b^2(\mathbb{R})$, we have

$$\mathrm{Ent}_\gamma(f^2)\le 2\mathscr{E}(f,f)=2\int_{\mathbb{R}}|f'(x)|^2\,d\gamma(x).$$

Finally, to extend this to all functions

$$g\in\mathscr{D}(\mathscr{E})=\{f\in L^2(\gamma):f'\in L^2(\gamma)\}$$

it suffices to notice that $C_b^2(\mathbb{R})$ is dense in $\mathscr{D}(\mathscr{E})$, so for any $g\in\mathscr{D}(\mathscr{E})$ we can construct a sequence so smooth, compactly supported $g_k\in C_b^2(\mathbb{R})$ such that $g_k\to g\in L^2(\gamma)$ and $g_k'\to g'\in L^2(\gamma)$. Then, the desired inequality holds for each of our approximating functions and we pass to the limit as $k\to\infty$. Finally, this completes the proof! □

Now that we've gone to a lot of trouble to establish this log-Sobolev inequality for $\mathbb{R}$, we can get a host of nice results for $\mathbb{R}^n$ almost for free with the help of Lemma 2.2.7.

Standard Gaussian on $\mathbb{R}^n$

Proposition 4.2.4 (Gaussian Measures on $\mathbb{R}^n$ Satisfy an LS(1) Inequality)**.** *Let $\gamma_n=\gamma\otimes\gamma\cdots\otimes\gamma$ be the product measure on $\mathbb{R}^n$ where $\gamma\sim\mathcal{N}(0,1)$. Then, for all $f\in\mathscr{D}(\mathscr{E})$, γ_n satisfies the following Log-Sobolev inequality*

$$Ent_{\gamma_n}(f^2)\le 2\mathscr{E}(f,f)=2\int_{\mathbb{R}^n}|\nabla f|^2d\gamma_n(x)$$

Before we state the relatively short proof of this fact, it is worth remarking on the fact that the constant C in Proposition 4.2.4 and in Theorem 4.2.1 are the same. This means that this inequality is 'dimension-free' in a very nice sense. This property, that the constant of the Log-Sobolev inequality is independent of the dimension of the measure is one reason why Log-Sobolev inequalities are so useful.

Moreover, as we saw in Chapter 3, it is in general hard to determine the optimal couplings in dimension $n\ge 1$, especially for more complicated cost functions. The fact that our Log-Sobolev inequality is dimension free will allow us to infer transport-entropy inequalities in higher dimensions without any additional work.

Proof of Proposition 4.2.4. To see this, recall the tensorization of entropy lemma from section 2.2. In particular, Lemma 2.2.7 says that measures μ_i on measure spaces $\mathscr{X}_i$ for $i \in \{1,2,\dots,n\}$ and any function $f : \mathscr{X}_1 \times \cdots \times \mathscr{X}_n \to \mathbb{R}$, such that $f \in \mathscr{D}(\mathscr{E})$, we have

$$\mathrm{Ent}_{\mu_1 \otimes \mu_2 \cdots \otimes \mu_n}(f) \leq \sum_{i=1}^{n} \mathbb{E}[\mathrm{Ent}_{\mu_i}(f_i)].$$

Now, applying this to an arbitrary test function, we find that

$$\mathrm{Ent}_{\gamma_n}(f^2) \leq \sum_{i=1}^{n} \mathbb{E}[\mathrm{Ent}_{\mu_i}(f_i)] \leq \sum_{i=1}^{n} (2 \int_{\mathbb{R}^n} (f_i'(x_i))^2 d\gamma_n) = 2 \int_{\mathbb{R}^n} |\nabla f|^2 d\gamma_n(x)$$

This completes the proof. □

Gaussian with Finite Covariance

We can even go one step further – working with a standard Gaussian for this many pages of a thesis may bore some readers, so we'll also show that an *LS* inequality holds for a multivariate gaussian vector with mean $\vec{m}$ and non-singular covariance matrix Σ.

Corollary 4.2.5 (Non-Standard Gaussian Log-Sobolev Inequality, compare with Proposition 1.6 in [11] and a remark on page 258 in [3])**.** *Suppose $\mu \sim \mathcal{N}_n(\vec{m}, \Sigma)$ is the measure corresponding multivariate normal random variable in $\mathbb{R}^n$. Then, for all $f \in \mathscr{D}(\mathscr{E})$, μ satisfies the following Log-Sobolev Inequality*

$$Ent_{\mu}(f^2) \leq 2 \int_{\mathbb{R}^n} \Sigma \nabla f \cdot \nabla f \, d\mu(x)$$

Proof. The proof follows from simply applying the change of variables formula. In particular, we can represent $X = \Sigma^{1/2} Z + m$ where $Z \sim \gamma_n$. In particular, define

$$g(Z) = f(m + \Sigma^{1/2} Z) = f(X)$$

Then, for each g, we must have

$$\begin{aligned} \mathrm{Ent}_{\mu}(f^2) &= \mathrm{Ent}_{\gamma_n}(g^2) \\ &\leq 2 \int_{\mathbb{R}^n} |\nabla g|^2 d\gamma_n = 2 \int_{\mathbb{R}^n} |\Sigma^{1/2} \nabla f(m + \Sigma^{1/2} z)|^2 d\gamma_n \\ &= 2 \int_{\mathbb{R}^n} \Sigma \nabla f \cdot \nabla f \, d\mu \end{aligned}$$

□

4.2.2 Strongly Log-Concave Measures

A natural next question is about non-Gaussian measures. After all, we've already established transport-entropy inequalities for the simple one-dimensional Gaussian case, so while the above will allow us to move into $n-$dimensions and infer TE-inequalities for non-standard Gaussians, if we couldn't extend our claims beyond the Gaussian case after introducing all this new machinery, the above might seem a little pointless. However, there is a very nice, natural extension of this for non-Gaussian measures that will prove very powerful.

Theorem 4.2.6 (LS Inequality for Convex Potentials (Corr. 5.7.2 in [3]))**.** *Consider the measure μ on $\mathbb{R}^n$, written in potential form as $e^{-W}dx$, where $W:\mathbb{R}^n\to\mathbb{R}$ is a smooth, scalar-valued potential. Denote the Hessian of W as $\mathscr{H}W$. If W is uniformly $\rho-$ convex, that is*

$$\mathscr{H}W(x)\geq \rho I_n \quad \text{for all } x\in\mathbb{R}^n$$

for some $\rho>0$, then μ satisfies the LS$(\frac{1}{\rho})$ inequality,

$$Ent_\mu(f^2)\leq \frac{2}{\rho}\int_{\mathbb{R}^n}|\nabla f|^2 d\mu(x)$$

for all $f\in\mathscr{D}(\mathscr{E})$.

Proof. We'll provide an outline of the proof of this statement, here, but a full proof is presented in [3], pages 268-269.

Consider the Markov semi-group $(P_t)_{t\geq 0}$ which has μ as its stationary measure and is generated by $\mathscr{L}=\Delta-\nabla W\cdot\nabla$. We'll use this semi-group to understand how the curvature of W relates to Log-Sobolev inequality we're interested in proving.

First, recall that by Proposition 4.1.10, we have for any positive, sufficiently smooth function h

$$\frac{d}{dt}\text{Ent}_\mu(P_t h)=-I_\mu(P_t h)$$

Equivalently, by integrating both sides we have

$$\text{Ent}_\mu(P_t h)=\int_0^\infty I_\mu(P_t h)dt.$$

Our job, then is to prove that

$$\int_0^\infty I_\mu(P_t h)dt\leq\frac{2}{\rho}I_\mu(h).$$

If we can show this, we are done, since we will just take $h=f^2$ and recover

$$\text{Ent}_\mu(f^2)\leq\frac{2}{\rho}\int_{\mathbb{R}^n}|\nabla f|^2 d\mu(x).$$

How, then, do we show $\int_0^\infty I_\mu(P_t f)\,dt \leq \frac{2}{\rho} I_\mu(h)$?
First notice if we can show that

$$I_\mu(P_t h) \leq e^{-2\rho t} I_\mu(h)$$

we're done, since integrating the RHS of this gives exactly the desired bound.

To see

$$I_\mu(P_t h) \leq e^{-2\rho t} I_\mu(h)$$

involves comparing the first and second derivatives of

$$\int_{\mathbb{R}^n} P_t h \log(P_t h)\, d\mu$$

and applying the Hessian bound. For details, see 5.7.2 and Theorem 5.5.2 in [3]. The theorem follows immediately from this bound. □

4.2.3 Product Measures, Convolution of Measures and Tensorization

Finally, now that we know how to prove individual Log-Sobolev inequalities, it would be very nice to know how to combine them.

We've already seen how to extend Log-Sobolev inequalities to product measures via Lemma 2.2.7.

However, we might also be interested in convolutions of measures. In particular, if we have two measures μ_1 and μ_2 which individually satisfy $LS(C_1)$ and $LS(C_2)$ inequalities respectively, what can we say about the convolution $\mu_1 * \mu_2$. The extremely handy Lemma 2.2.7 allows us to answer that question with just a bit more work.

Theorem 4.2.7 (Log-Sobolev Inequalities for Convolution of Measures, an extended version of Proposition 1.3 in [11])**.** *Suppose $\mu_1, \ldots, \mu_n$ are measures on $\mathbb{R}^d$ each of which satisfy a log-sobolev inequality with constant C_i. Then for any $f \in \mathscr{D}(\mathscr{E})$, we have*

$$Ent_{\mu_1 * \mu_2 * \cdots * \mu_n}(f^2) \leq 2\left(\sum_{i=1}^n C_i\right) \int_{\mathbb{R}^d} |\nabla f|^2 d(\mu_1 * \mu_2 * \cdots * \mu_n).$$

Proof. We'll just show this for two measures μ_1 and μ_2, since the result for n measures follows immediately by induction.
First, recall that to sample from the convolution of two measures μ_1 and μ_2, we sample from μ_1 and μ_2 independently and sum our observations. This means that

$$\mu_1 * \mu_2 = (x+y)_\#(\mu_1 \otimes \mu_2)$$

This allows us to connect the convolution of μ_1 and μ_2 to their product measure. In particular if we let $g(x,y) := f(x+y)$, then

$$\mathrm{Ent}_{\mu_1 * \mu_2}(f^2) = \mathrm{Ent}_{\mu_1 \otimes \mu_2}(g^2).$$

Then, Lemma 2.2.7, combined with the individual Log-Sobolev Inequalities gives

$$\begin{aligned}\mathrm{Ent}_{\mu_1 * \mu_2}(f^2) &= \mathrm{Ent}_{\mu_1 \otimes \mu_2}(g^2) \\ &\leq 2C_1 \iint |\nabla_x g(x,y)|^2 \, d\mu_1(x) d\mu_2(y) + 2C_2 \iint |\nabla_y g(x,y)|^2 \, d\mu_1(x) d\mu_2(y)\end{aligned}$$

Then, since $\nabla_x g(x,y) = \nabla_y g(x,y) = \nabla f(x+y)$, we have

$$\mathrm{Ent}_{\mu_1 * \mu_2}(f^2) = \mathrm{Ent}_{\mu_1 \otimes \mu_2}(g^2) \leq 2(C_1 + C_2) \int |\nabla f(z)|^2 d(\mu_1 * \mu_2)(z)$$

exactly as desired. □

4.3 Log-Sobolev Inequalities and Transport-Entropy Inequalities

There are rich connections between transport-entropy inequalities and log-Sobolev inequalities as hinted at by the introduction to this chapter. The following section introduces two major theorems that connect these perspectives, on the one hand showing that Log-Sobolev inequalities are exactly the right tool to imply T_2 transport-entropy inequalities, and on the other connecting the key ingredients of a log-Sobolev inequality (namely, entropy and information) with the key ingredients of a transport-entropy inequality (namely Wasserstein distance and entropy).

4.3.1 Otto-Villani Theorem

Now, we've developed a series of techniques for proving Log-Sobolev inequalities, it is finally time to see how these inequalities imply transport-entropy inequalities. The key tool here is the Otto-Villani theorem.

Theorem 4.3.1 (Otto-Villani)**.** *[26] Suppose μ is a probability measure on $\mathbb{R}^n$. Then if μ satisfies a Log-Sobolev inequality with constant C, then μ also satisfies a T_2 inequality with the same constant C.*

Remark 15 (Extensions of the Otto-Villani Theorem). As mentioned in [18] this theorem is actually much more general. It was first proved on a smooth complete Manifold in 2000 by Otto and Villani in [26]. The proof presented there is difficult and simpler, more general versions both of the statement and of the proof have been developed since. It turns out that this theorem is true not just on $\mathbb{R}^n$, but also on many other (relatively nice) Polish spaces. These extensions are beyond the scope of this thesis, but are an active area of research. Much more detail is available in [16].

As with many such important theorems, there are many different proofs of this fact available. We'll follow the proof presented in [16] and presented again for measures μ on $\mathbb{R}^n$ in [18], though we'll change their normalization conventions to match those used in the rest of the thesis and add detail where their proof is compact.

Proof of Theorem 4.3.1. The structure of this proof is quite neat and has the virtue of making use of a couple of the results we've shown previously in the thesis and in reiterating the connection between Log-Sobolev Inequalities, transport-entropy inequalities, and concentration inequalities. In particular, we'll do the following:

$$\text{LS(C)} \implies T_1(C) \implies \text{Dimension-Free Gaussian Concentration} \implies T_2(C)$$

We begin with the $LS(C)$ inequality, which states that for all $f \in \mathscr{D}(\mathscr{E})$

$$\text{Ent}_\mu(f^2) \leq 2C \int_{\mathbb{R}^n} |\nabla f|^2 d\mu(x).$$

Now take $f^2 = e^{tg}$, which is the right substitution since we want to use concentration-inequality style arguments and this exponential form fits nicely with those tools.
Then, substituting gives

$$\text{Ent}_\mu(e^{tg}) \leq 2C \int_{\mathbb{R}^n} |\nabla e^{tg/2}|^2 d\mu(x).$$

The LHS gives

$$\text{Ent}_\mu(e^{tg}) = \int e^{tg} \log \frac{e^{tg}}{\int e^{tg} d\mu} d\mu$$

Define

$$Z_t := \int e^{tg} d\mu$$

as in [18]. Then, we have that

$$\text{Ent}_\mu(e^{tg}) = tZ_t' - Z_t \log Z_t.$$

Then, we have on the RHS

$$\int_{\mathbb{R}^n} |\nabla e^{tg/2}|^2 d\mu(x) = \frac{t^2}{4} \int_{\mathbb{R}^n} |\nabla g|^2 e^{tg} d\mu$$

Then, since g is $1-$Lipschitz by construction, we have $|\nabla g(x)| \leq 1$ for almost every x. Therefore, we have

$$\int_{\mathbb{R}^n} |\nabla g|^2 e^{tg} d\mu \leq \int_{\mathbb{R}^n} e^{tg} d\mu = Z_t$$

Combining the LHS and RHS of the $LS(C)$ inequality gives

$$tZ_t' - Z_t \log Z_t \leq \frac{Ct^2}{2}$$

Then, dividing both sides by $t^2 Z_t$ gives

$$\frac{tZ_t' - Z_t \log Z_t}{t^2 Z_t} \leq \frac{C}{2}$$

Recognizing that the LHS looks exactly like the quotient rule, we'll find that

$$\frac{d}{dt}\left(\frac{\log(Z_t)}{t}\right) \leq \frac{C}{2}$$

Now we can integrate both sides with respect to t and find that

$$\int e^{tg} d\mu \leq e^{t\int g d\mu + \frac{C}{2}t^2}$$

This says exactly that Z_t is sub-gaussian. This is true for all $1-$Lipschitz g.

Lemma 4.3.2 (W_1 and 1-Lipschitz Functions)**.** *Suppose μ is a probability measure on $\mathbb{R}^n$. Then μ satisfies*

$$W_1(\mu,\nu)^2 \leq 2CH(\nu|\mu) \quad \forall \nu$$

if and only if for every $1-$ Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ and all $s \in \mathbb{R}$, we have

$$\int_{\mathbb{R}^n} e^{sf} d\mu \leq e^{s\int f d\mu + Cs^2/2}$$

This lemma is a very simple, special case of Corr. 3.4 in [18], where the more general case is proved, so the full proof is omitted here. The idea of the proof is to the use Kantrovich dual characterization of a transport distance and apply the definition of entropy.

Thus, we've successfully established that μ satisfies a $T_1(C)$ inequality. However, the claim is that μ satisfies a $T_2(C)$ inequality. As demonstrated in Theorem 3.2.5, unfortunately the implication only goes from $T_2(C) \implies T_1(C)$, so we have some more work to do.

The last step is to recognize that T_2 inequalities are special in the following sense.

Lemma 4.3.3 (T_2 and Gaussian Dimension Free Concentration Property, a specialization of Corollary 5.5 in [18] to $\mathbb{R}^n$)**.** *Let μ be a measure on $\mathbb{R}^n$. μ satisfies $T_2(C)$ if and only if there exists a constant K such that*

$$\mu^n(A^r) \geq 1 - Ke^{-r^2/(2C)}$$

for all $A \in \mathbb{R}^n$ such that $\mu^n(A) \geq 1/2$, and $r \geq 0$. We define A^r as the neighborhood of radius r in $\mathbb{R}^n$ where distance is the Euclidean distance. We call this a Gaussian Concentration inequality, because it implies that μ^n has sub-gaussian tails.

As above, this is proved in [18] and so we will not repeat the proof here. Instead, we refer interested readers to Theorem 5.4 and Corollary 5.5.

The key idea here is to notice that we can show that this inequality holds by applying two theorems we've already proved. First we'll invoke the ever-helpful Lemma 2.2.7.
Then, since we have by hypothesis that μ satisfies $LS(C)$ on $\mathbb{R}^n$. Then

$$\text{Ent}_\mu(f^2) \leq 2C \int_{\mathbb{R}^n} |\nabla f|^2 d\mu(x)$$

Then, Lemma 2.2.7 tells us that

$$\text{Ent}_{\mu_1 \otimes \mu_2 \cdots \otimes \mu_k}(g) \leq \sum_{i=1}^{k} \mathbb{E}[\text{Ent}_{\mu_i}(g_i)].$$

Thus, considering the $k-$times tensorization of μ, we have

$$\text{Ent}_{\mu^k}(f^2) \leq \sum_{i=1}^{k} \mathbb{E}[\text{Ent}_\mu(f_i^2)]$$

Now, applying the $LS(C)$ inequality in each coordinate we get

$$\text{Ent}_{\mu^k}(f^2) \leq \sum_{i=1}^{k} \mathbb{E}[\text{Ent}_\mu(f_i^2)] \leq 2C \sum_{i=1}^{k} \mathbb{E}_{\mu^k} \left[\int_{\mathbb{R}^n} |\nabla_i f|^2 d\mu \right]$$

Then, swapping the sum and integral and applying the fact that on the product space $|\nabla f|^2 = \sum_{i=1}^k |\nabla_i f|^2$, gives

$$\text{Ent}_{\mu^k}(f^2) \leq 2C \int |\nabla f|^2 d\mu^k$$

Notice that this $k-$fold inequality preserves the constant C. Then, recall Marton's Argument (Theorem 3.4.2), which says that if μ satisfies

$$\mathcal{T}_d(\mu, \nu) \leq \alpha^{-1}(H(\nu|\mu))$$

Then, for all measurable subsets $A \subseteq \mathcal{X}$ such that $\mu(A) \geq \frac{1}{2}$, the following concentration inequality holds

$$\mu(A^r) \geq 1 - e^{-\alpha(r - \alpha^{-1}(\log(2)))}$$

where $r \geq \alpha^{-1}\log(2)$. Here we apply Marton's argument to the tensorized measure μ^k taking the cost function which is the distance metric $d(x,y) = |x-y|$ and the function $\alpha(\cdot) = (\cdot)^2$. Therefore, $\alpha^{-1}(\cdot) = \sqrt{\cdot}$. Therefore, Theorem 3.4.2 gives exactly that there exists some constant K such that

$$\mu^k(A^r) \geq 1 - Ke^{-r^2/(2C)}.$$

where K collects the constant terms in $e^{-\alpha(r-\alpha^{-1}(\log(2)))}$. Then, we can notice that the conditions of Lemma 4.3.3 are satisfied exactly. Thus since μ^k satisfies a dimension-free concentration inequality, have that μ satisfies a $T_2(C)$ transport-entropy inequality, exactly as desired. □

4.3.2 HWI Inequalities

To conclude this chapter, we will also mention a key inequality which combines entropy (H), Wasserstein distance (W), and Donsker-Varadhan information (I).

Theorem 4.3.4 (HWI Inequality, Theorem 3 in [26])**.** *Consider the measure μ on $\mathbb{R}^n$, written in potential form as $e^{-W}dx$, where $W:\mathbb{R}^n\to\mathbb{R}$ is twice-differentiable scalar-valued potential. Denote the Hessian of W as $\mathscr{H}W$. If W is uniformly $\rho-$ convex, that is*

$$\mathscr{H}W(x)\succeq\rho I_n \quad \text{for all } x\in\mathbb{R}^n,$$

then, μ satisfies the HWI inequality for all non-negative, smooth, and compactly supported functions which are normalized to be densities on $\mathbb{R}^n$, that is

$$H(f\mu|\mu)\leq W_2(f\mu,\mu)\sqrt{I_\mu(f)}-\frac{\rho}{2}W_2^2(f\mu,\mu)$$

The proof of this theorem can be found in the original paper [26] and a nice expository version in [18].

Note the striking similarity of the conditions in Theorem 4.3.4 to the conditions of Theorem 4.2.6. In particular, we also have under these conditions that μ satisfies an $LS(\frac{1}{\rho})$ inequality, and thus by Theorem 4.3.1, that μ satisfies a $T_2(\frac{1}{\rho})$ inequality.

This theorem seems appropriate to end this thesis on because it unifies these three major tools for comparing distributions. It unites three key perspectives on how distributions differ and offers an extremely elegant way to compare them.

Chapter 5

Modern Applications and Conclusion

This thesis has traced the relatively short history of transport-entropy inequalities from Monge's original formulation of the optimal transport problem in the 1780s to the central theorems of Chapter 4 discovered in 2000 by Otto and Villani in [26]. This thesis has primarily (and intentionally) focused on transport-entropy inequalities as objects that are of intrinsic interest for anyone who wants to study and compare measures, as well as a beautiful unifying approach to understanding many seemingly disparate inequalities.

Historically, however ,transport-entropy inequalities have entered the literature primarily as means to an end. These tools became popular and well studied in the 1990s, when it became increasingly clear that these inequalities were intimately related to the concentration of measure phenomenon.

While I hope I've demonstrated over these last four chapters that these tools are of independent and intrinsic interest, in this final, very short chapter, I'll break slightly from the conceit of the thesis and admit that my underlying motivation has not just been to explain and present these fascinating and beautiful tools, but especially to present them in an accessible way to people interested in probabilistic machine learning.

In particular, I think that these tools should be of interest to many people currently studying diffusion models. While there has been an enormous amount of interest in the last few years in studying diffusion as an optimal transport problem, there has been comparatively much less work which applies transport-entropy inequalities to the diffusion model setting. In the following section I'll offer a short introduction to diffusion models, and in the final section of the thesis I'll discuss a few recent papers which succeed in connecting diffusion models to Log-Sobolev and transport-entropy inequalities as well as mention a few open questions, which I hope are of interest.

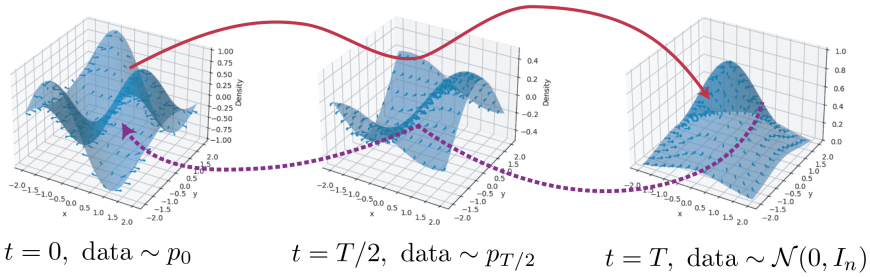


Figure 5.1: Intuition for a Diffusion Map [14]

5.1 Diffusion Models

Score-based generative models construct a generative process by learning to reverse a prescribed noising diffusion. The forward (data-destruction) process gradually perturbs samples from the data distribution according to a stochastic differential equation (SDE). A common choice is the Ornstein–Uhlenbeck or variance-preserving diffusion [36]:

$$d\mathbf{X}_t = \underbrace{f(\mathbf{X}_t, t)}_{\text{drift}}\, dt + \underbrace{g(t)}_{\text{diffusion}}\, dW_t \tag{5.1}$$

which defines a family of intermediate densities $(p_t)_{t\in[0,1]}$ that interpolate smoothly between the data distribution p_0 and a simple reference distribution (typically a standard Gaussian) at time $t = 1$.

The generative mechanism relies on the fact that this diffusion is approximately reversible. Under suitable regularity conditions, the reverse-time dynamics of the process satisfy another SDE:

$$d\mathbf{X}_t = \left[f(\mathbf{X}_t, t) - g(t)^2 \underbrace{\nabla_{\mathbf{X}_t} \log p_t(\mathbf{X}_t)}_{\text{score function}} \right] dt + g(t), d\bar{W}_t, \tag{5.2}$$

where the additional drift term depends on the score $\nabla_x \log p_t(x)$, i.e., the gradient of the log-density of the forward process at time t. Thus, generative sampling amounts to running this reverse-time SDE, provided the score function is known.

Since p_t is not available in closed form, we parametrize a neural network $s_\theta(\cdot, t)$ to approximate the score function. A natural training objective is the denoising score-matching loss:

$$\mathscr{L}(\theta) = \int_0^T \lambda(t) \mathbb{E}_{p_t(\mathbf{X})} \left[|s(\mathbf{X}_t, t) - s_\theta(\mathbf{X}_t, t)|_2^2 \right] dt, \tag{5.3}$$

where $\lambda(t)$ is a user-chosen weighting [14]. This is the classical perspective on diffusion models.

However, increasingly, people are formulating diffusion models as a transport problem [28]. In particular, we can reformulate the diffusion problem as follows. Suppose we have a reference distribution μ_0 which corresponds to a simple, easy to sample from distribution, like a multivariate Gaussian. Our target distribution is usually the true data distribution (which we'll call μ_1) from which we only have finitely many samples.

The key idea here is that we can understand $\mu_1 = T_{\#}\mu_0$, so if we can construct T explicitly, we can sample from μ_1 easily by drawing a sample from μ_0 and applying the transport map, ie $X_1 = T(X_0) \sim \mu_1$. The innovation of diffusion models in this framework is to introduce a time dependent interpolation of $(\mu_t)_{t\in[0,1]}$ which obeys certain nice continuity properties and then learn a vector field that represents how μ_t should change at that time step. For Gaussian μ_0, it can be shown that this vector field can be found as a conditional expectation, which diffusion models approximate by solving a least squares problem.

5.2 Recent Work and Open Problems

As discussed in [28], a detailed survey of optimal transport as it relates to diffusion models, the stochastic interpolates learned by modern diffusion models are crucially *not optimal in the transport sense*. This was proved in [21] only in 2022, and previously it had been conjectured that the Fokker-Planck equation provided the OT map between the target and initial densities. That such a fundamental question about the properties of diffusion models was open until relatively recently should give the reader some idea of the difficulty and interest of studying the properties of these models.

The question of in what sense the transport map implicitly constructed by diffusion models is optimal is very interesting, unsolved, and relevant to understanding how these models generalize. Understanding the geometric properties of the maps which diffusion models do learn is also an open and interesting question.

It is perhaps obvious from the structure of the diffusion problem (learn a map subject to some functional form constraints which sends p_0 to p_{data}) that transport-entropy inequalities might provide some insight. In the following section, I'll highlight a few recent papers that apply these tools and invite further study.

In [27], they derive a new bound on the W_2 distance between the true data distribution p_{true} and the learned distribution p_θ for their particular and new diffusion sampling algorithm when the data distribution is norm sub-gaussian (ie if $X \sim p_{\text{true}}$, $P(||X-\mathbb{E}[X]|| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$) and the score approximation s_θ is L_s-Lipschitz. Their key contribution is to provide a bound

$$W_2(p_{\text{true}}, p_\theta) \leq T_i + T_{ii} + T_{iii}$$

where T_i is a constant derived from the early-stopping of the reverse process (for computational reasons, diffusion models are never run all the way to $t = 0$, but instead to some $\epsilon > 0$), T_{ii} is a bound in terms of $\sqrt{\mathscr{L}(\theta)}$ where $\mathscr{L}(\theta)$ is the score-matching loss, which captures the approximation error between $s_\theta(t,\cdot)$ and ∇p_t. Finally T_{iii} comes from approximation error due to the Langevin dynamics. Interestingly their proof relies on the Log-Sobolev Inequality (and the Otto-Villani theorem) to derive T_{iii} in order to bound the Langevin sampling dynamics. This paper offers good evidence for the fact that TE inequalities and Log-Sobolev inequalities are extremely useful for analysis of diffusion models.

Another such paper [9] gives non-asymptotic estimates for the Wasserstein-2 distance between the p_{true} and p_θ under a series of assumptions, first that the optimization algorithm is sufficiently nice in the sense described in Assumption 1 in [9] and that the data distribution has finite second moment and is strongly log-concave, as well as $\int_\epsilon^T |\nabla \log p_t(0)|^2 dt < \infty$, the approximation for the score s_θ is continuously differentiable and Lipschitz. Then, $W_2(p_{\text{true}}, p_\theta)$ is bounded by four different constants, which are fairly involved and described in detail in Appendix E of [9].

Both these papers make very good use of these transport-entropy tools to bound the loss of interest. However, there is another, more tentative sense in which transport-entropy inequalities may be useful for understanding the geometry of the non-optimal stochastic interpolatant maps learned by diffusion models. This section is rather speculative, but I hope it provides a sense of what these tools have to offer to the machine learning world.

Two other recent papers provide independent bounds, one on the relative entropy and the other on the W_2 distance between the true data distribution p_{true} and the learned distribution p_θ. In [35], they show that (under some fairly mild regularity conditions) the KL-divergence between p_{true} and p_θ can be upper bounded by $\mathscr{L}(\theta) + K$ for a relatively simple constant K. This shows that training effectively decreases upper bound on the KL-divergence. In [20] they follow a similar process to demonstrate that (again under suitable regularity conditions) training a diffusion model by minimizing the score matching loss implicitly minimizes the Wasserstein-2 distance between p_{true} and p_θ. Both upper bounds are given in terms of functions of the score-matching loss. If the two bounds in these papers are compatible in the sense that $W_2(p_{\text{true}}, p_\theta) \leq K\alpha(D_{KL}(p_{true}||p_\theta)$ then that transport-entropy style inequality might be of value in understanding certain geometric properties of diffusion maps.

Bibliography

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows : In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1 edition, 2005. ISBN 9783764324285. 7

[2] Gennaro Auricchio, Gabriele Loli, and Marco Veneroni. On the computation of the infinity wasserstein distance and the wasserstein projection problem, 2025. URL https://arxiv.org/abs/2508.10589. 39, 42

[3] D. Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*. Grundlehren der mathematischen Wissenschaften, 348. Springer, Cham, 2014. ISBN 3319002260. 59, 63, 65, 67, 73, 74, 75

[4] Mathias Beiglböck and Markus Zona. Pinsker's inequality for adapted total variation, 2025. URL https://arxiv.org/abs/2506.22106. 48

[5] Patrick Billingsley. *Probability and Measure*. Wiley, New York, 1995. ISBN 9780471007104. 8

[6] Patrick Billingsley. *Convergence of probability measures*. Wiley series in probability and statistics. Probability and statistics. Wiley, New York, 2nd ed. edition, 1999. ISBN 0471197459. 8, 23, 24, 41

[7] Joseph Blitzstein and Carl Morris. Probability for statistical science. Unpublished textbook manuscript for Stat 210, 2024. 3, 12, 31, 70

[8] J. Bretagnolle, C. Huber, M. Weil, P. A. Meyer, and C. Dellacherie. Estimation des densités : Risque minimax. In *Séminaire de Probabilités XII*, Lecture Notes in Mathematics, pages 342–363. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 3540087613. 51

[9] Stefano Bruno, Ying Zhang, Dong-Young Lim, Ömer Deniz Akyildiz, and Sotirios Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates, 2025. URL https://arxiv.org/abs/2311.13584. 84

[10] Clément L. Canonne. A short note on an inequality between kl and tv, 2023. URL https://arxiv.org/abs/2202.07198. 48, 51

[11] Thomas A. Courtade and Edric Wang. Subadditivity of the log-sobolev constant on convolutions, 2025. URL https://arxiv.org/abs/2508.19648. 73, 75

[12] T. M. Cover and Joy A. Thomas. *Elements of Information Theory*. J. Wiley, Hoboken, N.J, 2nd ed. edition, 2006. ISBN 0471241954. 9, 10, 12

[13] Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019. 58

[14] Emma Finn, Binxu Wang, T. Anderson Keller, and Demba Ba. Where the score lives: A wavelet view of diffusion. In *NeurIPS Workshop on Structured Probabilistic Inference and Generative Modeling: Probabilistic Inference in the Era of Large Foundation Models*, 2025. 82

[15] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a wasserstein loss, 2015. URL https://arxiv.org/abs/1506.05439. 46

[16] Nathael Gozlan. A characterization of dimension free concentration in terms of transportation inequalities. *The Annals of Probability*, 37(6), 2009. ISSN 0091-1798. doi: 10.1214/09-aop470. URL http://dx.doi.org/10.1214/09-AOP470. 76, 77

[17] Nathael Gozlan. Transport-entropy inequalities on the line, 2012. URL https://arxiv.org/abs/1203.0326. 29

[18] Nathael Gozlan and Christian Léonard. Transport inequalities. a survey, 2010. URL https://arxiv.org/abs/1003.3852. 14, 30, 33, 34, 48, 53, 56, 64, 76, 77, 78, 79, 80

[19] Leonard Gross. Logarithmic sobolev inequalities. *American journal of mathematics*, 97(4):1061–1083, 1975. ISSN 0002-9327. 67, 68

[20] Dohyun Kwon, Ying Fan, and Kangwook Lee. Score-based generative modeling secretly minimizes the wasserstein distance, 2022. URL https://arxiv.org/abs/2212.06359. 84

[21] Hugo Lavenant and Filippo Santambrogio. The flow map of the fokker–planck equation does not provide optimal transport. *Applied mathematics letters*, 133: 108225–, 2022. ISSN 0893-9659. 83

[22] Michel Ledoux. Concentration of measure and logarithmic sobolev inequalities, 1997. URL https://www.math.univ-toulouse.fr/~ledoux/Berlin.pdf. Lecture notes, Berlin, 3–7 November 1997. 14, 15, 68, 69

[23] Jesús A. De Loera and Edward D. Kim. Combinatorics and geometry of transportation polytopes: An update, 2013. URL https://arxiv.org/abs/1307.0124. 19

[24] Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, with a foreword by Jean Picard. 44

[25] Fredrik Nilsson. Whose entropy is it anyway? (part 1: Boltzmann, shannon, and gibbs), 2014. URL https://adami.natsci.msu.edu/blog/2014/6/25/whose-entropy-is-it-anyway-part-1-boltzmann-shannon-and-gibbs-. Blog post. 9

[26] F. Otto and C. Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of functional analysis*, 173(2):361–400, 2000. ISSN 0022-1236. 76, 80, 81

[27] Francesco Pedrotti, Jan Maas, and Marco Mondelli. Improved convergence of score-based diffusion models via prediction-correction, 2024. URL https://arxiv.org/abs/2305.14164. 83

[28] Gabriel Peyré. Optimal and diffusion transports in machine learning, 2025. URL https://arxiv.org/abs/2512.06797. 83

[29] D. Revuz and Marc Yor. *Continuous Martingales and Brownian Motion*. Grundlehren der mathematischen Wissenschaften, A Series of Comprehensive Studies in Mathematics, 293. Springer Berlin Heidelberg, Berlin, Heidelberg, 3rd ed. 1999. edition, 1999. ISBN 3-662-06400-6. 62

[30] Siobhan Roberts. A mathematician who makes the best of things: A conversation with. *The New York times*, 2025. ISSN 1553-8095. 4

[31] Steven Roman. *Advanced Linear Algebra*, volume 135 of *Graduate Texts in Mathematics*. Springer Nature, New York, third edition, 2007. ISBN 9780387728315. 38

[32] Sheldon M. Ross. *Stochastic Processes*. Wiley Series in Probability and Statistics. Wiley, New York, 2 edition, 1996. ISBN 978-0-471-12062-9. 61

[33] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians : Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications, 87. Springer International Publishing, Cham, 1st ed. 2015. edition, 2015. ISBN 3-319-20828-4. 26, 28, 40

[34] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. 9

[35] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models, 2021. URL https://arxiv.org/abs/2101.09258. 84

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456. 82

[37] M. Talagrand. Transportation cost for gaussian and other product measures. *Geometric and functional analysis*, 6(3):587–600, 1996. ISSN 1016-443X. 34

[38] Terance Tao. 245a, notes 5: Differentiation theorems. https://terrytao.wordpress.com/2010/10/16/245a-notes-5-differentiation-theorems/, October 2010. Blog post, accessed March 2026. 35

[39] Matthew Thorpe. Introduction to optimal transport. Lecture notes, University of Cambridge, 2018. URL https://www.damtp.cam.ac.uk/research/cia/files/teaching/Optimal_Transport_Notes.pdf. 4, 6, 7, 17, 20, 21, 22, 25, 26, 28

[40] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats. 51

[41] Cédric Villani. *Optimal transport : old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. edition, 2009. ISBN 9783540710493. 18, 48

[42] Alex Williams. A short introduction to optimal transport and wasserstein distance. *Its Neuronal* (research blog), October 2020. URL https://alexhwilliams.info/itsneuronalblog/2020/10/09/optimal-transport/. Blog post. 21

[43] Yihong Wu. Lecture notes on: Information-theoretic methods for high-dimensional statistics. Online lecture notes, 2020. URL http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf. Accessed: 2026-03-08. 48