National Technical University of Athens
School of Electrical and Computer Engineering
School of Applied Mathematical and Physical Sciences
Department of Mathematics

Optimal Transport, Wasserstein Spaces and Applications to Machine Learning

Diploma Thesis

of

PANTELIS EMMANOUIL

Supervisor: Michalis Loulakis
Professor, NTUA

Athens, June 2025

National Technical University of Athens
School of Electrical and Computer Engineering
School of Applied Mathematical and Physical Sciences
Department of Mathematics

Optimal Transport, Wasserstein Spaces and Applications to Machine Learning

Diploma Thesis

of

PANTELIS EMMANOUIL

Supervisor: Michalis Loulakis
Professor, NTUA

Approved by the examination committee on 30th June 2025.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Michalis Loulakis	Dimitris Fotakis	Aris Pagourtzis
Professor, NTUA	Professor, NTUA	Professor, NTUA

Athens, June 2025

Copyright © – Pantelis Emmanouil, 2025
All rights reserved.

The copying, storage and distribution of this diploma thesis, exall or part of it, is prohibited for commercial purposes. Reprinting, storage and distribution for non - profit, educational or of a research nature is allowed, provided that the source is indicated and that this message is retained.

The content of this thesis does not necessarily reflect the views of the Department, the Supervisor, or the committee that approved it.

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....

Pantelis Emmanouil
Graduate of School of Electrical and Computer Engineering,
National Technical University of Athens

30th June 2025

to my parents

Περίληψη

Στην διπλωματική αυτή εργασία ασχολούμαστε με τη θεωρία και τις εφαρμογές του Optimal Transport (OT), με ιδιαίτερη έμφαση στον ρόλο της στην ανάλυση και την ποσοτικοποίηση της συμπεριφοράς γενίκευσης των βαθιών νευρωνικών δικτύων (DNN). Παρουσιάζουμε μια διπλή προσέγγιση: αρχίζουμε με μια αυστηρή θεωρητική ανάπτυξη του optimal transport που βασίζεται σε θεμελιώδεις μαθηματικές έννοιες (μετρικοί χώροι, θεωρία μέτρου και συναρτησιακή ανάλυση) και στη συνέχεια, διερευνούμε εμπειρικά τις αποστάσεις Wasserstein στο πλαίσιο της μηχανικής μάθησης. Το θεωρητικό μέρος καλύπτει τις διατυπώσεις του προβλήματος από τους Monge και Kantorovich, τη θεωρία δυϊκότητας και τη γεωμετρία του χώρου Wasserstein. Στο υπολογιστικό κομμάτι, μελετάμε τη σύγκλιση των εμπειρικών αποστάσεων Wasserstein για διάφορες περιπτώσεις δειγματοληψίας και συναρτήσεων κόστους, επιβεβαιώνοντας τα θεωρητικά αποτελέσματα και εξετάζοντας την εξάρτησή τους από τη διάσταση. Εξερευνούμε τη χρήση των pushforward αποστάσεων Wasserstein από τους Λουλάκη και Μακριδάκη για τη μελέτη του σφάλματος γενίκευσης στη βαθιά μάθηση και βλέπουμε ότι αυτές οι αποστάσεις προσφέρουν μια καλύτερη ποσοτικοποίηση του σφάλματος γενίκευσης από τα παραδοσιακά όρια. Είναι ενδιαφέρον ότι τα αποτελέσματά μας αποκαλύπτουν ότι η συμπεριφορά σύγκλισης των αποστάσεων pushforward συχνά αποκλίνει από την κλασική θεωρία OT, υποδηλώνοντας μια βαθύτερη αλληλεπίδραση μεταξύ της γεωμετρίας των δεδομένων και της δυναμικής μάθησης των νευρωνικών δικτύων. Τα ευρήματά μας αναδεικνύουν τις δυνατότητες του optimal transport στη σύγχρονη μελέτη της μηχανικής μάθησης και αναδεικνύουν νέες πολλά υποσχόμενες κατευθύνσεις στην κατανόηση του σφάλματος γενίκευσης.

Λέξεις Κλειδιά

Optimal Transport, Απόσταση Wasserstein, Εμπειρικό Μέτρο, Μηχανική Μάθηση, Στατιστική Σύγκλιση, Σφάλμα Γενίκευσης.

Abstract

This thesis explores the theory and applications of optimal transport (OT), with a particular emphasis on its role in analyzing and quantifying the generalization behavior of deep neural networks (DNNs). We present a dual approach: a rigorous theoretical development of optimal transport grounded in foundational mathematics (metric spaces, measure theory, and functional analysis), and a comprehensive empirical investigation of Wasserstein distances in machine learning contexts. The theoretical component covers the Monge and Kantorovich formulations, duality theory, and the geometry of the Wasserstein space. On the computational side, we study the convergence of empirical Wasserstein distances under various sampling schemes and cost functions, validating theoretical rates and examining their dependence on dimension. We explore the use of pushforward Wasserstein distances by Loulakis and Makridakis to study generalization error in deep learning. We demonstrate that these distances, induced by learned feature representations, offer a tighter quantification of generalization error than traditional bounds. Interestingly, our results reveal that the convergence behavior of pushforward distances often deviates from classical OT theory, hinting at a deeper interaction between data geometry and neural network learning dynamics. Our findings highlight the potential of optimal transport as a principled tool in modern machine learning and suggest promising directions for future theoretical work on the convergence of learned transport maps.

Keywords

Empirical Measure, Generalization error, Machine learning, Optimal transport, Statistical convergence, Wasserstein distance.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Michalis Loulakis, for his invaluable guidance throughout the thesis process. His support during my PhD application, his introduction to advanced mathematical concepts, and his inspiring mentorship in the field of optimal transport have been instrumental in shaping both this work and my academic path.

I am also deeply thankful to Professor Dimitris Fotakis for sharing with me his passion for mathematically grounded algorithms, as well as for his important guidance throughout my studies and during the PhD application period.

I would like to thank Professor Aris Pagourtzis for his support during my undergraduate studies and for his role in fostering a strong academic foundation.

Special thanks are due to Professor Vasilis Nakos for his continuous support, insightful advice, and thoughtful guidance throughout the thesis process.

Most importantly, I want to thank my family — my parents, Giannis and Kaiti, for their unconditional love, support, and belief in me, and my brothers, Dimitris and Christos, for their patience, encouragement, and help with understanding complex mathematical concepts.

Finally, I am deeply grateful to my friends, who provided motivation, balance, and a constant reminder to take breaks during this demanding period. A special mention goes to Asterios Tsiourvas for his help with the experimental section, including valuable discussions and access to additional computational resources.

Athens, June 2025

Pantelis Emmanouil

Contents

List of Figures

List of Tables

Κεφάλαιο 1

Εκτεταμένη Ελληνική Περίληψη

Στην περίληψη αυτή, αρχικά θα περιγράψουμε τα βασικότερα αποτελέσματα της θεωρίας του Optimal Transport, και στη συνέχεια θα εξετάσουμε τη θεωρία σύγκλισης των αποστάσεων Wasserstein μεταξύ εμπειρικών και πραγματικών μέτρων, επαληθεύοντας τα αποτελέσματα πειραματικά.
Η θεωρία του Optimal Transport βασίζεται σε έννοιες από διάφορους τομείς των μαθηματικών. Στην παρούσα περίληψη, υποθέτουμε ότι ο αναγνώστης διαθέτει εξοικείωση με τις βασικές έννοιες των παρακάτω περιοχών, οι οποίες περιγράφονται πιο αναλυτικά στο αγγλικό σκέλος της εργασίας:

- **Μετρικοί Χώροι και Τοπολογία:** Μετρικές, νόρμες, ανοιχτά/κλειστά σύνολα, σύγκλιση ακολουθιών, πληρότητα, συμπάγεια.
- **Θεωρία Μέτρου:** σ-άλγεβρες, μέτρα, μετρήσιμες συναρτήσεις, ολοκλήρωση Lebesgue, μέτρα πιθανότητας, push-forward μέτρα, ασθενής σύγκλιση μέτρων, μέτρα γινόμενο.
- **Συναρτησιακή Ανάλυση:** Χώροι Banach, δυϊκοί χώροι, ασθενείς τοπολογίες, θεώρημα Hahn-Banach, στοιχεία κυρτής ανάλυσης.

1.1 Optimal Transport

Η θεωρία του Optimal Transport εξετάζει τον πιο αποδοτικό τρόπο μετακίνησης μάζας από μια κατανομή $\mu \in \mathcal{P}(X)$ σε μια άλλη κατανομή $\nu \in \mathcal{P}(Y)$, δεδομένης μιας συνάρτησης κόστους $c : X \times Y \to \mathbb{R}$ που κοστολογεί τη μεταφορά μάζας από ένα σημείο $x \in X$ σε ένα σημείο $y \in Y$. Στόχος είναι η ελαχιστοποίηση του συνολικού κόστους μεταφοράς.

Διατυπώσεις Monge και Kantorovich Η αρχική διατύπωση του Monge (1781) αναζητά έναν μετασχηματισμό $T : X \to Y$ που μεταφέρει τη μ στη ν (δηλαδή $T_{\#}\mu = \nu$) και ελαχιστοποιεί το συνολικό κόστος:

$$\inf_{T_{\#}\mu=\nu} \int_X c(x, T(x))\, d\mu(x)$$

Ωστόσο, ο μετασχηματισμός T δεν υπάρχει πάντα. Ο Kantorovich (1942) εισάγει την πιο ευέλικτη έννοια του πλάνου μεταφοράς $\pi \in \Pi(\mu, \nu)$, δηλαδή ενός μέτρου στο $X \times Y$ με περιθώριες κατανομές μ και ν, και διατυπώνει το πρόβλημα ως:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X\times Y} c(x,y)\, d\pi(x,y)$$

Σε αυτήν τη διατύπωση του προβλήματος, αποδεικνύεται ότι, υπό κάποιες σχετικά ελαφριές προϋποθέσεις, το βέλτιστο πλάνο υπάρχει.

Θεώρημα: Έστω X, Y Polish χώροι και $c : X \times Y \to [0, +\infty]$ μια κάτω ημι-συνεχής συνάρτηση. Τότε, υπάρχει $\pi^* \in \Pi(\mu, \nu)$ που επιτυγχάνει το ελάχιστο κόστος μεταφοράς, δηλαδή:

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

Δυϊκή θεωρία του Kantorovich Το πρόβλημα βέλτιστης μεταφοράς του Kantorovich είναι πρόβλημα κυρτής βελτιστοποίησης: το σύνολο των επιτρεπτών σχεδίων μεταφοράς $\Pi(\mu, \nu)$ είναι κυρτό, και το κόστος $\int c\, d\pi$ εξαρτάται γραμμικά από τη μεταβλητή π. Σε αυτό το πλαίσιο, είναι φυσικό να εξετάσουμε τη σχετιζόμενη δυϊκή διατύπωση του προβλήματος.

Θεώρημα (Δυϊκό πρόβλημα Kantorovich):

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X\times Y} c(x,y)\, d\pi(x,y) = \sup_{\varphi,\psi \in \Phi_c} \left\{ \int_X \varphi(x)\, d\mu(x) + \int_Y \psi(y)\, d\nu(y) \right\}$$

όπου το σύνολο των επιτρεπτών δυϊκών μεταβλητών είναι:

$$\Phi_c = \left\{ (\varphi, \psi) \in L^1(\mu) \times L^1(\nu) \;\middle|\; \varphi(x) + \psi(y) \le c(x,y) \quad \forall x \in X,\ y \in Y \right\}.$$

Διαισθητικά, σκεφτείτε ότι ένας βιομήχανος θέλει να μεταφέρει προϊόντα από εργοστάσια (κατανομή $\mu \in \mathcal{P}(X)$) προς πελάτες (κατανομή $\nu \in \mathcal{P}(Y)$), με κόστος μετακίνησης $c(x, y)$. Μια εταιρεία μεταφορών προτείνει να παίρνει προϊόντα από το εργοστάσιο x χρεώνοντας $\varphi(x)$ και να τα παραδίδει στον πελάτη y χρεώνοντας $\psi(y)$, όπου $\varphi(x) + \psi(y) \le c(x, y)$ για όλα τα ζεύγη (x, y). Το παραπάνω θεώρημα δηλώνει ότι, επιλέγοντας κατάλληλες τιμές φ και ψ, η εταιρεία θα χρεώσει ακριβώς όσο είναι το βέλτιστο κόστος μεταφοράς στον βιομήχανο.

Όταν το κόστος είναι μια μετρική $c(x, y) = d(x, y)$, η δυϊκότητα λαμβάνει πιο απλή μορφή:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int d(x,y)\, d\pi(x,y) = \sup_{\varphi \in \mathrm{Lip}_1(X)} \int_X \varphi(x)\, d(\mu - \nu)(x)$$

όπου το supremum λαμβάνεται σε όλες τις 1-Lipschitz συναρτήσεις $\varphi : X \to \mathbb{R}$.
Η ισότητα του πρωτεύοντος και του δυϊκού προβλήματος (δηλαδή η απουσία δυϊκού χάσματος) ισχύει υπό ήπιες υποθέσεις, όπως η κάτω ημισυνεχεία του c και η ύπαρξη πλάνου με πεπερασμένο κόστος.

Αποστάσεις Wasserstein Όταν τα μέτρα μ και ν ορίζονται στον ίδιο χώρο X, το κόστος μεταφοράς είναι της μορφής $c(x,y) = d(x,y)^p$, για κάποια $p \geq 1$, ορίζουμε την απόσταση Wasserstein τάξης p ως:

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

Η ποσότητα αυτή επάγει μια απόσταση στον χώρο $P_p(X)$ των κατανομών πιθανοτήτων με πεπερασμένη p-οστή ροπή. Δηλαδή, η W_p ικανοποιεί όλες τις ιδιότητες μιας μετρικής: μη αρνητικότητα, συμμετρία, τριγωνική ανισότητα και $W_p(\mu,\nu) = 0 \iff \mu = \nu$.

Σε αντίθεση με τις σημειακές αποστάσεις μεταξύ κατανομών, όπως η ℓ^∞, η απόσταση Wasserstein ενσωματώνει τη γεωμετρική δομή του υποκείμενου χώρου X: δεν εξετάζει απλώς αν οι κατανομές διαφέρουν, αλλά πόσο μακριά βρίσκονται μεταξύ τους, μετρώντας το ελάχιστο έργο που απαιτείται για να μεταφερθεί η μάζα της μίας κατανομής στην άλλη.

Οι αποστάσεις Wasserstein σχετίζονται στενά με τη θεωρία ασθενούς σύγκλισης μέτρων:

- Αν ο X είναι συμπαγής μετρικός χώρος, τότε η σύγκλιση $W_p(\mu_n,\mu) \to 0$ ισοδυναμεί με ασθενή σύγκλιση $\mu_n \xrightarrow{w} \mu$.

$$W_p(\mu_n,\mu) \to 0 \quad \iff \quad \mu_n \xrightarrow{w} \mu \quad \iff \quad \forall f \in C_b(X),\ \int f\, d\mu_n \to \int f\, d\mu.$$

- Αν ο X είναι Polish αλλά όχι συμπαγής, τότε η ισοδυναμία απαιτεί επιπλέον σύγκλιση των p-οστών ροπών:

$$W_p(\mu_n,\mu) \to 0 \quad \iff \quad \mu_n \xrightarrow{w} \mu \quad \text{και} \quad \int d(x,x_0)^p \, d\mu_n(x) \to \int d(x,x_0)^p \, d\mu(x)$$

Μονοδιάστατη Περίπτωση Στην ειδική περίπτωση $X = Y = \mathbb{R}$ με κόστος $c(x,y) = |x-y|^p$, το πρόβλημα βέλτιστης μεταφοράς απλοποιείται σημαντικά και επιτρέπει ρητή λύση μέσω των αντίστροφων συναρτήσεων κατανομής.
Έστω $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$ και F_μ, F_ν οι συναρτήσεις κατανομής τους. Οι αντίστροφες συναρτήσεις ορίζονται ως

$$F_\mu^{-1}(t) := \inf\{x \in \mathbb{R} \mid F_\mu(x) \geq t\}, \quad t \in (0,1).$$

Τότε η απόσταση Wasserstein υπολογίζεται αναλυτικά ως

$$W_p^p(\mu,\nu) = \int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^p \, dt.$$

Στην περίπτωση αυτή, το βέλτιστο πλάνο μεταφοράς π^* είναι μονότονο: Αν τα $(x_1,y_1),(x_2,y_2)$ ανήκουν στο $supp(\pi^*)$ και $x_1 < x_2$ τότε $y_1 \leq y_2$. Το πλάνο αυτό είναι:

$$\pi^* = (F_\mu^{-1}, F_\nu^{-1})_{\#}\lambda$$

όπου λ είναι το μέτρο Lebesgue στο διάστημα $[0, 1]$.

Το πρόβλημα Monge έχει λύση στην περίπτωση που το μ δεν έχει άτομα (σημεία με θετική μάζα): $T = F_\nu^{-1} \circ F_\mu$.

Η παραπάνω μονοτονική ιδιότητα οδηγεί σε έναν απλό αλγόριθμο υπολογισμού της απόστασης Wasserstein στη μονοδιάστατη περίπτωση:
Για να κατασκευάσουμε το βέλτιστο πλάνο, ταξινομούμε τα δείγματα και κινούμενοι από το μικρότερο προς το μεγαλύτερο σημείο του μ, αναθέτουμε τη μάζα του στο μικρότερο διαθέσιμο στοιχείο του ν. Παράλληλα, προσθέτουμε το κόστος της μεταφοράς στο συνολικό μας βέλτιστο κόστος. Η πολυπλοκότητα του αλγορίθμου είναι $O(N \log N)$ λόγω της ταξινόμησης, ενώ αν τα δεδομένα είναι ήδη ταξινομημένα, είναι $O(N)$.

1.2 Σφάλμα γενίκευσης και Εμπειρικά Μέτρα

Εξετάζουμε τη σύγκλιση της αναμενόμενης απόστασης Wasserstein $\mathbb{E}[W_p(\hat{\mu}_n, \mu)]$ μεταξύ μιας κατανομής $\mu \in \mathbb{R}^d$ και του εμπειρικού μέτρου $\hat{\mu}_n$ που προκύπτει από n ανεξάρτητα και ισόνομα δείγματα $X_i \sim \mu$:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

Αποδεικνύεται ότι, $\hat{\mu}_n \xrightarrow{w} \mu$ σχεδόν σίγουρα.

Επιπλέον, αν $\mu \in \mathcal{P}_p(X)$ (δηλαδή έχει πεπερασμένη p-οστή ροπή), ισχύει σχεδόν σίγουρα:

$$W_p(\hat{\mu}_n, \mu) \to 0.$$

$$\mathbb{E}[W_p(\hat{\mu}_n, \mu)] \to 0.$$

Ο ρυθμός σύγκλισης της αναμενόμενης απόστασης $\mathbb{E}[W_p(\hat{\mu}_n, \mu)]$ είναι της τάξης $n^{-1/2}$ για $p = 1$ και της τάξης $O(n^{-1/d})$ για $p \neq 1$.

Για το αναμενόμενο κόστος μεταφοράς $\mathbb{E}[W_p^p(\hat{\mu}_n, \mu)]$, έχουμε ρυθμούς σύγκλισης της τάξης $n^{-p/2}$ για $p = 1$ και της τάξης $O(n^{-p/d})$ για $p \neq 1$.

Σφάλμα Γενίκευσης Νευρωνικών Δικτύων Η παρούσα ενότητα βασίζεται στο [6] και παρουσιάζει ένα νέο πλαίσιο ανάλυσης του σφάλματος γενίκευσης βαθιών νευρωνικών δικτύων με εργαλεία βέλτιστης μεταφοράς και αποστάσεων Wasserstein, χωρίς να απαιτούνται ισχυρές υποθέσεις για την αρχιτεκτονική τους.

Εξετάζουμε το πρόβλημα της παλινδρόμησης, όπου προσπαθούμε να εκπαιδεύσουμε ένα δίκτυο ώστε να προσεγγίσει μια συνάρτηση-στόχο $f : D \to \mathbb{R}$, έχοντας πρόσβαση μόνο στις τιμές της σε συγκεκριμένα δείγματα εκπαίδευσης που ακολουθούν την κατανομή μ. Έστω $\mathcal{N}$ η κλάση νευρωνικών δικτύων πάνω στα οποία εκπαιδεύουμε, η οποία παράγει τον χώρο συναρτήσεων $V_\mathcal{N} \subset L^p(D)$. Χρησιμοποιώντας ως μετρική σφάλματος την L^p απόσταση,

Θέλουμε να προσεγγίσουμε το

$$v^* = \arg\min_{v \in V_N} \int_D |f - v|^p d\mu.$$

Καθώς διαθέτουμε μόνο ένα περιορισμένο δείγμα $X(\omega) = (X_1(\omega), \ldots, X_n(\omega))$, ελαχιστοποιούμε το εμπειρικό σφάλμα

$$\mathcal{E}_{n,\omega}(v) = \frac{1}{n} \sum_{i=1}^{n} |f(X_i(\omega)) - v(X_i(\omega))|^p = \int_D |f - v|^p d\mu_{n,X(\omega)},$$

όπου $\mu_{n,X(\omega)} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}$. Έστω $u_X(\omega)$ η συνάρτηση που ελαχιστοποιεί το εμπειρικό αυτό σφάλμα.

Το συνολικό αναμενόμενο σφάλμα γράφεται:

$$\mathbb{E}\left[\|f - u_X\|_p\right],$$

και διαχωρίζεται σε δύο συνιστώσες:

- Σφάλμα προσέγγισης: προκύπτει από τους εκφραστικούς περιορισμούς της κλάσης V_N, και δίνεται από
$$\|f - v^*\|_p.$$
- Σφάλμα γενίκευσης: αποδίδεται στη διαφορά ανάμεσα στην εμπειρική και την πραγματική ελαχιστοποιούσα συνάρτηση.
$$\|u_X - v^*\|_p.$$

Για να ελεγχθεί το $\|u_X - f\|_p$, η ανάλυση στηρίζεται σε εκτίμηση του τύπου:

$$\|u_X(\omega) - f\|_p \leq \|v^* - f\|_p + W_p\left[(u_X(\omega) - f)_{\#}\mu,\ (u_X(\omega) - f)_{\#}\mu_{N,X(\omega)}\right].$$

Η παραπάνω εκτίμηση συνδέει το σφάλμα γενίκευσης με την απόσταση Wasserstein μεταξύ των push-forward μέτρων μέσω του σφάλματος πρόβλεψης $(u_X - f)$.

Αυτή η απόσταση μπορεί να εκτιμηθεί από πάνω μέσω της Lipschitz σταθεράς $L_X(\omega)$ της συνάρτησης $(u_X(\omega) - f)$, οδηγώντας στην πιο χαλαρή αλλά χρήσιμη ανισότητα:

$$\|u_X(\omega) - f\|_p \leq \|v^* - f\|_p + L_X(\omega) \cdot W_p\left(\mu,\ \mu_{N,X(\omega)}\right).$$

Αν υποθέσουμε ακόμη ότι σχεδόν σίγουρα ισχύει $L_X(\omega) \leq L_N$, τότε λαμβάνουμε την αναμενόμενη εκτίμηση:

$$\mathbb{E}\left[\|u_X - f\|_p\right] \leq \|v^* - f\|_p + L_N \cdot \mathbb{E}\left[W_p(\mu, \mu_{N,X})\right].$$

Η παραπάνω ανάλυση εδραιώνει έναν διαχωρισμό του σφάλματος σε δύο ερμηνεύσιμα μέρη: (α) την ικανότητα έκφρασης του μοντέλου και (β) την επίδραση της δειγματοληψίας. Επιπλέον, παρακινεί τη χρήση της απόστασης Wasserstein μεταξύ των push-forward μέτρων

στην πράξη, καθώς τα φράγματα που παρέχουν στο σφάλμα γενίκευσης είναι πιο αυστηρά.

Εκτίμηση Εμπειρικής Απόστασης Wasserstein Για την επαλήθευση των παραπάνω, διεξήχθησαν αριθμητικά πειράματα για τη μελέτη της σύγκλισης της αναμενόμενης εμπειρικής απόστασης Wasserstein $\mathbb{E}[W_p(\hat{\mu}_n, \mu)]$ σε κατανομές στον $\mathbb{R}^d$ για $d = 1, \ldots, 10$ και $p = 1, \ldots, 10$. Χρησιμοποιήθηκαν δύο βασικές κατανομές, η ομοιόμορφη στο $[0, 1]^d$ και η κανονική $\mathcal{N}(0, I_d)$. Η εμπειρική απόσταση υπολογίστηκε με χρήση του αλγορίθμου network simplex (POT βιβλιοθήκη), με πολλαπλές επαναλήψεις για εκτίμηση του μέσου όρου.

Τα αποτελέσματα επιβεβαιώνουν τη θεωρητική σύγκλιση $O(N^{-p/d})$ για την ομοιόμορφη κατανομή, ενώ για την κανονική παρατηρείται πιο αργή σύγκλιση λόγω μη φραγμένου φορέα του μέτρου.

Εκτίμηση Γενίκευσης μέσω Απόστασης Wasserstein των push-forward μέτρων Μελετήθηκε επίσης η σύγκλιση του σφάλματος γενίκευσης εκτιμώντας τις αναμενόμενες Wasserstein αποστάσεις μεταξύ των push-forward μέτρων $e_{\#}\mu_n$ και $e_{\#}\mu$, όπου $e(x) = |(u_X(\omega)(x) - f(x)|$ είναι το σημειακό σφάλμα εκπαίδευσης του δικτύου.

Οι μετρήσεις δείχνουν ότι οι αποστάσεις αυτές συγκλίνουν με ρυθμούς συνήθως πιο αργούς από το αναμενόμενο $n^{-1/2}$, πιθανώς επειδή η συνάρτηση e εξαρτάται άμεσα από το ίδιο το εμπειρικό μέτρο μ_n, το οποίο αποτελεί και σύνολο εκπαίδευσης του μοντέλου.

Για πιο λεπτομερή παρουσίαση των πειραμάτων και ανάλυση των αποτελεσμάτων τους, βλέπε την Ενότητα 5.5.

1.3 Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Η παρούσα εργασία μελέτησε σε βάθος τη θεωρία της βέλτιστης μεταφοράς και των αποστάσεων Wasserstein, με τελικό στόχο τη χρήση της στην κατανόηση και ποσοτικοποίηση του σφάλματος γενίκευσης σε βαθιά νευρωνικά δίκτυα. Όπως είδαμε, οι αποστάσεις Wasserstein, μεταξύ push-forward μέτρων παρέχουν αυστηρότερα φράγματα του σφάλματος γενίκευσης σε σχέση με τη χρήση σταθερών Lipschitz.
Ένα σημαντικό εύρημα της μελέτης ήταν ότι οι ρυθμοί σύγκλισης των αποστάσεων μεταξύ push-forward μέτρων συχνά αποκλίνουν από τις προβλέψεις της κλασικής θεωρίας για μονοδιάστατα εμπειρικά μέτρα, γεγονός που αποδίδουμε στη χρήση του εμπειρικού μέτρου τόσο για την εκπαίδευση του μοντέλου, όσο και για την αξιολόγησή του μέσω των αποστάσεων. Μελλοντική έρευνα θα μπορούσε να ασχοληθεί με τη θεωρητική διατύπωση του ακριβούς ρυθμού σύγκλισης αυτών των αποστάσεων. Παράλληλα, θα μπορούσε να μελετηθεί και η επίδραση διαφορετικών αρχιτεκτονικών και σχημάτων δειγματοληψίας σε αυτούς τους ρυθμούς σύγκλισης, συμβάλλοντας στον σχεδιασμό πιο αξιόπιστων και θεωρητικά τεκμηριωμένων αλγορίθμων βαθιάς μάθησης.

Chapter 2

Introduction

Optimal transport is a profound mathematical theory that seeks to answer a deceptively simple question: given two distinct configurations of "mass" or "resources," what is the most efficient way to transform one into the other? This fundamental inquiry dates back to the 18th century with Gaspard Monge's pioneering work on moving piles of earth while minimizing effort. Centuries later, Leonid Kantorovich revolutionized the field by rephrasing Monge's challenging non-linear problem into a more tractable linear programming framework, a contribution for which he was awarded the Nobel Prize in Economic Sciences. In its modern incarnation, optimal transport provides not only deep theoretical insights into disparate fields such as geometry, probability theory, and analysis, but also powerful practical tools with broad applicability across science and engineering.

In recent years, the optimal transport framework has experienced a remarkable resurgence, establishing itself as an increasingly indispensable tool in applied mathematics, data science, and machine learning. Its ability to quantify dissimilarity between complex probability distributions in a geometrically meaningful way sets it apart from traditional statistical divergences. For instance, while classical metrics might declare two spatially separated but otherwise identical distributions (like two non-overlapping piles of sand) as infinitely distant, optimal transport measures the actual cost required to shift one pile to match the other. This inherent sensitivity to the underlying geometry of data has led to its successful employment in diverse tasks, including image processing (e.g., image registration and color transfer), domain adaptation (aligning data from different sources), and generative modeling (like Wasserstein GANs).

This thesis offers a dual perspective on optimal transport, bridging fundamental theory with cutting-edge applications. The first part provides a comprehensive theoretical exposition of optimal transport, beginning from foundational mathematical concepts. Recognizing that optimal transport draws upon a rich tapestry of advanced mathematics, we construct this theoretical groundwork carefully, assuming little specialized prior knowledge beyond a standard university-level background. We progressively establish the necessary mathematical foundations, including the topology of metric spaces, essential concepts from measure theory (such as σ-algebras, measurable functions, integration, and properties of probability distributions), and key elements of functional analysis (includ-

ing normed spaces, dual spaces, and weak topologies). Building upon this, we formally define the classical Monge problem and its more flexible Kantorovich relaxation. We then delve into crucial theoretical results, exploring the existence of optimal transport plans, the powerful duality theory (including Kantorovich-Rubinstein theorem that reveals the role of Lipschitz functions), the explicit solution in the one-dimensional case, and the formal definition and properties of Wasserstein metrics and their associated spaces. Our theoretical treatment extensively draws upon seminal works in the field, including those by Villani [8] and Santambrogio [7], ensuring a rigorous yet accessible foundation.

The second part of this thesis transitions from theoretical foundations to practical investigations, conducting a series of numerical experiments to explore the empirical behavior and applications of optimal transport in high-dimensional settings. We begin by examining the convergence rates of empirical Wasserstein distances, studying how well finite samples approximate the underlying true distributions. This investigation is informed by theoretical bounds derived from the works of researchers such as Bobkov, Ledoux [1]. Subsequently, we explore various efficient computational methods for solving optimal transport problems, ranging from general algorithms like Sinkhorn's algorithm and the Network Simplex Algorithm to specialized algorithms tailored for one-dimensional cases. Finally, we apply these insights to a critical challenge in modern machine learning: the analysis of generalization error in deep learning. Here, we leverage recent theoretical developments put forth by Loulakis and Makridakis in [6], which utilize pushforward Wasserstein distances to quantify the error arising from training deep neural networks on finite samples rather than the entire data distribution. Understanding how this generalization error depends on factors like the number of training samples (N) is paramount. It not only deepens our comprehension of why and how deep learning algorithms generalize but also provides crucial guidance for practical considerations, such as determining the optimal number of training samples required to achieve a desired level of accuracy.
By meticulously navigating both the rigorous theoretical landscape and the practical computational challenges of optimal transport, this thesis aims to contribute to a deeper understanding of its mathematical underpinnings and its growing relevance in addressing complex problems in data science and machine learning.

Chapter 3

Mathematical Background

3.1 Metric Spaces and Topology

3.1.1 Metric and Normed Spaces

Definition 3.1 (Metric Space)**.** *Let X be a nonempty set. A function $d : X \times X \to \mathbb{R}$ is called a* ***metric*** *on X if it has the following properties:*

- ***Non-negativity:*** *$d(x, y) \geq 0$ for all $x, y \in X$, and $d(x, y) = 0$ if and only if $x = y$*
- ***Symmetry:*** *$d(x, y) = d(y, x)$ for all $x, y \in X$*
- ***Triangle inequality:*** *$d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$*

If d is a metric on X, the pair (X, d) is called a ***metric space****.*

For any set X, a trivial example of a metric space is (X, d), where d is the discrete metric:

$$d(x, y) = \begin{cases} 0, & \text{if } x = y \\ 1, & \text{if } x \neq y \end{cases}$$

Definition 3.2 (Normed Space)**.** *Let X be a vector space over K ($\mathbb{R}$ or $\mathbb{C}$). A* ***norm*** *on X is a function $\|\cdot\| : X \to \mathbb{R}$ that has the following properties:*

- ***Non-negativity:*** *$\|x\| \geq 0$ for all $x \in X$, and $\|x\| = 0$ if and only if $x = 0$*
- ***Absolute homogeneity:*** *$\|\lambda x\| = |\lambda| \|x\|$ for all $x \in X, \lambda \in K$*
- ***Triangle inequality:*** *$\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$*

If $\|\cdot\|$ is a norm on X, the pair $(X, \|\cdot\|)$ is called a ***normed space****.*

A norm $\|\cdot\|$ induces the metric $d(x, y) = \|x - y\|$.

We will highlight some important examples of normed (metric) spaces.

- $\mathbb{R}^m$ **with the Euclidean norm and metric:** For $x, y \in \mathbb{R}^m$, the Euclidean norm is:

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_m^2},$$

and the induced metric (Euclidean distance) is:

$$d_2(x, y) = \|x - y\|_2 = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}.$$

- $\mathbb{R}^m$ **with the p-norm and metric:** For $1 \leq p < \infty$, the p-norm is:

$$\|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p \right)^{\frac{1}{p}},$$

and the induced metric is:

$$d_p(x, y) = \|x - y\|_p = \left(\sum_{i=1}^{m} |x_i - y_i|^p \right)^{\frac{1}{p}}.$$

- $\mathbb{R}^m$ **with the infinity norm and metric:** The infinity norm is:

$$\|x\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_m|\},$$

and the corresponding metric is:

$$d_\infty(x, y) = \|x - y\|_\infty = \max_i |x_i - y_i|.$$

- **Sequence space ℓ_p with the p-norm:** The space

$$\ell_p = \left\{ x = (x_1, x_2, \ldots) : \|x\|_{\ell_p} = \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{\frac{1}{p}} < \infty \right\}$$

has the metric:

$$d_{\ell_p}(x, y) = \left(\sum_{i=1}^{\infty} |x_i - y_i|^p \right)^{\frac{1}{p}}.$$

- **Sequence space ℓ_∞ with the supremum norm:**

$$\ell_\infty = \left\{ x = (x_1, x_2, \ldots) : \|x\|_\infty = \sup_{i \in \mathbb{N}} |x_i| < \infty \right\},$$

with the metric:

$$d_\infty(x, y) = \sup_{i \in \mathbb{N}} |x_i - y_i|.$$

- **Function space $L^p(\Omega)$ with the p-norm:** For a measure space (Ω, μ), the norm is:

$$\|f\|_p = \left(\int_\Omega |f(x)|^p \, d\mu(x) \right)^{\frac{1}{p}}.$$

  The induced metric is:

$$d_p(f, g) = \left(\int_\Omega |f(x) - g(x)|^p \, d\mu(x) \right)^{\frac{1}{p}}.$$

- **Space of continuous functions** $C(A)$ on a compact set A, with the supremum norm:

$$\|f\|_\infty = \sup_{x \in A} |f(x)|$$

  The induced metric is:

$$d_\infty(f, g) = \sup_{x \in A} |f(x) - g(x)|$$

3.1.2 Topological Concepts

We will now introduce some important topological concepts, such as convergence, open and closed sets.
An open set is a general topological concept, but in this analysis, we will limit ourselves to metric spaces.
We will call the set $B_d(x_0, \epsilon) = \{x \in X : d(x, x_0) < \epsilon\}$ an **open ball** centered at x_0 with radius ϵ.

Definition 3.3 (Open set)**.** *Let (X, d) be a metric space and $G \subset X$. G is called **open** if for every $x \in G$, there exists $\epsilon_x > 0$ such that $B_d(x, \epsilon_x) \subset G$.*

Definition 3.4 (Convergence of sequence)**.** *Let (x_n) be a sequence in metric space (X, d). We say that (x_n) **converges** to $x \in X$ and denote $x_n \to x$ if:*

$$\forall \epsilon > 0 \; \exists n_0 \in \mathbb{N} : n \geq n_0 \Rightarrow d(x_n, x) < \epsilon$$

It can be proven that the limit is unique and that if $x_n \to x$ and $y_n \to y$, then $d(x_n, y_n) \to d(x, y)$.

Definition 3.5 (Closed set)**.** *A subset $F \subset X$ is called **closed** if its complement F^c is open. In a metric space (X, d), F is closed if and only if for every sequence (x_n) in F that converges to $x \in X$: $x_n \to x$, we have that $x \in F$.*

Some important results about open and closed sets state that:

- A finite intersection or union of open (closed) sets is open (closed).
- An infinite union of open sets is an open set.
- An infinite intersection of closed sets is a closed set.

Definition 3.6 (Interior)**.** *Let A be a subset of a metric space (X, d). The* ***interior*** *of A, denoted as $\text{int}(A)$ or A°, is the set of all interior points of A. A point $x \in A$ is called an* ***interior point*** *if there exists $r > 0$ such that the open ball $B(x, r)$ is entirely contained in A:*

$$B(x, r) \subseteq A.$$

The interior of A is the ***largest open set*** *contained within A.*

Definition 3.7 (Closure)**.** *The* ***closure*** *of a set A, denoted $\overline{A}$, is the smallest closed set containing A. It can be defined equivalently as:*

$$\overline{A} = A \cup A',$$

where A' is the set of all ***limit points*** *of A, meaning points x such that there exists a sequence (x_n) in A with $x_n \to x$. Alternatively, $\overline{A}$ is the intersection of all closed sets containing A.*

Definition 3.8 (Boundary)**.** *The* ***boundary*** *of a set A, denoted as ∂A, consists of all points that are* ***neither purely interior nor purely exterior*** *to A. Formally,*

$$\partial A = \overline{A} \setminus \text{int}(A).$$

Equivalently, a point x belongs to ∂A if every open ball $B(x, r)$ intersects both A and A^c.

Definition 3.9 (Dense set)**.** *A subset $A \subset X$ is called* ***dense*** *in X if its closure is the entire space:*

$$\overline{A} = X.$$

This means that every open set in X contains at least one point of A. Equivalently, for every point $x \in X$, there exists a sequence $(a_n) \subset A$ such that $a_n \to x$ in the metric d.

Definition 3.10 (Separable space)**.** *A metric space (X, d) is called* ***separable*** *if it contains a countable dense subset. That is, there exists a countable set $D \subset X$ such that*

$$\overline{D} = X.$$

This means that every point of X can be approximated arbitrarily well by points from D, i.e., for every $x \in X$, there exists a sequence $(d_n) \subset D$ such that $d_n \to x$.

Examples:

- *$\mathbb{R}^n$ with the standard Euclidean metric is separable, since the set of rational points $\mathbb{Q}^n$ is a countable dense subset.*
- *$\mathbb{R}^n$ with the discrete metric $d(x, y) = 1$ for $x \neq y$ is not separable. In this case, no countable subset can be dense, since every singleton $\{x\}$ is an open set, meaning that the closure of any countable set is just itself and cannot cover the whole space.*

3.1.3 Functions

Continuity is one of the most important properties of functions, ensuring that small changes in the input lead to small changes in the output.

Definition 3.11 (Continuous function)**.** *A function $f : X \to Y$ is called **continuous** if it the inverse image of open sets is open, that is, for every open set $U \subset Y$, the set $f^{-1}(U)$ is open in X.*

When the function is defined between metric spaces, the above definition is equivalent to the $\epsilon - \delta$ definition we are familiar with.

Definition 3.12 (Epsilon-delta definition)**.** *A function $f : X \to Y$ between metric spaces (X, d_X) and (Y, d_Y) is **continuous at a point** $x \in X$ if for every $\epsilon > 0$, there exists $\delta > 0$ such that for all $y \in X$,*

$$d_X(x, y) < \delta \quad \Rightarrow \quad d_Y(f(x), f(y)) < \epsilon.$$

*If this holds for all $x \in X$, we say that f is **continuous**.*

The following important theorem links continuity of functions to sequence convergence.

Theorem 3.1. *A function $f : X \to Y$ is continuous if and only if it preserves the limit of sequences: whenever $x_n \to x$ in X, we have $f(x_n) \to f(x)$ in Y.*

The above $\epsilon - \delta$ definition of continuity ensures that small δ changes of the input result in small ϵ changes of the output. However, δ may depend on x. Uniform continuity is a stronger condition, that ensures that δ is independent of x.

Definition 3.13 (Uniform continuity)**.** *A function $f : X \to Y$ is **uniformly continuous** if for every $\epsilon > 0$, there exists $\delta > 0$ such that for all $x, y \in X$,*

$$d_X(x, y) < \delta \quad \Rightarrow \quad d_Y(f(x), f(y)) < \epsilon.$$

Unlike standard continuity, δ is chosen independently of x.

It is apparent that a uniformly continuous function is continuous. The opposite is not generally true: Take for example $f(x) = x^2$. While it is continuous, for large x values, even small changes of x can cause arbitrarily large changes of $f(x)$.
There exists a similar theorem linking uniform continuity to Cauchy sequences. We will first define Cauchy sequences as sequences whose terms get arbitrarily close to each other as n increases.

Definition 3.14 (Cauchy sequence)**.** *A sequence (x_n) in a metric space (X, d) is called **Cauchy** if for every $\epsilon > 0$, there exists an integer N such that for all $m, n \geq N$,*

$$d(x_m, x_n) < \epsilon.$$

Theorem 3.2. *A function $f : X \to Y$ is uniformly continuous if and only if it preserves Cauchy sequences, i.e., if (x_n) is a Cauchy sequence in X, then $(f(x_n))$ is a Cauchy sequence in Y.*

The final notion of continuity we will explore is Lipschitz continuity, where there is a bound L on how fast a function can change.

Definition 3.15 (Lipschitz continuity)**.** *A function $f : X \to Y$ is **Lipschitz continuous** if there exists a constant $L \geq 0$ such that for all $x, y \in X$,*

$$d_Y(f(x), f(y)) \leq L d_X(x, y).$$

.

3.1.4 Completeness

Now, we will focus on convergent and Cauchy sequences. It can be easily shown that every convergent sequence is Cauchy. However, the opposite is not always true. Take for example the sequence (a_n) of decimal approximations of $\sqrt{2}$ on $\mathbb{Q}$. While the elements of (a_n) get closer to each other as $n \to \infty$, thus it is Cauchy, it does not converge to any point of $\mathbb{Q}$.

An important result states that if a Cauchy sequence has a convergent subsequence, then the sequence itself is convergent.

Theorem 3.3. *Let (X, d) be a metric space, and let (x_n) be a Cauchy sequence in X. If (x_{n_k}) is a subsequence of (x_n) that converges to some limit $L \in X$, then (x_n) converges to L.*

Definition 3.16 (Complete Metric Space)**.** *A metric space (X, d) is **complete** if every **Cauchy sequence** (x_n) in X has a limit in X. That is, if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $m, n \geq N$,*

$$d(x_m, x_n) < \epsilon,$$

then there exists some $x \in X$ such that $x_n \to x$ as $n \to \infty$.

Examples of complete metric spaces include $\mathbb{R}^n$ with the Euclidean or the p-metric, all discrete metric spaces and the space of continuous functions with the supremum norm. Any closed subset of a complete metric space is also complete.

Every incomplete metric space X has a **completion**, which is a larger complete space $\hat{X}$ containing X as a dense subset. Intuitively, for each Cauchy sequence in X, $\hat{X}$ contains its limit; that is, it contains all points of X, as well as all "gaps" X might have. For example, $\mathbb{R}$ is the completion of $\mathbb{Q}$.

3.1.5 Compactness

We first present the topological definition of compactness.

Definition 3.17 (Compact Set)**.** *A subset K of a metric space (X, d) (it could be $K = X$) is* ***compact*** *if, for every open cover $\{U_i\}$ of K ($K \subseteq \bigcup_i U_i$, where each U_i is open), there exists a finite subcover $\{U_{i_1}, U_{i_2}, \ldots, U_{i_k}\}$ such that $K \subseteq \bigcup_{j=1}^{k} U_{i_j}$.*

An important result by Bolzano and Weierstrass links compactness with subsequence convergence.

Theorem 3.4 (Bolzano-Weierstrass)**.** *A subset K of a metric space (X, d) is compact if and only if every sequence in K has a convergent subsequence: If (x_n) is a sequence in K, there exists a subsequence (x_{n_k}) that converges to some $x \in K$:*

$$\lim x_{n_k} = x$$

We now extend the usual definition of bounded sets to metric spaces.

Definition 3.18 (Bounded set)**.** *A subset A of a metric space (X, d) is called* ***bounded*** *if there exists a point $x_0 \in X$ and a real number $M > 0$ such that*

$$d(x, x_0) \leq M, \quad \forall x \in A.$$

This means that all points of A lie within some ball of finite radius centered at x_0.

Let (X, d) be a metric space. Every compact subset of X is **closed and bounded**. If K was not closed, we could take a sequence converging in $X \setminus K$, thus every subsequence it has would also converge outside the set. If K was unbounded, we could take a sequence escaping every bounded set, thus every subsequence would be non convergent.
When X is $\mathbb{R}^n$ equipped with the Euclidean norm, the above relation goes both ways:

Theorem 3.5 (Heine-Borel)**.** *A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded. This theorem extends to all finite dimension spaces.*

It is also simple to show that every compact subset K of a metric space is complete. If it weren't complete, there would exist a Cauchy sequence that does not converge in K, thus every subsequence it has would not converge in K, thus it would not be compact.
An important theorem links compactness with completeness and total boundedness.

Definition 3.19 (Totally bounded set)**.** *A subset K of a metric space (X, d) is called totally bounded if for every $\epsilon > 0$, there exist $m \in \mathbb{N}$ and $x_1, \ldots, x_m \in X$ such that:*

$$X \subset \bigcup_{i=1}^{m} B(x_i, \epsilon)$$

To give some intuition, a bounded set can fit inside a large ball, while a totally bounded set can be covered by finitely many small balls of any given radius. In finite dimension spaces, like $\mathbb{R}^n$, the two concepts are equivalent; in infinite dimension spaces, however, this is not the case. While a totally bounded set is always bounded, there exist sets, like the unit ball in $C([0,1])$ (or any infinite dimension space) that are bounded but not totally bounded.
An important result states that a subset K of a metric space (X,d) is totally bounded if and only if every sequence (x_n) in K has a Cauchy subsequence. This leads to the following theorem.

Theorem 3.6. *A subset of a metric space is compact if and only if it is totally bounded and complete.*

An important property of totally bounded spaces is that they are separable (take the sequence of balls with radius $\frac{1}{n}$ that cover the space for every n). This immediately leads to the following result:

Theorem 3.7. *Every compact space is separable.*

Now, we will highlight some interesting properties of continuous functions defined on compact sets.

- Every continuous function on a compact metric space is uniformly continuous.
- If $f : X \to Y$ is continuous and X is compact, then $f(X)$ is compact in Y. In general, a continuous function maps compact sets to compact sets.
- A continuous function on a compact space attains its maximum and minimum values.
- If $f : X \to Y$ is a continuous bijection ($1-1$ and onto) from a compact space X to Y, then its inverse f^{-1} is also continuous.

3.2 Measure Theory

3.2.1 Sigma Algebra and Measures

Let X be a set. We would like to assign a probability (or, in more general cases, a measure) to subsets of X. We will define the σ-algebras as the collections of sets we can assign probabilities to. Ideally, we would like to be able to assign a probability to every subset; however, as we will see later, this is not always possible.

Definition 3.20 (σ-algebra)**.** *A σ-algebra on a set X is a subset of the power set of X $\mathcal{F} \subset \mathcal{P}(X)$ that satisfies the following properties:*

- *It contains the empty set: $\emptyset \in \mathcal{F}$.*
- *It is closed under complements: If $A \in \mathcal{F}$ then $X \setminus A \in \mathcal{F}$.*
- *It is closed under countable unions: If $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.*

Some examples of σ-algebras are:

- The trivial σ-algebra on X: $\{\emptyset, X\}$.
- The set $\mathcal{P}(X)$ of all subsets of X.
- Let C be a collection of subsets of X. The σ-algebra generated by C, $\sigma(C)$, is the smallest σ-algebra containing C.
- The Borel σ-algebra on a metric space X, $\mathcal{B}(X)$, is the σ-algebra generated by the collection of open sets on X. $\mathcal{B}(\mathbb{R})$ is of particular interest to us, as it contains all intervals of $\mathbb{R}$. For metric spaces, the Borel σ-algebra can also be generated by the collection of all closed sets, or by ϵ-balls.

Let X be a set and $\mathcal{A}$ a σ-algebra on X. We call $(X, \mathcal{A})$ a measurable space. Now, we will define measures on measurable spaces.

Definition 3.21 (Measure)**.** *Let $(X, \mathcal{A})$ be a measurable space. We define a measure on $(X, \mathcal{A})$ a function $\mu : \mathcal{A} \to [0, \infty]$ that satisfies the following:*

- *Null empty set: $\mu(\emptyset) = 0$.*
- *Countable additivity: If (A_n) is a sequence of disjoint elements in $\mathcal{A}$: $\mu(\bigcup_{n=1}^{\infty}) = \sum_{n=1}^{\infty} \mu(A_n)$,*

Then, we call $(X, \mathcal{A}, \mu)$ a measure space.

A **signed** measure is a measure that can also take negative values.

A measure space where $\mu(X) < \infty$ is called a **finite measure space**.

Some examples of measures include:

- The counting measure, where $\mu(A)$ is the number of elements in A (∞ if they are infinite).

- The Dirac measure δ_{x_0}, where we select an element $x_0 \in X$ and let $\delta_{x_0}(A) = 1$ if $x_0 \in A$ and $\delta_{x_0}(A) = 0$ else.
- The Lebesgue measure λ on $\mathbb{R}^n$, which generalizes the notion of length/area/volume on intervals to $\mathcal{B}(\mathbb{R}^n)$.

The only requirement we have for the Lebesgue measure is that $\lambda(I) = volume(I)$ for every interval I. Ideally, we would like to define the Lebesgue measure on $\mathcal{P}(\mathbb{R}^n)$; however, this extension is not possible while the countable additivity property holds.
Some important properties of measures are:

- **Monotonicity:** If $A \subseteq B$, then $\mu(A) \leq \mu(B)$.
- **Finite Additivity**: If A, B are **disjoint** measurable sets, then

$$\mu(A \cup B) = \mu(A) + \mu(B).$$

- **Subadditivity:** For any countable collection $\{A_n\}$,

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

- **Continuity from Below:** If $A_n \subseteq A_{n+1}$, then

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

- **Continuity from Above:** If $A_n \supseteq A_{n+1}$ and $\mu(A_1) < \infty$, then

$$\mu\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n).$$

3.2.2 Measurable Functions

Definition 3.22. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, meaning $\mathcal{A}$ and $\mathcal{B}$ are σ-algebras on X and Y, respectively. A function $f : X \to Y$ is said to be $\mathcal{A}/\mathcal{B}$-**measurable** if for every $B \in \mathcal{B}$, the preimage $f^{-1}(B) \in \mathcal{A}$.*

We will often simply say that a function is measurable if the underlying σ-algebra is clear from context (e.g., the Borel σ-algebra on $\mathbb{R}^n$).

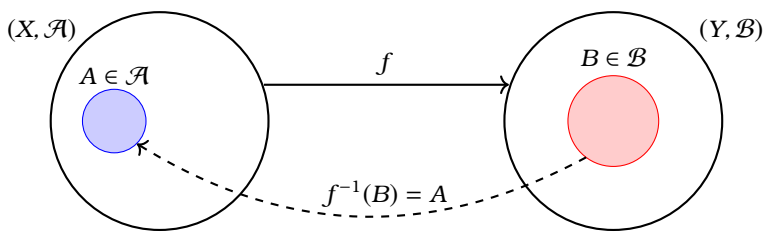


Figure 3.1. *Illustration of a measurable function: $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$*

The reason we require a function X to be measurable is that, when we define a measure μ on $\mathcal{F}$, we want to be able to assign values to sets of the form $\mu(X \in A)$ for some $A \subseteq \mathbb{R}$ (or more generally, for some subset of a metric space). For such expressions to make sense, the set $\{X \in A\} = X^{-1}(A)$ must belong to the domain of μ, that is, the σ-algebra $\mathcal{F}$. Hence, the measurability condition ensures that the preimages of Borel sets under X are measurable and thus compatible with the structure of the measure space.

Below are some examples of measurable functions.

Proposition 3.1. *Let $(X, \mathcal{A})$ be a measurable space. Then:*

1. *Any constant function $f(x) = c$ is measurable.*
2. *The identity function $f(x) = x$ is measurable.*
3. *Any continuous function is measurable, since the preimage of every open set is open and open sets generate the Borel σ-algebra.*
4. *If $f, g : X \to \mathbb{R}$ are measurable, then so are $f + g$, $f \cdot g$, $\max(f, g)$, $\min(f, g)$, and $|f|$.*
5. *If $(f_n)_{n\in\mathbb{N}}$ is a sequence of measurable functions, then the functions*

$$\sup f_n, \quad \inf f_n, \quad \limsup f_n, \quad \liminf f_n, \quad \lim f_n \text{ (if it exists)}$$

   are all measurable.
6. *If $f : X \to Y$ is $(\mathcal{A}, \mathcal{B})$-measurable and $g : Y \to Z$ is $(\mathcal{B}, C)$-measurable, then the composition $g \circ f : X \to Z$ is $(\mathcal{A}, C)$-measurable.*

Now, we will talk about σ-algebras generated by functions.
Given a function $f : \Omega \to E$ from a set Ω to a measurable space $(E, \mathcal{E})$, we define the *σ-algebra generated by f* as

$$\sigma(f) := \{f^{-1}(A) : A \in \mathcal{E}\} \subseteq \mathcal{F}.$$

This is the smallest σ-algebra on Ω with respect to which f is measurable. Intuitively, it captures all the information about Ω that is "visible" through f.
In the context of probability theory, if $X : \Omega \to \mathbb{R}$ is a random variable, then $\sigma(X)$ is the collection of events that can be determined by observing the value of X.

3.2.3 Integration and Radon-Nikodym's Theorem

Having defined measurable functions, we now turn to the notion of integration with respect to a measure. We will begin by defining the integral for non-negative simple functions, then extend it to general non-negative measurable functions, and finally to integrable functions. After establishing the properties of the Lebesgue integral, we will present one of the central results in measure theory: the Radon-Nikodym theorem, which, under certain conditions, allows us to express one measure as a density with respect to another.

We start by defining the integral for non-negative simple functions.

Definition 3.23. *Let $(X, \mathcal{F}, \mu)$ be a measure space. A function $f : X \to [0, \infty)$ is called a simple function if it can be written as*

$$f(x) = \sum_{i=1}^{n} a_i 1_{A_i}(x),$$

where $a_i \geq 0$ and $A_i \in \mathcal{F}$ are disjoint. The integral of f with respect to μ is defined as

$$\int_X f \, d\mu = \sum_{i=1}^{n} a_i \mu(A_i).$$

We extend the integral to non-negative measurable functions by approximation via simple functions.

Definition 3.24. *Let $f : X \to [0, \infty]$ be measurable. Define*

$$\int_X f \, d\mu = \sup \left\{ \int_X s \, d\mu : 0 \leq s \leq f, \ s \text{ simple} \right\}.$$

For a general real-valued measurable function f, we write $f = f^+ - f^-$ and define

$$\int_X f \, d\mu = \int_X f^+ \, d\mu - \int_X f^- \, d\mu,$$

provided at least one of the two integrals is finite.

Comparison with the Riemann Integral The Riemann and Lebesgue integrals both aim to assign a meaningful "area under the curve" for real-valued functions, but they do so in fundamentally different ways. The Riemann integral partitions the domain of the function (typically an interval) and sums up rectangles whose heights are determined by function values over each subinterval. Formally, it approximates the integral by

$$\int_a^b f(x) \, dx \approx \sum_{i=1}^{n} f(x_i^*)(x_i - x_{i-1}),$$

where $x_i^* \in [x_{i-1}, x_i]$ are sample points and the partition becomes finer.

In contrast, the Lebesgue integral partitions the range of the function and measures the size (under a measure μ) of the preimages of these range slices. For non-negative functions, this leads to an approximation of the form

$$\int f \, d\mu \approx \sum_{i=1}^{n} y_i \, \mu(f^{-1}([y_{i-1}, y_i))).$$

This shift in perspective offers several advantages. Lebesgue integration handles a broader class of functions, including those with too many discontinuities for Riemann integration, and it interacts better with limits, allowing the development of powerful convergence theorems.

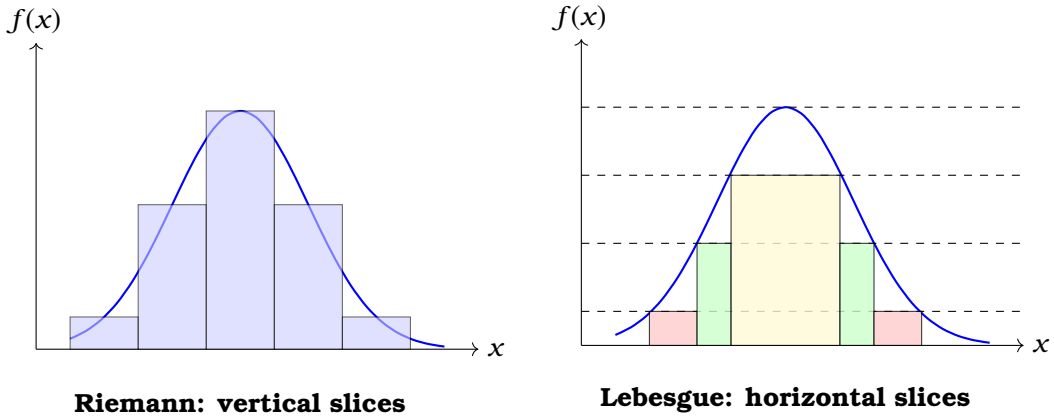


Figure 3.2. *Comparison of Riemann and Lebesgue integration. Riemann sums vertical slices under the graph, Lebesgue sums over horizontal level sets.*

Some important properties of Lebesgue integrals are:

- **Linearity.** For all $a, b \in \mathbb{R}$:

$$\int (af + bg) \, d\mu = a \int f \, d\mu + b \int g \, d\mu.$$

- **Monotonicity.** If $f \leq g$ almost everywhere, then:

$$\int f \, d\mu \leq \int g \, d\mu.$$

- **Absolute Integrability.** If f is integrable, then $|f|$ is integrable and:

$$\left| \int f \, d\mu \right| \leq \int |f| \, d\mu.$$

If we want to integrate over $A \subset X$, we write $\int_A f \, d\mu := \int f \cdot \mathbb{1}_A \, d\mu$.

A property $P(x)$ is said to hold *almost everywhere* if it fails only on a set of measure zero:

$$\mu(\{x \in X : P(x) \text{ fails}\}) = 0.$$

In measure theory, the behavior of a function is unaffected by changes on sets of measure zero. As a result, functions that are equal almost everywhere are considered equivalent.

Convergence Theorems A fundamental question in analysis is under what conditions we can interchange the limit and the Lebesgue integral. That is, given a sequence of measurable functions $\{f_n\}$ and a pointwise limit $f = \lim_{n\to\infty} f_n$, when does the following hold?

$$\lim \int f_n \, d\mu = \int \lim f_n \, d\mu$$

From the following example, we can see that the limit and the integral cannot always be interchanged.

Example 3.1. *Consider the sequence of functions $f_n(x)$ defined by:*

$$f_n(x) = \begin{cases} n, & 0 < x \leq \frac{1}{n}, \\ 0, & \textit{otherwise}. \end{cases}$$

This sequence converges pointwise to the zero function $f(x) = 0$. Now, let's examine both the limit of the integrals and the integral of the limit:
For each n, we compute:

$$\int_0^1 f_n(x) \, dx = n \cdot \frac{1}{n} = 1.$$

Thus,

$$\lim \int_0^1 f_n(x) \, dx = 1$$

.
Since $f_n(x) \to 0$ pointwise, the integral of the limit function is:

$$\int_0^1 0 \, dx = 0.$$

Thus, we have

$$\lim \int_0^1 f_n(x) \, dx = 1 \quad \textit{but} \quad \int_0^1 \lim f_n(x) \, dx = 0.$$

This shows that the limit and integral cannot be interchanged.

Unlike the Riemann integral, the Lebesgue integral offers powerful tools to answer this question.

Theorem 3.8 (Monotone Convergence Theorem (Beppo Levi)). *Let $\{f_n\}$ be a sequence of non-negative measurable functions such that $f_n(x) \uparrow f(x)$ for all x. Then,*

$$\int f_n \, d\mu \uparrow \int f \, d\mu.$$

Theorem 3.9 (Fatou's Lemma). *Let $\{f_n\}$ be a sequence of non-negative measurable functions. Then,*

$$\int \liminf f_n \, d\mu \leq \liminf \int f_n \, d\mu.$$

Theorem 3.10 (Dominated Convergence Theorem). *Let $\{f_n\}$ be a sequence of measurable functions such that $f_n \to f$ pointwise almost everywhere and there exists an integrable function g with $|f_n(x)| \leq g(x)$ for all n and a.e. x. Then,*

$$\int f_n \, d\mu \to \int f \, d\mu.$$

Theorem 3.11 (Bounded Convergence Theorem). *Let $\{f_n\}$ be a sequence of measurable functions such that $f_n \to f$ pointwise almost everywhere and $|f_n| \leq M$ for some constant $M < \infty$, and $\mu(X) < \infty$. Then,*

$$\int f_n \, d\mu \to \int f \, d\mu.$$

The L^p Spaces Let $(X, \mathcal{F}, \mu)$ be a measure space and f be a measurable function $X \to \overline{\mathbb{R}}$. For $1 \leq p < \infty$, we define the **p-norm** of f as follows:

$$\|f\|_p = \left(\int |f(x)|^p d\mu(x) \right)^{1/p}.$$

We define the L^p **space** as the normed vector space induced by the p-norm:

$$L^p(X, \mathcal{F}, \mu) = \left\{ f : X \to \overline{\mathbb{R}} \mid \|f\|_p < \infty \right\},$$

To properly extend the notions of supremum and infimum to the measure theoretic setting, we use the essential supremum and essential infimum, which disregard sets of measure zero and are well-defined for equivalence classes of measurable functions.

- The **essential supremum** of f is defined as

$$\operatorname{ess\,sup} f(x) := \inf \{ M \in \mathbb{R} : \mu(\{x \in X : f(x) > M\}) = 0 \}.$$

- The **essential infimum** of f is defined as

$$\operatorname{ess\,inf} f(x) := \sup\{m \in \mathbb{R} : \mu(\{x \in X : f(x) < m\}) = 0\}.$$

We can now define the **supremum norm** as:

$$\|f\|_\infty = \operatorname{ess\,sup} |f(x)|.$$

The L^∞ **space** is defined as:

$$L^\infty(X, \mathcal{F}, \mu) = \left\{ f : X \to \overline{\mathbb{R}} \mid \|f\|_\infty < \infty \right\},$$

In the case of *finite measure spaces*, the L^p spaces can be ordered as follows:

$$L^\infty \subseteq L^q \subseteq L^p \quad \text{for} \quad 1 \leq p \leq q \leq \infty.$$

This ordering reflects the fact that if $f \in L^q$, then $f \in L^p$ for $p \leq q$. Intuitively, higher powers impose stricter conditions on the function f, as the function needs to be more "well-behaved" in terms of its integrability.
Another important property of L^p spaces is that they are complete.

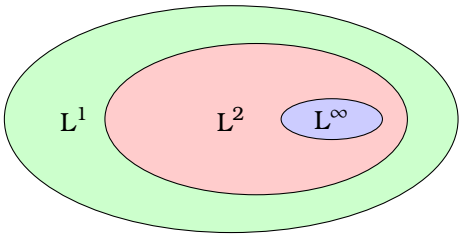


Figure 3.3. *Illustration of L^p spaces on a finite measure space*

The Radon-Nikodym Theorem In measure theory, we often encounter situations where two measures are related in such a way that one can be thought of as having a *density* with respect to the other. The Radon-Nikodym theorem formalizes this idea: if one measure is absolutely continuous with respect to another, then it can be expressed as an integral involving a suitable density function.

Definition 3.25 (Absolute Continuity)**.** *Let μ and ν be two measures on the same measurable space $(\Omega, \mathcal{F})$. We say that ν is **absolutely continuous** with respect to μ: $\nu \ll \mu$, if for every measurable set $A \in \mathcal{F}$,*

$$\mu(A) = 0 \quad \Rightarrow \quad \nu(A) = 0.$$

This means that ν does not assign positive measure to sets that are negligible under μ.

Theorem 3.12 (Radon-Nikodym)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space, and let μ and ν be σ-finite measures on $\mathcal{F}$ (countable sums of finite measures). If $\nu \ll \mu$, then there*

exists a unique (up to μ-almost everywhere equality) non-negative measurable function $f : \Omega \to [0, \infty)$ such that

$$\nu(A) = \int_A f \, d\mu \quad \textit{for all } A \in \mathcal{F}.$$

This function f is called the ***Radon-Nikodym derivative*** *of ν with respect to μ, and we write:*

$$f = \frac{d\nu}{d\mu}.$$

In the following section, we will see how the Radon-Nikodym derivative formalizes the idea of density functions from probability theory.

3.2.4 Probability Distributions

Let $(\Omega, \mathcal{F})$ be a measurable space. A probability measure $\mathbb{P}$ is a measure where $\mathbb{P}(\Omega) = 1$, and it is defined on a measurable space $(\Omega, \mathcal{F})$. A measurable function defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **random variable**.
For a random variable X defined on a probability space, we define its expected value (or mean) as:

$$\mathbb{E}(X) = \int X \, d\mathbb{P}.$$

The **variance** of a random variable X is defined as:

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2,$$

which gives a measure of how spread out the values of X are around its mean.

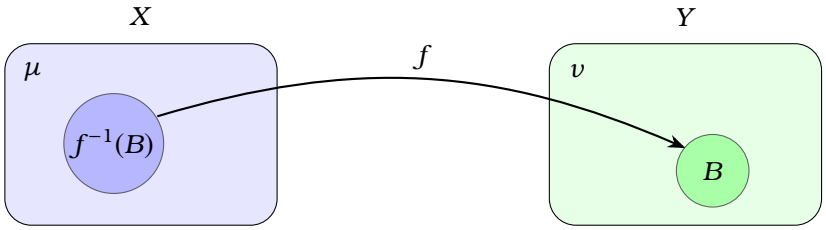
Distribution of a Random Variable We now explain the connection between probability density functions (pdfs) and the Radon-Nikodym theorem. To do this, we begin by introducing the notion of pushforward measures, which allow us to rigorously define the distribution of a random variable.
Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and let $f \colon X \to Y$ be a measurable function. If μ is a measure on $(X, \mathcal{A})$, the **pushforward measure** $f_\# \mu$ on $(Y, \mathcal{B})$ is defined by

$$f_\#\mu(B) := \mu(f^{-1}(B)), \quad \text{for all } B \in \mathcal{B}.$$

This defines a measure on Y that describes how μ is transported via f.
Let $X \colon \Omega \to \mathbb{R}$ be a real-valued random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then the **distribution** P_X of X is the pushforward measure $X_\#\mathbb{P}$, defined by

$$P_X(A) = X_\#\mathbb{P}(A) = \mathbb{P}(X \in A), \quad \text{for Borel sets } A \subseteq \mathbb{R}.$$

This measure $P_X := X_\#\mathbb{P}$ is a probability measure on $\mathbb{R}$.
If $P_X \ll \lambda$, where λ is the Lebesgue measure, then P_X admits a density $f \in L^1(\mathbb{R})$ such



$$f_{\#}\mu(B) = \mu(f^{-1}(B)) \quad \text{for } B \subseteq Y$$

Figure 3.4. *Illustration of a pushforward measure $\nu = f_{\#}\mu$.*

that

$$\mathbb{P}(X \in A) = \int_A f \, d\lambda.$$

The function f is called the **probability density function (pdf)** of X.

This construction generalizes naturally to random variables taking values in more general spaces.

Let $X\colon \Omega \to \mathbb{R}$ be a real-valued random variable with distribution $\mu = X_{\#}\mathbb{P}$, and suppose $\mu \ll \lambda$, where λ is the Lebesgue measure. Let $f = \frac{d\mu}{d\lambda}$ denote the corresponding probability density function (pdf). Then, for any measurable function $h\colon \mathbb{R} \to \mathbb{R}$ such that $h(X) \in L^1(\mathbb{P})$, the expected value of $h(X)$ is given by

$$\mathbb{E}[h(X)] = \int_\Omega h(X(\omega)) \, d\mathbb{P}(\omega) = \int_{\mathbb{R}} h(x) f(x) \, dx.$$

This identity shows how integration with respect to the original probability measure $\mathbb{P}$ can be expressed as integration against the pdf of the pushforward measure μ.

The **cumulative distribution function** (CDF) of a random variable X is defined as:

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t) \, dt,$$

where $f_X(t)$ is the probability density function (PDF), if it exists.

The CDF $F_X(x)$ gives the probability that the random variable X takes a value less than or equal to x. It is a non-decreasing, right-continuous function that converges to 0 as $x \to -\infty$ and to 1 as $x \to +\infty$.

We call two random variables X and Y **identically distributed** and write $X \overset{d}{=} Y$ if they have the same distribution. Equivalently, their CDFs should be equal:

$$F_X(x) = F_Y(x), \quad \forall x \in \mathbb{R}$$

Discrete Distributions A discrete random variable has a countable set of values, each with a certain probability. Let X be a discrete random variable with possible values $x_1, x_2, \ldots$, and associated probabilities $p_i = \mathbb{P}(X = x_i)$, where $\sum_i p_i = 1$. The probability mass function (PMF) of X is given by:

$$P_X(x) = \mathbb{P}(X = x).$$

The expected value is computed as:

$$\mathbb{E}(X) = \sum x_i P_X(x_i)$$

While this discrete formulation may initially appear different from the continuous setting, it also arises naturally from the Radon-Nikodym theorem, as we demonstrate below.
Let $\Omega = \{x_1, x_2, x_3, \ldots\}$ be a countable sample space equipped with the sigma-algebra $\mathcal{F} = 2^\Omega$. Suppose P is a discrete probability measure on $(\Omega, \mathcal{F})$. We consider the *counting measure* μ on Ω (remember that $\mu(A) = |A|$).
Since $\mu(A) = 0 \Rightarrow P_X(A) = 0$, the measure P_X is absolutely continuous with respect to μ. [Note that P_X is not absolutely continuous with respect to the Lebesgue measure λ, since it could give positive probability on sets like $\{0\}$, that have $\lambda(\{0\}) = 0$.] Therefore, by the Radon-Nikodym theorem, there exists a measurable function $p : \Omega \to \mathbb{R}_+$ such that:

$$P_X(A) = \int_A p \, d\mu = \sum_{x \in A} p(x), \quad \forall A \subseteq \Omega.$$

The function $p = \frac{dP_X}{d\mu}$ is the *Radon-Nikodym derivative* of P_X with respect to μ, and corresponds exactly to the *probability mass function* (PMF) of P. That is, for each $x \in \Omega$:

$$p(x) = P(\{x\}).$$

3.2.5 Convergence of Measures

In measure theory and probability, there are several different types of convergence for sequences of random variables and measures. These types vary in strength, and we summarize the relationships between them below.

Almost Sure Convergence A sequence of random variables (X_n) converges almost surely (or with probability 1) to a random variable X if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n(\omega) = X(\omega)\right) = 1.$$

This is the strongest form of convergence, as it implies pointwise convergence for almost every outcome.

Convergence in Probability We say that X_n converges to X in probability if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

Almost sure convergence implies convergence in probability.

Convergence in L^p A sequence (X_n) converges to X in L^p for $p \geq 1$ if

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

This implies convergence in probability (via Markov's inequality).

Weak Convergence of Measures A sequence of probability measures (μ_n) on a metric space $(\mathcal{X}, d)$ converges weakly to a measure μ if for all bounded continuous functions $f : \mathcal{X} \to \mathbb{R}$,

$$\lim_{n\to\infty} \int f \, d\mu_n = \int f \, d\mu.$$

This form of convergence is central in optimal transport, particularly in the convergence of empirical measures and in the definition of Wasserstein distances.

Convergence in Distribution Also known as convergence in law, when X_n and X are real-valued random variables $X_n \xrightarrow{d} X$ means that the cdf of X_n converges to the cdf of X:

$$\lim_{n\to\infty} F_n(x) = F(x) \qquad \forall x \in \mathbb{R}$$

In the general case, convergence in distribution is equivalent to weak convergence of the pushforward measures $\mu_n = X_{n\#}\mathbb{P}$ toward $\mu = X_\#\mathbb{P}$. Convergence in probability (which is implied by convergence in L^p or almost sure convergence) implies convergence in distribution.

3.2.6 Product Measures and Independence

Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be two measure spaces. The **product σ-algebra** $\mathcal{A}\otimes\mathcal{B}$ on $X\times Y$ is the smallest σ-algebra containing all measurable rectangles $A \times B$ with $A \in \mathcal{A}$ and $B \in \mathcal{B}$. A measure π on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ is called a **product measure** if $\pi(A \times B) = \mu(A)\nu(B)$ for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$. That is, the measure of each rectangle $A \times B$ is the product of the measures of each of its sides, as illustrated graphically below. The existence and uniqueness of the product measure $\mu \otimes \nu$ is guaranteed when both μ and ν are σ-finite.

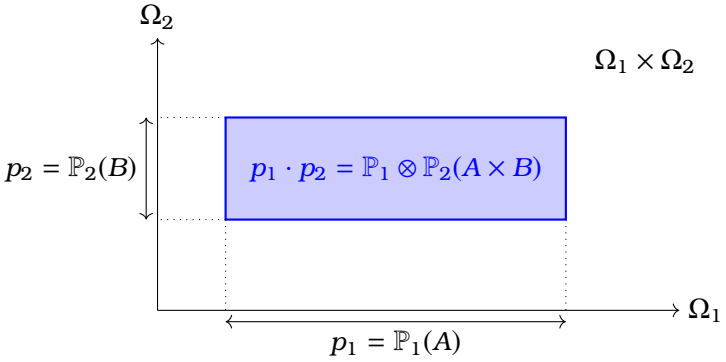


Figure 3.5. *Illustration of product measure*

Theorem 3.13 (Tonelli-Fubini's Theorem)**.** *For a nonnegative measurable function $f : X \times Y \to \mathbb{R}$ (or integrable if f is signed), the integral with respect to the product measure satisfies:*

$$\int_{X\times Y} f(x,y)\, d(\mu\otimes\nu)(x,y) = \int_X \left(\int_Y f(x,y)\, d\nu(y)\right) d\mu(x) = \int_Y \left(\int_X f(x,y)\, d\mu(x)\right) d\nu(y).$$

This theorem is fundamental in defining expectations over multivariate distributions.

Corollary 3.1. *If $X \geq 0$ is a random variable, then:*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\, dt,$$

a classic identity obtained by applying Tonelli's theorem to the function $\mathbf{1}_{\{X>t\}}$.

Independence

Definition 3.26 (Independent Events)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Two events $A, B \in \mathcal{F}$ are said to be independent if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

This notion extends to random variables, as illustrated below.

Definition 3.27 (Independent Random Variables)**.** *A collection of random variables $X_1, \ldots, X_d$ defined on a common probability space is said to be independent if for every choice of Borel sets $A_1, \ldots, A_d \subseteq \mathbb{R}$,*

$$\mathbb{P}(X_1 \in A_1, \ldots, X_d \in A_d) = \prod_{i=1}^{d} \mathbb{P}(X_i \in A_i).$$

That is, the probability of all the random variables simultaneously taking values in their respective sets factors into the product of the individual probabilities.
Equivalently, the push-forward measure of the joint distribution of $(X_1, \ldots, X_d)$ is the product of the marginal distributions:

$$\mathbb{P}_{(X_1,\ldots,X_d)} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_d}.$$

That is, X is independent if its joint law is the product of the marginal laws.
Connecting to classical probability, random variables $X_1, \ldots, X_n$ are independent if and only if:

$$\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x_i).$$

If they admit densities, then independence implies:

$$f_X(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

3.3 Functional Analysis

3.3.1 Normed and Banach Spaces

We begin our functional analysis background with the concept of Banach spaces. Recall that a normed space is a vector space equipped with a norm, that induces a metric $d(x, y) = \|x - y\|$, and thereby a topology on X.

Definition 3.28 (Banach Spaces)**.** *A normed space $(X, \|\cdot\|)$ is called a **Banach space** if it is complete: every Cauchy sequence (x_n) in X has a limit in X.*

Completeness ensures the space is robust under limits and infinite processes.
An important consequence of completeness concerns the convergence of series: if $\sum_{n=1}^{\infty} x_n$ is a series in X, it converges if and only if the sequence of partial sums $s_n = \sum_{k=1}^{n} x_k$ is a Cauchy sequence. Thus, the ability to define infinite linear combinations is tightly linked to completeness.
Conversely, if every absolutely convergent series in a normed space converges (i.e., $\sum \|x_n\| < \infty \Rightarrow \sum x_n$ converges), then the space must be complete. This makes Banach spaces the natural setting for series expansions.

Finite-Dimensional Spaces

Definition 3.29 (Equivalent norms)**.** *Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a space X are said to be **equivalent** if there exist constants $c, C > 0$ such that for all $x \in X$,*

$$c\|x\|_1 \leq \|x\|_2 \leq C\|x\|_1.$$

Equivalent norms induce the same topology, and thus the same notions of convergence, continuity, and compactness.

The theorem below highlights the most important properties of finite-dimensional normed spaces.

Theorem 3.14 (Characterization of Finite-Dimensional Normed Spaces)**.** *Let $(X, \|\cdot\|)$ be a normed vector space. The following statements are equivalent:*

1. *X is finite-dimensional.*
2. *The closed unit ball $B = \{x \in X : \|x\| \leq 1\}$ is compact in the norm topology.*
3. *All norms on X are equivalent, thus they induce the same topology.*
4. *X is Banach (complete) and has the Heine-Borel property (every closed and bounded subset of X is compact).*

This theorem tells us that every finite dimensional space behaves exactly as $\mathbb{R}^n$ with some norm, so we can use our geometric intuition. It highlights a key difference between finite and infinite-dimensional analysis: compactness, completeness, and topology are tightly coupled in finite dimensions, but diverge in infinite-dimensional settings.

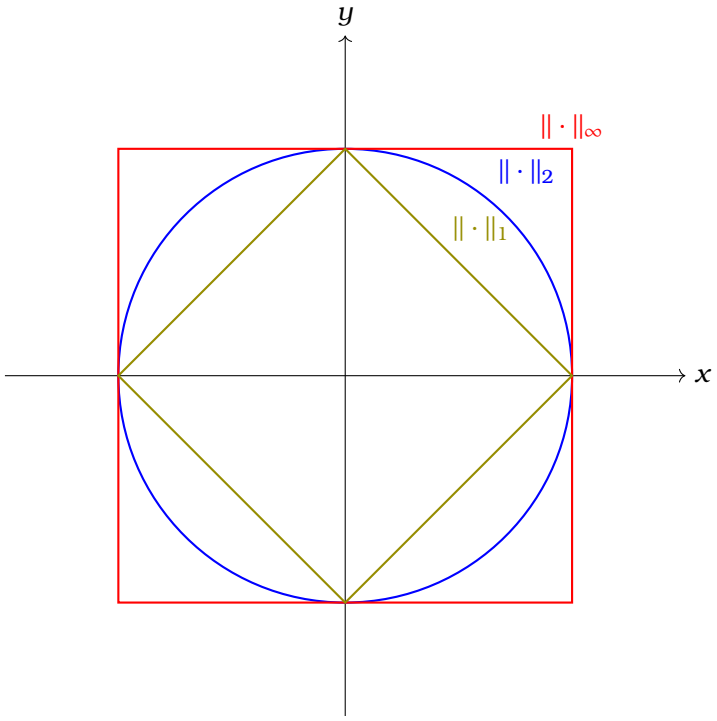


Figure 3.6. *Unit balls in $\mathbb{R}^2$ under different norms: Euclidean norm (ℓ^2), maximum norm (ℓ^∞), and Manhattan norm (ℓ^1). All define different shapes but induce the same topology, illustrating norm equivalence in finite dimensions.*

The figure above illustrates how different norms on $\mathbb{R}^2$ induce the same topology. The example below shows how in an infinite-dimensional Banach space like ℓ^2, the Heine-Borel property does not hold.

Example 3.2. *Consider the sequence (e_n) in the infinite-dimensional Hilbert space ℓ^2, where*

$$e_n = (0, 0, \ldots, 0, 1, 0, \ldots),$$

with the 1 in the n-th position and zeros elsewhere. This sequence lies in the unit ball but has no convergent subsequence since for any $m \neq n$,

$$\|e_n - e_m\|_2 = \sqrt{2}.$$

This example demonstrates the failure of compactness of the unit ball in infinite-dimensional spaces, contrasting the finite-dimensional case.

Schauder Bases and Separability In infinite-dimensional spaces, we generalize the notion of a basis to that of a **Schauder basis**. A sequence (x_n) in a Banach space X is a Schauder basis if every element $x \in X$ has a unique expansion

$$x = \sum_{n=1}^{\infty} a_n x_n,$$

where the series converges in norm. The existence of such a basis implies that X is **separable**, meaning it has a countable dense subset.

3.3.2 Dual Spaces

We begin by introducing bounded linear operators and functionals, which are essential for defining dual spaces and stating the Hahn–Banach theorem.

Definition 3.30 (Linear Operators)**.** *Let X and Y be vector spaces. A map $T : X \to Y$ is called a linear operator if for all $x_1, x_2 \in X$ and $\lambda \in \mathbb{R}$ (or $\mathbb{C}$), we have*

$$T(x_1 + x_2) = T(x_1) + T(x_2), \quad T(\lambda x) = \lambda T(x).$$

Definition 3.31 (Bounded Linear Operators)**.** *If X and Y are normed spaces, a linear operator $T : X \to Y$ is bounded if there exists a constant $C \geq 0$ such that*

$$\|T(x)\|_Y \leq C\|x\|_X \quad \text{for all } x \in X.$$

The smallest such C is called the operator norm of T, given by

$$\|T\| = \sup_{\|x\| \leq 1} \|T(x)\| = \sup_{\|x\| = 1} \|T(x)\|$$

Below are some important properties of linear operators:

- In normed spaces, boundedness of an operator is equivalent to continuity.
- Let $\mathcal{B}(X, Y)$ denote the space of all bounded linear operators from X to Y, equipped with the operator norm. If Y is a Banach space, then so is $\mathcal{B}(X, Y)$.
- If X is finite-dimensional, then all linear operators $T : X \to Y$ are automatically bounded, regardless of the norm.

Definition 3.32 (Linear Functionals)**.** *A linear functional is a linear map $f : X \to \mathbb{R}$ (or $\mathbb{C}$). It is bounded if it is bounded as an operator. That is,*

$$\|f\| = \sup_{\|x\| \leq 1} |f(x)| < \infty.$$

An example is given by integration: for a fixed $g \in L^2$, the map $f(\varphi) = \int \varphi(x)g(x)\,dx$ defines a bounded linear functional on L^2. Its norm is given by

$$\|f\| = \sup_{\|\varphi\|_2 = 1} \left| \int \varphi(x)g(x)\,dx \right| \leq \|g\|_2,$$

with equality achieved when $\varphi = g/\|g\|_2$, hence

$$\|f\| = \|g\|_2.$$

The Dual Space The *dual space* of a normed vector space X, denoted X^*, is the space of all bounded linear functionals on X:

$$X^* = B(X, \mathbb{R})$$

Equipped with the operator norm, X^* is a Banach space.

Some examples of dual spaces follow. We use the symbol $\cong$ to denote that two spaces are isometrically isomorphic; that is, they are isomorphic as vector spaces and the isomorphism preserves the norm (distance).

- For finite-dimensional $X \cong \mathbb{R}^n$, we have $X^* \cong \mathbb{R}^n$ again.
- If $1 < p < \infty$, then $(L^p)^* = L^q$, with $1/p + 1/q = 1$.
- If $C[a, b]$ is the space of continuous functions on $[a, b]$, its dual is isometrically isomorphic to the space of signed Borel measures.

Riesz Representation Theorem This last example is very important in the context of optimal transport. It is a special case of Riesz's Representation Theorem, which we will state now. First, we must make some definitions.

Definition 3.33 (Locally Compact Space)**.** *A topological space X is locally compact if every point $x \in X$ has a neighborhood whose closure is compact.*

Definition 3.34 (Radon Measure)**.** *A Borel measure μ on a locally compact space X is a Radon measure if it is*

- *locally finite: $\mu(K) < \infty$ for every compact $K \subseteq X$,*
- *inner regular: for every Borel set B, $\mu(B) = \sup\{\mu(K) : K \subseteq B, K \text{ compact}\}$, thus it can be approximated by measures on compact sets.*

Theorem 3.15 (Riesz–Markov–Kakutani Representation Theorem)**.** *Let X be a locally compact space and $C_c(X)$ the space of continuous functions on X with compact support. Then every bounded linear functional $L : C_c(X) \to \mathbb{R}$ is represented uniquely by a signed Radon measure μ on X such that*

$$L(f) = \int_X f \, d\mu \quad \text{for all } f \in C_c(X).$$

Thus, the dual space of $C_c(X)$ can be identified with the space of signed Radon measures on X:

$$(C_c(X))^* = M_r(X)$$

The norm on $C_c(X)$ is the supremum norm $\|f\|_\infty = \sup |f(x)|$, and the dual norm on $M_r(X)$ is the total variation norm:

$$\|\mu\|_{TV} = \sup_{\|f\|_\infty \leq 1} \left| \int_X f \, d\mu \right|.$$

Using the Hahn decomposition $X = A^+ \cup A^-$, where μ is non-negative on A^+ and non-positive on A^-, we have

$$\|\mu\|_{TV} = \mu(A^+) - \mu(A^-),$$

which sums the positive and negative masses of μ. For probability measures, $\|\mu\|_{TV} = 1$.

In the context of this thesis, we will mainly focus on locally compact Polish spaces (complete separable metric spaces) X. In these spaces, all finite Borel measures on them are Radon. So the above theorem tells us that the dual of $C_c(X)$ is the space of finite Borel measures $M(X)$.

With the extra assumption that X is compact, $C_c(X) = C_b(X)$. In this case, the Riesz Representation Theorem identifies the dual of $C_b(X)$ with the space of finite signed Radon measures on X.

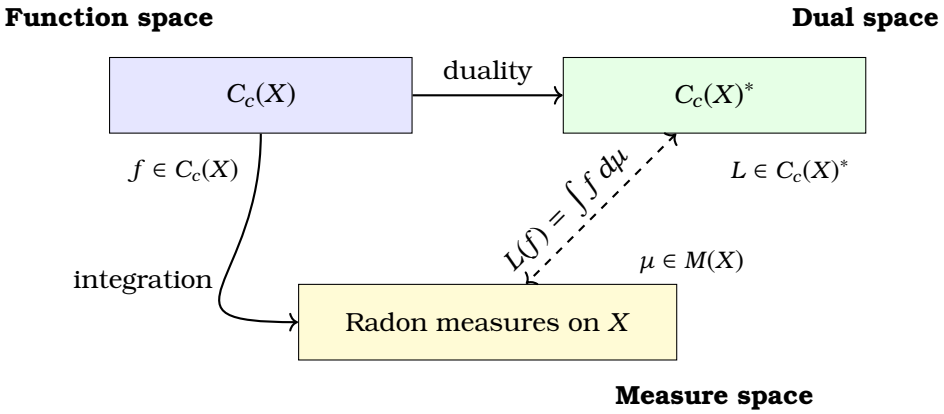


Figure 3.7. *Illustration of the Riesz-Markov-Kakutani theorem: positive linear functionals on $C_c(X)$ correspond uniquely to Radon measures on X.*

3.3.3 The Hahn–Banach Theorem

A foundational result in functional analysis, the Hahn–Banach theorem allows us to extend bounded linear functionals from subspaces to the whole space without increasing the norm.

Theorem 3.16 (Hahn–Banach)**.** *Let X be a normed vector space, $Y \subset X$ a linear subspace, and $f_0 : Y \to \mathbb{R}$ a bounded linear functional. Then there exists a bounded linear functional $f : X \to \mathbb{R}$ such that*

$$f|_Y = f_0 \quad \text{and} \quad \|f\| = \|f_0\|.$$

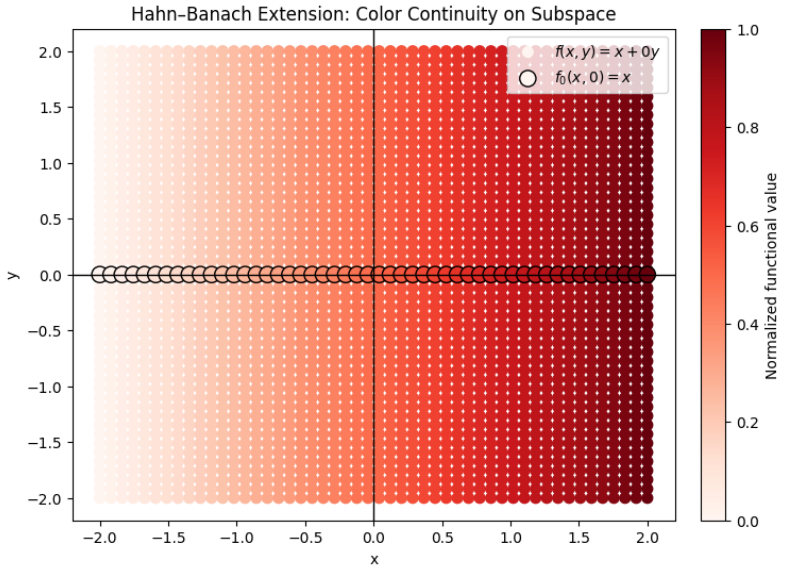


Figure 3.8. *Visualization of a Hahn–Banach extension. The original linear functional $f_0(x, 0) = x$ is defined on the x-axis (the subspace U), and extended to the whole plane via $f(x, y) = x + \lambda y$. We chose $\lambda = 0$ so that the extension preserves the operator norm of 1. For other values of λ, the norm of the extension becomes $\sqrt{1 + \lambda^2} > 1$, thus violating the norm-preserving requirement of the Hahn–Banach theorem.*

Geometric Interpretation. An alternative geometric version of the theorem states that for any closed convex set $C \subset X$ and point $x \notin C$, there exists a continuous linear functional that separates x from C, i.e.,

$$f(x) < \inf_{y \in C} f(y).$$

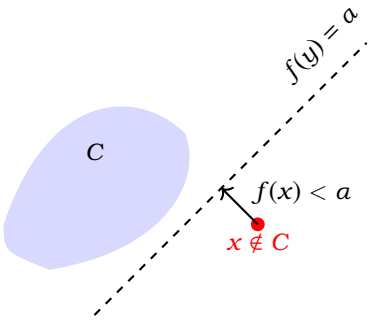


Figure 3.9. *Hahn-Banach separation: the point $x \notin C$ is separated from the convex set C by a hyperplane $f(y) = a$.*

Some important consequences of the Hahn-Banach theorem are:

- Dual spaces separate points: For $x \neq y$ in X, there exists $f \in X^*$ such that $f(x) \neq f(y)$.
- Any normed space embeds isometrically into its bidual X^{**} via the canonical map $x \mapsto \hat{x}$, where $\hat{x}(f) = f(x)$.
- The norm of an element $x \in X$ can be recovered from the dual space:

$$\|x\| = \sup_{\substack{f \in X^* \\ \|f\|=1}} |f(x)|.$$

Fundamental Theorems for Banach Spaces

The following theorems are cornerstones in the analysis of bounded linear operators between Banach spaces.

Uniform Boundedness Principle (Banach–Steinhaus). Let X be a Banach space and $\{T_i\} \subset \mathcal{B}(X, Y)$ a family of bounded linear operators such that for every $x \in X$, the set $\{\|T_i x\|_Y\}$ is bounded. Then, the operator norms $\|T_i\|$ are uniformly bounded.

$$\{\forall x \; \exists M_x : \forall i \; \|T_i x\|_Y \leq M_x\} \Rightarrow \{\exists M \; \forall i : \|T_i\| \leq M\}$$

Open Mapping Theorem. Let $T : X \to Y$ be a bounded linear operator between Banach spaces that is surjective. Then T maps open subsets of X to open subsets of Y.

Closed Graph Theorem. Let $T : X \to Y$ be a linear operator between Banach spaces. If the graph of T, defined as $\{(x, Tx) \in X \times Y\}$, is closed in $X \times Y$, then T is bounded.

3.3.4 Hilbert Spaces

Although not central to our main results, we include a short section on Hilbert spaces for the sake of completeness.

Definition 3.35 (Hilbert Space)**.** *An inner product space is a vector space H equipped with an inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\| = \sqrt{\langle x, x \rangle}$. If H with this norm is complete, we call H a* ***Hilbert space****. Thus, H is a Banach space whose norm comes from an inner product.*

Some examples of Hilbert spaces are:

- $\mathbb{R}^n$ with the standard dot product.
- The sequence space $\ell^2 = \{x = (x_n) : \sum |x_n|^2 < \infty\}$ with $\langle x, y \rangle = \sum x_n y_n$.
- The space $L^2(\Omega)$ of square-integrable functions with $\langle f, g \rangle = \int_\Omega f(x)g(x)dx$.

Hilbert spaces behave much like Euclidean spaces, with an inner product structure that allows notions of angles, orthogonality, and projections—making geometric intuition and techniques applicable in infinite-dimensional settings.

One of the most elegant features of Hilbert spaces is the intimate connection between vectors and functionals.

Theorem 3.17 (Riesz Representation Theorem (Hilbert Space Version))**.** *For every bounded linear functional $\varphi \in H^*$ on a Hilbert space H, there exists a unique vector $y \in H$ such that*

$$\varphi(x) = \langle x, y \rangle \quad \textit{for all } x \in H.$$

That is, every continuous linear functional can be written as an inner product with a fixed vector.

This result is powerful because it translates abstract functionals into concrete vectors: we can "represent" any functional just by finding the right vector to plug into the inner product. Moreover, this correspondence is *isometric* and *bijective*, meaning that the Hilbert space H is *isometrically isomorphic* to its dual H^*. Thus, $H^* \cong H$.

Geometrically, the theorem says that linear functionals "point" in the direction of the vector y. The action of the functional is fully determined by projecting any input x onto this direction via the inner product.

3.3.5 Weak Topologies

In normed vector spaces, there are important topologies weaker than the standard topology induced by the norm. Note that we call a topology $\mathcal{T}_1$ weaker than a topology $\mathcal{T}_2$ iff convergence of a sequence in $\mathcal{T}_2$ implies its convergence in $\mathcal{T}_1$.

Definition 3.36 (Weak Convergence)**.** *Let X be a Banach space. A sequence $(x_n) \subset X$ is said to converge weakly to $x \in X$ (written $x_n \rightharpoonup x$) if*

$$f(x_n) \to f(x) \quad \text{for all } f \in X^*.$$

That is, all bounded linear functionals behave continuously under the sequence. The weak topology is the coarsest topology that makes all functionals in X^* continuous.

Definition 3.37 (Weak-* Convergence)**.** *Let X^* be the dual of X. A sequence $(f_n) \subset X^*$ converges weak-* to $f \in X^*$ (written $f_n \overset{*}{\rightharpoonup} f$) if*

$$f_n(x) \to f(x) \quad \text{for all } x \in X.$$

The weak-* topology on X^* is weaker than the weak topology because it requires convergence only when evaluated at points $x \in X$, that is, $f_n(x) \to f(x)$ for all $x \in X$. In contrast, weak convergence in X^* requires $g(f_n) \to g(f)$ for all $g \in X^{**}$, an (in principle) larger set of test functionals that includes the evaluations $\hat{x} \in X^{**}$ defined by $\hat{x}(f) = f(x)$.

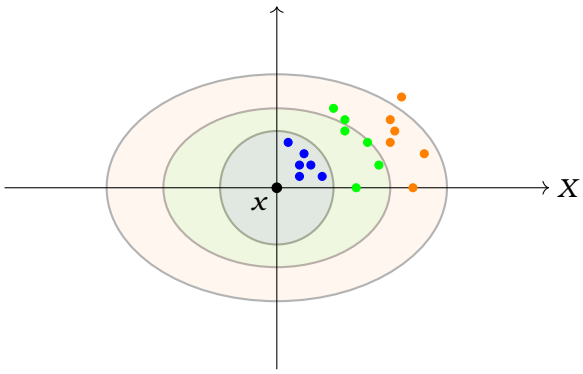


Figure 3.10. ***Neighborhood sizes and sequence convergence under different topologies:*** *The norm topology has the smallest neighborhoods (blue), the weak topology has larger neighborhoods (green), and the weak-* topology has the largest neighborhoods (orange). As the topology weakens, neighborhoods become bigger, so sequences can vary more and still converge. Hence, some sequences fail to converge under the norm but do converge weakly or weak-*. More specifically:*
- Blue points converge to x under ***norm****.*
- Blue and Green points converge to x under ***weak*** *convergence.*
- Blue, Green, and Orange points converge to x under ***weak-**** *convergence.*

Compactness plays a central role in analysis, since it ensures that all sequences have convergent subsequences. In infinite-dimensional normed spaces, compactness in the norm topology is rare, but weak compactness provides a useful substitute.

Theorem 3.18 (Banach–Alaoglu)**.** *Let X be a normed space. The closed unit ball in X^*, the dual space of X, is compact in the weak-* topology.*

This behavior is highlighted by the example below.

Example 3.3. *In ℓ^∞, consider the sequence $e_n = (0, \dots, 0, 1, 0, \dots)$ with 1 in the n-th position. It lies in the unit ball of ℓ^∞ and does not admit a norm-convergent subsequence, since $\|e_n - e_m\| = 1$ for $n \neq m$, thus the subsequences are not even Cauchy.*
We now view this sequence as functionals in the dual space $\ell^\infty = (\ell^1)^*$. Each f_n is the sequence $(0, 0, \dots, 0, 1, 0, \dots) \in \ell^\infty$, and is viewed as a functional on ℓ^1 via the standard dual pairing: $f_n(x) = x_n$ for $x \in \ell^1$. We can check that

$$f_n(x) = x_n \to 0 \quad \text{for all } x \in \ell^1.$$

This shows that $f_n \xrightarrow{*} 0$ in the weak- topology.*

Reflexive Spaces Recall that any normed space X embeds isometrically into its bidual X^{**} via the canonical map $J : X \to X^{**}$ defined by

$$J(x)(f) = f(x) \quad \text{for all } f \in X^*.$$

Definition 3.38 (Reflexive Space)**.** *A Banach space X is said to be reflexive if the canonical embedding J is surjective, i.e., every element of X^{**} arises as evaluation at some $x \in X$.*

A reflexive space X is obviously isometrically isomorphic to its bidual X^{**}. The main significance of reflexivity lies in its compactness properties:

Theorem 3.19 (Kakutani)**.** *If X is reflexive, then the closed unit ball in X is compact in the weak topology.*

This result complements the Banach–Alaoglu theorem, which guarantees weak-* compactness of the closed unit ball in the dual space X^*. Reflexivity ensures that compactness extends from the dual space back to the original space X, providing stronger convergence properties.
Below are some examples of reflexive and non-reflexive spaces.

- Every Hilbert space is reflexive.
- The Lebesgue spaces $L^p(\Omega)$ are reflexive for $1 < p < \infty$.
- If X is reflexive, then its dual X^* is also reflexive.
- The spaces $L^1(\Omega)$ and $L^\infty(\Omega)$ are *not* reflexive.
- The sequence space ℓ^1 is not reflexive, while ℓ^2 is.

Weak Convergence of Measures Recall that we defined weak convergence of measures (μ_n) on a topological space X by

$$\mu_n \rightharpoonup \mu \quad \text{if} \quad \int f\, d\mu_n \to \int f\, d\mu \quad \text{for all } f \in C_b(X).$$

By the Riesz Representation Theorem, we can interpret each function in C_b as a continuous linear functional on the measure space $M(X)$. Hence, the weak convergence of measures defined above—that is, $\mu_n \to \mu$ if and only if $\int f\, d\mu_n \to \int f\, d\mu$ for all $f \in C_b(X)$—can be interpreted as convergence with respect to all continuous linear functionals $f \in C_b(X)$, viewed as functionals acting on measures via

$$f(\mu_n) := \int f\, d\mu_n,$$

so that

$$f(\mu_n) \to f(\mu) \quad \text{for all } f \in C_b(X).$$

Note that according to the definitions above, the convergence we describe is the weak-* convergence of measures, as it corresponds to the weak-* topology on $\mathcal{M}(X)$, the dual of $C_b(X)$. However, for simplicity and in line with common terminology in the literature, we will refer to it simply as weak convergence of measures throughout this thesis. We will not consider the (true) weak topology on $\mathcal{M}(X)$, as the dual of $\mathcal{M}(X)$ is not tractable in our setting and plays no role in our analysis.

Topology on Probability Measures. If X is a Polish space (a complete separable metric space), the space of probability measures $\mathcal{P}(X)$, equipped with the weak (or narrow) topology, is metrizable (it can be described using a metric).
If we take a sequence of probability measures, under a condition called *tightness* — which means that for every small $\varepsilon > 0$, there is a compact set $K \subset X$ such that all measures in the sequence assign at least $1-\varepsilon$ probability to K — that sequence has a weakly convergent subsequence. This result is known as **Prokhorov's theorem**.

Example 3.4. *Let $x_n \to x$ in a Polish space X. Then the sequence of Dirac measures δ_{x_n} converges weakly to δ_x in $\mathcal{P}(X)$, since*

$$\int f\, d\delta_{x_n} = f(x_n) \to f(x) = \int f\, d\delta_x \quad \textit{for all } f \in C_b(X).$$

3.3.6 Convexity

We begin by introducing the basic concepts of convex sets and convex functions, which form the foundation of convex analysis. Because convex functions can be nondifferentiable or take infinite values, we also define lower semicontinuity, a key property for ensuring well-behaved functions.

Definition 3.39 (Convex Set)**.** *Let X be a vector space. A set $C \subset X$ is convex if for every $x, y \in C$ and $\lambda \in [0,1]$,*

$$\lambda x + (1-\lambda)y \in C.$$

Definition 3.40 (Convex Function)**.** *A function $f : X \to (-\infty, \infty]$ is convex if its epigraph*

$$\operatorname{epi}(f) := \{(x,t) \in X \times \mathbb{R} : f(x) \le t\}$$

is a convex set. Equivalently, for all $x, y \in X$ and $\lambda \in [0,1]$,

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y).$$

Definition 3.41 (Lower Semicontinuity)**.** *A function $f : X \to (-\infty, \infty]$ is lower semicontinuous (l.s.c.) if for every $x \in X$ and any sequence (x_n) in X converging to x,*

$$\liminf_{x_n \to x} f(x_n) \ge f(x).$$

Lower semicontinuity ensures that f does not jump downward abruptly and is critical in guaranteeing the existence of minimizers of f.

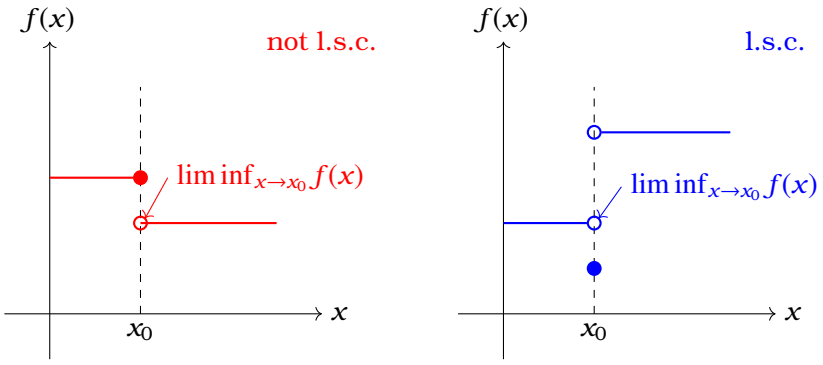


Figure 3.11. *Illustration of lower semicontinuity at x_0. Left: function with a jump down at x_0, failing lower semicontinuity since $\liminf_{x \to x_0} f(x) < f(x_0)$. Right: lower semicontinuous function where $\liminf_{x \to x_0} f(x) \ge f(x_0)$.*

Subdifferential As mentioned earlier, convex functions may fail to be differentiable everywhere. Nonetheless, they admit a powerful generalization of derivatives called *subgradient*.

Intuitively, a subgradient $p \in X^*$ at a point $x \in X$ provides a linear underestimate of the

function f near x. That is, the affine function

$$y \mapsto f(x) + p(y - x)$$

lies below the graph of f and touches it at x. The linear functional p belongs to the dual space X^* of continuous linear functionals on X.

Definition 3.42 (Subdifferential)**.** *The subdifferential of f at x, denoted $\partial f(x)$, is the set of all subgradients at x:*

$$\partial f(x) := \{p \in X^* : f(y) \geq f(x) + p(y - x) \text{ for all } y \in X\}.$$

The subdifferential generalizes the classical gradient: if f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.

Importantly, $\partial f(x)$ is always a closed, convex set. It may be empty at boundary points of the domain but is nonempty for points in the interior of $\mathrm{dom}(f)$. For proper, convex, lower semicontinuous functions, the subdifferential is nonempty at every point in the interior of $\mathrm{dom}(f)$. Intuitively, this reflects the fact that convex functions **always** admit supporting hyperplanes at interior points—a consequence of the separation theorems of convex analysis (related to Hahn-Banach). In contrast, at boundary points of the domain, $\partial f(x)$ may be empty.

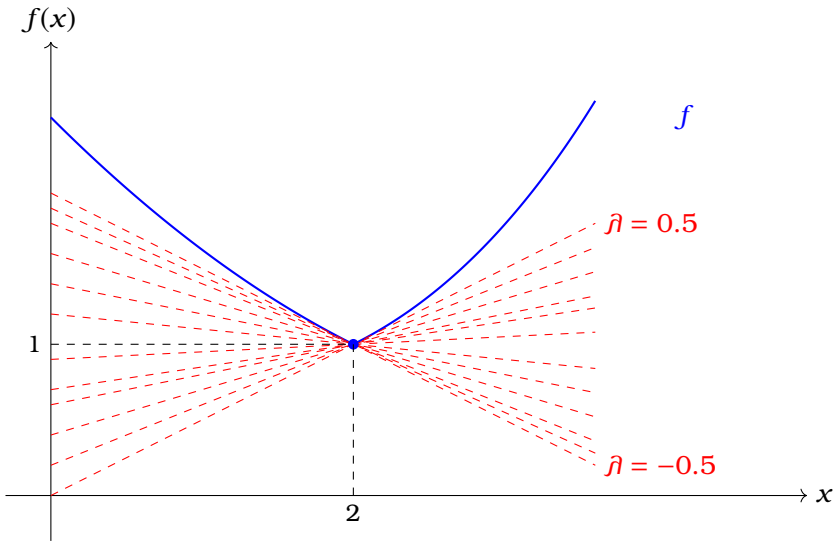


Figure 3.12. *Subdifferential of a convex function at the nondifferentiable point $x = 2$. All affine lines shown touch the graph at $(2, 1)$ and lie below it everywhere. The subdifferential $\partial f(2)$ consists of all slopes between -0.5 and 0.5, illustrating that the set of subgradients at a nondifferentiable point can be a nontrivial interval.*

Legendre–Fenchel Transform and Convex Duality The Legendre–Fenchel transform (or convex conjugate) is a key tool for representing convex functions in dual variables.

Definition 3.43 (Convex Conjugate)**.** *Given a function $f : X \to (-\infty, \infty]$, its convex conjugate $f^* : X^* \to (-\infty, \infty]$ is defined by*

$$f^*(p) := \sup_{x \in X} (p(x) - f(x)).$$

One way to understand the convex conjugate is to think about lines with a fixed slope p that lie below the graph of the function f. Among all such lines, we want to find the one with the highest possible offset.

More precisely, suppose we want a line with slope p and offset b such that for all x in the domain of f,

$$p(x) + b \leq f(x).$$

There may be many values of b satisfying this, but we want the largest such offset. Rearranging, we get

$$b \leq f(x) - p(x) \quad \text{for all } x.$$

The largest b that works must be less than or equal to the smallest value of $f(x) - p(x)$ over all x, so

$$b = \inf_x (f(x) - p(x)) = -\sup_x (p(x) - f(x)) = -f^*(y).$$

That is, for a fixed slope p, the intercept of the highest affine function lying below f is $-f^*(p)$, and thus the convex conjugate $f^*(y)$ encodes this offset information.

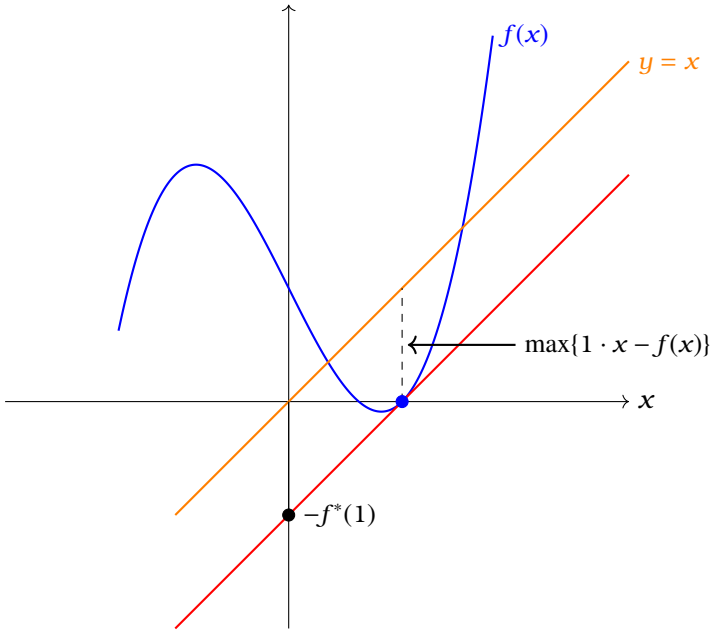


Figure 3.13. *Calculation of $f^*(1)$ by finding the point where the gap between the line $y = x$ and the function is maximized. At this point, the tangent to the function has slope 1, matching the slope of the line. The vertical offset of the tangent line is precisely $-f^*(1)$.*

Regardless of whether f is convex or not, f^* is a convex function.

From the definition of convex conjugate functions, we can easily derive the Fenchel-Young inequality: For every $x \in X$ and $p \in X^*$:

$$f(x) + f^*(p) \geq p(x).$$

Taking the convex conjugate twice yields the **biconjugate** $f^{**} := (f^*)^*$, which satisfies:

$$f^{**} \leq f.$$

This inequality holds because the biconjugate f^{**} is constructed as the supremum over all affine functions that underestimate f. More concretely, for any $p \in X^*$,

$$p(x) - f^*(p) \leq f(x)$$

by the definition of f^* as a supremum. Since $f^{**}(x)$ is the supremum of all such affine functions, it follows that

$$f^{**}(x) = \sup_{p \in X^*} \{p(x) - f^*(p)\} \leq f(x).$$

In fact, f^{**} is the greatest convex, lower semicontinuous function that does not exceed f. It is also called the *convex lower semicontinuous envelope* of f.
The Fenchel–Moreau theorem states that if f is proper (not infinite everywhere), convex, and lower semicontinuous, then

$$f = f^{**}.$$

Concave functions are defined as functions whose negative is convex. All the above concepts and results extend naturally to the concave setting by working with $-f$.

Chapter 4

Optimal Transport

4.1 Introduction and Motivation

Optimal transport is the mathematical theory concerned with finding the most efficient way to move mass from one probability distribution to another, given a specified cost of transportation. Its origins trace back to the work of Gaspard Monge in 1781, who posed the problem in the context of civil engineering: how can one move a pile of soil to a desired configuration while minimizing the total effort? In Monge's formulation, each unit of mass from the source must be transported to a single destination, reflecting a physical, one-to-one reallocation of material.

This question, while natural in physical and economic settings, poses nontrivial analytical challenges. Monge's original formulation does not always admit a solution, especially when the mass must be split or redistributed in more flexible ways. In the 20th century, Leonid Kantorovich introduced a relaxed formulation based on transport *plans* rather than maps, which allowed for splitting and led to a convex optimization problem. This relaxation made the theory more robust and opened the door to deep analytical and geometric insights.

Beyond its origins, optimal transport has become a powerful tool in modern mathematics, with connections to analysis, geometry, partial differential equations, and probability theory. In recent years, it has also gained significant traction in machine learning and statistics, where comparing and manipulating probability distributions is a central task. The Wasserstein distances, which arise from optimal transport costs, provide a meaningful geometry on the space of probability measures and are now widely used in generative modeling, domain adaptation, and robust statistical inference.

This section develops the mathematical foundations of optimal transport. We begin by introducing the Monge and Kantorovich formulations and the associated cost minimization problems. We then present key theoretical results, including the existence of optimal transport plans and the duality theory that characterizes them. We introduce Wasserstein distances and explore their topological and geometric properties. These results will provide the theoretical basis for the experiments in the next part of this thesis, where Wasserstein distances are used to study empirical distributions.

The presentation is based primarily on Villani's Topics in Optimal Transport [8] and Santambrogio's Optimal Transport for Applied Mathematicians [7],

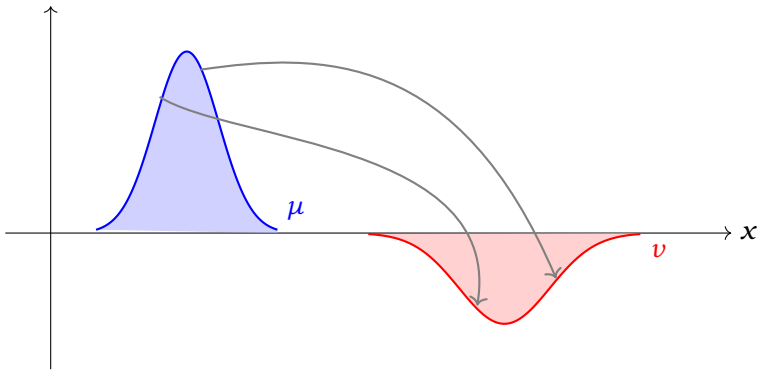


Figure 4.1. *A conceptual illustration of optimal transport: the goal is to move mass from the source distribution μ (blue, left) to the target distribution υ (red, right) while minimizing transport cost.*

4.2 Formulation of the problem

We formalize the problem of transporting mass efficiently from a source to a target distribution, minimizing an associated cost function. Recall that a Polish space is a complete separable metric space. Below, we will restrict our discussion to Borel probability measures on Polish spaces. Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \upsilon)$ be the source and target probability spaces. Let $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$ be a measurable cost function, which tells us what the cost of transporting a unit of mass from x to y is.

Monge's Formulation In the classical Monge problem, we seek a measurable map $T : \mathcal{X} \to \mathcal{Y}$ that pushes the measure μ onto υ. Using the notation of push-forward measures, we can state that:

$$T_{\#}\mu = \upsilon$$

Equivalently, for any measurable set $B \in \mathcal{Y}$:

$$\upsilon(B) = \mu(T^{-1}(B))$$

Using a standard approximation argument (see Appendix A for details), this condition is equivalent to requiring:

$$\int_{\mathcal{X}} f \, dT_{\#}\mu = \int_{\mathcal{Y}} f \, d\upsilon \qquad \forall f \in C_b(\mathcal{Y})$$

$$\int_{\mathcal{X}} (f \circ T) \, d\mu = \int_{\mathcal{Y}} f \, d\upsilon \qquad \forall f \in C_b(\mathcal{Y})$$

We can now formulate Monge's problem:

Monge's Problem: Find a measurable map $T : \mathcal{X} \to \mathcal{Y}$ which minimizes

$$M[T] = \int_{\mathcal{X}} c(x, T(x))\, d\mu(x).$$

Formally, solve

$$\inf_{T} \{M[T] : T_{\#}\mu = \nu\}.$$

The map T can be interpreted as transporting mass located at point x to the location $T(x)$. This interpretation reveals a fundamental limitation of Monge's problem: it does not allow for *splitting mass*.

For example, consider the case where $\mu = \delta_0$, meaning all the mass is concentrated at point 0, and $\nu = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$, meaning we wish to transport half of the mass to point 1 and the other half to point 2. Under Monge's formulation, this is impossible: the entire mass at 0 must be transported to a *single point*, since T is a function.

This illustrates why *existence of solutions* is not guaranteed in the Monge problem.

Kantorovich's Relaxation To resolve this issue, Kantorovich proposed a relaxed formulation of the problem. Instead of searching for transport maps T, he considered *transport plans*, also known as *couplings*, between the measures μ and ν.
A coupling $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is a joint probability measure on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν, meaning that for all measurable sets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$, we have:

$$\pi(A \times \mathcal{Y}) = \mu(A), \qquad \pi(\mathcal{X} \times B) = \nu(B).$$

Let $\Pi(\mu, \nu)$ denote the set of all such couplings. Using the standard characterization via integration (see Appendix A), a measure $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ belongs to $\Pi(\mu, \nu)$ if and only if, for all $(\phi, \psi) \in L^1(\mu) \times L^1(\nu)$, we have:

$$\int_{\mathcal{X}\times\mathcal{Y}} [\phi(x) + \psi(y)]\, d\pi(x, y) = \int_{\mathcal{X}} \phi(x)\, d\mu(x) + \int_{\mathcal{Y}} \psi(y)\, d\nu(y).$$

Kantorovich's optimal transport problem then consists of minimizing the total transportation cost over all admissible couplings:

Kantorovich's Problem: Find a transport plan $\pi \in \Pi(\mu, \nu)$ that minimizes the total transport cost:

$$I[\pi] = \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\pi(x, y).$$

Formally, solve

$$\inf_{\pi} \{I[\pi] : \pi \in \Pi(\mu, \nu)\}$$

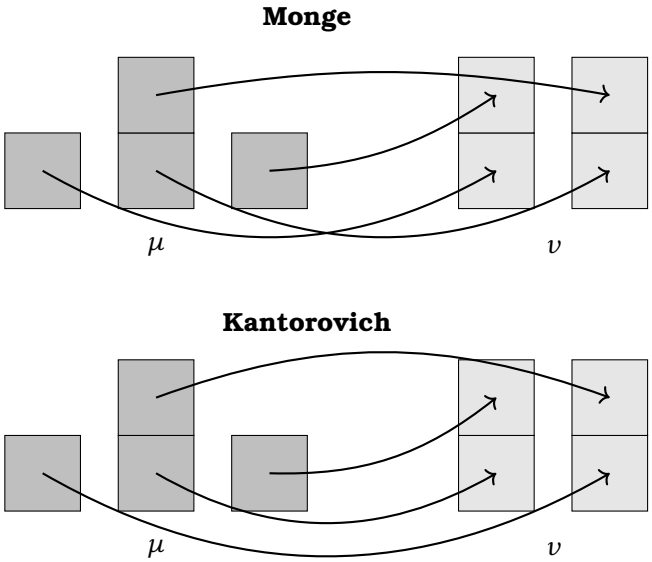


Figure 4.2. *Illustration of Monge's transport map (top) and Kantorovich's transport plan (bottom), starting from a discrete distribution with masses 1, 2, and 1 (represented as vertical stacks) and ending at a distribution with two equal stacks of mass 2. In Monge's formulation, mass from each source point must be moved entirely to a single target, which requires the middle heap (mass 2) to be transported as a whole. In contrast, Kantorovich's plan allows splitting: the mass from a single source can be distributed across multiple targets.*

Convexity and Existence of Solutions Compared to Monge's problem, Kantorovich's formulation exhibits significantly better mathematical behavior. One of its main strengths is that, as we will see in the next section, it always admits a solution under very mild assumptions.

A key reason for this improved behavior is that **the admissible set $\Pi(\mu, \nu)$ is convex**. Intuitively, this means that if two different transport plans π_1 and π_2 satisfy the marginal constraints, then so does any weighted average of them.

Proof. Let $\pi_1, \pi_2 \in \Pi(\mu, \upsilon)$ and let $\lambda \in [0, 1]$. Define the convex combination

$$\pi := \lambda\pi_1 + (1 - \lambda)\pi_2.$$

Then $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and is easily seen to have the correct marginals. Indeed, for any measurable set $A \subset \mathcal{X}$,

$$\pi(A \times \mathcal{Y}) = \lambda\pi_1(A \times \mathcal{Y}) + (1 - \lambda)\pi_2(A \times \mathcal{Y}) = \lambda\mu(A) + (1 - \lambda)\mu(A) = \mu(A),$$

and similarly, $\pi(\mathcal{X} \times B) = \upsilon(B)$ for all $B \subset \mathcal{Y}$. Hence, $\pi \in \Pi(\mu, \upsilon)$, proving that $\Pi(\mu, \upsilon)$ is convex. □

In contrast, Monge's problem can be seen as minimizing the transport cost $I[\pi]$ over the subset of transport plans $\pi \in \Pi(\mu, \upsilon)$ that are induced by transport maps. More precisely, these are measures π supported entirely on the graph of some measurable map

$T : \mathcal{X} \to \mathcal{Y}$, meaning

$$\pi(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : y = T(x)\}) = 1.$$

This subset of transport plans is generally *not convex*. To see why, consider two such plans π_1 and π_2, induced by maps T_1 and T_2, respectively. Their convex combination $\pi = \frac{1}{2}(\pi_1 + \pi_2)$ typically assigns positive mass to points $(x, T_1(x))$ and $(x, T_2(x))$ simultaneously, thus "splitting" mass from x between two locations. As a result, π cannot be represented as a transport plan induced by a single map T, because it is not concentrated on the graph of any function.

This lack of convexity — in stark contrast to the structure of $\Pi(\mu, \nu)$ — is a key reason Monge's problem often fails to admit minimizers.

4.3 Existence of Optimal Transport Plans

In this section, we will formally prove that Kantorovich's problem admits minimizers.

We will first focus on the case where the underlying spaces are compact, and the cost function is continuous.

Theorem 4.20 ([7, Theorem 1.4])**.** *Let X and Y be compact metric spaces, and let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$. Assume the cost function $c : X \times Y \to \mathbb{R}$ is continuous. Then there exists at least one optimal transport plan $\pi^* \in \Pi(\mu, \nu)$ such that*

$$\int_{X\times Y} c(x, y)\, d\pi^*(x, y) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} c(x, y)\, d\pi(x, y).$$

Proof. We want to show that the space $\Pi(\mu, \nu)$ is compact and metrizable under the weak-* topology on measures. This ensures that every sequence (π_n) in $\Pi(\mu, \nu)$ has a subsequence converging weakly-* to some measure π^*, meaning that for every $f \in C_b(X \times Y)$,

$$\int f\, d\pi_{n_k} \to \int f\, d\pi^*.$$

This guarantees the existence of a minimizer for the problem.

Since X and Y are compact metric spaces, their product $X \times Y$ is also a compact metric space. By the Riesz Representation Theorem, the dual of $C_b(X \times Y)$ can be identified as the space of signed Borel measures $M(X \times Y)$.

Because $C_b(X \times Y)$ is a Banach space, the Banach–Alaoglu theorem implies that the unit ball in $M(X \times Y)$ is compact under the weak-* topology. Moreover, since $X \times Y$ is compact, the weak-* topology on $M(X\times Y)$ is metrizable [1], and hence weak-* compactness is equivalent to sequential compactness. Since every $\pi \in \Pi(\mu, \nu)$ is a probability measure with norm 1, it lies in this unit ball. Therefore, any sequence π_n in $\Pi(\mu, \nu)$ has a weak-* convergent subsequence $\pi_{n_k} \to \pi^*$.

[1] This follows from Prokhorov's theorem, which states that the weak-* topology on probability measures is metrizable if the underlying space is Polish (separable and complete metric).

It remains to verify that π^* also belongs to $\Pi(\mu, \nu)$ (that the set is closed). For any $f \in C_b(X)$, we have by the marginal constraint

$$\int_{X\times Y} f(x)\, d\pi_{n_k}(x, y) = \int_X f(x)\, d\mu(x).$$

Passing to the limit, using the weak-* convergence, we get

$$\int_{X\times Y} f(x)\, d\pi^*(x, y) = \lim_{k\to\infty} \int_{X\times Y} f(x)\, d\pi_{n_k}(x, y) = \int_X f(x)\, d\mu(x).$$

Similarly, for any $g \in C_b(Y)$,

$$\int_{X\times Y} g(y)\, d\pi^*(x, y) = \int_Y g(y)\, d\nu(y).$$

Thus, π^* has marginals μ and ν, so $\pi^* \in \Pi(\mu, \nu)$.
This shows $\Pi(\mu, \nu)$ is weak-* compact, proving the existence of a solution.

□

While the existence of optimal transport plans is classical for compact metric spaces with continuous cost functions, the result extends naturally to Polish spaces.

Theorem 4.21 (Existence of Optimal Transport Plan in Polish Spaces)**.** *Let X, Y be Polish metric spaces, and let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ be probability measures. Assume the cost function $c : X \times Y \to [0, +\infty]$ is continuous. Then there exists a transport plan $\pi^* \in \Pi(\mu, \nu)$ minimizing the Kantorovich problem*

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} c(x, y)\, d\pi(x, y).$$

Proof. When X and Y are not compact, the dual of $C_0(X \times Y)$ is the space of Radon measures $M(X \times Y)$, but $C_c(X \times Y)$, its predual in the general setting, is not a Banach space. Hence, the Banach–Alaoglu theorem does not directly yield compactness. Instead, we apply Prokhorov's theorem, which guarantees relative sequential compactness under the condition of tightness.

We will show that $\Pi(\mu, \nu)$ is tight. Recall that tightness means that for every small $\varepsilon > 0$, there exists a compact set $K \subset X \times Y$ such that all measures π in the set satisfy

$$\pi(K^c) < \varepsilon.$$

Since X and Y are Polish spaces, every finite Borel measure on them is Radon, and thus inner regular (i.e., can be approximated by measures supported on compact sets). In particular, this applies to μ and ν, so there exist compact sets $K_X \subset X$ and $K_Y \subset Y$ such that

$$\mu(X \setminus K_X) < \frac{\varepsilon}{2} \quad \text{and} \quad \nu(Y \setminus K_Y) < \frac{\varepsilon}{2}.$$

Now consider the product set $K_X \times K_Y \subset X \times Y$, which is compact. For any $\pi \in \Pi(\mu, \nu)$, we estimate:

$$\pi\big((X \times Y) \setminus (K_X \times K_Y)\big) \leq \pi(X \times (Y \setminus K_Y)) + \pi((X \setminus K_X) \times Y) = \nu(Y \setminus K_Y) + \mu(X \setminus K_X) < \varepsilon.$$

Therefore, $\Pi(\mu, \nu)$ is tight, and by Prokhorov's theorem, relatively compact in the topology of weak convergence of measures. Since $\Pi(\mu, \nu)$ is closed (under weak convergence), it is sequentially compact.
Now, let $(\pi_n) \subset \Pi(\mu, \nu)$ be a minimizing sequence. By compactness, there exists a subsequence (π_{n_k}) that converges weakly to some $\pi^* \in \Pi(\mu, \nu)$. Hence,

$$\int c \, d\pi^* = \lim_{k \to \infty} \int c \, d\pi_{n_k} = \inf_{\pi \in \Pi(\mu,\nu)} \int c \, d\pi,$$

and so the infimum is attained at π^*. □

The theorem remains valid if the cost function $c : X \times Y \to [0, +\infty]$ is merely *lower semicontinuous* and bounded from below. In that case, the map

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

is lower semicontinuous with respect to weak convergence of measures. A full proof of this fact can be found in standard references such as [7, Theorem 1.7].

4.4 Kantorovich duality

The Kantorovich formulation of optimal transport is a linear program over the space of probability measures. It is natural to ask whether there exists a corresponding dual problem, and whether strong duality holds. The answer is yes, under general assumptions — a remarkable fact taking into account the infinite-dimensional nature of the problem.

Theorem 4.22 (Kantorovich duality [8, Thm 1.3])**.** *Let $(X, \mu), (Y, \nu)$ be probability Polish spaces, $c : X \times Y \to [0, +\infty]$ a lower semi-continuous cost function. Define $I[\pi] = \int_{X\times Y} c(x, y)\, d\pi(x, y)$ as above and $J(\phi, \psi) = \int_X \phi(x)d\mu(x) + \int_Y \psi(y)d\nu(y)$. Define the set of admissible plans $\Pi(\mu, \nu)$ as above and Φ_c as the set of all measurable functions $(\phi, \psi) \in L^1(d\mu) \times L^1(d\nu)$ such that $\phi(x) + \psi(y) \leq c(x, y)$ $\mu \otimes \nu$-almost surely. Then,*

$$\inf_{\Pi(\mu,\nu)} I[\pi] = \sup_{\Phi_c} J(\phi, \psi) \tag{4.1}$$

Proof. This proof sketch relies on a minimax principle – Fenchel–Rockafellar duality – valid due to convexity and lower semi-continuity assumptions. Recall that $M_+(X \times Y)$ is the space of nonnegative Borel measures on $X \times Y$.

$$\inf_{\Pi(\mu,\nu)} I[\pi] = \inf_{M_+(X\times Y)} \left\{ I[\pi] + \begin{cases} 0, & \text{if } \pi \in \Pi(\mu, \nu) \\ +\infty, & \text{else} \end{cases} \right\}$$

Notice that the indicator function of the constraint $\pi \in \Pi(\mu, \nu)$ can be expressed as:

$$\sup_{(\phi,\psi)} \left\{ \int_X \phi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) - \int_{X\times Y} [\phi(x) + \psi(y)]\, d\pi(x, y) \right\}$$

This dual expression arises from enforcing the marginal constraints via Lagrange multipliers (ϕ, ψ), which play the role of pricing functions on the source and target spaces. Substituting into the expression above and using the min-max principle, we obtain:

$$\begin{aligned}
\inf_{\Pi(\mu,\nu)} I[\pi] &= \inf_{\pi\in M_+(X\times Y)} \sup_{(\phi,\psi)} \left\{ \int_X \phi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) - \int_{X\times Y} [\phi(x) + \psi(y) - c(x, y)]d\pi(x, y) \right\} \\
&= \sup_{(\phi,\psi)} \inf_{\pi\in M_+(X\times Y)} \left\{ \int_X \phi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) - \int_{X\times Y} [\phi(x) + \psi(y) - c(x, y)]d\pi(x, y) \right\} \\
&= \sup_{(\phi,\psi)} \left\{ \int_X \phi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) - \sup_{\pi\in M_+(X\times Y)} \int_{X\times Y} [\phi(x) + \psi(y) - c(x, y)]d\pi(x, y) \right\}
\end{aligned}$$

Notice that if $c(x, y) \geq \phi(x) + \psi(y)$ almost surely, the right supremum is achieved for $\pi = 0$. On the other hand, if there exists a point where $c(x, y) < \phi(x) + \psi(y)$, choosing π be a Dirac mass at this point with very large mass, we see that the supremum is infinite. Thus,

$$\begin{aligned}
\inf_{\Pi(\mu,\nu)} I[\pi] &= \sup_{(\phi,\psi)} \left\{ \int_X \phi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) - \begin{cases} 0, & \text{if } (\phi, \psi) \in \Phi_c \\ +\infty, & \text{else} \end{cases} \right\} \\
&= \sup_{\Phi_c} \left\{ \int_X \phi(x)d\mu(x) + \int_Y \psi(y)d\nu(y) \right\} = \sup_{\Phi_c} J(\phi, \psi)
\end{aligned}$$

□

Now, we build intuition behind the dual problem formulation. Imagine you are an industrialist who needs to transport goods from factories to customers, where the transportation cost from factory x to customer y is $c(x, y)$. Your goal is to minimize the total transportation cost.

A friend offers to handle the transportation, charging a fee $\phi(x)$ per product taken from factory x and a fee $\psi(y)$ for delivering a product to customer y. He guarantees that for every factory-customer pair (x, y), the combined fee does not exceed the transportation cost:

$$\phi(x) + \psi(y) \leq c(x, y).$$

Kantorovich's duality theorem states that by carefully choosing these fees (which can be positive or negative), the total fees collected are equal to the minimal transportation cost:

$$\min_{\Pi(\mu,\nu)} \int_{X\times Y} c(x, y)\, d\pi(x, y) = \max_{\phi(x)+\psi(y)\leq c(x,y)} \int \phi(x)\, d\mu(x) + \int \psi(y)\, d\nu(y).$$

When equality holds for a pair (x, y), the transport along that route is *tight* and forms part of the optimal transport plan. Routes where the inequality is strict are not used. Thus, the dual potentials ϕ and ψ act as *prices* or *potentials* that reveal the structure of optimal transportation.

This dual perspective transforms the primal problem into a maximization over pricing functions, and strong duality ensures these two problems have the same value.

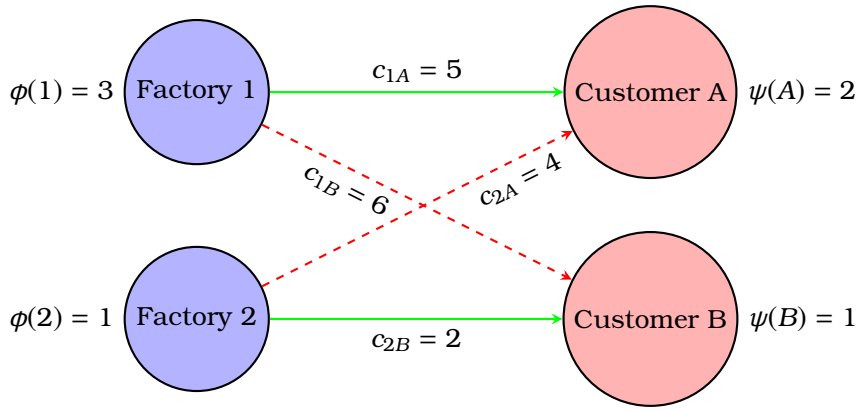


Figure 4.3. *Illustration of Kantorovich duality with dual potentials ϕ, ψ representing fees assigned to factories and customers. Arrows with "tight" constraints (green) correspond to active transport routes where $\phi(x)+\psi(y) = c(x, y)$, such as $\phi(1)+\psi(A) = 3+2 = 5 = c_{1A}$ and $\phi(2)+\psi(B) = 1+1 = 2 = c_{2B}$. "Slack" edges (dashed gray) satisfy $\phi(x)+\psi(y) < c(x, y)$ and do not appear in the optimal plan. For example, $\phi(1) + \psi(B) = 3 + 1 < 6 = c_{1B}$ and $\phi(2) + \psi(A) = 1 + 2 < 4 = c_{2A}$. The dual problem maximizes total fees $\sum \phi(x)\, d\mu(x) + \sum \psi(y)\, d\nu(y)$ under these constraints.*

The Metric case By limiting our analysis to specific cost function classes, we can gain more structure on the dual problem. Below, we will focus on cost functions on $X \times X$ that are metrics $c(x, y) = d(x, y)$. This case arises often in practical problems, when we want

to transport probability distributions living on the same space, since properties such as the triangle inequality and symmetry typically hold.

Let X be a Polish space. Recall that a function $f : X \to \mathbb{R}$ is Lipschitz if there exists $L > 0$ such that $\forall x, y \in X : |f(x) - f(y)| < Ld(x, y)$. The smallest such constant is called the Lipschitz constant of f, and we denote it by:

$$\|f\|_{\text{Lip}} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)}.$$

The space of all Lipschitz functions on X is denoted by $\text{Lip}(X)$. We also denote the unit ball of $\text{Lip}(X)$ as $\text{Lip}_1(X)$:

$$\text{Lip}_1(X) = \left\{ \varphi \in \text{Lip}(X) : \|\varphi\|_{Lip} \leq 1 \right\}.$$

Theorem 4.23 (Kantorovich-Rubinstein theorem [8, Thm 1.3])**.** *Let X be a Polish space, d a lower semi-continuous metric on X. Let $\mathcal{T}_d(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} I[\pi]$ be the optimal transportation cost defined as above. Then,*

$$\mathcal{T}_d(\mu, \nu) = \inf_{\Pi(\mu,\nu)} I[\pi] = \sup_{\text{Lip}_1(X)} \int_X \varphi d(\mu - \nu) = \sup \left\{ \int \varphi \, d(\mu - \nu) : \|\varphi\|_{\text{Lip}} \leq 1 \right\} \tag{4.2}$$

Remark 1. *The Kantorovich-Rubinstein distance defines a norm on signed measures with finite first moment. Let $\tilde{M}(X)$ denote the set of all finite signed Borel measures σ on X such that $\int d(x, x_0) d|\sigma|(x) < \infty$ for some (hence any) $x_0 \in X$. Then the KR norm of $\sigma \in \tilde{M}(X)$ is defined as:*

$$\|\sigma\|_{KR} = \sup_{\text{Lip}_1(X)} \int_X \varphi \, d\sigma$$

Kantorovich-Rubinstein's theorem then yields:

$$\mathcal{T}_d(\mu, \nu) = \|\mu - \nu\|_{KR}.$$

This implies mass invariance of the transportation cost when the cost is a metric: adding common mass to both measures does not affect their transportation cost. Formally, for any finite Borel measure σ:

$$\mathcal{T}_d(\mu + \sigma, \nu + \sigma) = \mathcal{T}_d(\mu, \nu)$$

Equivalently,

$$\mathcal{T}_d(\mu, \nu) = \mathcal{T}_d([\mu - \nu)]_+, [\nu - \mu]_+)$$

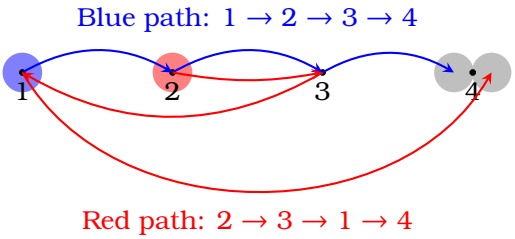
The above observation seems intuitively clear; if we want to transport some mass from $\mu + \sigma$ to $\nu + \sigma$, we can just transport μ to ν and leave the σ mass in place. However, for this property to hold, it is crucial for the cost function to be a metric and for the triangle inequality to hold. Conversely, suppose that $\mu = \delta_{-1}$, $\nu = \delta_1$, $\sigma = \delta_0$ and $c(x, y) = (x - y)^2$. Then, $\mathcal{T}_d(\mu, \nu) = ((-1) - 1)^2 = 4$, while to transport $\mu + \sigma$ to $\nu + \sigma$, we can transport the mass from -1 to 0 and the mass from 0 to 1, leading to a smaller cost

of $\mathcal{T}_d(\mu+\sigma, \nu+\sigma) = 1^2 + 1^2 = 2$.

In the case that the cost function is a metric, the total cost depends only on the difference between the measures μ and ν. Then, Kantorovich's optimal transportation problem is equivalent to the **Kantorovich-Rubinstein transshipment problem**.
In the transshipment problem, we want to minimize the transportation cost but the set of admissible plans changes from the distributions with marginals μ and ν to those that satisfy a flow conservation constraint, ensuring that in the end, the change in mass at each point is the difference between the target and the initial distribution:

$$\mathcal{T}_d(\mu, \nu) = \inf\{I[\pi] : \pi[A \times X] - \pi[X \times A] = (\mu - \nu)[A]\} \tag{4.3}$$

Intuitively, in this problem, we do not have to directly transport mass from a point x of the initial distribution to a point y of the target distribution. Instead, we are allowed to use any number of intermediate nodes for transshipment. The transshipment plan π now encompasses all mass movements between sources, transshipment points and destinations. Assuming for simplicity that the distributions μ and ν are disjoint, the condition in (4.3) guarantees that if x is a source point, the mass $\mu(x)$ leaves this point, while any mass transshipped at this point has a net zero effect on the balance. At transshipment points, the mass entering the point equals the mass leaving it. A destination point y has a net mass gain of $\nu(y)$, while some other mass may get transshipped there.
This is highlighted in the example of a transshipment plan below.



$$\pi = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 4.4. *Example of a transshipment plan transporting $\mu = \delta_1 + \delta_2$ to $\nu = 2\delta_4$. The nodes represent points in the network, where nodes 1 and 2 are sources each with one unit of mass, and node 4 is the sink receiving two units. The blue path shows mass moving from node 1 through nodes 2 and 3 to node 4, while the red path represents a cycle moving mass from node 2 through nodes 3 and 1 before reaching node 4. The matrix π encodes the amount of mass transported from node i to node j, where the entry π_{ij} corresponds to this transported quantity.*

In general, this is a strongly relaxed version of the optimal transportation problem. However, in the case where the cost function is a distance, the two problems are equivalent. That is because when the triangle inequality holds, there is no benefit in gradually transporting mass from x to y through $x_1, \ldots, x_n$ over transporting it directly: $c(x, y) \leq c(x, x_1) + \cdots + c(x_n, y)$. This equivalence highlights the geometric role of the metric: when the cost satisfies the triangle inequality, no detour via intermediate points can reduce transport cost.

4.5 Wasserstein Distances

The optimal transport formulation naturally gives rise to a family of metrics between probability measures known as Wasserstein distances. To appreciate their significance, consider two distributions with disjoint supports, such as indicator functions $f(x) = \mathbb{1}_{[0,1]}$ and $g_{\lambda}(x) = \mathbb{1}_{[\lambda,\lambda+1]}$ for $\lambda > 1$. Traditional function distances, like L^p or supremum distance, yield a constant value ($2^{1/p}$ for L^p or 1 for supremum distance), regardless of how far apart λ makes the supports. These metrics fail to account for the actual spatial displacement between the distributions on the horizontal axis.

In contrast, the Wasserstein distance is defined as the solution to an optimal transport problem, uniquely quantifying the minimum cost of physically moving mass from one distribution to another, thereby directly incorporating the spatial arrangement and distance between their supports. In particular, the p-Wasserstein distance W_p uses as a cost function the p-th power of the ground distance. Thus, in our example, the optimal transport cost W_p^p between the distributions would be λ^p, as the entire distribution is effectively shifted by λ units. Consequently, the p-Wasserstein distance $W_p(\mu, \nu)$ itself would be λ.

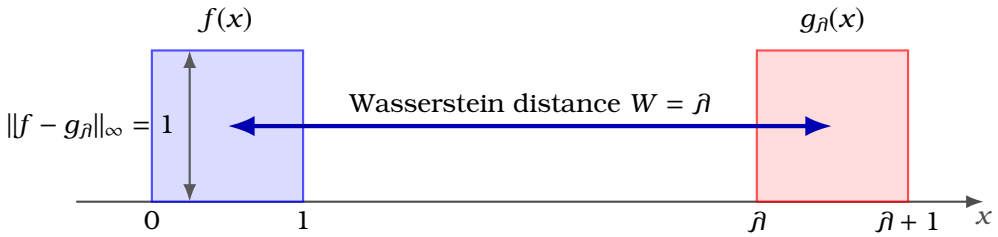


Figure 4.5. *Comparison of L^∞ and Wasserstein distances. The L^∞ distance reflects the maximum pointwise difference between the functions and remains constant, regardless of how far apart the distributions are. In contrast, the Wasserstein distance reflects the spatial cost of transporting the mass — increasing linearly with separation.*

Now, let us formally define the p-Wasserstein distance.

Definition 4.44 (p-Wasserstein Distance)**.** *Let (X, d) be a Polish (complete separable) space, and let $\mathcal{P}(X)$ denote the space of all Borel probability measures on X. Let $\mu, \nu \in \mathcal{P}(X)$ and $p \geq 1$. The p-**Wasserstein distance** is defined as*

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d(x, y)^p \, d\pi(x, y) \right)^{1/p},$$

*where $\Pi(\mu, \nu)$ denotes the set of all couplings of μ and ν (all probability measures on $X \times X$ with marginals μ and ν). The 1-Wasserstein distance is also called **Earth Mover's Distance**.*

In order for $W_p(\mu, \nu)$ to be finite, both μ and ν must have finite p-th moments. Indeed, if $\pi \in \Pi(\mu, \nu)$ is any coupling,

$$\begin{aligned} W_p(\mu, \nu) &= \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d(x, y)^p \, d\pi(x, y) \right)^{1/p} \\ &\leq \left(\int_{X \times X} (d(x, x_0) + d(x_0, y))^p \, d\pi(x, y) \right)^{1/p} \quad \text{(Triangle inequality)} \\ &\leq \left(\int_{X \times X} d(x, x_0)^p \, d\pi(x, y) \right)^{1/p} + \left(\int_{X \times X} d(x_0, y)^p \, d\pi(x, y) \right)^{1/p} \quad \text{(Minkowski)} \\ &= \left(\int_X d(x, x_0)^p \, d\mu(x) \right)^{1/p} + \left(\int_X d(y, x_0)^p \, d\nu(y) \right)^{1/p}. \end{aligned}$$

We define the Wasserstein space:

$$\mathcal{P}_p(X) := \left\{ \mu \in \mathcal{P}(X) \;\middle|\; \int_X d(x_0, x)^p \, d\mu(x) < \infty \text{ for some (hence any) } x_0 \in X \right\},$$

on which W_p is a well-defined metric.
The quantity

$$W_p^p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d(x, y)^p \, d\pi(x, y),$$

represents the minimal total cost of transporting the mass of μ to match that of ν under the cost function d^p. The $1/p$ exponent ensures that W_p has the same units as the ground metric d, and is essential for W_p to be a metric.

We will now prove that the Wasserstein distance is a metric.

Theorem 4.24. *[7, Proposition 5.1] Let (X, d) be a Polish space and $p \geq 1$. Then W_p defines a metric on the space $\mathcal{P}_p(X)$ of probability measures with finite p-th moment.*

Proof. Let $\mu, \nu, \sigma \in \mathcal{P}_p(X)$.
(1) Non-negativity and identity of indiscernibles: By definition, $W_p(\mu, \nu)$ is the infimum of a non-negative function, so $W_p(\mu, \nu) \geq 0$. If $\mu = \nu$, the diagonal coupling

$$\pi(x, y) = \delta_x(y)\mu(x)$$

yields zero cost, since $x = y$ π-almost everywhere. Conversely, if $W_p(\mu, \nu) = 0$, then there exists an optimal coupling $\pi \in \Pi(\mu, \nu)$ such that $x = y$ π-almost everywhere. Hence, for any test function $f : X \to \mathbb{R}$,

$$\int_X f(x) \, d\mu(x) = \int_{X \times X} f(x) \, d\pi(x, y) = \int_{X \times X} f(y) \, d\pi(x, y) = \int_X f(y) \, d\nu(y).$$

Since this holds for all $f \in C_b(X)$, we conclude that $\mu = \nu$ (from the Riesz Representation Theorem). Therefore, $W_p(\mu, \nu) = 0 \iff \mu = \nu$, as required.

(2) Symmetry: If $\pi \in \Pi(\mu, \nu)$, then the measure π^{rev}, defined by $\pi^{\text{rev}}(A \times B) = \pi(B \times A)$,

belongs to $\Pi(\nu, \mu)$, and since $d(x, y)^p = d(y, x)^p$, the cost is unchanged. Hence, $W_p(\mu, \nu) = W_p(\nu, \mu)$.

(3) Triangle inequality:

Before proving the triangle inequality, we state an important lemma that enables the construction of a joint coupling of three measures from pairwise couplings.

Theorem 4.25 (Gluing Lemma)**.** *Let $\pi \in \Pi(\mu, \sigma)$ and $\eta \in \Pi(\sigma, \nu)$. Then there exists a probability measure $\gamma \in \mathcal{P}(X \times X \times X)$ such that the marginal of γ on (x, y) is π and the marginal of γ on (y, z) is η.*

Proof. The existence of $\gamma \in \mathcal{P}(X \times X \times X)$ with the desired marginals follows from the Disintegration Theorem. Given $\pi \in \Pi(\mu, \sigma)$ and $\eta \in \Pi(\sigma, \nu)$, there exist families of probability measures $\{\pi_y\}_{y \in X}$ and $\{\eta_y\}_{y \in X}$ such that

$$\pi(A \times B) = \int_B \pi_y(A)\, d\sigma(y), \quad \eta(B \times C) = \int_B \eta_y(C)\, d\sigma(y),$$

for measurable sets $A, B, C \subseteq X$.

Define γ on measurable rectangles by

$$\gamma(A \times B \times C) := \int_B \pi_y(A)\eta_y(C)\, d\sigma(y).$$

This extends uniquely to a probability measure on X^3.

Its marginals satisfy

$$\gamma(A \times B \times X) = \int_B \pi_y(A) \cdot \underbrace{\eta_y(X)}_{=1}\, d\sigma(y) = \int_B \pi_y(A)\, d\sigma(y) = \pi(A \times B),$$

and

$$\gamma(X \times B \times C) = \int_B \underbrace{\pi_y(X)}_{=1} \cdot \eta_y(C)\, d\sigma(y) = \int_B \eta_y(C)\, d\sigma(y) = \eta(B \times C),$$

since π_y and η_y are probability measures for σ-a.e. y.

Thus, γ has marginals π and η as required. □

Now, we prove that W_p satisfies the triangle inequality.

Let $\pi \in \Pi(\mu, \sigma)$ and $\eta \in \Pi(\sigma, \nu)$ be optimal couplings, i.e., minimizers for $W_p(\mu, \sigma)$ and $W_p(\sigma, \nu)$, respectively. By the gluing lemma, there exists a measure γ on $X \times X \times X$ with marginals $\pi(x, y)$ and $\eta(y, z)$. Let ∂ be the marginal of γ on (x, z); then $\partial \in \Pi(\mu, \nu)$. Applying Minkowski's inequality:

$$
\begin{aligned}
W_p(\mu,\nu) &\le \left(\int_{X\times X} d(x,z)^p \, d\partial(x,z)\right)^{1/p} \quad (W_p \text{ is the infimum}) \\
&= \left(\int_{X\times X\times X} d(x,z)^p \, d\gamma(x,y,z)\right)^{1/p} \\
&\le \left(\int_{X\times X\times X} (d(x,y)+d(y,z))^p \, d\gamma(x,y,z)\right)^{1/p} \quad \text{(Triangle inequality)} \\
&\le \left(\int_{X\times X} d(x,y)^p \, d\pi(x,y)\right)^{1/p} + \left(\int_{X\times X} d(y,z)^p \, d\eta(y,z)\right)^{1/p} \quad \text{(Minkowski)} \\
&= W_p(\mu,\sigma) + W_p(\sigma,\nu).
\end{aligned}
$$

Hence, all metric properties are satisfied. □

Relation between W_1 and W_p Now, we will explore the comparison between W_1 and W_p for $p > 1$.

Theorem 4.26. *Let (X, d) be a metric space. For every $p \ge 1$ and any $\mu, \nu \in \mathcal{P}_p(X)$,*

$$W_p(\mu,\nu) \ge W_1(\mu,\nu).$$

In the case where X is bounded, with diameter $D = \sup_{x,y\in X} d(x,y) < \infty$, we also have that

$$W_p(\mu,\nu)^p \le D^{p-1} W_1(\mu,\nu).$$

Proof. By Jensen's inequality applied to the convex function $t \mapsto t^p$ (with $p \ge 1$), for any coupling π:

$$\left(\int d(x,y)\, d\pi(x,y)\right)^p \le \int d(x,y)^p \, d\pi(x,y).$$

Taking the infimum over all couplings π, we get

$$W_1(\mu,\nu)^p \le W_p(\mu,\nu)^p \iff W_1(\mu,\nu) \le W_p(\mu,\nu).$$

In the case where the space is bounded, we have

$$d(x,y)^p \le D^{p-1} d(x,y),$$

so for any coupling $\pi \in \Pi(\mu,\nu)$

$$\int d(x,y)^p \, d\pi(x,y) \le D^{p-1} \int d(x,y)\, d\pi(x,y).$$

Taking the infimum over π, we get

$$W_p(\mu,\nu)^p \le D^{p-1} W_1(\mu,\nu).$$

□

We illustrate this behavior in the following example.

Example 4.5. *Let $\mu = \delta_0$. We consider two different target measures.*

Case 1 (Symmetric): *Let $\nu_1 = \frac{1}{2}\delta_{-r} + \frac{1}{2}\delta_r$ for some $r > 0$. Then, the unique optimal transport plan moves half of the mass from 0 to $-r$ and half to r. For any $p \geq 1$, we compute:*

$$W_1(\mu, \nu_1) = \frac{1}{2}r + \frac{1}{2}r = r,$$

$$W_p(\mu, \nu_1) = \sqrt[p]{\frac{1}{2}r^p + \frac{1}{2}r^p} = r.$$

So $W_1 = W_p$. This equality holds because all transported mass moves the same distance.

Case 2 (Asymmetric): *Now let $\nu_2 = (1 - \varepsilon)\delta_0 + \varepsilon\delta_D$, with $0 \ll D$ and $\varepsilon \ll 1$. The transport plan sends a small amount of the mass to a distant point:*

$$W_1(\mu, \nu_2) = D \cdot \varepsilon,$$

$$W_p(\mu, \nu_2) = \sqrt[p]{\varepsilon D^p} = D \cdot \sqrt[p]{\varepsilon}.$$

For small ε, we have $W_1 \approx 0$, while $W_p = D \cdot \varepsilon^{1/p}$, which can be much larger than W_1.

We observe that $W_1 = W_p$ only in very symmetric configurations where all mass travels equal distances. In general, higher-order Wasserstein distances penalize long-range mass transport more, causing $W_p \gg W_1$ when even a small portion of mass is transported far.

Remark 2. *As seen in the section on duality, the Wasserstein-1 distance W_1 admits the Kantorovich–Rubinstein dual representation (since its cost function is a distance metric):*

$$W_1(\mu, \nu) = \sup\left\{\int f \, d(\mu - \nu) : f \in \mathrm{Lip}_1\right\}.$$

This duality implies that $W_1(\mu, \nu)$ only depends on the signed measure $\mu - \nu$. That is, if $\mu_1 - \nu_1 = \mu_2 - \nu_2$, then $W_1(\mu_1, \nu_1) = W_1(\mu_2, \nu_2)$. Moreover, this dual formulation often makes W_1 easier to compute or approximate than higher-order Wasserstein distances W_p with $p > 1$, since it avoids the explicit search over transportation plans and can be approached via optimization over Lipschitz functions.

Example 4.6. *Let $\mu_1 = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, $\nu_1 = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$, and $\mu_2 = \delta_0, \nu_2 = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_2$. We can easily check that $\mu_1 - \nu_1 = \mu_2 - \nu_2 = \frac{1}{2}(\delta_0 - \delta_2)$, so, by Kantorovich–Rubinstein duality, we expect that $W_1(\mu_1, \nu_1) = W_1(\mu_2, \nu_2)$. Indeed, one can easily verify that they are both equal to 1.*

However, computing the W_p distance reveals a difference.

For (μ_1, ν_1), one optimal plan moves mass:

$$0.5 : 0 \to 1, \quad 0.5 : 1 \to 2 \quad \textit{so} \quad W_p(\mu_1, \nu_1)^p = 0.5 \cdot 1^p + 0.5 \cdot 1^p = 1.$$

For (μ_2, ν_2), an optimal plan moves mass:

$$0.5 : 0 \to 0, \quad 0.5 : 0 \to 2 \quad \textit{so} \quad W_p(\mu_2, \nu_2)^p = 0.5 \cdot 2^p \implies W_p(\mu_2, \nu_2) = 2^{\frac{p-1}{p}} > 1.$$

For instance, $W_2 = \sqrt{2} \approx 1.41$, $W_4 = 2^{3/4} \approx 1.68$, and $W_{100} \approx 1.99$.
This example illustrates that although W_1 depends only on the difference $\mu - \nu$, higher-order Wasserstein distances W_p for $p > 1$ are sensitive to the specific transportation plan and penalize long-range transport more heavily.

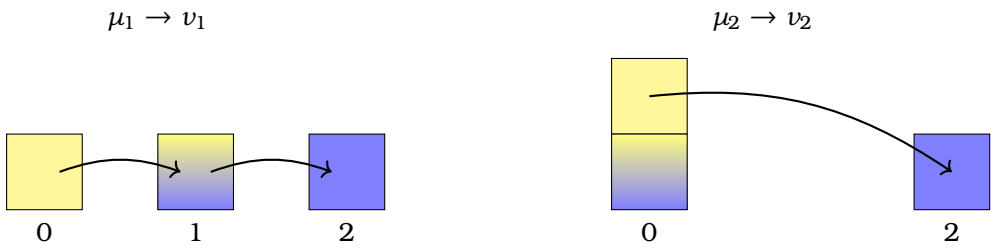


Figure 4.6. *Visualization of two transport problems with equal W_1 but different W_p values. Yellow boxes represent the mass of μ, blue boxes represent the mass of ν, and the multicolor boxes represent overlap. The first transport moves mass between adjacent points, while the second involves mass at 0 moving to point 2, resulting in a higher cost for $p > 1$.*

4.6 Optimal Transport in One Dimension

We now focus on the special case of optimal transport between one-dimensional probability measures, where the problem admits an elegant and explicit solution.
Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$, i.e., probability measures on $\mathbb{R}$. Denote their cumulative distribution functions (CDFs) by

$$F_\mu(x) := \mu((-\infty, x]), \quad F_\nu(x) := \nu((-\infty, x]),$$

and define their *quantile functions* by

$$F_\mu^{-1}(t) := \inf\{x \in \mathbb{R} : F_\mu(x) \geq t\}, \quad t \in [0, 1],$$

and similarly for F_ν^{-1}.

Theorem 4.27 (Optimal Transport in One Dimension)**.** *[7, Sections 2.1-2.2] Let μ, ν be probability measures on $\mathbb{R}$ with cumulative distribution functions F_μ, F_ν, and let λ denote the Lebesgue measure on $[0, 1]$. Then the unique optimal plan π^* for the cost $c(x, y) = d(x - y)$, with d continuous and convex, is*

$$\pi^* = (F_\mu^{-1}, F_\nu^{-1})_\# \lambda, \qquad H(x, y) := \pi^*((-\infty, x] \times (-\infty, y]) = \min\{F_\mu(x), F_\nu(y)\},$$

and the optimal transport cost is

$$\min_{\pi \in \Pi(\mu,\nu)} \int d(x - y)\, d\pi(x, y) = \int_0^1 d(F_\mu^{-1}(t) - F_\nu^{-1}(t))\, dt.$$

Proof. We begin by introducing the notion of **cyclical monotonicity**.

Definition 4.45 (Cyclical Monotonicity)**.** *A subset $\Gamma \subset \mathbb{R}^2$ is said to be cyclically monotone with respect to a cost function if for all $(x_1, y_1), (x_2, y_2) \in \Gamma$, the following inequality holds:*

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

This property captures the idea that swapping destinations between matched points cannot reduce the total transport cost.

An important theorem in optimal transport theory states that every optimal transport plan is supported on a cyclically monotone set. That is, if π is an optimal coupling, then the set of points (x, y) that it moves mass between must satisfy the inequality above. The full proof of this result is technical and can be found in [5, Theorem 2.3].

Intuitively, if both (x_1, y_1) and (x_2, y_2) belong to the support of a transport plan, this means that mass is being transported from x_1 to y_1 and from x_2 to y_2. If the inequality does not hold, we could instead move as much mass as possible from x_1 to y_2 and from x_2 to y_1, which would lower the total transport cost. In that case, at least one of the original pairs should not appear in the support of an optimal plan. Therefore, optimality requires the support of the plan to satisfy cyclical monotonicity.

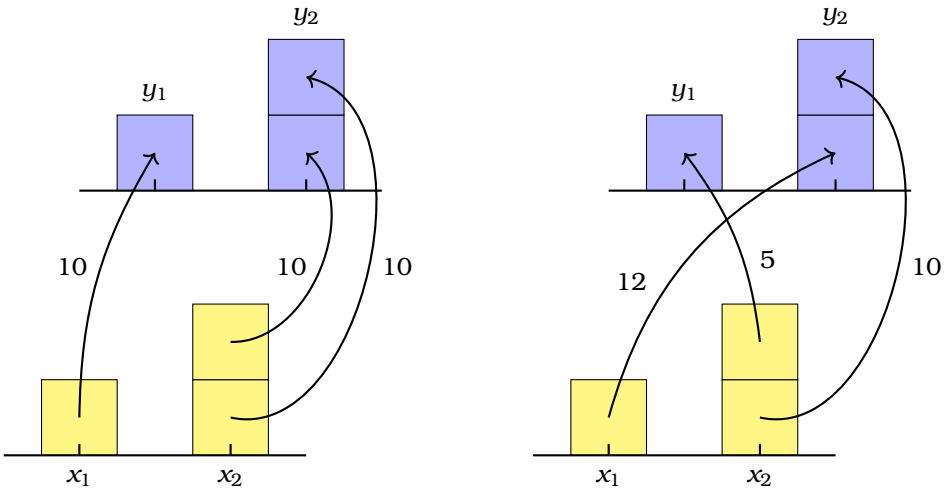


Figure 4.7. *An illustration of a transport plan that violates cyclical monotonicity. The source measure μ consists of one unit of mass at x_1 and two units at x_2, while the target measure ν consists of one unit of mass at y_1 and two units at y_2. The transport costs are: $c_{11} = 10$, $c_{12} = 12$, $c_{21} = 5$, and $c_{22} = 10$.* ***Left:*** *In the original plan, x_1 sends one unit to y_1, and x_2 sends its two units to y_2, for a total cost of $1 \cdot c_{11} + 2 \cdot c_{22} = 1 \cdot 10 + 2 \cdot 10 = 30$.* ***Right:*** *In the swapped plan, one unit from x_1 is redirected to y_2 and one unit from x_2 is redirected to y_1, while the second unit from x_2 still goes to y_2. The new total cost becomes $1 \cdot 12 + 1 \cdot 5 + 1 \cdot 10 = 27$, which is strictly lower. In the new plan, the pair (x_1, y_1) is no longer in the support.*

We will first focus on the case where d is **strictly convex**.
We know from Section 3.3 that the Kantorovich problem admits a minimizer π^*. From the discussion above, we know that its support $\Gamma = \text{supp}(\pi^*)$ is a cyclically monotone set. In our setting with cost $c(x, y) = d(x - y)$, where d is strictly convex, the cyclical monotonicity condition can be shown to imply the following:

Claim: For every $(x_1, y_1), (x_2, y_2) \in \Gamma$, if $x_1 < x_2$, then $y_1 \leq y_2$. This follows from a simple convexity argument.

Proof. Assume $x_1 < x_2$ and $y_1 > y_2$, let $a = x_1 - y_1$, $b = x_2 - y_2$, $\delta = x_2 - x_1 > 0$, $\Delta = y_1 - y_2 > 0$ and define $t = \frac{\Delta}{\delta+\Delta} \in (0, 1)$. Then we can write the cross differences as convex combinations:

$$x_1 - y_2 = (1 - t)a + tb, \quad x_2 - y_1 = ta + (1 - t)b.$$

Using the strict convexity of d, we have:

$$\begin{aligned} d(x_1 - y_2) + d(x_2 - y_1) &= d((1 - t)a + tb) + d(ta + (1 - t)b) \\ &< (1 - t)d(a) + td(b) + td(a) + (1 - t)d(b) \\ &= d(a) + d(b) \\ &= d(x_1 - y_1) + d(x_2 - y_2), \end{aligned}$$

This contradicts the cyclical monotonicity condition for $(x_1, y_1), (x_2, y_2)$, and thus such a configuration cannot exist in the support of an optimal transport plan. Therefore, if $x_1 < x_2$, it must be that $y_1 \leq y_2$. □

We now define the sets:

$$A_{xy} = (-\infty, x] \times (y, +\infty), \quad B_{xy} = (x, +\infty) \times (-\infty, y].$$

Claim: For every $(x, y) \in \mathbb{R}^2$, $\pi^*(A_{xy})$ and $\pi^*(B_{xy})$ cannot be nonzero at the same time.

Proof. Suppose for contradiction that both $\pi^*(A_{xy}) > 0$ and $\pi^*(B_{xy}) > 0$. Then there exist pairs $(x_1, y_1) \in A_{xy} \cap \Gamma$ and $(x_2, y_2) \in B_{xy} \cap \Gamma$. By the definition of A_{xy} and B_{xy}, we must have:

$$x_1 \leq x < x_2, \quad y_1 > y \geq y_2.$$

Thus, $x_1 < x_2$ and $y_1 > y_2$, which contradicts the monotonicity of Γ established earlier. Therefore, both sets cannot carry mass simultaneously. □

Consider now the cumulative distribution function of π^*, denoted $H(x, y)$. From the previous claim, we know that $\pi^*(A_{xy}) \cdot \pi^*(B_{xy}) = 0$. Therefore, at least one of the sets

$$(-\infty, x] \times (-\infty, y] \cup A_{xy} = (-\infty, x] \times \mathbb{R}, \qquad (-\infty, x] \times (-\infty, y] \cup B_{xy} = \mathbb{R} \times (-\infty, y]$$

has the same π^*-measure as $(-\infty, x] \times (-\infty, y]$, while the other set may have larger measure. Hence,

$$H(x, y) = \pi^*((-\infty, x] \times (-\infty, y]) = \min\{\mu((-\infty, x]),\ \nu((-\infty, y])\} = \min\{F_\mu(x), F_\nu(y)\}.$$

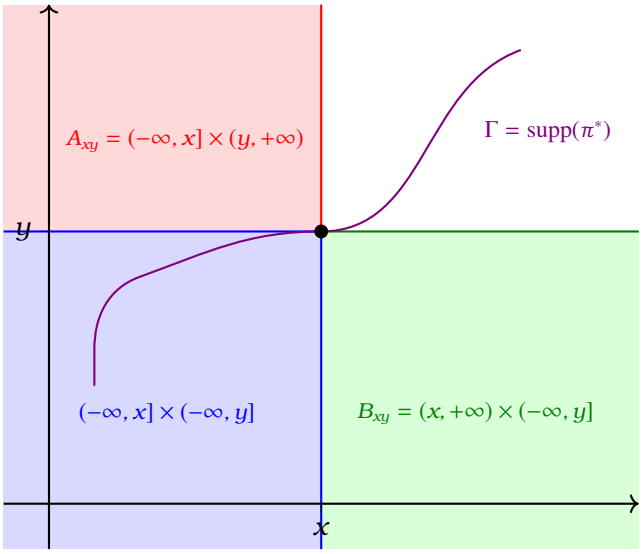


Figure 4.8. *Illustration of the sets involved in the proof that the optimal coupling π^* in one dimension satisfies $H(x, y) = \pi^*((-\infty, x] \times (-\infty, y]) = \min\{F_\mu(x), F_\nu(y)\}$. The rectangle $(-\infty, x] \times (-\infty, y]$ (blue) is extended by the sets A_{xy} (red), containing all points with $x' \le x$ and $y' > y$, and B_{xy} (green), containing all points with $x' > x$ and $y' \le y$. The support $\Gamma = \mathrm{supp}(\pi^*)$ (violet curve) avoids at least one of these sets.*

In the case where c is convex but not strictly convex, we can approximate it by a sequence of strictly convex functions $c_\varepsilon \to c$, whose corresponding optimal transport plans π_ε converge to an optimal plan for c. This justifies extending the result to the general convex case (see, e.g., [7, Theorem 2.9]).

Now, we will show that $\pi^* = (F^{-1}, G^{-1})_{\#}\lambda$, where λ denotes the Lebesgue measure on $[0, 1]$. By construction, the marginals of π^* are μ and ν. We just have to show that its cdf is $\min\{F(x), G(y)\}$.

$$\begin{aligned}
(F^{-1}, G^{-1})_{\#}\lambda((-\infty, x] \times (-\infty, y]) &= \lambda\Big((F^{-1}, G^{-1})^{-1}((-\infty, x] \times (-\infty, y])\Big) \\
&= \lambda\Big(\big\{t \in [0, 1] : F^{-1}(t) \le x \text{ and } G^{-1}(t) \le y\big\}\Big) \\
&= \lambda(\{t \in [0, 1] : t \le F(x) \text{ and } t \le G(y)\}) \\
&= \lambda([0, \min\{F(x), G(y)\}]) \\
&= \min\{F(x), G(y)\}.
\end{aligned}$$

Finally, we calculate the optimal cost. The last equality follows from the change of variables formula and standard approximation arguments (see Appendix A).

$$\begin{aligned}\int_{\mathbb{R}^2} d(x-y)\, d\pi^*(x,y) &= \int_{\mathbb{R}^2} d(x-y)\, d\left((F_\mu^{-1}, F_\nu^{-1})_\# \lambda\right)(x,y) \\ &= \int_0^1 d\left(F_\mu^{-1}(t) - F_\nu^{-1}(t)\right)\, dt.\end{aligned}$$

□

The above theorem is very useful in calculating Wasserstein distances between 1d measures.

Corollary 4.2. *Let μ, ν be probability measures on $\mathbb{R}$ with finite p-th moments, and let $F_\mu^{-1}, F_\nu^{-1} : [0,1] \to \mathbb{R}$ denote their quantile functions. Then*

$$W_p^p(\mu,\nu) = \int_0^1 \left|F_\mu^{-1}(t) - F_\nu^{-1}(t)\right|^p \, dt.$$

Proposition 4.2. *Let π^* be an optimal coupling for $W_p(\mu,\nu)$ in dimension one. Then* $\operatorname{supp}(\pi^*)$ *is monotone:*

$$x_1 < x_2 \quad \Rightarrow \quad y_1 \le y_2 \quad \text{for all } (x_1,y_1),(x_2,y_2) \in \operatorname{supp}(\pi^*).$$

The above statements are especially useful for numerical computations, since for empirical measures $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, sorted increasingly, the optimal coupling simply pairs x_i with y_i, and

$$W_p^p(\mu_n,\nu_n) = \frac{1}{n}\sum_{i=1}^n |x_i - y_i|^p.$$

Theorem 4.28 (Existence of Monge solution)**.** *Let μ be an atomless (assigns zero probability to points) probability measure on $\mathbb{R}$, and let ν be any probability measure on $\mathbb{R}$. Then the Monge problem*

$$\min_{T_\#\mu=\nu} \int_{\mathbb{R}} |x - T(x)|^p \, d\mu(x)$$

admits a unique solution given by the monotone transport map

$$T(x) = F_\nu^{-1}(F_\mu(x)).$$

When μ is atomless, the map $T(x) = F_\nu^{-1}(F_\mu(x))$ is well-defined and provides an optimal transport from μ to ν. The induced plan $(\mathrm{id}, T)_\#\mu$ coincides with the Kantorovich coupling $(F_\mu^{-1}, F_\nu^{-1})_\#\lambda$, since $\mu = (F_\mu^{-1})_\#\lambda$ and for almost every $t \in [0,1]$,

$$x = F_\mu^{-1}(t) \implies T(x) = F_\nu^{-1}(t).$$

This construction relies on the continuity and monotonicity of F_μ, which together ensure that F_μ^{-1} is a bijection, thus associating each t with a unique x. However, if μ has atoms, this bijection fails due to jumps in F_μ, resulting in ambiguities in defining T and potentially causing the Monge problem to have no solution.

Example 4.7. *Consider the probability measures on* $\mathbb{R}$*:*

$$\mu = \tfrac{1}{2}\delta_0 + \tfrac{1}{2}\mathrm{Uniform}(1,2), \quad \nu = \mathrm{Uniform}(0,1),$$

and a cost function of the form $c(x,y) = d(x-y)$*.*
Note that in this case, the optimal transport plan depends only on the source and target distributions μ *and* ν*.*
We first compute the cumulative distribution functions (CDFs) F_μ, F_ν *and their respective quantile functions* F_μ^{-1}, F_ν^{-1}*.*

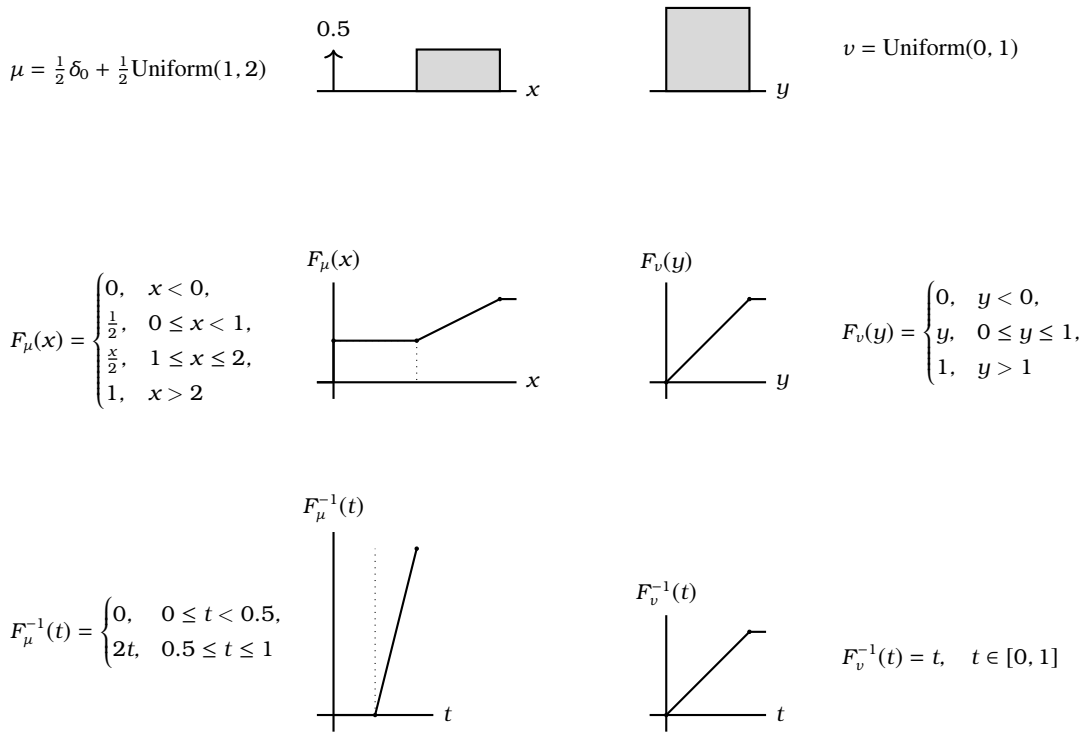


Figure 4.9. *Top: measures* μ *and* ν *with atom and uniform parts. Middle: their CDFs* F_μ *and* F_ν*. Bottom: their quantile functions* F_μ^{-1} *and* F_ν^{-1}*. Formulas for all are placed around the corresponding plots.*

We will now solve Kantorovich's problem. The optimal plan π^* is induced by the coupling

$$\pi^* = (F_\mu^{-1}, F_\nu^{-1})_{\#}\lambda,$$

and pairs points according to

$$(F_\mu^{-1}(t), F_\nu^{-1}(t)) = \begin{cases} (0, t), & 0 \le t < 0.5, \\ (2t, t), & 0.5 \le t \le 1. \end{cases}$$

We will now try to solve Monge's problem. We try to define the transport map

$$T(x) = F_\nu^{-1}(F_\mu(x)) = \begin{cases} F_\nu^{-1}(0.5) = 0.5, & x = 0, \\ F_\nu^{-1}\left(\frac{x}{2}\right) = \frac{x}{2}, & 1 \le x \le 2. \end{cases}$$

However, this is not a valid transport map since it sends the entire mass at $x = 0$ to a single point 0.5, contradicting the requirement that a measurable map push μ exactly to ν. Because μ has an atom of mass $1/2$ at 0, a transport plan must send this mass to a set of positive measure in ν. Since T assigns only one image to each point, it cannot spread the mass over the interval. Thus, no Monge map exists, even though the Kantorovich plan does.

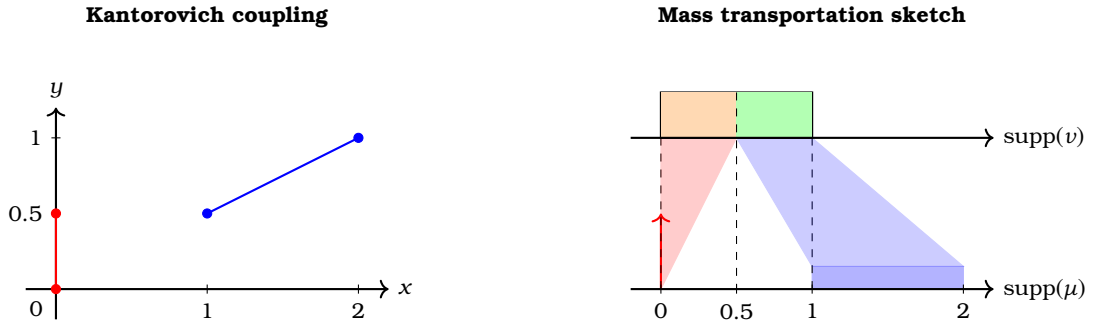


Figure 4.10. *(a) Kantorovich optimal coupling in the (x, y)-plane: red segment for the atom at $x = 0$ and blue segment for the uniform part $x \in [1, 2]$. (b) Mass transport sketch from μ (bottom line) to ν (top line). The red triangle shows transport of the atom mass $\frac{1}{2}\delta_0$ to $y \in [0, 0.5]$, and the blue polygon shows transport of the uniform mass to $y \in [0.5, 1]$. The top distribution is shaded to highlight these intervals.*

Wasserstein distance calculation:

Using the quantile functions, the p-Wasserstein distance satisfies

$$W_p^p(\mu, \nu) = \int_0^1 \left|F_\mu^{-1}(t) - F_\nu^{-1}(t)\right|^p dt = \int_0^{0.5} |0 - t|^p dt + \int_{0.5}^1 |2t - t|^p dt$$
$$= \int_0^1 t^p dt = \left[\frac{t^{p+1}}{p+1}\right]_0^1 = \frac{1}{p+1}.$$

Therefore,

$$W_p(\mu, \nu) = \left(\frac{1}{p+1}\right)^{\frac{1}{p}}.$$

In one dimension, we can calculate the W_1 distance using an even simpler formula.

Theorem 4.29 (1D Wasserstein–CDF Identity)**.** *[7, Proposition 2.17] Let μ and ν be probability measures on $\mathbb{R}$ with cumulative distribution functions F. Then:*

$$W_1(\mu, \nu) = \int_{\mathbb{R}} |F(x) - G(x)| \, dx.$$

Proof. Note that we just have to prove that

$$\int_0^1 \left|F^{-1}(t) - G^{-1}(t)\right| \, dt = \int_{\mathbb{R}} |F(x) - G(x)| \, dx.$$

Define the set $A \subset \mathbb{R}^2$ by

$$A := \{(x, t) \in \mathbb{R} \times [0, 1] : \min(F(x), G(x)) \leq t \leq \max(F(x), G(x))\}.$$

This region lies vertically between the two graphs $F(x)$ and $G(x)$, depending on which is larger.
Geometrically, one can reinterpret the set A as:

$$A = \{(x, t) \in \mathbb{R} \times [0, 1] : \min(F^{-1}(t), G^{-1}(t)) \leq x \leq \max(F^{-1}(t), G^{-1}(t))\}.$$

That is, for fixed t, the set of x for which t lies between $F(x)$ and $G(x)$ is exactly the interval from $\min(F^{-1}(t), G^{-1}(t))$ to $\max(F^{-1}(t), G^{-1}(t))$.
We compute the Lebesgue measure of this set $L(A)$ in two ways using Fubini's Theorem.

$$L(A) = \int_{\mathbb{R}} \int_{\min(F(x),G(x))}^{\max(F(x),G(x))} dt \, dx = \int_{\mathbb{R}} |F(x) - G(x)| \, dx.$$

$$L(A) = \int_0^1 \int_{\min(F^{-1}(t),G^{-1}(t))}^{\max(F^{-1}(t),G^{-1}(t))} dx \, dt = \int_0^1 |F^{-1}(t) - G^{-1}(t)| \, dt.$$

This proves the identity. □

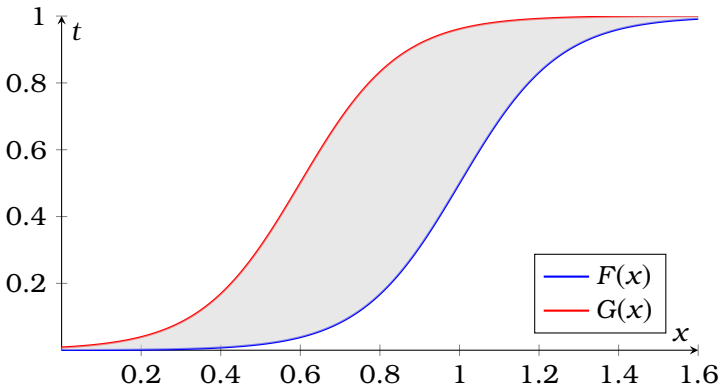


Figure 4.11. *Shaded area A between F and G equals $\int |F(x) - G(x)| \, dx$, which matches $\int |F^{-1}(t) - G^{-1}(t)| \, dt$ using Fubini's Theorem.*

4.7 Optimal plans and quadratic cost functions

Recall from the Kantorovich duality theorem (Chapter 3.4) that the optimal transport cost can be expressed as

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c\, d\pi = \sup_{\substack{\phi,\psi \\ \phi(x)+\psi(y)\le c(x,y)}} \left(\int \phi\, d\mu + \int \psi\, d\nu \right).$$

The functions ϕ and ψ that achieve the supremum are known as *Kantorovich potentials*. These potentials represent a pricing scheme assigning values to mass at points $x \in X$ and $y \in Y$, constrained by the transport cost function $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$.
To analyze their structure, the notion of *c-concavity* plays a central role.

Definition 4.46 (c-Concavity)**.** *Let $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ be a cost function. A function $\phi : X \to \mathbb{R} \cup \{-\infty\}$ is called c-concave if there exists a function $\psi : Y \to \mathbb{R} \cup \{-\infty\}$ such that*

$$\phi(x) = \inf_{y\in Y} [c(x, y) - \psi(y)].$$

Equivalently, ϕ can be expressed as the c-transform of ψ:

$$\phi = \psi^c, \quad \text{where} \quad \psi^c(x) := \inf_{y\in Y} [c(x, y) - \psi(y)].$$

Similarly, the c-transform of ϕ is given by

$$\phi^c(y) := \inf_{x\in X} [c(x, y) - \phi(x)].$$

This definition generalizes the classical notion of concavity and the Legendre-Fenchel transform from convex analysis, adapting it to the cost structure c.

Some important properties regarding c-concave functions are:

- For every function ϕ,
$$\phi^{cc} = (\phi^c)^c \geq \phi,$$
with equality if and only if ϕ is c-concave.

  This shows that the double c-transform acts as a c-concave envelope of ϕ - ϕ^{cc} is the smallest c-concave function that is larger than ϕ.

- Regardless of whether ϕ is c-concave, ϕ^c is always c-concave.

Kantorovich Potentials A fundamental result in optimal transport theory states that the *optimal potentials* ϕ, ψ solving the Kantorovich dual problem can be chosen to be c-concave, satisfying

$$\phi = \psi^c \quad \text{and} \quad \psi = \phi^c.$$

This characterization ensures that ϕ and ψ are tightly coupled through the cost function and encode the intrinsic geometry imposed by c.

The dual constraint

$$\phi(x) + \psi(y) \leq c(x, y), \quad \forall (x, y) \in X \times Y,$$

forces this c-transform relationship (by taking $\psi(y)$ to the other side and taking the infimum over y. Moreover, on the support of any optimal transport plan π, equality holds:

$$\phi(x) + \psi(y) = c(x, y), \quad \text{for } (x, y) \in \text{supp}(\pi).$$

The dual potentials are unique up to additive constants; that is, if (ϕ, ψ) is optimal, then so is $(\phi + a, \psi - a)$ for any $a \in \mathbb{R}$.

Example 4.8 (Quadratic cost)**.** *Let $X = Y = \mathbb{R}^d$ and consider the quadratic cost function*

$$c(x, y) = \frac{1}{2}\|x - y\|^2.$$

In this setting, the c-transform of a function $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$ is

$$\psi^c(x) = \inf_{y \in \mathbb{R}^d} \left\{ \frac{1}{2}\|x - y\|^2 - \psi(y) \right\}.$$

By expanding the squared norm, we have

$$\psi^c(x) = \frac{1}{2}\|x\|^2 - \sup_{y \in \mathbb{R}^d} \left\{ \langle x, y \rangle - \left(\psi(y) + \frac{1}{2}\|y\|^2 \right) \right\}.$$

Define the function

$$u(y) := \psi(y) + \frac{1}{2}\|y\|^2,$$

which is a proper function on $\mathbb{R}^d$. Then,

$$\psi^c(x) = \frac{1}{2}\|x\|^2 - u^*(x),$$

where u^* denotes the Legendre–Fenchel conjugate of u.
Since u^* is convex by definition, ψ^c is a semiconvex function (a quadratic term minus a convex function).
Thus, in the quadratic cost case, c-concave functions are precisely functions of the form

$$\phi(x) = \frac{1}{2}\|x\|^2 - u^*(x),$$

where u^* is convex. This characterization is fundamental for Brenier's theorem, which states that the optimal transport map for the quadratic cost is given by the gradient of a convex function.

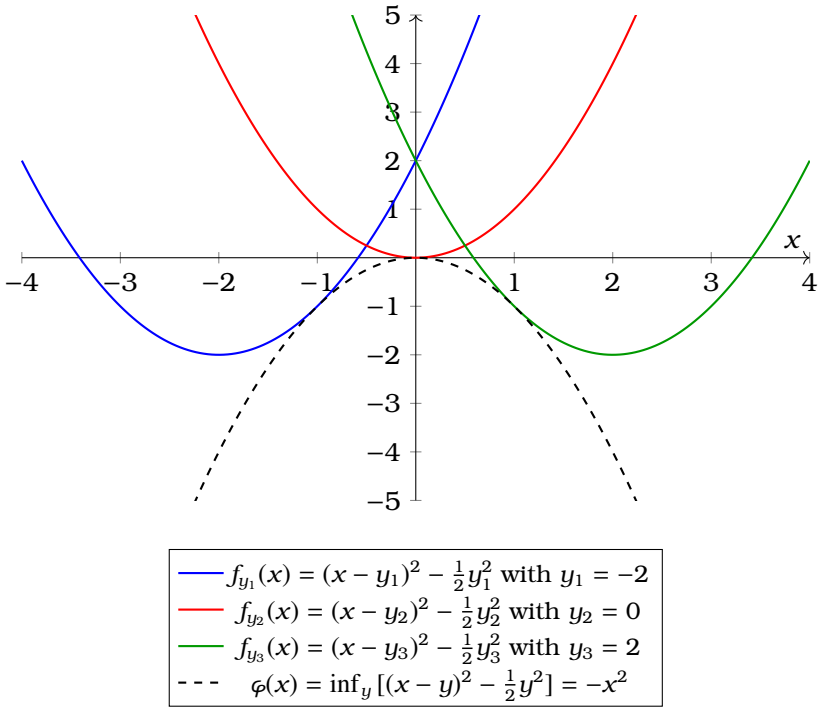


Figure 4.12. *Construction of the Kantorovich dual potential $\varphi(x)$ as the pointwise infimum over the functions $f_y(x) = c(x, y) - \psi(y)$, for $c(x, y) = (x-y)^2$ and $\psi(y) = \frac{1}{2}y^2$. This illustrates how φ can be expressed as $\frac{1}{2}x^2$ minus a convex function.*

c-superdifferential We define the *c-superdifferential* of a c-concave function ϕ as the set

$$\partial^c \phi := \{(x, y) \in X \times Y \mid \phi(x) + \phi^c(y) = c(x, y)\}.$$

This generalizes the classical subdifferential of a convex function and characterizes the set of optimal transport pairs.
For every $x \in X$ such that $\phi(x) > -\infty$, the c-superdifferential at x,

$$\partial^c \phi(x) := \{y \in Y \mid \phi(x) + \phi^c(y) = c(x, y)\},$$

is nonempty.

Plan Optimality The following result shows that a valid transport plan is optimal if and only if its support lies in the c-superdifferential of a c-concave function.

Theorem 4.30. *Let X, Y be Polish spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and let $c : X \times Y \to [0, +\infty]$ be a lower semicontinuous cost function. A transport plan $\pi \in \Pi(\mu, \nu)$ is optimal for the Kantorovich problem if and only if there exists a c-concave function $\phi : X \to \mathbb{R} \cup \{-\infty\}$ such that*

$$\operatorname{supp}(\pi) \subset \partial^c \phi.$$

Moreover, the functions ϕ and ϕ^c are Kantorovich potentials, i.e., they attain the supremum in the dual problem.

Proof. We divide the proof into two parts.

($\Longrightarrow$) **Optimality implies support on $\partial^c\varphi$:**

Assume $\pi \in \Pi(\mu, \nu)$ is an optimal transport plan for the cost c. By Kantorovich duality, there exists a pair of c-concave functions $\varphi : X \to \mathbb{R} \cup \{-\infty\}$ and $\varphi^c : Y \to \mathbb{R} \cup \{-\infty\}$ such that

$$\varphi(x) + \varphi^c(y) \leq c(x, y) \quad \text{for all } (x, y) \quad \text{(since they are c-concave)},$$

and

$$\int_X \varphi(x)\, d\mu(x) + \int_Y \varphi^c(y)\, d\nu(y) = \int_{X\times Y} c(x, y)\, d\pi(x, y) \iff$$

$$\int_{X\times Y} \{\varphi(x) + \varphi^c(y) - c(x, y)\}\, d\pi(x, y) = 0.$$

Since $\varphi(x) + \varphi^c(y) - c(x, y) \leq 0$ always, equality in the integral implies that

$$\varphi(x) + \varphi^c(y) = c(x, y) \quad \text{for } \pi\text{-almost every } (x, y).$$

By definition of the c-superdifferential, this means $(x, y) \in \partial^c\varphi$ for π-almost every (x, y):

$$\operatorname{supp}(\pi) \subset \partial^c\varphi.$$

($\Longleftarrow$) **Support on $\partial^c\varphi$ implies optimality:**

Suppose that $\pi \in \Pi(\mu, \nu)$ is such that $(x, y) \in \partial^c\varphi$ for π-almost every (x, y), where φ is c-concave. Then, by definition of the c-superdifferential,

$$\varphi(x) + \varphi^c(y) = c(x, y) \quad \text{for } \pi\text{-almost every } (x, y).$$

Therefore,

$$\int_{X\times Y} (\varphi(x) + \varphi^c(y))\, d\pi(x, y) = \int_{X\times Y} c(x, y)\, d\pi(x, y) \iff$$

$$\int_X \varphi(x)\, d\mu(x) + \int_Y \varphi^c(y)\, d\nu(y) = \int_{X\times Y} c(x, y)\, d\pi(x, y).$$

Thus, since (φ, φ^c) is an admissible pair in the dual formulation, the above equality shows that π is optimal and φ, φ^c are Kantorovich potentials. □

When $\partial^c\varphi(x)$ is a singleton for μ-almost every x, the optimal plan is induced by a measurable map $T : X \to Y$, with

$$T(x) = y \quad \text{such that} \quad (x, y) \in \partial^c\varphi.$$

The Quadratic Cost Case [8, Theorem 2.12]

We now focus on the case where the cost function is the squared Euclidean distance:

$$c(x, y) = \frac{1}{2}\|x - y\|^2,$$

with $x, y \in \mathbb{R}^d$. This setting enjoys remarkable geometric and analytic structure, and enables us to identify optimal transport plans as being supported on subdifferentials of convex functions. We state below two fundamental theorems: the *Knott–Smith optimality criterion*, which characterizes a plan's optimality similarly to the previous theorem, and *Brenier's theorem*, which shows that optimal transport maps in this setting are gradients of convex functions.

Theorem 4.31 (Knott–Smith Optimality Criterion)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be probability measures with finite second moments. A transport plan $\pi \in \Pi(\mu, \nu)$ is optimal for the Kantorovich problem with cost $c(x, y) = \frac{1}{2}\|x - y\|^2$ if and only if there exists a convex, lower semicontinuous function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that*

$$\operatorname{supp}(\pi) \subset \operatorname{Graph}(\partial\varphi),$$

that is, $y \in \partial\varphi(x)$ for π-almost every (x, y). Moreover, the pair (φ, φ^*) minimizes the dual Kantorovich problem.

Theorem 4.32 (Brenier's Theorem)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and suppose that μ is absolutely continuous with respect to Lebesgue measure. Then, there exists a unique optimal transport plan π, and it is induced by a transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ of the form*

$$T(x) = \nabla\varphi(x),$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a convex function. That is, the optimal plan is $\pi = (\mathrm{id}, T)_{\#}\mu$.

These results rely on the fact that for the quadratic cost, the c-concave potentials are closely related to convex conjugate functions; in fact, they are convex conjugates up to a quadratic term. More precisely, as we saw, any c-concave function φ can be written as

$$\varphi(x) = \frac{1}{2}\|x\|^2 - u^*(x),$$

where u^* is the convex conjugate of a proper function $u : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. The associated c-transform φ^c then satisfies

$$\varphi^c(y) = \frac{1}{2}\|y\|^2 - u(y),$$

so that φ and φ^c form a pair of convex conjugates (modulo the quadratic term). The Knott–Smith theorem follows by exploiting this structure, while Brenier's theorem uses the differentiability properties of convex functions and uniqueness of subgradients almost everywhere. *For detailed proofs of both theorems, see* [8, Villani, Theorem 2.12].

4.8 Wasserstein Spaces

Let (X, d) be a metric space. For any $p \geq 1$, we define the Wasserstein space $\mathcal{P}_p(X)$ as the set of Borel probability measures μ on X with finite p-th moment:

$$\mathcal{P}_p(X) := \left\{ \mu \in \mathcal{P}(X) : \int_X d(x_0, x)^p \, d\mu(x) < \infty \right\}$$

for some (and hence all) $x_0 \in X$. The p-Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(X)$ is defined as

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d(x, y)^p \, d\pi(x, y) \right)^{1/p}.$$

The Compact Case When X is compact, every Borel probability measure automatically has finite p-th moment, so $\mathcal{P}_p(X) = \mathcal{P}(X)$. In this case, convergence in Wasserstein distance is equivalent to weak convergence of measures:

Theorem 4.33. *Let X be compact. Then, for $\mu_n, \mu \in \mathcal{P}(X)$,*

$$W_p(\mu_n, \mu) \to 0 \quad \iff \quad \mu_n \rightharpoonup \mu.$$

Proof. It suffices to prove the result for W_1, since for any $p \geq 1$, the inequalities

$$W_1(\mu, \nu) \leq W_p(\mu, \nu) \leq \operatorname{diam}(X)^{\frac{p-1}{p}} \, W_1(\mu, \nu)^{1/p}$$

imply that convergence in W_1 is equivalent to convergence in W_p on compact X.

($\Rightarrow$) Suppose $W_1(\mu_n, \mu) \to 0$. From the dual formulation of W_1,

$$W_1(\mu_n, \mu) = \sup_{\varphi \in \mathrm{Lip}_1(X)} \left\{ \int \varphi \, d\mu_n - \int \varphi \, d\mu \right\},$$

so for every 1-Lipschitz function φ, we have

$$\int \varphi \, d\mu_n \to \int \varphi \, d\mu.$$

By linearity, this extends to all functions in $\mathrm{Lip}(X)$. By the Portmanteau theorem, this is enough to show that the convergence holds for all $f \in C(X)$. Hence, $\mu_n \rightharpoonup \mu$.

($\Leftarrow$) We omit the full proof of this implication but sketch the main idea. To show convergence in W_1, consider a subsequence along which $W_1(\mu_n, \mu)$ converges to its limsup. By the dual formulation of W_1, for each n there exists a 1-Lipschitz function φ_n such that

$$W_1(\mu_n, \mu) = \int \varphi_n \, d(\mu_n - \mu).$$

After some normalization and using the Arzelà–Ascoli theorem, there exists a uniformly convergent subsequence, $\varphi_{n_k} \to \varphi \in \mathrm{Lip}_1(X)$. Using the uniform convergence of φ_{n_k} and

the weak convergence $\mu_{n_k} \rightharpoonup \mu$, we conclude that

$$W_1(\mu_{n_k}, \mu) = \int \varphi_{n_k} \, d(\mu_{n_k} - \mu) \to \int \varphi \, d(\mu_{n_k} - \mu) \to 0,$$

Thus, $W_1(\mu_{n_k}, \mu) \to 0 \implies \limsup W_1(\mu_n, \mu) \leq 0$, thus $W_1(\mu_n, \mu) \to 0$. For a complete proof, see [8, Theorem 7.12]. □

The General Polish Case If X is a Polish space, then $\mathcal{P}_p(X)$ is also Polish under the Wasserstein metric. In this more general setting, convergence in W_p is stronger than weak convergence — it also requires convergence of p-th moments.

Theorem 4.34. *Let X be Polish, and $\mu_n, \mu \in \mathcal{P}_p(X)$. Then:*

$$W_p(\mu_n, \mu) \to 0 \quad \Longleftrightarrow \quad \begin{cases} \mu_n \rightharpoonup \mu, \\ \int_X d(x_0, x)^p \, d\mu_n(x) \to \int_X d(x_0, x)^p \, d\mu(x). \end{cases}$$

This tells us that W_p metrizes a stronger topology: the topology of weak convergence plus convergence of p-th moments.
A complete proof can be found in [8, Theorem 7.12].

Remark 3. *In the compact case, the function $x \mapsto d(x_0, x)^p$ is continuous and bounded, hence lies in $C_b(X)$. Therefore, weak convergence already implies convergence of p-th moments, and no additional condition is needed.*

The example below highlights how, in the non-compact case, weak convergence is not sufficient to guarantee convergence in Wasserstein distance. The key issue is that weak convergence alone does not control the tails of the distribution — that is, the p-th moments may fail to converge.

Example 4.9. *Let $X = \mathbb{R}$, and consider the sequence of measures*

$$\mu_n = \left(1 - \frac{1}{n}\right) \delta_0 + \frac{1}{n} \delta_n, \qquad \mu = \delta_0.$$

Then $\mu_n \rightharpoonup \mu$, since for any bounded continuous function f,

$$\int f \, d\mu_n = \left(1 - \frac{1}{n}\right) f(0) + \frac{1}{n} f(n) \to f(0) = \int f \, d\mu.$$

However, the Wasserstein distance does not converge to zero. For any $p \geq 1$, we compute:

$$W_p(\mu_n, \mu)^p = \frac{1}{n} |n - 0|^p = n^{p-1} \to \infty.$$

Hence $W_p(\mu_n, \mu) \not\to 0$, and in fact diverges. This is because the p-th moments do not converge:

$$\int |x|^p \, d\mu(x) = 0, \quad \textit{while}$$

$$\int |x|^p \, d\mu_n(x) = \left(1 - \frac{1}{n}\right) \cdot 0 + \frac{1}{n} \cdot |n|^p = n^{p-1} \to \infty.$$

Chapter 5

Empirical Measures and DNN Generalization Error

5.1 Introduction

In this section, we investigate how the Wasserstein distance between a reference probability measure and its empirical approximation depends on the number of samples used. Specifically, we examine the convergence behavior of $W_p(\mu_n, \mu)$, where μ_n is the empirical measure obtained from n i.i.d. samples drawn from a target distribution μ. This analysis is carried out for various dimensions d and Wasserstein exponents p, and later extended to include pushforward measures arising from neural network outputs.

The structure of this section is as follows. We first present theoretical results on the convergence rates of empirical measures in Wasserstein distance, which serve as a benchmark for interpreting our experimental findings. We then review classical and modern computational algorithms for computing Wasserstein distances, with a focus on both exact methods (such as network simplex and linear programming) and scalable approximate techniques (such as the Sinkhorn algorithm).

Next, we explore how the empirical prediction error of neural networks, in classical supervised regression tasks, is related to the Wasserstein distance between the true output distribution and the empirical distribution of the network's outputs. This provides a probabilistic view of model accuracy beyond traditional loss metrics.

Finally, we present a range of experiments illustrating these phenomena. We show how Wasserstein distances behave across varying sample sizes, dimensions, cost exponents p, and network architectures. Our results provide insight into both the computational and statistical aspects of Wasserstein metrics in practical settings.

5.2 One-Dimensional Empirical Measures and Order Statistics

In this section, we introduce the notion of empirical measures in one dimension and recall key results on their convergence. Our focus lies on their convergence in Wasserstein distance, both in the almost sure sense and in expectation. A key tool in this analysis is the classical Glivenko–Cantelli theorem, which provides a strong form of convergence for the empirical distribution function and underpins much of the theory surrounding empirical optimal transport.

5.2.1 Empirical Measures

Let $(X_i)_{i=1}^n$ be an i.i.d. sample of random variables taking values in a Polish space $\mathcal{X}$, with common distribution μ. The **empirical measure** μ_n associated with the sample is defined as

$$\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

where δ_x denotes the Dirac measure concentrated at the point $x \in \mathcal{X}$.
The empirical measure $\hat{\mu}_n$ is a **random** probability measure supported on the sample $\{X_1, \dots, X_n\}$. It provides a concrete realization of the law μ based on observed data and plays a fundamental role in sampling-based methods. In particular, empirical measures are essential in numerical approximations of probability distributions, as they are naturally suited for implementation on computers.

An empirical measure $\hat{\mu}_n$ almost surely weakly converges to μ:

$$\mathbb{P}(\hat{\mu}_n \xrightarrow{w} \mu) = 1.$$

Proof. Fix a countable dense subset $\{f_k\} \subset C_b(X)$ in the topology of uniform convergence. By the strong law of large numbers (SLLN), we have for each k,

$$\int f_k \, d\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} f_k(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}[f_k(X_1)] = \int f_k \, d\mu.$$

Since the set $\{f_k\}$ is countable, the convergence $\int f_k \, d\hat{\mu}_n \to \int f_k \, d\mu$ holds almost surely for all k.
Since convergence a.s. holds on a countable dense set $\{f_k\} \subset C_b(X)$, and the functionals $v \mapsto \int f_k \, dv$ determine the weak topology, we conclude that $\hat{\mu}_n \xrightarrow{w} \mu$ almost surely.
See Varadarajan's theorem in [3, Section 11.4.1]. □

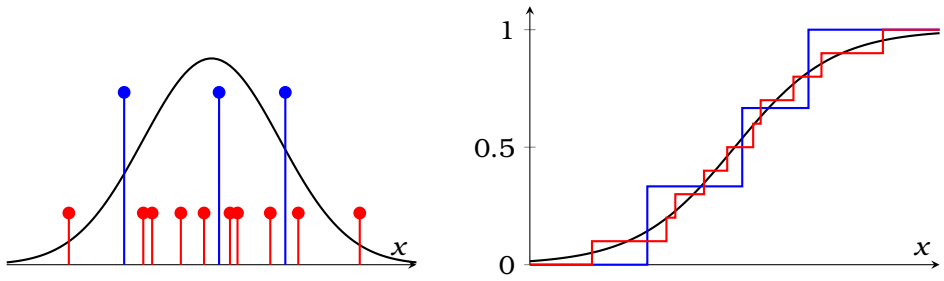


Figure 5.1. *Left: True measure μ (black curve) and empirical measures $\hat{\mu}_3$ (blue spikes) and $\hat{\mu}_{10}$ (red spikes). Right: Corresponding CDFs F_μ (black smooth curve), $F_{\hat{\mu}_3}$ (blue step function), and $F_{\hat{\mu}_{10}}$ (red step function).*

When the underlying measure is supported on the real line $\mathbb{R}$, we have stronger convergence guarantees for the empirical distribution function. The Glivenko–Cantelli theorem states that the empirical distribution function converges uniformly almost surely to the true distribution function.

Theorem 5.35 (Glivenko–Cantelli)**.** *[3, Theorem 11.4.2] Let $\{X_i\}_{i=1}^{\infty}$ be i.i.d. random variables with common distribution μ on $\mathbb{R}$ and distribution function F. Denote by F_n the empirical distribution function associated with the first n samples. Then, almost surely,*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0 \quad \text{as } n \to \infty.$$

A detailed proof of this result can be found in classical probability textbooks such as [3].

5.2.2 Wasserstein Convergence to Zero

Let (X, d) be a Polish metric space. We denote by $\mathcal{P}_p(X)$ the space of all probability measures on X with finite p-th moment:

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

Let μ a probability measure on $\mathcal{P}_p(X)$ and $\hat{\mu}_n$ be the empirical measure associated with i.i.d. samples $\{X_i\}_{i=1}^n$ drawn from μ.

Theorem 5.36. *The empirical measures $\hat{\mu}_n$ converge to μ in the p-Wasserstein metric almost surely:*

$$W_p(\hat{\mu}_n, \mu) \xrightarrow{a.s.} 0.$$

Proof. As highlighted before, for probability measures on a Polish space, convergence in W_p is equivalent to the combination of weak convergence and convergence of the p-th moments. We have already established weak convergence in the previous subsection, so it remains to show that

$$\int d(x_0, x)^p \, d\hat{\mu}_n(x) \to \int d(x_0, x)^p \, d\mu(x) \quad \text{almost surely.}$$

By the Strong Law of Large Numbers applied to the function $f(x) = d(x_0, x)^p$, we have

$$\int d(x_0, x)^p \, d\hat{\mu}_n(x) = \frac{1}{n} \sum_{i=1}^{n} d(x_0, X_i)^p \xrightarrow{a.s.} \mathbb{E}[d(x_0, X)^p] = \int d(x_0, x)^p \, d\mu(x).$$

Therefore, we conclude that

$$W_p(\hat{\mu}_n, \mu) \xrightarrow{a.s.} 0.$$

□

Now, we focus on the convergence of the average (expected) value of the Wasserstein distance W_p to zero. In the following, we restrict our attention to probability measures with finite p-moments in the space $\mathcal{P}_p(\mathbb{R})$.

Theorem 5.37. *[1, Theorem 2.14] Let $\mu \in \mathcal{P}_p(\mathbb{R})$, and $\hat{\mu}_n$ the associated empirical measure. Then,*

$$\mathbb{E}[W_p(\hat{\mu}_n, \mu)] \to 0.$$

Proof. The result follows from the special structure of the Wasserstein distance in one dimension. Recall that in that case, the Wasserstein distance admits the explicit formula:

$$W_p^p(\hat{\mu}_n, \mu) = \int_0^1 \left| F_n^{-1}(t) - F^{-1}(t) \right|^p \, dt,$$

where F_n^{-1} and F^{-1} denote the empirical and true quantile functions, respectively.
Using an analogue of the Glivenko–Cantelli theorem for quantile functions, one obtains that $F_n^{-1} \to F^{-1}$ almost surely, and after a non-trivial argument which is omitted, we get:

$$\mathbb{E}\left[W_p^p(\hat{\mu}_n, \mu) \right] \to 0.$$

Finally, applying Jensen's inequality gives:

$$\mathbb{E}[W_p(\hat{\mu}_n, \mu)] \leq \left(\mathbb{E}[W_p^p(\hat{\mu}_n, \mu)] \right)^{1/p} \to 0.$$

For a detailed proof of this result, see [1, Theorem 2.14]. □

Now that we have established the convergence of the expected Wasserstein distance to zero, we turn our attention to quantitative bounds. Specifically, we aim to understand how fast the empirical measure $\hat{\mu}_n$ converges to the true measure μ in Wasserstein distance as the number of samples n increases.

We will focus on deriving upper bounds for $\mathbb{E}[W_p(\hat{\mu}_n, \mu)]$ in terms of n, with particular emphasis on the case where μ is a probability measure on $\mathbb{R}$ with finite p-th moment.

5.2.3 Bounds for Expected Wasserstein Distance

We now present several non-asymptotic results on the convergence of empirical measures in Wasserstein distance, focusing on the one-dimensional case.

Let μ be a probability measure on $\mathbb{R}$ with finite 1-st moment, cumulative distribution function F and density function f.

We define the *median* of a random variable X as a value $m \in \mathbb{R}$ such that:

$$\mathbb{P}(X \leq m) \geq \frac{1}{2} \quad \text{and} \quad \mathbb{P}(X \geq m) \geq \frac{1}{2}.$$

This definition allows for non-uniqueness, especially when the distribution function has a flat or discontinuous region around the 0.5 level.

We first establish lower and upper bounds for $\mathbb{E}[W_1(\hat{\mu}_n, \mu)]$.

Theorem 5.38. *[1, Theorem 3.1] Let m be a median of X. There exists a constant $c > 0$, such that*

$$\mathbb{E}[W_1(\hat{\mu}_n, \mu)] \geq \frac{c}{\sqrt{n}} \mathbb{E}[|X - m|].$$

Theorem 5.39. *[1, Theorem 3.2] Define:*

$$J_1(\mu) := \int_{-\infty}^{\infty} \sqrt{F(x)(1 - F(x))}\, dx.$$

Then,

$$\mathbb{E}[W_1(\hat{\mu}_n, \mu)] \leq \frac{1}{\sqrt{n}} J_1(\mu).$$

From these bounds, we conclude that the expected Wasserstein-1 distance converges to zero at the rate of order $1/\sqrt{n}$, provided the underlying measure has finite $(2 + \epsilon)$-th moment for some $\epsilon > 0$.

Now, we will give an upper bound for $\mathbb{E}[W_p(\hat{\mu}_n, \mu)]$, assuming that $\mu \in \mathcal{P}_p(X)$ has finite $(2p + \epsilon)$-th moment for some $\epsilon > 0$.

Theorem 5.40. *[1, Theorem 5.3] Define:*

$$J_p(\mu) := \int_{-\infty}^{\infty} \frac{[F(x)(1 - F(x))]^{p/2}}{f(x)^{p-1}}\, dx.$$

Then,

$$\mathbb{E}[W_p^p(\hat{\mu}_n, \mu)] \leq \left(\frac{5p}{\sqrt{n+2}} \right)^p J_p(\mu).$$

$$\mathbb{E}[W_p(\hat{\mu}_n, \mu)] \leq \frac{5p}{\sqrt{n}}\, J_p^{1/p}(\mu).$$

Thus, we obtain a convergence rate of order $n^{-p/2}$ for $\mathbb{E}[W_p^p(\hat{\mu}_n, \mu)]$.

For detailed proofs of the above results, see [1, Theorems 3.1, 3.2, 5.3], where the authors develop these bounds in the one-dimensional setting with sharp constants.

These results are also highlighted in the paper *On the Rate of Convergence in Wasserstein Distance of the Empirical Measure* by Fournier and Guillin [4, Theorem 1].

Theorem 5.41. *Let μ be a probability measure on $\mathbb{R}^d$ with finite q-th moment for some $q > p \geq 1$, and let $\hat{\mu}_n$ denote the empirical measure based on n i.i.d. samples from μ. Define*

$$M_q^q(\mu) := \int_{\mathbb{R}^d} \|x\|^q \, d\mu(x).$$

Then there exists a constant $C = C(d, p, q)$ such that:

$$\mathbb{E}\left[W_p^p(\hat{\mu}_n, \mu)\right] \leq C \, M_q^{p/q}(\mu) \begin{cases} n^{-1/2} + n^{-(q-p)/q} & \text{if } p > d/2 \text{ and } q \neq 2p, \\ n^{-1/2} \log(1+n) + n^{-(q-p)/q} & \text{if } p = d/2 \text{ and } q \neq 2p, \\ n^{-p/d} + n^{-(q-p)/q} & \text{if } p \in (0, d/2) \text{ and } q \neq \frac{d}{d-p}. \end{cases}$$

Note that when μ admits sufficiently high moments, the second term $n^{-(q-p)/q}$ becomes negligible compared to the main rate.

For the special case where μ is the uniform distribution on $[0,1]^d$, sharper asymptotic rates for $\mathbb{E}[W_p^p(\hat{\mu}_N, \mu)]$ have been derived using methods from statistical physics. In particular, from the scaling relation (4) in [2], we have:

$$N^{p/d} \, \mathbb{E}[W_p^p(\hat{\mu}_N, \mu)] = \begin{cases} O(N^{p/2}) & \text{if } d = 1, \\ O((\log N)^{p/2}) & \text{if } d = 2, . \\ e_d^{(p)} + O(N^{-\gamma}) & \text{if } d > 2, \end{cases}$$

Therefore, the convergence rate of $\mathbb{E}[W_p^p(\hat{\mu}_N, \mu)]$ is approximately:

$$\mathbb{E}[W_p^p(\hat{\mu}_N, \mu)] \approx \begin{cases} O(N^{-p/2}) & \text{if } d = 1, \\ O(N^{-p/d}) & \text{if } d \geq 2. \end{cases}$$

We shall verify these theoretical convergence rated in the conducted experiments.

5.3 Computational Algorithms for Optimal Transport

Solving the optimal transport problem in its full generality is analytically and computationally challenging, especially for continuous probability measures supported on high-dimensional spaces. Since direct numerical optimization over such measures is infinite-dimensional and typically infeasible, practical approaches rely on discrete approximations derived from empirical samples, reducing the problem to a finite-dimensional linear program that is amenable to specialized algorithms.

Let $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$ be discrete probability measures on spaces $\mathcal{X}$ and $\mathcal{Y}$, supported on points $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ and $\{y_1, \ldots, y_m\} \subset \mathcal{Y}$. Given a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, the discrete optimal transport problem becomes:

$$\min_{\pi \in \mathbb{R}^{n \times m}} \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij}\, \pi_{ij} \quad \text{subject to} \quad \sum_{j=1}^{m} \pi_{ij} = a_i,\ \sum_{i=1}^{n} \pi_{ij} = b_j,\ \pi_{ij} \geq 0.$$

This is a classical linear programming problem with $n \times m$ variables and $n + m$ equality constraints. The solution π^* is the optimal transport plan, and the optimal value of the objective function defines the optimal transport cost between μ and ν.

Trying to solve the discrete optimal transport problem using generic linear programming algorithms, such as the simplex or interior point methods, quickly becomes computationally prohibitive for large-scale problems due to the high dimensionality and complexity of the constraints. To address this, more efficient and application-specific algorithms have been developed. These include specialized linear programming solvers, such as the network simplex algorithm, and modern iterative approximations like the Sinkhorn algorithm. In the special case of balanced discrete measures with equal cardinality and uniform weights, the optimal transport problem reduces to a linear assignment problem, for which the Hungarian algorithm provides an exact polynomial-time solution. Moreover, when the measures are supported on the real line, highly efficient exact solvers based on sorting and cumulative distribution functions become available, offering near-linear complexity. In the following subsections, we examine these algorithms in more detail.

5.3.1 Sinkhorn Algorithm

The Sinkhorn algorithm provides an efficient way to approximate the discrete optimal transport problem by introducing an entropic regularization term. This regularization not only ensures the uniqueness and smoothness of the solution, but also leads to a scalable iterative algorithm.

Let $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$ be two discrete probability measures supported on finite sets $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$. Let $C \in \mathbb{R}^{n \times m}$ be the cost matrix with entries $C_{ij} = c(x_i, y_j)$. The entropically regularized optimal transport problem is defined as

$$\begin{aligned} \min_{\pi \in \mathbb{R}_+^{n \times m}} \quad & \langle C, \pi \rangle - \varepsilon H(\pi) \\ \text{subject to} \quad & \pi \mathbf{1}_m = a, \\ & \pi^\top \mathbf{1}_n = b, \end{aligned}$$

where $H(\pi) := -\sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \log \pi_{ij}$ denotes the entropy of the transport plan π, $\mathbf{1}_n$ denotes the n-dimensional vector of ones, and $\varepsilon > 0$ is the regularization parameter. The entropic regularization encourages smoother, more diffuse transport plans by promoting higher entropy. Since the entropy function is strictly concave, the regularized problem is strictly convex, ensuring a unique optimal solution.

Lagrangian formulation: Introducing dual variables $\kappa \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m$ to enforce the constraints, the Lagrangian reads

$$\begin{aligned} \mathcal{L}(\pi, \kappa, \lambda) &= \langle C, \pi \rangle - \varepsilon H(\pi) + \kappa^\top (a - \pi \mathbf{1}_m) + \lambda^\top (b - \pi^\top \mathbf{1}_n) \\ &= \sum_{i,j} \left(C_{ij} \pi_{ij} + \varepsilon \pi_{ij} \log \pi_{ij} - \kappa_i \pi_{ij} - \lambda_j \pi_{ij} \right) + \kappa^\top a + \lambda^\top b \end{aligned}$$

Differentiating with respect to π_{ij} and setting to zero gives

$$\begin{aligned} & C_{ij} + \varepsilon(1 + \log \pi_{ij}) - \kappa_i - \lambda_j = 0 \\ \iff & \log \pi_{ij} = \frac{\kappa_i + \lambda_j - C_{ij}}{\varepsilon} - 1 \\ \iff & \pi_{ij} = e^{-1} \exp\left(\frac{\kappa_i}{\varepsilon}\right) \exp\left(\frac{\lambda_j}{\varepsilon}\right) \exp\left(-\frac{C_{ij}}{\varepsilon}\right). \end{aligned}$$

Define the Gibbs kernel matrix

$$K := \exp\left(-\frac{C}{\varepsilon}\right) \in \mathbb{R}_+^{n \times m}$$

where the exponential is taken elementwise, and positive scaling vectors

$$u := e^{-1} \exp\left(\frac{\kappa}{\varepsilon}\right) \in \mathbb{R}_+^n, \quad v := \exp\left(\frac{\lambda}{\varepsilon}\right) \in \mathbb{R}_+^m.$$

Then the optimal transport plan can be expressed compactly as

$$\pi^\varepsilon = \operatorname{diag}(u)\, K \operatorname{diag}(v).$$

The constraints become

$$\pi^\varepsilon \mathbb{1}_m = \operatorname{diag}(u) K v = a, \quad \text{and} \quad (\pi^\varepsilon)^\top \mathbb{1}_n = \operatorname{diag}(v) K^\top u = b.$$

Elementwise, these read

$$u_i (K v)_i = a_i, \quad \forall i, \quad \text{and} \quad v_j (K^\top u)_j = b_j, \quad \forall j.$$

Sinkhorn iterations: The marginal constraints can be enforced by iteratively updating u and v as

$$u^{(k+1)} = \frac{a}{Kv^{(k)}}, \quad v^{(k+1)} = \frac{b}{K^\top u^{(k+1)}},$$

where the divisions are elementwise. Starting from an initial guess (e.g., $u^{(0)} = \mathbf{1}_n$), the iterations alternate updates of u and v until convergence.

The resulting $\pi^\varepsilon = \mathrm{diag}(u)K\mathrm{diag}(v)$ approximates the entropic regularized optimal transport plan. The algorithm converges under mild conditions, such as strictly positive entries in K, which is ensured when ε is not too small and the cost matrix C does not have extremely large values. The quality of approximation depends on ε: smaller values give a plan closer to the true optimal transport plan but can cause numerical instability.
Sinkhorn's simplicity, differentiability, and ease of parallelization make it highly suitable for large-scale optimal transport computations. However, in our setting, we require high-accuracy solutions, which correspond to very small values of the regularization parameter ε. In this regime, the Sinkhorn algorithm becomes slow to converge and suffers from numerical instability due to the vanishing entries in the Gibbs kernel, making it less practical for our purposes.

5.3.2 Network Simplex Algorithm

The network simplex algorithm is a specialized variant of the classical simplex method, optimized for solving minimum-cost flow problems on graphs. In discrete optimal transport, the problem can be viewed as finding a feasible flow $\pi \in \mathbb{R}_+^{n\times m}$ on a bipartite graph $G = (V, E)$, where the vertex set consists of n supply nodes and m demand nodes, and edges connect every supply node i to every demand node j with cost C_{ij}. The marginal constraints translate directly to flow conservation: each supply node i must send exactly a_i units of mass, and each demand node j must receive exactly b_j units.
This structured flow problem enables the network simplex algorithm to efficiently navigate feasible solutions by exploiting the sparsity and topology of the bipartite network.

In this network setting, feasible solutions correspond to flows that satisfy the mass conservation constraints. A *basic feasible solution* to the linear program has at most $n+m-1$ non-zero entries in π, corresponding to a spanning tree of the bipartite graph. These tree structures play a central role in the network simplex algorithm: the algorithm maintains a current tree solution and explores adjacent trees by pivoting along cycles. The tree is embedded in the residual graph induced by the current flow, and each pivot step temporarily adds a non-tree edge, forming a unique cycle along which flow can be redistributed to reduce cost.

At each iteration, the algorithm attempts to insert a non-tree edge (an edge with zero flow) into the current tree, which induces a unique cycle. It then adjusts the flows along the cycle in a direction that reduces the total transport cost. This adjustment respects the capacity constraints (non-negativity of π) and results in a new feasible basic solution corresponding to a different tree.

To decide which non-tree edge to enter (i.e., which pivot to perform), the algorithm computes *reduced costs* using the current dual variables (a_i) for supply nodes and (β_j) for demand nodes. For an edge (i,j), the reduced cost is

$$\bar{C}_{ij} = C_{ij} - a_i - \beta_j.$$

Only edges with negative reduced cost are considered for pivoting, ensuring descent in objective value.

The algorithm maintains dual feasibility throughout, and terminates when no negative reduced-cost edge remains, at which point the current tree solution is optimal. The dual variables a, β can be interpreted as potentials (or prices) associated with the supply and demand nodes, and the reduced cost expresses the gain or loss from rerouting flow.

The network simplex algorithm is exact and efficient for moderate-sized instances. It leverages the sparse and combinatorial structure of transport problems, and its performance can be further improved using advanced data structures (e.g., dynamic trees) and warm starts (starting from a previous solution with slightly perturbed data). However, its worst-case complexity is exponential, and it is not used for very large-scale or noisy problems.

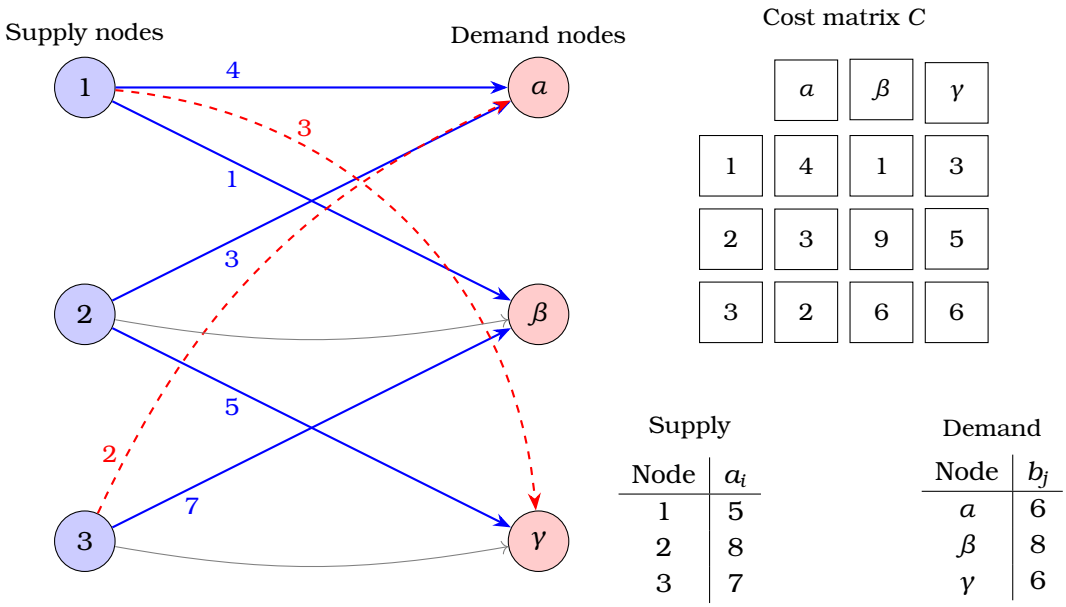


Figure 5.2. *Bipartite graph for discrete optimal transport with cost matrix C. Solid blue edges correspond to the current basic feasible solution with associated flow values. Red dashed edges are candidate edges with strictly lower cost, advantageous to enter the basis and potentially reduce total transport cost. This figure illustrates the original graph and feasible flow; the residual graph used internally by the network simplex algorithm is not shown here.*

All computations of Wasserstein distances between multidimensional empirical distributions were carried out using the network simplex algorithm.

5.3.3 Algorithm for One-Dimensional Optimal Transport

In the one-dimensional case, we leverage a specialized approach that exploits the structure of the problem to efficiently and exactly solve the empirical optimal transport (OT) problem.

In this subsection, we describe the algorithm implemented by functions such as `emp_1d_pot`, which computes the Wasserstein distance and optimal transport plan between weighted empirical measures.

We consider the cost function $c(x, y) = d(x - y)$, where $d : \mathbb{R} \to \mathbb{R}_+$ is a convex function. From the theoretical results in the previous section, the optimal transport plan π^* between two probability measures μ and ν supported on $\mathbb{R}$ is induced by the coupling

$$\pi^* = (F_\mu^{-1}, F_\nu^{-1})_{\#}\lambda,$$

where F_μ^{-1} and F_ν^{-1} are the generalized inverse (quantile) functions of μ and ν, and λ is the Lebesgue measure on $[0, 1]$.

A key property of this coupling is its *monotonicity*: it matches the mass in increasing order without crossings. This structure enables a highly efficient algorithm to compute the optimal transport cost between discrete empirical measures.

We first sort the support locations $\{x_i\}$ and $\{y_j\}$ in increasing order (if they are not already sorted). Then, starting from the smallest indices, we iteratively transfer mass $a = \min(w_i, w_j)$ from x_i to y_j, accumulate cost $a \cdot d(x_i - y_j)$, and subtract a from the available masses w_i and w_j. When the mass at either point is exhausted, we advance the corresponding index and proceed.

ALGORITHM 5.1: *Greedy algorithm for one-dimensional optimal transport*

1. Sort the point locations $x[1..n]$ and $y[1..m]$ in increasing order along with their weights $u[1..n]$, $v[1..m]$.
2. Initialize indices $i \leftarrow 1, j \leftarrow 1$.
3. Initialize remaining masses $w_i \leftarrow u[i]$, $w_j \leftarrow v[j]$.
4. Initialize `cost` $\leftarrow 0$.
5. Initialize empty list $\Pi \leftarrow []$ to store transport plan tuples (x_i, y_j, a).
6. **While** $i \leq n$ and $j \leq m$:
   a) Set $a \leftarrow \min(w_i, w_j)$.
   b) Update `cost` $\leftarrow$ `cost` $+ a \cdot d(x[i] - y[j])$ and append $(x[i], y[j], a)$ to Π.
   c) Update $w_i \leftarrow w_i - a$, $w_j \leftarrow w_j - a$.
   d) **If** $w_i = 0$, increment $i \leftarrow i + 1$. If $i \leq n$, set $w_i \leftarrow u[i]$.
   e) **If** $w_j = 0$, increment $j \leftarrow j + 1$. If $j \leq m$, set $w_j \leftarrow v[j]$.
7. **Return** `cost` and transport plan Π.

This procedure constructs the optimal coupling π^* exactly and computes the Wasserstein distance with complexity $O(n \log n + m \log m)$ due to sorting, or $O(n + m)$ if the inputs are already sorted. Its simplicity and computational efficiency make it particularly well-suited for empirical applications involving cost functions such as $|x - y|^p$.

5.4 Generalization Error in Neural Networks via Optimal Transport

This chapter explores a novel framework for analyzing the generalization error of Deep Neural Networks (DNNs), drawing upon the paper "A new approach to generalisation error of machine learning algorithms: Estimates and convergence" by Loulakis and Makridakis (2023) [6]. Unlike many traditional approaches that rely on specific structural assumptions about the neural network architectures, this work provides a more general methodology by leveraging the powerful tools of optimal transport theory and Wasserstein distances.

5.4.1 Deep Neural Networks and Error

Deep Neural Networks have become ubiquitous in various machine learning tasks, including function approximation and regression. In this chapter, we focus on the regression problem of learning an unknown target function $f : D \to \mathbb{R}$ from data, where $D \subset \mathbb{R}^d$.

Assume that we have a class of Deep Neural Networks $\mathcal{N}$, which induces a corresponding function space $V_N \subset L^p(D)$. Our goal is to find a function $v^* \in V_N$ that minimizes the true L_p risk given by the integral

$$v^* = \arg\min_{v \in V_N} \mathcal{E}, \quad \text{where} \quad \mathcal{E}(v) = \int_D |f(x) - v(x)|^p \, dx,$$

assuming such a minimizer exists.

However, since the target function f is unknown except at a finite set of sampled points $X = (X_1, \ldots, X_N)$, we do not have direct access to the integral above. Instead, we minimize the empirical risk

$$u_X(\omega) = \arg\min_{v \in V_N} E_{N,\omega}(v), \quad \text{where} \quad \mathcal{E}_{N,\omega}(v) := \frac{1}{N} \sum_{i=1}^{N} |f(X_i(\omega) - v(X_i(\omega))|^p,$$

which approximates the integral by averaging the L_p error over the training data. The resulting function $u_X(\omega)$, known as the probabilistic deep neural network interpolant, is the data-driven approximation to f constructed from the available samples.

Note that the true risk $\mathcal{E}(v)$ can be written more compactly using a probability measure μ on D, representing the (unknown) distribution of the input data. Specifically,

$$\mathcal{E}(v) = \int_D |f(x) - v(x)|^p \, d\mu(x).$$

This expresses the expected L_p error with respect to the true distribution of inputs.

Similarly, for a given sample $X = (X_1, \ldots, X_N)$, we define the empirical measure

$$\mu_{N,X}(\omega) := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i(\omega)},$$

where δ_x denotes the Dirac delta at point x. Using this, the empirical risk becomes

$$\mathcal{E}_{N,\omega}(v) = \int_D |f(x) - v(x)|^p \, d\mu_{N,X}(\omega)(x),$$

which approximates the true energy functional by integrating over the empirical distribution of the training data.

In practice, our goal is to bound the total error between the unknown target function f and the learned interpolant $u_X(\omega)$, measured in the L_p norm:

$$\mathbb{E}\left\{\|f - u_X(\cdot)\|_p^p\right\}.$$

This error can be naturally decomposed into two distinct contributions:

- The **approximation error**, which stems from the expressive limitations of the function class V_N. Even with full knowledge of f, we may not be able to represent it exactly within V_N. This component is quantified by the best possible approximation of f within V_N:
$$\inf_{v \in V_N} \|f - v\|_p = \|f - v^*\|.$$
- The **generalization error**, which arises from the fact that we only have access to finitely many training samples $X_1, \ldots, X_N$, and therefore must rely on the empirical minimizer $u_X(\omega)$ instead of the ideal minimizer v^* of the full risk. This term captures the discrepancy introduced by replacing the true data distribution μ with the empirical measure $\mu_{N,X}(\omega)$.

5.4.2 Bounding the Generalization Error

A key challenge in analyzing the generalization error is that the interpolant $u_X(\omega)$ is itself a *random function*, determined by the random sample X. As such, the map $\omega \mapsto u_X(\omega)$ introduces additional complexity, making standard concentration techniques difficult to apply directly. This motivates the development of alternative analytical tools that can handle this dependence more robustly. We will describe a framework introduced by Loulakis and Makridakis in [6], which leverages tools from optimal transport theory to

provide meaningful bounds on the generalization error in terms of Wasserstein distances between the empirical and true data distributions.

Theorem 5.42 (Estimate of the Generalization Error [6, Theorem 4.1])**.** *Consider for each $\omega \in \Omega$, the empirical risk minimization problem $\min_{v \in V_N} \mathcal{E}_{N,\omega}(v)$, and denote its solution by $u_X(\omega)$. Assume that f is Lipschitz, and let us denote the Lipschitz constant of $u_X(\omega) - f$ by $L_X(\omega)$. Then, for each $\omega \in \Omega$, and for any $\varphi \in V_N$, the following bound holds:*

$$\|u_X(\omega) - f\|_p \leq \left(\frac{1}{N} \sum_{i=1}^{N} |\varphi(X_i(\omega)) - f(X_i(\omega))|^p \right)^{1/p} + L_X(\omega) W_p(\mu, \mu_{N,X}(\omega)),$$

where W_p denotes the p-Wasserstein distance between the true data distribution μ and the empirical measure $\mu_{N,X}(\omega)$.

Furthermore, if $L_X \leq L_N$ almost surely for some constant $L_N > 0$, then we have the following expectation bound:

$$\begin{aligned} \mathbb{E}\left[\|u_X(\cdot) - f\|_p\right] &\leq \inf_{\varphi \in V_N} \|\varphi - f\|_p + L_N \, \mathbb{E}\left[W_p(\mu, \mu_{N,X})\right] \\ &= \|v^* - f\|_p + L_N \, \mathbb{E}\left[W_p(\mu, \mu_{N,X})\right] \end{aligned}$$

Proof. Notice that

$$\begin{aligned} \left(\int_D |u_X(\omega)(y) - f(y)|^p \, d\mu(y) \right)^{1/p} &= W_p\left((u_X(\omega) - f)_{\#}\mu, \delta_0\right) \\ &\leq W_p\left((u_X(\omega) - f)_{\#}\mu_{N,X(\omega)}, \delta_0\right) + W_p\left((u_X(\omega) - f)_{\#}\mu, (u_X(\omega) - f)_{\#}\mu_{N,X(\omega)}\right) \\ &= \left(\frac{1}{N} \sum_{i=1}^{N} |u_X(\omega)(X_i(\omega)) - f(X_i(\omega))|^p \right)^{1/p} + W_p\left((u_X(\omega) - f)_{\#}\mu, (u_X(\omega) - f)_{\#}\mu_{N,X(\omega)}\right) \\ &\leq \left(\frac{1}{N} \sum_{i=1}^{N} |\varphi(X_i(\omega)) - f(X_i(\omega))|^p \right)^{1/p} + W_p\left((u_X(\omega) - f)_{\#}\mu, (u_X(\omega) - f)_{\#}\mu_{N,X(\omega)}\right). \end{aligned}$$

The first inequality uses the triangle inequality, and the last uses the fact that $u_X(\omega)$ minimizes the empirical risk.

To bound the second term, note that for any $g \in \mathrm{Lip}_1$, the map $x \mapsto u_X(\omega)(x) - f(x)$ is $L_X(\omega)$-Lipschitz, so

$$\begin{aligned} W_p\left((u_X(\omega) - f)_{\#}\mu,\ (u_X(\omega) - f)_{\#}\mu_{N,X(\omega)}\right) &= \left(\inf_{\pi \in \Pi(\mu, \mu_{N,X(\omega)})} \int_{D \times D} |(u_X(\omega)(x) - f(x)) - (u_X(\omega)(y) - f(y))|^p \ d\pi(x,y) \right)^{1/p} \\ &\leq \left(\inf_{\pi \in \Pi(\mu, \mu_{N,X(\omega)})} \int_{D \times D} L_X(\omega)^p |x - y|^p \, d\pi(x,y) \right)^{1/p} \\ &= L_X(\omega) W_p(\mu, \mu_{N,X(\omega)}). \end{aligned}$$

This completes the proof of the pointwise estimate.

To obtain the expectation bound, observe that for any fixed $\varphi \in V_N$,

$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}|\varphi(X_i(\omega)) - f(X_i(\omega))|^p\right)\right] = \int_D |\varphi(x) - f(x)|^p \, d\mu(x),$$

since $X_i(\omega)$ are i.i.d. with law μ. Using this together with the assumption $L_X(\omega) \leq L_N$, we obtain

$$\mathbb{E}\left[\|u_X(\cdot) - f\|_p\right] \leq \|\varphi - f\|_p + L_N \, \mathbb{E}\left[W_p(\mu, \mu_{N,X})\right].$$

Taking the infimum over $\varphi \in V_N$ completes the proof. □

This result provides a decomposition of the expected error $\mathbb{E}\left[\|u_X(\cdot) - f\|_p\right]$ into two interpretable contributions. The first term, $\|f - v^*\|_p$, corresponds to the *approximation error*, which reflects the expressiveness of the function class V_N, while the second term, $L_N \, \mathbb{E}\left[W_p(\mu, \mu_{N,X})\right]$, captures the *generalization error*.
Together, these two terms isolate the sources of error in data-driven learning with neural networks: one intrinsic to the model class, and the other induced by sampling.

5.4.3 Motivation for Wasserstein Distances of Pushforward Measures

The motivation behind our experimental study of pushforward Wasserstein distances stems directly from the structure of the generalization error estimate. As highlighted in Remark 4 of [6], the Monte Carlo integration error is ultimately governed by the quantity

$$W_p\left((u_X(\omega) - f)_{\#}\mu, \ (u_X(\omega) - f)_{\#}\mu_{N,X(\omega)}\right),$$

rather than by the more commonly used upper bound involving the Lipschitz constant,

$$L_X(\omega) W_p(\mu, \mu_{N,X(\omega)}).$$

While the latter offers a general bound, it can be extremely loose in practice due to the difficulty of tightly estimating the Lipschitz constant $L_X(\omega)$ of the error function $u_X(\omega) - f$. Instead, directly computing the Wasserstein distance between the pushforward measures provides a more refined and realistic measure of the empirical error. This motivates our decision to investigate and estimate this quantity directly in the experiments. We expect these pushforward Wasserstein distances to be significantly smaller and to yield tighter and more informative control over the generalization error.

5.5 Experiments

5.5.1 Estimating the Empirical Wasserstein Distance

We conduct numerical experiments to investigate the convergence behavior of empirical Wasserstein distances in high dimensions. Specifically, we compute the Wasserstein distance between an empirical measure $\hat{\mu}_n$ and the underlying true measure μ for varying

sample sizes n. Our goal is to observe how this distance decreases as n increases and to compare the empirical convergence rates with theoretical predictions.

Setup We consider two types of underlying distributions:

- Uniform: $\mu = \mathcal{U}([0,1]^d)$
- Gaussian: $\mu = \mathcal{N}(0, I_d)$

For each distribution, we fix a reference measure (the "true" distribution), represented by 1000 i.i.d. samples. We verified that increasing the number of reference samples beyond 1000 does not significantly affect the computed Wasserstein distance, justifying this choice.

To compute empirical Wasserstein distances, we generate N i.i.d. samples from the same distribution to construct an empirical measure $\hat{\mu}_N$, with N ranging up to 800. For each N, we repeat the experiment 100 times and average the results to estimate the expected empirical Wasserstein distance. This Monte Carlo average converges well with this number of iterations.

The experiments are performed for dimensions $d = 1, \ldots, 10$ and Wasserstein exponents $p = 1, \ldots, 10$. For the uniform case, we generate the "true" distribution using a Halton sequence for better uniformity:

$$\text{Halton}(i) = \left[\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \ldots\right]$$

Empirical samples are drawn using standard pseudo-random generators. The Wasserstein distance is computed using the `emd2` function from the Python `POT` library, which uses the network simplex algorithm and provides an exact solution to the discrete optimal transport problem.

Results and Analysis Below, we present log-log plots of the averaged Wasserstein distances $\mathbb{E}[W_p(\hat{\mu}_N, \mu)]$ versus sample size N, for various dimensions d and exponents p. We include both the uniform and Gaussian cases.

The theoretical rate for convergence of empirical measures under Wasserstein distance is $O(N^{-p/d})$ for the uniform distribution. For the Gaussian case, the decay is typically slightly slower due to the unbounded support.

To estimate the empirical convergence rates, we perform a linear regression on the $\log_{10}(N)$ vs. $\log_{10}(W_p)$ data and extract the slope.

We first present the uniform case results. They confirm the expected convergence behavior of empirical Wasserstein distances. The empirical slopes closely match the theoretical rate of $-p/d$. For the Gaussian distribution, the convergence is slower, due to the distribution's unbounded support and heavier tails.

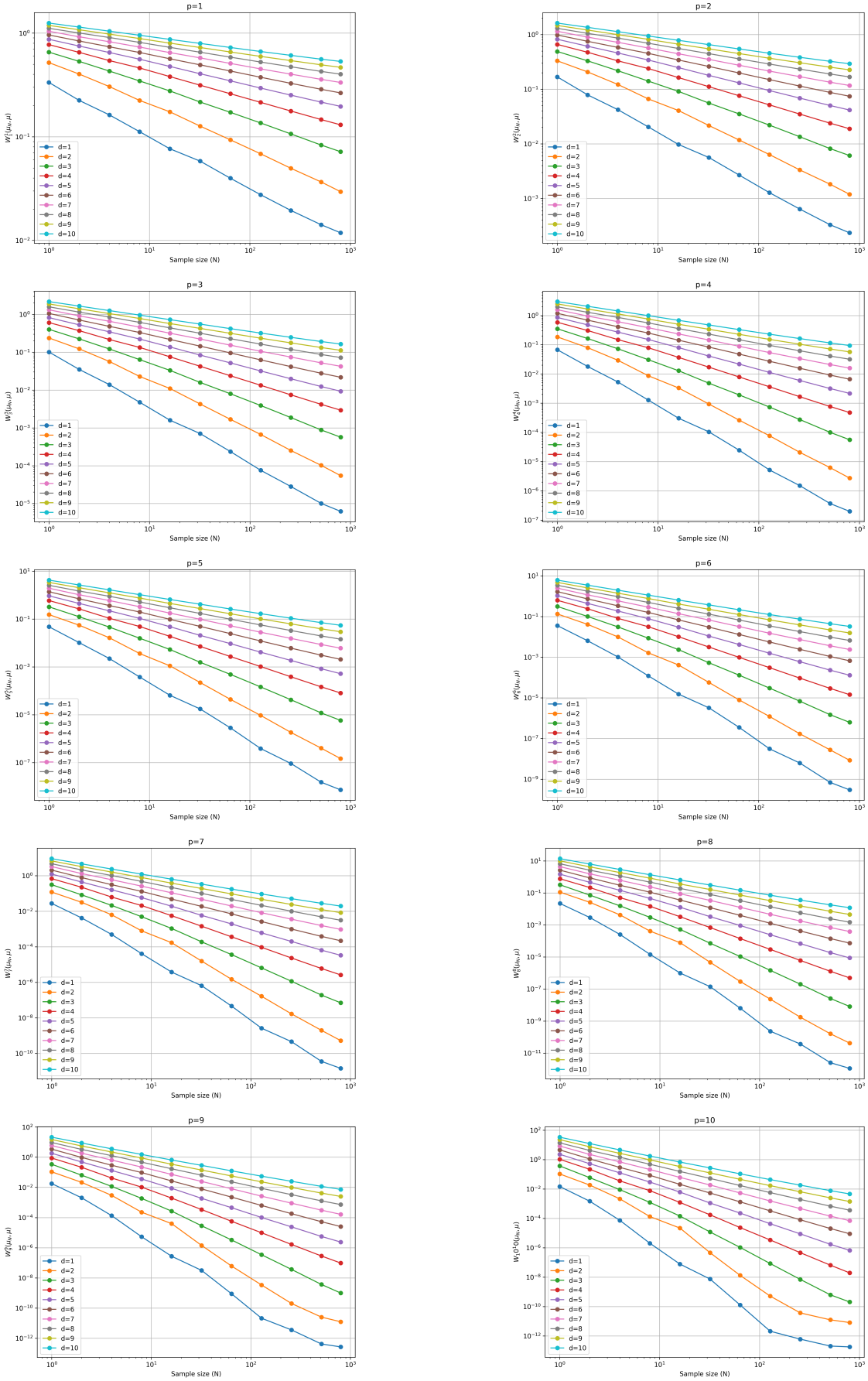


Figure 5.3. *Log-log plots of $W_p^p(\mu_n, \mu)$ vs. sample size n for $\mu \sim \mathcal{U}([0,1]^d)$ and various (p, d).*

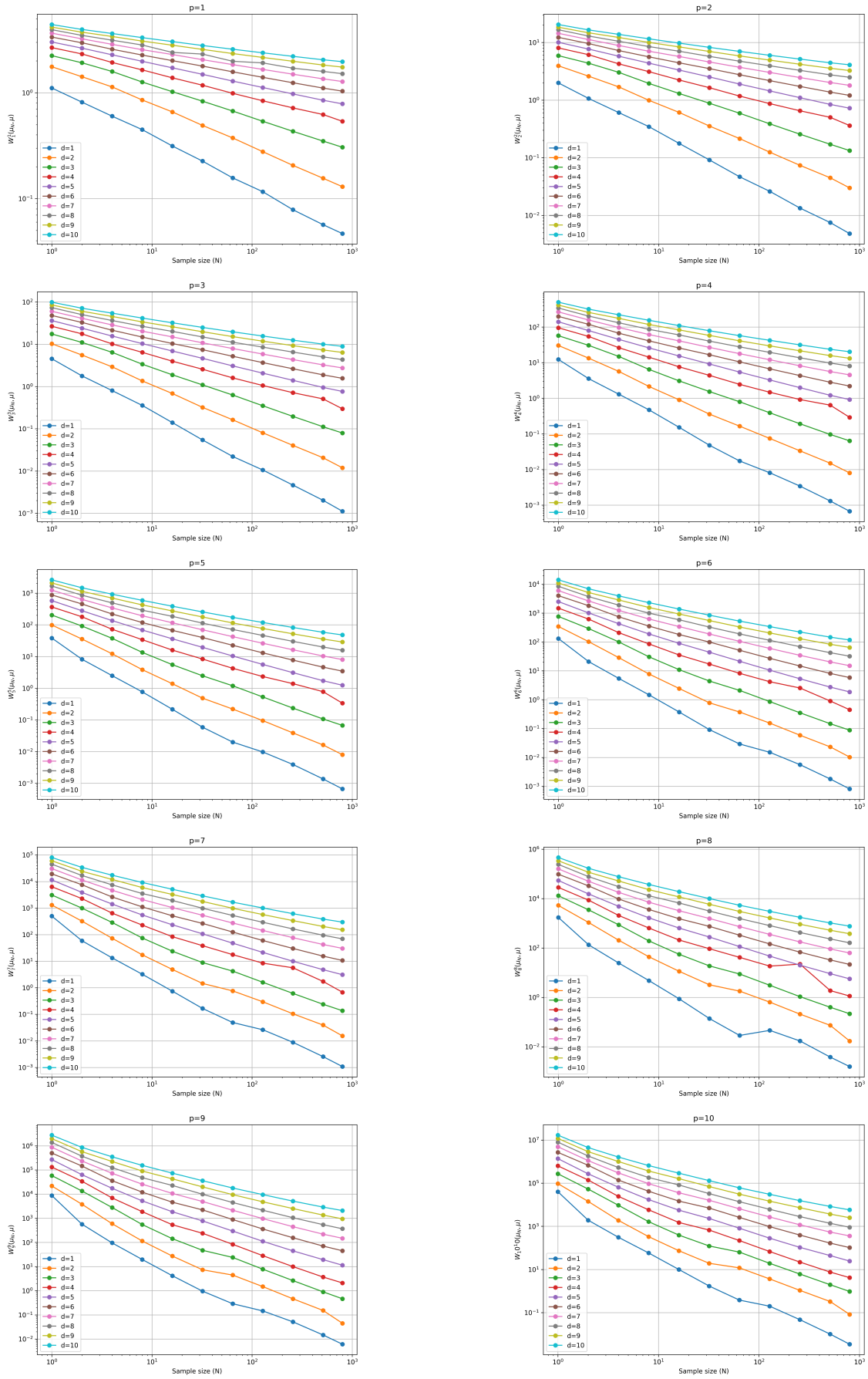


Figure 5.4. *Log-log plots of $W_p^p(\mu_n, \mu)$ vs. sample size n for $\mu \sim \mathcal{N}(0, I_d)$ and various (p, d).*

Table 5.1. *Estimated convergence rates (slopes) of* $\log W_p^p(\mu_n, \mu)$ *vs.* $\log n$ *for* $\mu \sim \mathcal{U}([0,1]^d)$ *and different* d *and* p *values.*

Dimension d	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$	$p=7$	$p=8$	$p=9$	$p=10$
1	-0.50	-0.99	-1.47	-1.93	-2.39	-2.85	-3.29	-3.69	-3.96	-4.09
2	-0.43	-0.85	-1.27	-1.70	-2.12	-2.54	-2.97	-3.38	-3.67	-3.83
3	-0.33	-0.66	-0.99	-1.32	-1.66	-1.99	-2.33	-2.66	-2.99	-3.28
4	-0.27	-0.54	-0.81	-1.08	-1.35	-1.62	-1.89	-2.16	-2.43	-2.69
5	-0.22	-0.45	-0.68	-0.91	-1.13	-1.36	-1.59	-1.82	-2.05	-2.28
6	-0.19	-0.39	-0.58	-0.78	-0.98	-1.18	-1.38	-1.58	-1.77	-1.97
7	-0.17	-0.34	-0.52	-0.69	-0.87	-1.04	-1.22	-1.40	-1.58	-1.75
8	-0.15	-0.31	-0.47	-0.62	-0.78	-0.94	-1.10	-1.26	-1.42	-1.58
9	-0.14	-0.28	-0.42	-0.57	-0.71	-0.86	-1.01	-1.15	-1.30	-1.45
10	-0.13	-0.26	-0.39	-0.52	-0.65	-0.79	-0.92	-1.06	-1.19	-1.33

Table 5.2. *Estimated convergence rates (slopes) of* $\log W_p^p(\mu_n, \mu)$ *vs.* $\log n$ *for* $\mu \sim \mathcal{N}(0, I_d)$ *and different* d *and* p *values.*

Dimension d	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$	$p=7$	$p=8$	$p=9$	$p=10$
1	-0.48	-0.91	-1.24	-1.45	-1.60	-1.73	-1.86	-1.95	-1.98	-2.28
2	-0.40	-0.74	-1.01	-1.23	-1.39	-1.52	-1.63	-1.77	-1.84	-1.95
3	-0.30	-0.58	-0.82	-1.03	-1.20	-1.35	-1.49	-1.62	-1.71	-1.82
4	-0.24	-0.46	-0.66	-0.83	-1.00	-1.17	-1.29	-1.42	-1.62	-1.75
5	-0.20	-0.40	-0.58	-0.75	-0.92	-1.07	-1.21	-1.35	-1.47	-1.59
6	-0.18	-0.35	-0.51	-0.67	-0.82	-0.97	-1.11	-1.23	-1.36	-1.48
7	-0.16	-0.31	-0.46	-0.60	-0.74	-0.88	-1.01	-1.14	-1.26	-1.38
8	-0.14	-0.28	-0.42	-0.55	-0.69	-0.82	-0.94	-1.06	-1.19	-1.32
9	-0.13	-0.26	-0.38	-0.51	-0.63	-0.75	-0.88	-0.99	-1.11	-1.23
10	-0.12	-0.24	-0.36	-0.47	-0.59	-0.71	-0.82	-0.93	-1.05	-1.16

5.5.2 Estimating Generalization via Pushforward Wasserstein Distances

Setup To empirically investigate the convergence of the generalization error through the lens of pushforward Wasserstein distances, we consider a function $f(x)$ and a data distribution μ. We train a fully connected neural network with residual connections to learn this function using SGD with a learning rate of 0.01, up to 2000 epochs or until the empirical loss drops below 0.02. The loss we use is the empirical L_p distance:

$$L(\hat{f}, f) = \left(\frac{1}{n} \sum_{i=1}^{n} |\hat{f}(x_i) - f(x_i)|^p \right)^{1/p}.$$

Pushforward Error Distributions After training, we compute the pointwise error $e(x) = |\hat{f}(x) - f(x)|$. This induces two probability measures:

- The empirical pushforward measure $e_{\#}\mu_n$, based on the training sample $\{x_i\}_{i=1}^n$.
- The true pushforward $e_{\#}\mu$, which can be approximated using a large empirical measure.

Since both measures live on $\mathbb{R}$, we can compute $W_p^p(e_{\#}\mu_n, e_{\#}\mu)$ exactly in 1D using `ot.emd2_1d` from the POT library.

We conduct 3 different experiments:

- We consider $f(X) = \|x\|_2^2$ and $\mu \sim \mathcal{U}([0,1]^d)$, using a Halton distribution of $d \cdot 2^{15}$ points to simulate it. We vary the empirical/training sample size $n \in \{32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384\} \cdot d$, and repeat the experiment for dimensions $d \in \{1, \dots, 10\}$ and Wasserstein exponents $p \in \{1, 2, 5\}$. Each setting is averaged over 1000 Monte Carlo trials to reduce variance.
- We consider $f(X) = \|x\|_2^2$ and $\mu \sim \mathcal{N}(0, I_d)$, using an empirical distribution of $d \cdot 2^{20}$ points to simulate it. We vary the empirical/training sample size $n \in \{4000, 8000, 16000, 32000, 64000, 128000\} \cdot d$, and repeat the experiment for dimensions $d \in \{1, \dots, 4\}$ and Wasserstein exponents $p \in \{1, 2, 5\}$. Each setting is averaged over 200 Monte Carlo trials.
- We consider $f(X) = \|x\|_2^{-1/8}$ and $\mu \sim \mathcal{U}([0,1]^d)$, using the Halton distribution of 2^{16} points to simulate it. We vary the training sample size $n \in \{250, 500, 1000, 2000, 4000, 8000, 16000\}$, and repeat the experiment for pairs of dimensions and Wasserstein exponents $(d, p) \in \{(1,1), (1,2), (1,3), (1,5), (2,1), (2,2), (2,7), (2,9), (2,12), (2,15)\}$. Each setting is averaged over 200 Monte Carlo trials.

Results The average values of $W_p^p(e_{\#}\mu_n, e_{\#}\mu)$ are plotted against n in log-log scale, and linear regression is applied to estimate convergence rates. The results are presented on the following graphs and tables.

Table 5.3. *Empirical convergence slopes for* $f(x) = \|x\|_2^2$, $\mu \sim \mathcal{U}([0,1]^d)$.

p	$d=1$	$d=2$	$d=3$	$d=4$
1	-0.53	-0.52	-0.56	-0.58
2	-0.52	-0.52	-0.54	-0.55
5	-0.52	-0.54	-0.46	-0.39

Since the pushforward measures are supported on the real line, classical results predict that W_p^p between an n-point empirical measure and the corresponding true distribution should scale as $n^{-p/2}$. However, our experiments do not always exhibit this rate. A

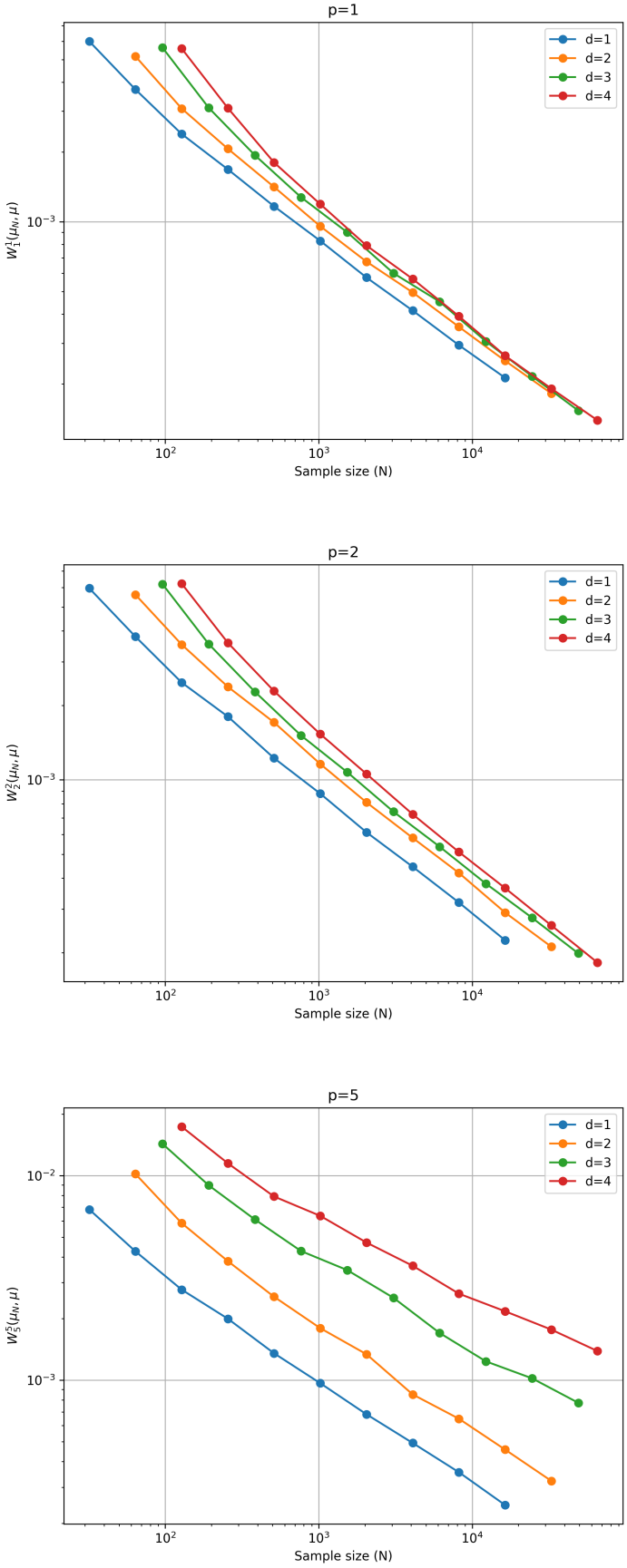


Figure 5.5. *Log-log plots of* $W_p^p(e_\# \mu_n, e_\# \mu)$ *vs.* n *for* $f(x) = \|x\|_2^2$, $\mu \sim \mathcal{U}([0,1]^d)$, *for various* p, d *values.*

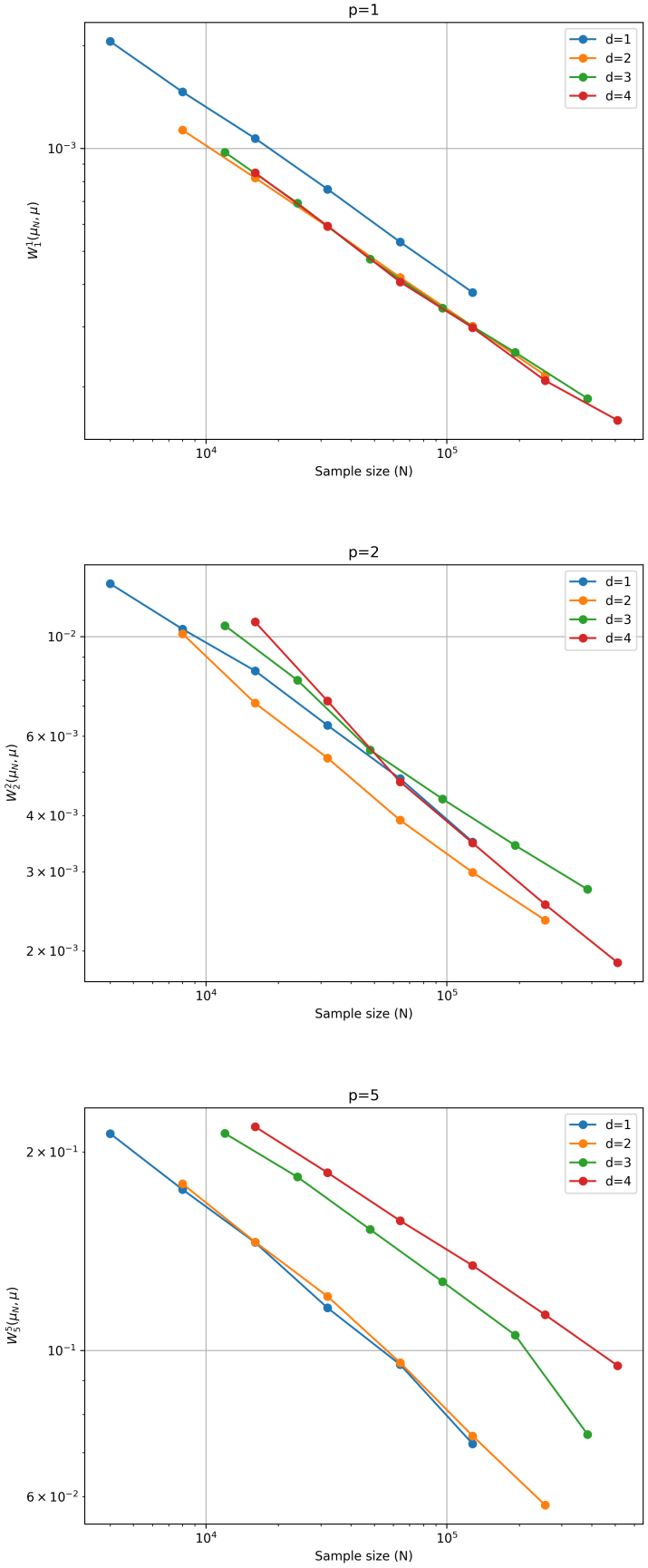


Figure 5.6. *Log-log plots of* $W_p^p(e_\#\mu_n, e_\#\mu)$ *vs.* n *for* $f(x) = \|x\|_2^2$, $\mu \sim \mathcal{N}(0, I_d)$, *for various* p, d *values.*

Table 5.4. *Empirical convergence slopes for* $f(x) = \|x\|_2^2$, $\mu \sim \mathcal{N}(0, I_d)$.

p	$d = 1$	$d = 2$	$d = 3$	$d = 4$
1	-0.49	-0.48	-0.48	-0.49
2	-0.38	-0.42	-0.39	-0.50
5	-0.31	-0.32	-0.29	-0.24

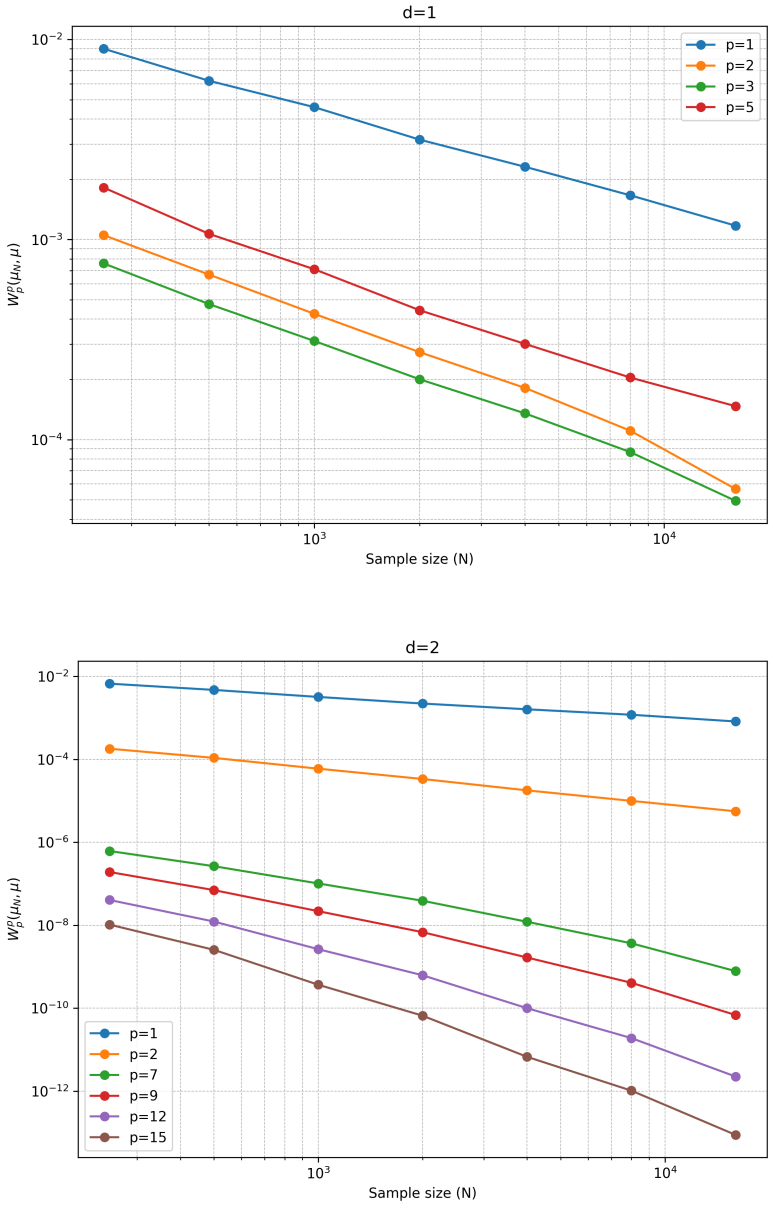


Figure 5.7. *Log-log plots for* $f(x) = \|x\|_2^{-1/8}$, $\mu \sim \mathcal{U}([0,1]^d)$, *for selected* p *values and* $d = 1, 2$.

Table 5.5. *Empirical convergence slopes for* $f(x) = \|x\|_2^{-1/8}$, $\mu \sim \mathcal{U}([0,1]^d)$, *and selected* d, p *values.*

$d = 1$	$p = 1$	$p = 2$	$p = 3$	$p = 5$
	-0.49	-0.68	-0.64	-0.60

$d = 2$	$p = 1$	$p = 2$	$p = 7$	$p = 9$	$p = 12$	$p = 15$
	-0.50	-0.85	-1.58	-1.89	-2.35	-2.82

plausible explanation is that the functions defining the pushforward measures are not fixed in advance; instead, they are themselves learned from the empirical data. Consequently, the error distributions—and hence the pushforward measures—are coupled with the empirical measure, breaking the independence assumptions typically required for classical convergence results. This dependence likely alters the effective complexity of the pushforward measure and slows down the convergence compared to the theoretical $n^{-p/2}$ rate.

These findings validate the refined view proposed in [6], namely that the pushforward Wasserstein distance $W_p(e_\# \mu_n, e_\# \mu)$ provides a sharper and more stable measure of generalization than rough bounds involving global Lipschitz constants. Additionally, the empirical rates confirm that generalization improves predictably with increasing sample size, and that optimal transport tools are not only theoretically sound but also practically computable and insightful.

Chapter 6

Conclusion

This thesis rigorously explored the applications of optimal transport theory, specifically Wasserstein distances, to advance our understanding of generalization in deep learning, complementing theoretical foundations with empirical validation.

Our experimental investigations first confirmed the expected convergence rates of empirical Wasserstein distances, demonstrating their reliable behavior in high-dimensional settings across various distributions. More significantly, we applied the optimal transport framework to analyze the generalization error of deep neural networks, building upon recent theoretical developments by Loulakis and Makridakis. Our results underscore that pushforward Wasserstein distances offer a remarkably sharper and more stable measure of generalization error compared to traditional Lipschitz-based bounds. A key empirical observation was that the observed convergence rates for these pushforward measures frequently deviated from classical theoretical predictions for one-dimensional empirical measures. We posited that this discrepancy arises from the intrinsic coupling between the learned error function, which defines the pushforward measure, and the finite empirical data from which it is derived.

This work not only validates the practical utility of optimal transport in quantifying complex aspects of deep learning but also opens compelling avenues for future research. Foremost among these is the critical need to theoretically formulate the precise convergence rates for pushforward Wasserstein distances when the underlying map is dynamically learned by a neural network. Such a theoretical breakthrough would explain the empirically observed behaviors and provide deeper insights into the interplay between model complexity, training data characteristics, and generalization performance. Further work could also explore the impact of specific neural network architectures and sampling methodologies on these novel convergence rates, ultimately contributing to the design of more theoretically grounded and robust deep learning algorithms.

Appendix A: The Standard Machine — From Sets to Functions

In this thesis, many measure-theoretic properties are first stated for sets and then equivalently reformulated for integrals against functions. The following 4-step procedure, which we call the *standard machine*, formalizes this transition from set-level conditions to functional formulations.

The Standard Machine Suppose we want to verify a property $\mathcal{P}$ involving a measure μ that is initially given as a statement about measurable sets, e.g., for all measurable sets $A \subseteq \mathcal{X}$,

$$\mathcal{P}(A) \text{ holds.}$$

To extend $\mathcal{P}$ to hold for integrals against measurable functions, we proceed as follows:

1. **Indicator functions:** By assumption, $\mathcal{P}$ holds for characteristic functions $\mathbb{1}_A$ of measurable sets $A \subseteq \mathcal{X}$. That is,

$$\mathcal{P}(\mathbb{1}_A) \quad \text{holds for all measurable } A.$$

2. **Simple functions:** Since any simple function $s : \mathcal{X} \to \mathbb{R}$ can be written as a finite linear combination of indicator functions,

$$s = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}, \quad \alpha_i \in \mathbb{R}, A_i \subseteq \mathcal{X} \text{ measurable},$$

   and $\mathcal{P}$ is linear (or otherwise compatible with finite sums and scalar multiplication), it follows that $\mathcal{P}(s)$ holds.

3. **Positive measurable functions:** Every positive measurable function $f : \mathcal{X} \to [0, \infty]$ can be approximated pointwise by an increasing sequence of simple functions $\{s_n\}$ with

$$s_n \uparrow f \quad \text{pointwise as } n \to \infty.$$

   By monotone convergence arguments (e.g., Monotone Convergence Theorem), $\mathcal{P}$ extends to all positive measurable functions f.

4. **General measurable functions:** For any measurable function $g : \mathcal{X} \to \mathbb{R}$, write

$$g = g^+ - g^-,$$

where $g^+ = \max(g, 0)$ and $g^- = \max(-g, 0)$ are positive measurable functions. Since $\mathcal{P}$ holds for g^+ and g^-, linearity (or suitable compatibility) ensures $\mathcal{P}(g)$ holds for all measurable functions g.

Depending on the context, the class of test functions g may be restricted to bounded continuous functions $C_b(\mathcal{X})$ or other subclasses, but the approximation idea remains essentially the same.

Example 6.10. *Pushforward Measures*
Recall the pushforward condition:

$$T_{\#}\mu(B) = \mu(T^{-1}(B)) \quad \forall B \subseteq \mathcal{Y} \text{ measurable.}$$

Using the standard machine, this implies that:

$$\int_{\mathcal{Y}} f(y)\, d(T_{\#}\mu)(y) = \int_{\mathcal{X}} f(T(x))\, d\mu(x) \quad \forall f \in C_b(\mathcal{Y}).$$

This equivalence follows by applying the four-step standard machine to extend from characteristic functions $\mathbf{1}_B$ to all bounded continuous functions f.

Bibliography

[1] Sergey Bobkov and Michel Ledoux. *One-Dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances*, volume 261 of *Memoirs of the American Mathematical Society*. American Mathematical Society, Providence, RI, 2019.

[2] Sergio Caracciolo, Claudio Lucibello, Giorgio Parisi, and Giorgio Sicuro. Scaling hypothesis for the euclidean bipartite matching problem. *Physical Review E*, 90(1):012118, 2014.

[3] R. M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2 edition, 2002.

[4] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[5] Wilfrid Gangbo and Robert J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[6] Michail Loulakis and Charalambos G. Makridakis. A new approach to generalisation error of machine learning algorithms: Estimates and convergence. *arXiv preprint arXiv:2306.13784*, 2023.

[7] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, 2015.

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