Université Côte d'Azur
École doctorale Sciences Fondamentales et Appliquées

Habilitation à Diriger des Recherches

Transport optimal pour l'apprentissage statistique

Rémi FLAMARY

Laboratoire Lagrange, UMR CNRS 7293
Observatoire de la Côte d'Azur

Soutenue le 29 Novembre 2019

Devant le jury composé de :

Rapporteurs :	Massimiliano PONTIL	-	Professeur University College London
	Florence D'ALCHÉ BUC	-	Professeur Télécom Paristech
	Gabriel PEYRÉ	-	Directeur de Recherche, CNRS, ENS Paris
Examinateurs :	Francis BACH	-	Directeur de Recherche, INRIA, ENS Paris
	Stéphane CANU	-	Professeur INSA Rouen
	Cédric RICHARD	-	Professeur Université Côte d'Azur
	André FERRARI	-	Professeur Université Côte d'Azur

Contents

List of Figures

Curriculum vitae

FLAMARY Rémi

Associate Professor, 34 years old
Université Côte d'Azur
Parc Valrose, 06000 NICE
Phone : +33 4 92 07 63 80
E-mail : remi.flamary@unice.fr
Web : https://remi.flamary.com

Education

2008 - 2011 **PhD in Computer Science**, *Université de Rouen, Laboratoire LITIS EA 4108.*
PhD advisor: **Alain Rakotomamonjy**.
PhD Title :
Apprentissage statistique pour le signal: applications aux interfaces cerveau-machine
(Machine learning for signal processing: Application to Brain Computer Interfaces)
Defense on December 6, 2011 at UFR de Sciences et Techniques de l'Université de Rouen in front of the jury :

- *Reviewers*: **Jalal Fadili** (PR ENSICAEN) and **Michele Sebag** (DR LRI CNRS).
- *Examiners*: **Liva Ralaivola** (PR Univ. Aix-Marseille), **Jocelyn Chanussot** (PR INP Grenoble) et **Stéphane Canu** (PR INSA Rouen).
- *Advisor*: **Alain Rakotomamonjy** (PR Univ. Rouen) .

2007 - 2008 **Master's degree in Electrical Engineering** *INSA de Lyon.*
Option: **Images and Systems**.

2003 - 2008 **Engineer's degree in Electrical Engineering**, *INSA de Lyon.*
Option: **Signal and image processing**.

Teaching experience

2012 - 2019 **Associate Professor**, *Departement of Electronics, Université Côte d'Azur*
Coordinator of the Bachelor (License 3) in Electronics for 2017-2019.

- **Signals and Systems**, Licence 3 Elec., 18h, until 2017 (5 years).
  Co-coordinator for signal part.
  Creator of courses and practical sessions.
- **Random processes**, Master 1 Elec., 33h (7 years).
  Coordinator.
- **Numerical methods in C**, Licence 3 Elec., 38.5h (6 years since 2013).
  Coordinator.
  Creator of courses and practical sessions.
- **Statistical learning and Brain Computer Interfaces**, Master 1 GBM, 60h (7 years).
  Coordinator.
  Creator of courses and practical sessions.
- **Signal processing and applications**, Master 1 Elec., 48h (7 years).
  Co-coordinator.
  Creator of the practical sessions.
- **Theory of Machine Learning**, Master 1 Data science, 30h (1 year in 2019).
  Coordinator.
  Creator of courses and practical sessions.

2011 - 2012 **Assistant Professor (Demi ATER)**, *Université de Rouen*, 88h.

- **Mathematical tools for signal processing**, Licence 3 EEA, 15h.
- **Algorithmic, numerical methods**, Licence 3 EEA, 16h.
- **Pattern recognition for brain computer interfaces**, M2 IBIOM, 30h.

2008 - 2011 **Teaching Assistant (Monitorat)**, *INSA de Rouen*, 192h (3×64h).

- **Principles of information processing. (M8)**, 2nd year INSA, 84h.
- **Algorithmic (I3)**, 2nd year INSA, 42h.
- **Signal processing (TdS)**, 3rd year INSA, 42h.
- **Statistics (Stat)**, 3rd year INSA, 21h.

Administrative and teaching responsibilities

Coordinator of License 3 Electronics, 2017-2019,
Department of Electronics, Université Côte d'Azur.
Resp: Planning, Jury, Admission in L3.

Coordinator of Competency-based learning since 2017,
Department of Electronics, Université Côte d'Azur.
Resp: Define, write and evaluate competencies for Licence and Masters degrees in Electronics.

Research themes

My research activities can be split between fundamental research in machine learning (ML) and numerical optimization and applications of those methods in domains such as biomedical data processing, remote sensing and astronomy.

Machine learning and numerical optimization

Large scale sparse numerical optimization for machine learning An important part of our research work investigate large scale numerical optimization. I proposed with collaborators several works investigating solving sparse linear models using convex [Ferrari 2014, Ammanouil 2017, Flamary 2015] and non convex [Rakotomamonjy 2011, Boisbunon 2014, Laporte 2014b] sparsity promoting regularization terms. Among the family of algorithms for sparse optimization we proposed several active set methods [Boisbunon 2014, Rakotomamonjy 2013, Flamary 2011] that are particularly efficient when the solution of the problem is very sparse. We also proposed algorithms using proximal gradient descent in [Rakotomamonjy 2013, Mourya 2017a] and provided one of the first quasi-Newton accelerations for non-convex non-smooth problems in [Rakotomamonjy 2016].

In recent years, we investigated distributed optimization on graphs, especially the problem of privacy [Harrane 2016] and energy efficiency/communication load [Harrane 2018a, Harrane 2018b] during the PhD co-direction of Khalil Haranne.

Multi-task and transfer learning We proposed in [Rakotomamonjy 2011] to generalize the seminal Multiple Kernel Learning (MKL) approach to a multi-task learning by performing joint kernel selections across tasks. More complex relations between individual tasks was modeled by a graph that was estimated with bi-level optimization in [Flamary 2014b].

More recently we have investigated transfer learning and more precisely domain adaptation where the objective is to train a predictor on one dataset so that is predicts well on a new dataset only partially known (no labels for instance). In this domain we proposed seminal works in [Courty 2014, Courty 2016a] that used the theory of optimal transport to transfer information between domains. This work had an important impact on the community (more than 150 citations in June 2019) and we published several extensions [Courty 2017, Damodaran 2018b, Redko 2019, Perrot 2016] and theoretical justifications [Flamary 2019].

Optimal transport for machine learning Following our original works in domain adaptation we proposed the ANR OATMIL project (for OptimAl Transport for MachIne Learning) to investigate new ML approaches that can benefit from the theory of optimal transport (OT). We investigated OT for musical spectrum unmixing [Flamary 2016], discriminant subspace estimation denoted Wasserstein Discriminant Analysis [Flamary 2018], and robust learning in the presence of label noise [Damodaran 2018a, Damodaran 2019].

Another aspect of our project was to investigate the use of common ML tools to solve fundamental numerical problems related to OT. We proposed a stochastic optimization [Seguy 2018] that allowed an efficient estimation of large scale regularized OT solution and mapping estimation. We also used deep learning to estimate an embedding that mimic the geometry of the Wasserstein space in [Courty 2018] leading to dramatic speedup for common data science applications (OT computation, Wasserstein barycenters, Principal Geodesic Analysis). We investigated the statistical properties of linear OT Monge mapping in [Flamary 2019] that lead to a theoretical generalization bound for the domain adaptation problem.

Finally we investigated the use of OT for modeling and measuring similarity between graphs. To this end we extended the Gromov-Wasserstein distance [Vayer 2018, Vayer 2019a] which allows for an explicit modeling of graphs in data science applications such as clustering, denoising or classification. We also proved a very efficient closed form solution for solving the Gromov-Wasserstein optimization problem in 1D in [Vayer 2019b].

Applications of machine learning and numerical optimization

Biomedical data processing During my PhD I applied sparse multitask learning techniques for sensor selection to Brain Computer Interface data for the classification of mental tasks from EEG recordings [Flamary 2011, Laporte 2014b, Jrad 2011, Flamary 2014a]. I also participated to the BCI competition on movement prediction and achieved second place with switching linear models [Flamary 2012a]. We collaborated with the CREATIS Lab. to design novel SVM classifiers that can encode uncertainty in the labels [Niaf 2014, Niaf 2014] and represent MRI data on sparse dictionary [Lehaire 2014] with application to Computer Aided Diagnosis of prostate cancer. More recently I started a collaboration with Lab. Dieudonné, that has taken the form of a PhD co-direction of Laurent Dragoni. The objective of this collaboration is to investigate efficient solvers based on active set [Boisbunon 2014] for Lasso estimator with application to spike sorting of neurons activations for large signals.

Remote sensing During my PhD, I began a collaboration with researchers in the remote sensing community. These collaborations lead to design of new SVM classifiers for remote sensing images classification [Flamary 2012b, Tuia 2010, Tuia 2014b]. We also proposed a novel algorithm to perform complex feature selection based on active set in [Tuia 2014a, Tuia 2015b] and the use of non convex regularization in remote sensing in [Tuia 2015a, Tuia 2016]. More recent work have been focused on training in the presence of label noise that is a fundamental problem plaguing all remote sensing datasets [Damodaran 2018a, Damodaran 2019].

Astronomy As part of the Observatoire de la Côte d'Azur, I have been collaborating with astronomers to apply some of the most recent optimization techniques to fundamental estimation problems in astronomy. First we have been investigating astronomical image reconstruction that is a numerical challenge due to the size of the data [Ferrari 2014, Ammanouil 2017, Mourya 2017b, Mourya 2017b]. An original contribution in this area was to propose the first use of neural network to perform fast astronomical image reconstruction in [Flamary 2017a]. Another application of ML to astronomy has been a collaboration for automatic detection of strong gravitational lenses with the University of Manchester [Hartley 2017, Metcalf 2019].

Finally I have also contributed to the solar and spatial coronography community that aim at observing directly the solar corona or exo-planets. To this end the neighboring star (the sun or the star of the exoplanet) needs to be attenuated so that the small signal around the star can be observed. We proposed an optimization of the shape of an external solar occulter for exoplanet direct observation in [Flamary 2014c]. I am also co-directing the PhD of Raphael Rougeot from ESA on the topic of hybrid solar coronograph [Rougeot 2017, Rougeot 2018, Rougeot 2019] where several studies have estimated performance prior to a launch of two satellites in space.

Research activity

Research awards

- **Chair in Artificial Intelligence**, 3IA Côte d'Azur.
- **Outstanding reviewer**, *International Conf. in Machine Learning (ICML) 2018.*
- **Helava award of best paper 2012-2015** in the journal *ISPRS Journal of Photogrammetry and Remote Sensing* for paper [J10].
- **Best paper** at conference Photogrammetric Computer Vision (PCV) 2014 at Zurich for paper [C25].

Research activities

- **Area chair**, *Neural Information Processing Systems (**NeurIPS ex NIPS**) 2019.*

- **Journal Reviewer** for Journal of Machine Learning (**JMLR**), IEEE Trans. on Pattern Analysis and Machine Intelligence (**TPAMI**), IEEE Trans. Signal Processing (**TSP**), IEEE Trans. Neural Network and Learning Systems (**TNNLS**). Signal Processing (**SP**), IEEE Trans. on Geoscience and Remote Sensing (**TGRS**), ISPRS Journal of Photogrammetry and Remote Sensing (**JPRS**), IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (**JSTARS**), IEEE Geoscience and Remote Sensing Letters (**GRSL**), IEEE Signal Processing Letters (**SPL**).

- **Conference Reviewer** for Neural Information Processing Systems (**NIPS**), International Conference in Machine Learning (**ICML**), International conference on Learning Representations (**ICLR**), Machine Learning for Signal Processing Workshop (**MLSP**).

Current Research projects

- **OATMIL**, ANR Project 2017-2020, *Optimal transport for machine learning*, with Nicolas Courty (PI, IRISA, Vannes), and Alain Rakotomamonjy (Local PI, LITIS, Rouen).

  Local Principal Investigator

- **Magellan**, ANR project, *Machine learning methods for very large arrays in radio astronomy, with André Ferrari (PI, Lagrange, Nice), Pascal Larzabal (Local PI, SATIE, Paris), Pascal Bianchi (Local PI, LTCI, Paris).*

  Collaborator

- **ON FIRE**, Young researcher project GDR ISIS 2018-2019, *Calibration of future large interferometers* With Mohammed Nabil El Korso (PI, LEME Lab, Ville d'Avray) and Franck Iutzeler (LJK, Grenoble).

  Collaborator

Past research projects

- **DESTOPT**, CNRS PEPS Project 2018-2019, *Deep Semi-Supervised and Transfer Learning with Optimal Transport*, with Yves Grandvalet (lead, HEUDIASIC, Compiegne), Alain Rakotomamonjy (LITIS, Rouen), Nicolas Courty (IRISA, Vannes).

  Collaborator

- **TOPASE**, CNRS PEPS Project 2015-2016, *Deep Semi-Supervised and Transfer Learning with Optimal Transport*, with Alain Rakotomamonjy (lead, LITIS, Rouen) Nicolas Courty (IRISA, Vannes).

  Collaborator

- **AMOR**, Young researcher project GDR ISIS 2013-2014, *Analysis of multi-temporal remote sensing images*, with Mathieu Fauvel (DYNAFOR, Toulouse), Mauro Dalla Mura (GIPSA-lab, Grenoble) and Silvia Valero-Valbuena (CESBIO, Toulouse).

  Principal Investigator

- **HYPANEMA**, ANR Blanc *HYPerspectral data Analysis with NonlinEar unMixing Algorithm. Principal Investogator: Cédric Richard.*

  Collaborator and webmaster `http://www.hypanema.fr/`

Open source and reproducible research

- Python Optimal Transport (POT) (100 000+ downloads, `https://github.com/rflamary/POT`).
- SVMRank with non convex regularization [J16] (1500+ downloads).
- SVM with general regularization [J13] (2000+ downloads).
- Multi-task Multiple Kernel Learning [J20] (1700+ downloads).
- Large Margin Filtering [J19] (1700+ downloads).
- SVM with label uncertainty [J14] (2400+ downloads).

Students supervision

PhD students

- **Kilian Fatras**, *Optimal Transport and deep learning*, with Nicolas Courty, Université Bretagne Sud, 2018-2021. Submitted publications: [O5],[O3].
- **Laurent Dragoni**, Spike sorting for massive neurophysiological data sets, with Karim Lounici and Patricia Bouret, Université Côte d'Azur, 2017-2020. Submitted publication: [O2].
- **Raphael Rougeot**, Modeling and computation of diffraction effects for end-to-end performance of hight-contrast space optical instruments, with David Mary and Claude Aime, Université Côte d'Azur / European Spatial Agency (ESA), 2017-2020. Publications: [J6],[C6],[J1].
- **Ibrahim El Khalil Harrane**, Distributed estimation over multitask networks, with Cédric Richard, Université Côte d'Azur, 2015-2019, defended on June 21 2019. Publications: [J3], [C15], [C4], [C18].

Note that my first three PhD students were working on large scale optimization and machine learning that are some of my major research themes as discussed in the previous section. But they did not work on Optimal Transport for Machine learning that is the main theme of this document. This comes from the fact that at the time, our work on OT for ML was only beginning and I felt not yet confident enough to supervise a PhD on this theme. That is why I waited until the PhD of Kilian Fatras in 2018 to supervise a PhD student on this domain.

Master students, internship supervision

- **Adam Hessas**, Coronography optimization for direct exo-planet imaging, Ecole Centrale de Lyon, 2019.
- **Mircea Moscu**, Deep learning for astronomical image reconstruction, Université Côte d'Azur, 2017.
- **Ioana Boian**, Analyzing the granulation in red supergiant stars, with Andrea Chiavassa, Master final project, Erasmus, University of Glasgow, 2014-2015.
- **Ibrahim El Khalil Harrane**, Periodic discrimination for multitemporal data, Master student, Université Côte d'Azur, 2015.

International Competitions

- **MLSP Competition 2010**, Ranking : 3/35.

  Evoked potential classification in brain computer interface.

- **BCI Competition IV**, 2008, Ranking : 2/5, Method published in [J18].

  Prediction of finger movement from recorded ECoG signals.

Invited conferences

- **Domain adaptation with optimal transport : from mapping to learning with joint distribution**

  Invited conference, Istituto Italiano di Tecnologia, Genoa, Italy, 2019.

  Invited conference, NIPS Workshop, OTML 2017.

- **Optimal transport for machine learning and signal processing**

  Invited conference, General assembly of GDR ISIS, 2017.

- **Joint distribution optimal transportation for domain adaptation**,

  Invited conference, Banff International Research Station, Optimal Transport meets Probability Statistics and Machine Learning, may 2017, Oaxaca, Mexico

- **Domain adaptation with optimal transport**

  Invited conference, Statlearn 2016, Vannes.

- **Learning with infinitely many features**,

  Meeting of GDR ISIS, Paris

- **Mixed-norm regularization for Event-Related Potential based BCI**,

  BSCB, Cornell University, Ithaca, NY and LORIA, Nancy

International tutorials

- **Optimal Transport for Machine Learning and Signal Processing**

  Tutorial for ISBI 2019 at Venice, Italy,

  Support available at `https://remi.flamary.com/cours/tuto_otml.html`.

- **Optimal Transport for Machine Learning**

  Tutorial for Statlearn 2018, with Nicolas Courty, Nice, France, 2019.

  Tutorial for DataScience Summer School (DS3) 2018 with Nicolas Courty and Marco Cuturi, École Polytechnique, Palaiseau, France, 2019.

Organization of scientific events

- **Optimal tranport for Machine learning workshop at NeurIPS**, co-organized with A. Suvorikova, M. Cuturi and G. Peyré, December 2019 (to come).

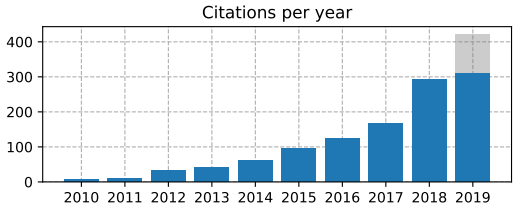
- **Optimal tranport for Machine learning**, GDR ISIS meeting, co-organized with N. Courty, N. Papadakis, A. Rakotomamonjy, 90 participants, July 2019.

- **CNRS Summer school Basmati 2 2018.** Mathematical basis for statistical learning for applications in astronomy and astrophysics, co-organized with C. Theys, D. Mary and C. Aime, Porquerolles.
- **CNRS Summer school Basmati 2015.** Mathematical basis for statistical learning for applications in astronomy and astrophysics, co-organized with C. Theys, D. Mary and C. Aime, Porquerolles.
- **Representation learning, methodology and applications**, GDR ISIS meeting, co-organized with A. Rakotomamonjy, 90 participants, October 2016.
- **Seminar 'Mathématiques pour l'analyse de données' (MAD)**, 18 seminars from 2013 to 2016 at Parc Valrose, co-organized with C. Févotte and D. Mary, Nice.

Publications

Citations and bibliometry

	All	Since 2014
Citations	1150	1057
h-index	18	17
i10	30	28



Citation information has been obtained from Google Scholar on 26/9/2019. Citations per references are provided below only for references with more than 10 citations.

Journal articles

[J1] R. Rougeot, R. Flamary, D. Mary, C. Aime, *Influence of surface roughness on diffraction in the externally occulted Lyot solar coronagraph*, Astronomy and Astrophysics, 2019.

[J2] R. B. Metcalf, M. Meneghetti, C. Avestruz, F. Bellagamba, C. R. Bom, E. Bertin, R. Cabanac, E. Decencière, R. Flamary, R. Gavazzi, others, *The Strong Gravitational Lens Finding Challenge*, Astronomy and Astrophysics, Vol. 625, pp A119, 2019.

[J3] I. Harrane, R. Flamary, C. Richard, *On reducing the communication cost of the diffusion LMS algorithm*, IEEE Transactions on Signal and Information Processing over Networks (SIPN), 2018.

[J4] R. Flamary, M. Cuturi, N. Courty, A. Rakotomamonjy, *Wasserstein Discriminant Analysis*, Machine learning , 2018 (19 citations).

[J5] P. Hartley, R. Flamary, N. Jackson, A. S. Tagore, R. B. Metcalf, *Support Vector Machine classification of strong gravitational lenses*, Monthly Notices of the Royal Astronomical Society (MNRAS), 2017 (10 citations).

[J6] R. Rougeot, R. Flamary, D. Galano, C. Aime, *Performance of hybrid externally occulted Lyot solar coronagraph, Application to ASPIICS*, Astronomy and Astrophysics, 2017.

[J7] N. Courty, R. Flamary, D. Tuia, A. Rakotomamonjy, *Optimal transport for domain adaptation*, Pattern Analysis and Machine Intelligence, IEEE Transactions on , 2016 (180 citations).

[J8] D. Tuia, R. Flamary, M. Barlaud, *Non-convex regularization in remote sensing*, Geoscience and Remote Sensing, IEEE Transactions on, 2016 (19 citations).

[J9] A. Rakotomamonjy, R. Flamary, G. Gasso, *DC Proximal Newton for Non-Convex Optimization Problems*, Neural Networks and Learning Systems, IEEE Transactions on, Vol. 27, N. 3, pp 636-647, 2016 (18 citations).

[J10] D. Tuia, R. Flamary, N. Courty, *Multiclass feature learning for hyperspectral image classification: sparse and hierarchical solutions*, ISPRS Journal of Photogrammetry and Remote Sensing, 2015 (62 citations).

[J11] R. Flamary, M. Fauvel, M. Dalla Mura, S. Valero, *Analysis of multi-temporal classification techniques for forecasting image times series*, Geoscience and Remote Sensing Letters (GRSL), Vol. 12, N. 5, pp 953-957, 2015.

[J12] R. Flamary, C. Aime, *Optimization of starshades: focal plane versus pupil plane*, Astronomy and Astrophysics, Vol. 569, N. A28, pp 10, 2014.

[J13] R. Flamary, N. Jrad, R. Phlypo, M. Congedo, A. Rakotomamonjy, *Mixed-Norm Regularization for Brain Decoding*, Computational and Mathematical Methods in Medicine, Vol. 2014, N. 1, pp 1-13, 2014 (13 citations).

[J14] E. Niaf, R. Flamary, O. Rouvière, C. Lartizien, S. Canu, *Kernel-Based Learning From Both Qualitative and Quantitative Labels: Application to Prostate Cancer Diagnosis Based on Multiparametric MR Imaging*, Image Processing, IEEE Transactions on, Vol. 23, N. 3, pp 979-991, 2014 (23 citations).

[J15] D. Tuia, M. Volpi, M. Dalla Mura, A. Rakotomamonjy, R. Flamary, *Automatic Feature Learning for Spatio-Spectral Image Classification With Sparse SVM*, Geoscience and Remote Sensing, IEEE Transactions on, Vol. 52, N. 10, pp 6062-6074, 2014 (65 citations).

[J16] L. Laporte, R. Flamary, S. Canu, S. Déjean, J. Mothe, *Nonconvex Regularizations for Feature Selection in Ranking With Sparse SVM*, Neural Networks and Learning Systems, IEEE Transactions on, Vol. 25, N. 6, pp 1118-1130, 2014 (73 citations).

[J17] A. Rakotomamonjy, R. Flamary, F. Yger, *Learning with infinitely many features*, Machine Learning, Vol. 91, N. 1, pp 43-66, 2013 (14 citations).

[J18] R. Flamary, A. Rakotomamonjy, *Decoding finger movements from ECoG signals using switching linear models*, Frontiers in Neuroscience, Vol. 6, N. 29, 2012 (48 citations).

[J19] R. Flamary, D. Tuia, B. Labbé, G. Camps-Valls, A. Rakotomamonjy, *Large Margin Filtering*, IEEE Transactions Signal Processing, Vol. 60, N. 2, pp 648-659, 2012 (15 citations).

[J20] A. Rakotomamonjy, R. Flamary, G. Gasso, S. Canu, *lp-lq penalty for sparse linear and sparse multiple kernel multi-task learning*, IEEE Transactions on Neural Networks, Vol. 22, N. 8, pp 1307-1320, 2011 (72 citations).

[J21] N. Jrad, M. Congedo, R. Phlypo, S. Rousseau, R. Flamary, F. Yger, A. Rakotomamonjy, *sw-SVM: sensor weighting support vector machines for EEG-based brain–computer interfaces*, Journal of Neural Engineering, Vol. 8, N. 5, pp 056004, 2011 (44 citations).

International conferences

[C1] T. Vayer, R. Flamary, R. Tavenard, L. Chapel, N. Courty, *Sliced Gromov-Wasserstein*, Neural Information Processing Systems (NeurIPS), 2019.

[C2] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, N. Courty, *Optimal Transport for structured data with application on graphs*, International Conference on Machine Learning (ICML), 2019 (11 citations).

[C3] I. Redko, N. Courty, R. Flamary, D. Tuia, *Optimal Transport for Multi-source Domain Adaptation under Target Shift*, International Conference on Artificial Intelligence and Statistics (AISTAT), 2019.

[C4] I. Harrane, R. Flamary, C. Richard, R. Couillet, *Random matrix theory for diffusion LMS analysis.*, Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 2018.

[C5] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, N. Courty, *DeepJDOT: Deep Joint distribution optimal transport for unsupervised domain adaptation*, European Conference in Computer Visions (ECCV), 2018 (24 citations).

[C6] R. Rougeot, C. Aime, C. Baccani, S. Fineschi, R. Flamary, D. Galano, C. Galy, V. Kirschner, F. Landini, M. Romoli, others, *Straylight analysis for the externally occulted Lyot solar coronagraph ASPIICS*, Space Telescopes and Instrumentation 2018: Optical, Infrared, and Millimeter Wave, Vol. 10698, pp 106982T, 2018.

[C7] V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, M. Blondel, *Large-Scale Optimal Transport and Mapping Estimation*, International Conference on Learning Representations (ICLR), 2018 (39 citations).

[C8] N. Courty, R. Flamary, M. Ducoffe, *Learning Wasserstein Embeddings*, International Conference on Learning Representations (ICLR), 2018 (13 citations).

[C9] N. Courty, R. Flamary, A. Habrard, A. Rakotomamonjy, *Joint Distribution Optimal Transportation for Domain Adaptation*, Neural Information Processing Systems (NIPS), 2017 (53 citations).

[C10] R. Mourya, A. Ferrari, R. Flamary, P. Bianchi, C. Richard, *Distributed Approach for Deblurring Large Images with Shift-Variant Blur*, European Conference on Signal Processing (EUSIPCO), 2017.

[C11] R. Flamary, *Astronomical image reconstruction with convolutional neural networks*, European Conference on Signal Processing (EUSIPCO), 2017.

[C12] R. Ammanouil, A. Ferrari, R. Flamary, C. Ferrari, D. Mary, *Multi-frequency image reconstruction for radio-interferometry with self-tuned regularization parameters*, European Conference on Signal Processing (EUSIPCO), 2017.

[C13] R. Flamary, C. Févotte, N. Courty, V. Emyia, *Optimal spectral transportation with application to music transcription*, Neural Information Processing Systems (NIPS), 2016 (10 citations).

[C14] M. Perrot, N. Courty, R. Flamary, A. Habrard, *Mapping estimation for discrete optimal transport*, Neural Information Processing Systems (NIPS), 2016 (32 citations).

[C15] I. Harrane, R. Flamary, C. Richard, *Doubly partial-diffusion LMS over adaptive networks*, Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 2016.

[C16] S. Nakhostin, N. Courty, R. Flamary, D. Tuia, T. Corpetti, *Supervised planetary unmixing with optimal transport*, Whorkshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS), 2016.

[C17] N. Courty, R. Flamary, D. Tuia, T. Corpetti, *Optimal transport for data fusion in remote sensing*, International Geoscience and Remote Sensing Symposium (IGARSS), 2016.

[C18] I. Harrane, R. Flamary, C. Richard, *Toward privacy-preserving diffusion strategies for adaptation and learning over networks*, European Conference on Signal Processing (EUSIPCO), 2016.

[C19] R. Flamary, A. Rakotomamonjy, G. Gasso, *Importance Sampling Strategy for Non-Convex Randomized Block-Coordinate Descent*, IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015.

[C20] R. Flamary, I. Harrane, M. Fauvel, S. Valero, M. Dalla Mura, *Discrimination périodique à partir d'observations multi-temporelles*, GRETSI, 2015.

[C21] D. Tuia, R. Flamary, A. Rakotomamonjy, N. Courty, *Multitemporal classification without new labels: a solution with optimal transport*, International Workshop on the Analysis of Multitemporal Remote Sensing Images (Multitemp), 2015.

[C22] D. Tuia, R. Flamary, M. Barlaud, *To be or not to be convex? A study on regularization in hyperspectral image classification*, International Geoscience and Remote Sensing Symposium (IGARSS), 2015.

[C23] A. Boisbunon, R. Flamary, A. Rakotomamonjy, A. Giros, J. Zerubia, *Large scale sparse optimization for object detection in high resolution images*, IEEE Workshop in Machine Learning for Signal Processing (MLSP), 2014.

[C24] E. Niaf, R. Flamary, A. Rakotomamonjy, O. Rouvière, C. Lartizien, *SVM with feature selection and smooth prediction in images: application to CAD of prostate cancer*, IEEE International Conference on Image Processing (ICIP), 2014 (14 citations).

[C25] D. Tuia, N. Courty, R. Flamary, *A group-lasso active set strategy for multiclass hyperspectral image classification*, Photogrammetric Computer Vision (PCV), 2014.

[C26] J. Lehaire, R. Flamary, O. Rouvière, C. Lartizien, *Computer-aided diagnostic for prostate cancer detection and characterization combining learned dictionaries and supervised classification*, IEEE International Conference on Image Processing (ICIP), 2014 (11 citations).

[C27] A. Ferrari, D. Mary, R. Flamary, C. Richard, *Distributed image reconstruction for very large arrays in radio astronomy*, IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM), 2014 (10 citations).

[C28] N. Courty, R. Flamary, D. Tuia, *Domain adaptation with regularized optimal transport*, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2014 (68 citations).

[C29] A. Boisbunon, R. Flamary, A. Rakotomamonjy, *Active set strategy for high-dimensional non-convex sparse optimization problems*, International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014.

[C30] W. Gao, J. Chen, C. Richard, J. Huang, R. Flamary, *Kernel LMS algorithm with Forward-Backward splitting for dictionnary learning*, International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2013 (22 citations).

[C31] R. Flamary, A. Rakotomamonjy, *Support Vector Machine with spatial regularization for pixel classification*, International Workshop on Advances in Regularization, Optimization, Kernel Methods and Support Vector Machines : theory and applications (ROKS), 2013.

[C32] D. Tuia, M. Volpi, M. Dalla Mura, A. Rakotomamonjy, R. Flamary, *Create the relevant spatial filterbank in the hyperspectral jungle*, IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2013.

[C33] D. Tuia, R. Flamary, M. Volpi, M. Dalla Mura, A. Rakotomamonjy, *Discovering relevant spatial filterbanks for VHR image classification*, International Conference on Pattern Recognition (ICPR), 2012.

[C34] E. Niaf, R. Flamary, S. Canu, O. Rouvière, C. Lartizien, *Handling learning samples uncertainties in SVM : application to MRI-based prostate cancer Computer-Aided Diagnosis*, IEEE International Symposium on Biomedical Imaging , 2012.

[C35] R. Flamary, F. Yger, A. Rakotomamonjy, *Selecting from an infinite set of features in SVM*, European Symposium on Artificial Neural Networks, 2011.

[C36] R. Flamary, X. Anguera, N. Oliver, *Spoken WordCloud: Clustering Recurrent Patterns in Speech*, International Workshop on Content-Based Multimedia Indexing, 2011 (21 citations).

[C37] E. Niaf, R. Flamary, C. Lartizien, S. Canu, *Handling uncertainties in SVM classification*, IEEE Workshop on Statistical Signal Processing , 2011 (17 citations).

[C38] R. Flamary, B. Labbé, A. Rakotomamonjy, *Large margin filtering for signal sequence labeling*, International Conference on Acoustic, Speech and Signal Processing 2010, 2010.

[C39] R. Flamary, B. Labbé, A. Rakotomamonjy, *Filtrage vaste marge pour l'étiquetage séquentiel de signaux*, Conference en Apprentissage CAp, 2010.

[C40] D. Tuia, G. Camps-Valls, R. Flamary, A. Rakotomamonjy, *Learning spatial filters for multispectral image segmentation*, IEEE Workshop in Machine Learning for Signal Processing (MLSP), 2010.

[C41] R. Flamary, A. Rakotomamonjy, G. Gasso, S. Canu, *Selection de variables pour l'apprentissage simultanée de tâches*, Conférence en Apprentissage (CAp'09), 2009.

[C42] R. Flamary, J. Rose, A. Rakotomamonjy, S. Canu, *Variational Sequence Labeling*, IEEE Workshop in Machine Learning for Signal Processing (MLSP), 2009.

Book chapters

[CH1] S. Canu, R. Flamary, D. Mary, *Introduction to optimization with applications in astronomy and astrophysics*, Mathematical tools for instrumentation and signal processing in astronomy, 2016.

[CH2] R. Flamary, A. Rakotomamonjy, M. Sebag, *Apprentissage statistique pour les BCI*, Les interfaces cerveau-ordinateur 1, fondements et méthodes, pp 197-215, 2016.

[CH3] R. Flamary, A. Rakotomamonjy, M. Sebag, *Statistical learning for BCIs*, Brain Computer Interfaces 1: Fundamentals and Methods, pp 185-206, 2016.

[CH4] R. Flamary, A. Rakotomamonjy, G. Gasso, *Learning Constrained Task Similarities in Graph-Regularized Multi-Task Learning*, Regularization, Optimization, Kernels, and Support Vector Machines, 2014.

Books as editor

[B1] D. Mary, R. Flamary, C. Theys, C. Aime, *Mathematical Tools for Instrumentation and Signal Processing in Astronomy*, 2016.

Other communications

[O1] R. Flamary, K. Lounici, A. Ferrari, *Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation*, 2019.

[O2] L. Dragoni, R. Flamary, K. Lounici, P. Reynaud-Bouret, *Large scale Lasso with windowed active set for convolutional spike sorting*, 2019.

[O3] B. Bhushan Damodaran, K. Fatras, S. Lobry, R. Flamary, D. Tuia, N. Courty, *Pushing the right boundaries matters! Wasserstein Adversarial Training for Label Noise*, 2019.

[O4] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, N. Courty, *Fused Gromov-Wasserstein distance for structured objects: theoretical foundations and mathematical properties*, 2018.

[O5] B. B. Damodaran, R. Flamary, V. Seguy, N. Courty, *An Entropic Optimal Transport Loss for Learning Deep Neural Networks under Label Noise in Remote Sensing Images*, 2018.

[O6] A. Rakotomamonjy, A. Traore, M. Berar, R. Flamary, N. Courty, *Distance Measure Machines*, 2018.

[O7] R. Mourya, A. Ferrari, R. Flamary, P. Bianchi, C. Richard, *Distributed Deblurring of Large Images of Wide Field-Of-View*, 2017.

[O8] A. Rakotomamonjy, R. Flamary, N. Courty, *Generalized conditional gradient: analysis of convergence and applications*, 2015.

[O9] R. Flamary, N. Courty, D. Tuia, A. Rakotomamonjy, *Optimal transport with Laplacian regularization: Applications to domain adaptation and shape matching*, NIPS Workshop on Optimal Transport and Machine Learning OTML, 2014.

[O10] R. Flamary, *Apprentissage statistique pour le signal: applications aux interfaces cerveau-machine*, Laboratoire LITIS, Université de Rouen, 2011.

[O11] R. Flamary, B. Labbé, A. Rakotomamonjy, *Large margin filtering for signal segmentation*, NIPS Workshop on Temporal Segmentation NIPS Workshop in Temporal Segmentation, 2009.

[O12] R. Flamary, A. Rakotomamonjy, G. Gasso, S. Canu, *SVM Multi-Task Learning and Non convex Sparsity Measure*, The Learning Workshop The Learning Workshop (Snowbird), 2009.

[O13] R. Flamary, *Filtrage de surfaces obtenues à partir de structures M-Rep (M-Rep obtained surface filtering)*, Laboratoire CREATIS-LRMN, INSA de Lyon, 2008.

Chapter 1

Introduction

Contents

This document focuses on the part of my research about the relations between Optimal Transport (OT) and Machine Learning (ML). This choice has been made in order to provide a self content document on a major part of our recent results. For a short discussion about my other research projects I refer the reader to the Curriculum Vitae at the beginning of the document.

Note that I tried to keep this chapter very general but it still requires to use the terminology of optimal transport. I refer the reader to the next chapter for a short introduction to optimal transport with all necessary definitions.

1.1 Optimal transport for Machine Learning

In the following I provide a quick history of the introduction to OT in the machine learning community followed by a discussion of some aspects of OT that can be used in ML applications.

1.1.1 A brief history of OT for ML

Optimal transport (OT) Optimal Transport aims at finding the solution of least effort to move mass from one distribution to another. It is a fundamental problem strongly related to physics and has been investigated by mathematicians since the introduction of the problem by Monge [Monge 1781]. When a solution of the problem exists, OT also provides a measure of similarity between the two distributions under the form of the optimal transport cost. This similarity, also called Wasserstein distance, can be used in practice on any distribution of mass and encodes the geometry of the space through the optimization problem.

Despite all those nice properties, OT has only relatively recently been used in data science. One of the reasons was the numerical complexity of solving the OT problem on large datasets. For instance computing a Wasserstein distance between empirical distributions having n samples has a time complexity of $O(n^3)$.

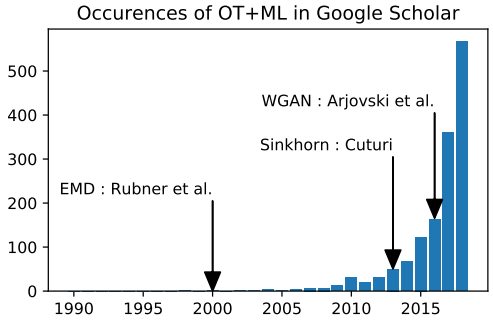


Figure 1.1: Number of references in google scholar that contain Optimal Transport and Machine Learning in recent years. We also report the years corresponding to the works of Rubner (EMD [Rubner 2000]), Cuturi (Entropic OT [Cuturi 2013]) and Arjovski (WGAN [Arjovsky 2017]).

Early applications to image processing and ML The image processing community has investigated the use of OT before its introduction to ML. The use of Earth Mover's Distance (EMD), a special case of Wasserstein distance has been proposed to measure similarity between gray images in [Peleg 1989] and between color histograms of images in [Rubner 1998, Rubner 2000]. OT has been investigated also in the image processing community for color adaptation and interpolation [Rabin 2011, Bonneel 2011]. It has also been used to reconstruct the mass distribution the early universe in [Frisch 2002]. A good introduction to applications in image and signal processing is available in [Kolouri 2016]. One of the first application of OT in ML is its use as a divergence in matrix factorization [Sandler 2011]. It was also investigated for semi supervised learning in [Solomon 2014b]. One difficulty that greatly limited the use of OT in ML on large datasets was numerical complexity.

Regularized OT In 2013, Marco Cuturi proposed to solve an approximation of the original OT problem in [Cuturi 2013] by adding an entropic regularization term to the optimization problem. The resulting problem is strictly convex and easier to solve with the Sinkhorn-Knopp algorithm, that can be implemented efficiently in parallel on GPU. It was later extended thanks to Bregman projections to the fast computation of regularized Wasserstein barycenter [Benamou 2015] and can be greatly accelerated [Solomon 2015] on images. Those new computational solvers for OT problems opened the door for larger datasets and the statistical properties of regularized OT have been investigated since [Cazelles 2018, Genevay 2018].

One major feature of entropic regularized OT is that one can express its dual or semi-dual (introduced more in detail in Chapter 1) without constraints. The objective value in the dual or semi-dual is an expected value *w.r.t.* the source and target distributions, so [Genevay 2016] proposed to solve it with stochastic gradient. This allowed for fast iterative methods that can scale to a large number of samples and also to solve semi-discrete OT when samples from the continuous distribution can be drawn. Finally the work that cemented the introduction of OT in ML was the Wasserstein Generative Adversarial Network proposed in [Arjovsky 2017]. In this work, Arjovski and Bottou proposed to minimize the Wasserstein distance as an objective value for a Generative network. To this end they proposed to solve the problem in the dual and to use neural networks to estimate dual potentials.

Optimal Transport provides in the ML context a unique set of tools that can estimate mappings or similarities between distributions that take into account the geometry of the space. Those tools are meaningful even when the distributions do not share the same support which is the case when working with empirical distributions. Since its reintroduction to the ML community in 2014, research in OT for ML has been steadily increasing as illustrated in Figure 1.1. There have been several NeurIPS workshops focusing on this theme with increasing number of participants in 2015, 2017 and we will co-organize the

NeurIPS OTML workshop[1] in 2019.

1.1.2 Four aspects of optimal transport

The following present four aspects of OT that are of particular interest for ML applications. These aspects will be used to provide a structure for the rest of the document as discussed in the next section.

OT for mapping between distributions One major aspect of Optimal Transport is the estimation of a transportation. The solution of OT problems always provides an optimal pairwise relation between the support of two distributions. This relation can take the form of Monge mapping or a more "fuzzy" representation as a joint source/target distribution. A nice property of this mapping is that it is the one corresponding to the least effort or moving the mass between the two distributions, making it a sensible and natural choice when no other pairwise relations are available in the data. The Monge mapping, or an approximation, can be used in ML to move samples between distributions. I discuss in the following how OT cab be used for domain adaptation, *i.e.* to adapt samples from different distributions/datasets.

OT for similarity between histograms A second aspect of OT that is a key element to its use in ML is the Wasserstein distance. This distance allows a measure of similarity between distributions that take into account the geometry of the space and it can be used to measure similarity between histograms. We define more in detail the difference between what we call histograms and empirical distribution in the next chapter but keep in mind that we suppose that histograms have a fixed support and encode their information on the weights of their bins. An increasing amount of data available today can be represented as histograms. Classical divergences on histograms are separable in the sense that they treat every bins of the histograms separately. While this is very efficient, it limits the information provided by the divergence and especially when there is a complex geometry defining the relations between the bins. Wasserstein distance can (at a numerical cost) encode this geometry and provide better modeling of the histogram data.

OT for similarity between empirical distributions When the distribution is an empirical distribution, it can be expressed as a weighted sum of Diracs (often with uniform weights). In this case the information on the distribution is not in the weights but in the position of those Diracs. Most of the datasets in ML can be expressed as empirical distributions. One extra difficulty when learning from those distributions is that their support never overlap (you never have two Diracs exactly at the same position). One approach consists in using Kernel density estimation with Maximum Mean Discrepancy (MMD) [Gretton 2012]. But this approach can sometimes lose information due to the kernel smoothing and becomes non-informative when the distributions support are far away. In this case Wasserstein distance can still provide a meaningful similarity measure that despite being non-smooth provides usefully sub-gradients for model fitting.

OT on structured objects OT has been recently extended to a similarity measure between distributions that do not share a common space with the Gromov-Wasserstein distance [Mémoli 2011]. It has been used with success on structured objects seen as distributions such as graphs. This similarity, has seen an increasing interest in the ML community since it opens the door for template based classifiers and computation of barycenters of structured objects. Interestingly, using OT in this setting provides a transport matrix between parts of the structured objects (the nodes for graphs) and allows for nice interpretation of the relations between the objects.

[1] https://sites.google.com/view/otml2019/

1.2 Manuscript outline and contributions

In this section, we provide a short description of the chapters of the manuscript and of our contributions that are presented in these chapters.

Chapter 1 : Introduction This is the current chapter. It acts as a short introduction to the manuscript and a positioning into the young history of OT for ML. It also provides an outline of the manuscript and of the contributions that will not be detailed so as to avoid an infinite recursion.

Chapter 2 : Optimal Transport tools and algorithm In the next chapter I introduce the main tools of OT theory that will be used in the rest of the manuscript. We provide an introduction to the optimization problems of OT and the resulting Wasserstein distance. The following section focuses particularly on discrete distributions that are often encountered in ML and introduce regularized OT. Finally I discuss the algorithms that can be used to solve those problems and present generic solvers we proposed to solve regularized OT.

Chapter 3 : Mapping with OT This chapter discusses the use of OT mapping in ML applications. In its first part I introduce several approaches that have been proposed to estimate an OT mapping between distributions. I also define the "barycentric" mapping followed by some of our contributions in this domain for estimating continuous OT mappings [Perrot 2016, Seguy 2018, Flamary 2019].

The second part of the chapter presents our first foray into the application of OT for ML problems with the proposition of using OT to align in an unsupervised way (no known relations between source and target samples) distributions in Domain Adaptation (DA) problem. In this problem the objective is to estimate a predictor on an unlabeled target distribution using information from a labeled but different source distribution. We proposed in [Courty 2014, Courty 2016a] to use OT to transfer samples (and their labels) with OT so as to be able to train a predictor in the target domain (OTDA). We discuss the assumptions of OTDA and the generalization properties of the estimated predictors. Finally we discuss some extensions of OTDA and several successful applications found in the literature.

Chapter 4 : OT between histograms This chapter discusses the use of OT and Wasserstein distance to process data that can be expressed as histograms. The first section of the chapter is a short state of the art of methods proposed in this domain such as Geodesic PCA in the Wasserstein space and supervised learning with the Wasserstein distance.

Next I introduce two of our contributions in this domain. The first one is an application of Wasserstein distance on musical audio signals [Flamary 2016]. We proposed a novel ground metric for OT in the frequency domain that can encode some robustness to the variability of the audio spectrum of musical notes (magnitude of the harmonics for instance). The second contribution denoted as Deep Wasserstein Embedding (DWE) aims at learning an embedding that can emulate the properties of the Wasserstein space [Courty 2018]. The learned embedding allows for a very fast computation of approximate Wasserstein distance and opens the door for a fast implementation of a large family of data mining approaches (PCA, Kmeans) in the Wasserstein space.

Chapter 5 : OT between empirical distributions This chapter discusses the use of OT on empirical distributions (for instance datasets) or between an empirical distribution and a parametrized distribution (semi-discrete case). The first part of the chapter is again a state of the art in ML, that will discuss both unsupervised learning through Generative models (WGAN) and supervised learning using robust optimization procedures.

The second part of the chapter presents two contributions in this domain. The first one is denoted Wasserstein Discriminant Analysis (WDA) and can be seen as a generalization of the Fisher discriminant analysis when the distributions are not linearly separable. The main idea is to optimize a ratio similar to the one in the Fisher Discriminant but using Regularized OT to measure inter and intra- class similarities.

The second contribution, denoted as Joint Distribution Domain Adaptation or JDOT, is an approach that aims at transferring information between source and target distributions to find directly a predictor in the target domain that aligns the joint feature/label distribution. I will also discuss extensions of JDOT to deep learning where the feature extraction is estimated simultaneously with the predictor.

Chapter 6 : OT for structured data This chapter describes an extension of OT when measuring similarity between distributions lying in different spaces. I will first introduce the Gromov-Wasserstein distance that has been proposed recently to measure similarity between structured objects and discuss its optimization problem.

Next I present the Fused Gromov-Wasserstein distance that we proposed recently to measure similarity between structured objects such as labeled graphs. I illustrate how this new metric allows for comparison and estimation of graph approximations and barycenters in the last section of the chapter.

Chapter 7 : Concluding remarks This last chapter is a short conclusion for the document. The first section discusses current ongoing works and what I believe are pertinent research directions. The second section is a more general discussion about what I believe are the major questions that OT is facing in future ML applications.

Chapter 2

Optimal Transport tools and algorithms

Contents

This chapter will shortly introduce some definitions about OT in general with a focus on OT between discrete distributions. The last section will discuss more in details the numerical aspects of the optimization problems for OT and regularized OT.

Notations. Let Ω be a set of $\mathbb{R}^d$ and μ, μ_s and μ_t three probability measures on Ω, $\Omega_s \subset \Omega$ and $\Omega_t \subset \Omega$ respectively. $P(\Omega)$ is the set of probability distributions on Ω.

2.1 Optimal transport theory

Optimal transport (OT) is a fascinating problem that has been studied by mathematicians for hundred of years. A very good introduction with simple notations is available in the book by Santambrogio in [Santambrogio 2014]. For a more detailed report of major results (and proofs) we refer the reader to the impressive books by Villani [Villani 2009, Villani 2003]. While [Villani 2003] is out of print, it is probably more accessible for ML practitioners. Finally the most complete document to date about numerical aspect of OT is the book by Peyré and Cuturi [Peyré 2017].

2.1.1 Monge and Kantorovitch problems

Monge problem The OT problem has been historically introduced in a mémoire by Gaspard Monge [Monge 1781]. The objective was to move dirt from one place (déblais) to another (remblais) in the most efficient way possible. To this end, he seeks a mapping T that will displace the mass between the source and target mass distributions (μ_s and μ_t). But he also wants this mapping to be optimal with respect to a given cost function c that gives the effort necessary for moving a unit of mass between two positions in the space Ω.

The problem can be expressed formally as follows. Provided two probability measures μ_s and μ_t and a cost function $c : \Omega \times \Omega \to [0, +\infty]$, the Monge formulation of optimal transport [Monge 1781] aims at

finding a mapping $m : \Omega_s \to \Omega_t$ such that

$$\inf_{m\#\mu_s=\mu_t} \quad \int_{\Omega_s} c(\mathbf{x}, m(\mathbf{x}))\mu_s(\mathbf{x})d\mathbf{x} \tag{2.1}$$

where # is the mass preserving push forward operator. This operator displaces the mass from a given distribution using mapping m such that that for any measurable Borel subset $A \in \Omega_t$, the mass is preserved through mapping: $\mu_t(A) = \mu_s(m^{-1}(A)) = m\#\mu_s(A)$.

The optimization problem in Equation (2.2) is non convex because of the constraints on the mapping m and a solution for a Monge map might not even exist. For instance the fact that T is a mapping means that one cannot split the mass from a single point, which means that a solution might not exist for discrete distributions with no density. In the general case, there is also no unicity in the solution of the problem. For these reasons the Monge problem remained an open question for many years. Note that in 1884 the Académie des Sciences proposed the Prix Bordin [Bordin 1884], an award of 3000 Francs, to the first work proposing general solution to the Monge problem. The prize was not awarded in the following years, but one can note the impressive mémoire submitted by Appell that discussed several discrete and continuous cases [Appell 1887]. Results about the existence and unicity of the Monge map were limited to special cases until the works of Brenier [Brenier 1991]. He proved that when distributions μ_s and μ_t have densities and the cost is the squared euclidean distance $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$, the Monge map exists and is unique as discussed more in details later.

Kantorovitch Primal formulation In the 1940s, Kantorovitch proposed a relaxation of the problem with applications in optimal resource allocation [Kantorovich 1942]. The main idea is that instead of seeking a mapping, one can seek a joint distribution between the source and target that defines how the mass is allocated. For a given symmetric cost function $c : \Omega \times \Omega \to [0, +\infty]$ the primal OT problem can be expressed as the following problem known as the Kantorovitch formulation

$$\min_{\gamma \in \Pi(\mu_s, \mu_t)} \quad \left\{ \int_{\Omega \times \Omega} c \; d\gamma = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}[c(\mathbf{x}, \mathbf{y})] \right\}. \tag{2.2}$$

This is a constrained linear program where the constraints are defined as the polytope of the so-called transport plans defined as

$$\Pi(\mu_s, \mu_t) = \left\{ \gamma \in P(\Omega, \Omega) : \int \gamma(\mathbf{x}, \mathbf{y})d\mathbf{y} = \mu_s(\mathbf{x}), \int \gamma(\mathbf{x}, \mathbf{y})d\mathbf{x} = \mu_t(\mathbf{y}) \right\}, \tag{2.3}$$

where we can clearly see that γ is constrained to have μ_s and μ_t as left and right marginals respectively. In other words, we seek for the joint distribution γ with μ_s and μ_t as marginals that minimizes the expected transportation cost. This optimization problem is a linear program (linear objective and linear constraints). It is convex and always has a solution for a semi lower continuous c because the independent distribution $\gamma(\mathbf{x}, \mathbf{y}) = \mu_s(\mathbf{x})\mu_t(\mathbf{y})$ respects the constraints.

Kantorovitch dual formulation The Kantorovitch primal formulation (2.2) is a linear program. Its dual by the Rockafellar-Fenchel theorem is:

$$\max_{\phi \in \mathcal{C}(\Omega_s), \psi \in \mathcal{C}(\Omega_t)} \quad \left\{ \int \phi d\mu_s + \int \psi d\mu_t \;\middle|\; \phi(\mathbf{x}) + \psi(\mathbf{y}) \leq c(\mathbf{x}, \mathbf{y}) \right\} \tag{2.4}$$

where $\mathcal{C}(\Omega)$ is the set of continuous functions on Ω. The two scalar functions ϕ and ψ (also known as Kantorovich potentials) are the dual variables of the optimization problem.

c-transform and semi-dual The *c-transform* (or *c-conjugate*) H^c is a formulation that appears naturally in the dual formulation of OT. It is defined as

$$\phi^c =: H^c(\phi) =: \inf_{\mathbf{x}} \quad c(\mathbf{x}, \mathbf{y}) - \phi(\mathbf{x}) \tag{2.5}$$

and can be seen as a generalization of the Legendre transform. Using the *c-transform* one can reformulate the dual problem (2.4) in its semi dual form

$$\max_{\phi \in \mathcal{C}(\Omega_s)} \quad \int \phi d\mu_s + \int \phi^c d\mu_t \tag{2.6}$$

where the problem now depends only on the first dual potential and is the supremum of a linear function *w.r.t.* μ_s and μ_t.

Case when $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ In this case there exists a solution but it is not unique. But one can show that the optimal dual potential $\phi \in \text{Lip}^1$ is a 1-Lipschitz function and we have a close form solution of the c-transform $\phi^c(x) = -\phi(x)$. The optimal transport problem then amounts to finding $\phi \in \text{Lip}^1$ maximizing

$$\sup_{\phi \in \text{Lip}^1} \int \phi d(\mu_s - \mu_t) = \sup_{\phi \in \text{Lip}^1} \mathbb{E}_{\mathbf{x} \sim \mu_s}[\phi(\mathbf{x})] - \mathbb{E}_{\mathbf{y} \sim \mu_t}[\phi(\mathbf{y})] \tag{2.7}$$

This formulation is very nice due to the very simple *c-transform*, the optimization problem depends only on one dual variable. The main difficulty in this optimization problem is to optimize over the set of 1-Lipschitz functions. Some relaxations of this problem were proposed in the context of Generative models as discussed in the next section.

Case when $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2/2$ Whenever the cost is quadratic and the distributions μ_s and μ_t have a density, then **Brenier's Theorem** [Brenier 1991] states that the optimal transport mapping $m(x)$ exists and is unique. More remarkably, it is a gradient of a convex function $\Phi(\mathbf{x})$:

$$m(\mathbf{x}) = \mathbf{x} - \nabla\phi(\mathbf{x}) = \nabla\left(\frac{\|\mathbf{x}\|^2}{2} - \phi(\mathbf{x})\right) = \nabla(\Phi(\mathbf{x})) \tag{2.8}$$

Note that this result can be generalized to any strictly convex loss of the form $c(\mathbf{x}, \mathbf{y}) = h(\mathbf{x} - \mathbf{y})$.

Case between Gaussian distributions when $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2/2$ A well known special case of OT is when $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2/2$ and $\mu_s = \mathcal{N}(\mathbf{m}_s, \Sigma_s)$ and $\mu_t = \mathcal{N}(\mathbf{m}_t, \Sigma_t)$. When Σ_s and Σ_t are both strictly positive definite, the Monge mapping can be expressed as

$$m(\mathbf{x}) = \mathbf{m}_s + \mathbf{A}(x - \mathbf{m}_t) \tag{2.9}$$

with

$$\mathbf{A} = \Sigma_s^{-\frac{1}{2}} \left(\Sigma_s^{\frac{1}{2}} \Sigma_t \Sigma_s^{\frac{1}{2}}\right)^{\frac{1}{2}} \Sigma_s^{-\frac{1}{2}} = \mathbf{A}^T \tag{2.10}$$

This result has been discussed and proven several times in the Optimal Transport literature [Givens 1984, McCann 1997, Takatsu 2011, Bhatia 2018, Malagò 2018]. Note that we have investigated the quality of the Monge mapping estimator $\tilde{m}$ when the covariances and means are estimated from a finite number of IID samples in [Flamary 2019]. In this work we proved a mapping error $E_{\mathbf{x} \sim \mu_s}[\|m(\mathbf{x}) - \tilde{m}(\mathbf{x})\|]$ is $O(n^{-1/2})$ that is remarkably independent from the dimensionality but of course limited to linear Monge mapping.

2.1.2 Wasserstein distance

From the optimal transport optimization problem one can define the Wasserstein distance over the set of distributions as

$$W_p(\mu_s, \mu_t) = \min_{\gamma \in \Pi(\mu_s, \mu_t)} \quad \left\{ \int_{\Omega_s \times \Omega_t} c(\mathbf{x}, \mathbf{y}) \; d\gamma(\mathbf{x}, \mathbf{y}) \right\}^{\frac{1}{p}} \tag{2.11}$$

where $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^p$ and $p \geq 1$. This distance is also known in the computer vision community as the Earth Mover's Distance (W_1^1) when $p = 1$ [Rubner 1998]. It can encode the geometry of the space through c and always gives a meaningful value even when the two distributions have no overlapping support as discussed in the following examples.

Example 2.1.1 (Wasserstein distance between 1D Gaussian distributions)**.** We illustrate the Wasserstein distance on a simple example between two 1D Gaussian distributions $\mathcal{N}(m_s, \sigma^2)$ and $\mathcal{N}(m_t, \sigma^2)$. The value of the W_1 distance in this case is $|m_s - m_t|$, which obviously increases with the separation $|m_s - m_t|$ of the two distributions. As a comparison to another classical divergence, the total variation converges to 2 when $|m_s - m_t| \to \infty$.

Example 2.1.2 (Wasserstein distance between 1D uniform distributions)**.** We illustrate the Wasserstein distance on a simple example between two 1D uniform distributions $\mathcal{U}(m_s - \frac{1}{2}, m_s + \frac{1}{2})$ and $\mathcal{U}(m_t - \frac{1}{2}, m_t + \frac{1}{2})$. The value of the W_1 distance in this case is also $|m_s - m_t|$. But in this case the well known Kullback Leibler (KL) is not defined when $|m_s - m_t| > 1$, and the total variation is exactly equal to 2 and hence has a null gradient. This property has been a key reason in the recent interest in ML since a lot of learning methods rely on gradient descent for fitting empirical distributions.

Special case between 1D distributions As the two examples above suggest the Wasserstein distance in 1D can be easily solved when one have access to its cumulative distribution functions. When $c(x, y)$ is a strictly convex and increasing function of $|x - y|$ the OT plan respects the ordering of the elements and the solution is given by the monotone rearrangement of μ_s onto μ_t. The value of the W_1 Wasserstein distance with ground cost c is

$$W_1(\mu_s, \mu_t) = \int_0^1 c(F_{\mu_s}^{-1}(q), F_{\mu_t}^{-1}(q))dq \tag{2.12}$$

where F_μ is the cumulative distribution function of μ and $F_\mu^{-1}(q)$, $q \in [0, 1]$ is the quantile function such that $F_\mu^{-1}(q) = \inf\{x \in \mathbb{R} : F_\mu(x) \geq q\}$. This close form can be easily approximated when using empirical distributions and can be computed with a $O(n \log(n))$ sorting. This very efficient solution led to the proposition of Sliced Radon Wasserstein in [Bonneel 2014] that consists in computing the expected value of the Wasserstein distance over the 1D projections integrated on the unit sphere $\mathbb{S}^{d-1}$ (often from finite random directions). The Sliced Radon Wasserstein has been used as a measure of fit for generative networks thanks to its efficient computation [Kolouri 2018, Deshpande 2018, Liutkus 2018].

Bures-Wasserstein distance between Gaussien distributions When $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2/2$, $\mu_s = \mathcal{N}(\mathbf{m}_s, \Sigma_s)$ and $\mu_t = \mathcal{N}(\mathbf{m}_t, \Sigma_t)$ the Wasserstein distance can be expressed as:

$$W_2^2(\mu_s, \mu_t) = ||\mathbf{m}_s - \mathbf{m}_t||_2^2 + \mathcal{B}(\Sigma_s, \Sigma_t)^2 \tag{2.13}$$

where $\mathbb{B}(,)$ is the so-called Bures metric:

$$\mathcal{B}(\Sigma_s, \Sigma_t)^2 = \text{trace}(\Sigma_s + \Sigma_t - 2(\Sigma_s^{1/2}\Sigma_t\Sigma_s^{1/2})^{1/2}). \tag{2.14}$$

This distance is also discussed in [Peyré 2017, Remark 2.29]. Note that when only empirical estimates (for means and covariances) for the distributions are available, using these estimates in the equation above gives an approximation with a sample complexity of $O(n^{-1/2})$ as shown in the supplementary material of [Rakotomamonjy 2018].

Wasserstein Barycenters One fascinating aspect of the Wasserstein distance is the geometrical space that is induced by the distance. This geometry has been investigated by [McCann 1997] who proposed an interpolation between two distributions minimizing the Wasserstein distance. This interpolation can be generalized to what is known as Wasserstein barycenter [Agueh 2011] between distributions $\{\mu_i\}_i$ expressed as follows

$$\bar{\mu} = \arg\min_{\mu} \quad \sum_{i=1}^{n} \lambda_i W_p^p(\mu_i, \mu) \tag{2.15}$$

where $\lambda_i > 0$ and $\sum_i^n \lambda_i = 1$. The barycenter can be seen as a Frèchet mean with respect to the Wasserstein distance. The McCann interpolant is a special case of (2.15) where n=2 and $\boldsymbol{\lambda} = [1 - t, t]$ with $0 \leq t \leq 1$ [McCann 1997].

Example 2.1.3 (Wasserstein barycenter between 1D Gaussians)**.** The W_2 barycenter between the two Gaussian distributions $\mathcal{N}(m_s, \sigma^2)$ and $\mathcal{N}(m_t, \sigma^2)$ from Example 2.1.1 has a very simple form. For interpolation time t as defined above, the barycenter is the Gaussian distribution $\mathcal{N}(tm_t + (1-t)m_s, \sigma^2)$. We can see that the Gaussian is displaced along tey axis. The fact that the distribution interpolates implies that only one mode will be in the barycenter as opposed to the Euclidean barycenter that can have two modes depending on the variance σ^2.

2.2 Discrete distribution and entropic regularization

This section discusses the numerical aspects of OT computation for discrete distributions. It also introduces entropic regularization of OT and the corresponding optimization algorithms. Finally we talk about more general regularizations that can encode different information and discuss the recent stochastic optimization approaches.

2.2.1 Discrete Optimal Transport

In this subsection we focus on the OT problem for discrete distributions since it is the most common situation in machine learning.

Discrete distributions In the following we will use discrete distributions of the form

$$\mu = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}, \quad \mu_s = \sum_{i=1}^{n_s} a_i \delta_{\mathbf{x}_i^s}, \quad \mu_t = \sum_{j=1}^{n_t} b_j \delta_{\mathbf{x}_j^t} \tag{2.16}$$

where $\mathbf{x}_i^s, \mathbf{x}_j^t \in \Omega^2, \forall i, j$ and $\mathbf{a} \in \Sigma_{n_s}$ $\mathbf{b} \in \Sigma_{n_s}$ and $\Sigma_n = \{(a_i)_i \geq 0; \sum_{i=1}^n a_i = 1\}$ is the simplex polytop. This formulation can represent both the Lagrangian formulation (that is called empirical distribution in this document) where both the support x_i and the weights a_i are free (quotient space: Ω^n, Σ_n) and the Eulerian formulation (called histogram) where the support x_i is fixed (such as a regular grid for instance) and the information is encoded in $\mathbf{a}$ (Quotient space: Σ_n). Note that for a ML dataset with IID samples, the Lagrangian formulation with uniform weights $a_i = \frac{1}{n}, \forall i$ is often used. Finally the samples $\mathbf{x}_i$ can be stored into matrices $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$ (resp. $\mathbf{X}_s, \mathbf{X}_t$).

Primal problem The optimal transport problem between distributions μ_s and μ_t can be expressed as the following linear program

$$\mathbf{T}_0 = \underset{\mathbf{T} \in \mathcal{P}(\mathbf{a}, \mathbf{b})}{\arg\min} \quad \langle \mathbf{T}, \mathbf{C} \rangle_F \tag{2.17}$$

where $\mathbf{C}$ is a cost matrix with $C_{i,j} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$, $\langle \mathbf{T}, \mathbf{C} \rangle_F = \sum_{i,j} T_{i,j} C_{i,j}$ is the Frobenius scalar product and the marginal linear constraints are

$$\mathcal{P}(\mathbf{a}, \mathbf{b}) = \left\{ \mathbf{T} \in (\mathbb{R}^+)^{n_s \times n_t} \,|\, \mathbf{T}\mathbf{1}_{n_t} = \mathbf{a}, \mathbf{T}^T \mathbf{1}_{n_s} = \mathbf{b} \right\} \tag{2.18}$$

The solution is sparse with at most $n_s + n_t - 1$ non-zero coefficients in the transport matrix $\mathbf{T}_0$. The matrix is actually very interpretable because each line i describes how the mass from bin a_i is transported onto the bins b_j of the target distribution. Optimization problem (2.17) is usually solved in the equivalent dual problem.

Example 2.2.1. This problem can be seen as the search for the optimal transport between production sites i producing an amount a_i and the stores j selling an amount b_j when the cost of moving a unit of mass from i to j is $C_{i,j}$. The constraints imply that all the production from the production sites is transported to the stores and each store receives exactly the required amount b_i. The transport company wants to maximize its profit by minimizing the transport cost.

Dual formulation The dual formulation of the convex discrete OT problem (2.17) is

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^{n^s},\boldsymbol{\beta}\in\mathbb{R}^{n^t}} \quad \boldsymbol{\alpha}^T\mathbf{a}+\boldsymbol{\beta}^T\mathbf{b} \tag{2.19}$$

$$\text{s.t.} \quad \alpha_i+\beta_j \leq C_{i,j} \quad \forall i,j \tag{2.20}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the dual potentials (n_s+n_t coefficients with $n_s n_t$ constraints). The problem above can be cast as a Network Flow problem and solved with a Network Simplex algorithm [Peyré 2017, Section 3.5] such as the one implemented in [Bonneel 2011] of complexity $O(n^3\log(n))$ when $n=n_s=n_t$. In the particular case where the weights in $\mathbf{a}$ and $\mathbf{b}$ are uniform and the number of samples is the same, the OT matrix is a permutation matrix and other approaches such as the Auction Algorithm [Bertsekas 1981] can also be used. This complexity is clearly a limit to ML applications that sometimes require to learn from very large datasets.

Example 2.2.2. This dual can be interpreted as switching from the transport company view (where the mass is moved to minimize cost) to the view of the individuals that sell and buy some amount of mass and want to maximize their profit. The potentials can be seen as the unit price where the goods are sold ($\boldsymbol{\alpha}$) and bought ($-\boldsymbol{\beta}$) at the source and target. Intuitively both the seller and buyer want to maximize (max in (2.19)) their profit but the difference between the selling α_i and buying $-\beta_j$ cannot be more than the transport price $C_{i,j}$ (constraints in (2.20)).

c-transform and semi-dual Similarly to the continuous case, one can define the *c-transform* for the dual potential $\boldsymbol{\alpha}$ as

$$(\boldsymbol{\alpha}^c)_j = \min_i \quad C_{i,j}-\alpha_i \tag{2.21}$$

The optimization problem can then be expressed only as a function of $\boldsymbol{\alpha}$ in the semi-dual formulation

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^{n^s}} \quad \boldsymbol{\alpha}^T\mathbf{a}-\sum_j b_j \max_i(\alpha_i-C_{i,j}) \tag{2.22}$$

This formulation can also be interpreted as the fact that for a given selling price α the buyer j will always buy at the minimum price ensuring that the transport price is covered (the *c-transform*).

Discrete Wasserstein distance For discrete distributions, we will use the following notation for the Wasserstein distance

$$W_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \langle \mathbf{T}_0, \mathbf{C}\rangle_F \tag{2.23}$$

where $\mathbf{C}$ is the cost matrix and $\mathbf{T}_0$ is the solution of optimization problem (2.17). Note that this is a slight abuse of notation since we can use any cost matrix $\mathbf{C}$ and we discard the parameter p used in the continuous case (2.11). This formulation will also be used for regularized OT but using $\mathbf{T}$ as the solution of the regularized OT problem.

Sub-gradients for the discrete Wassersein distance The Wasserstein distance is the solution of a linear program and is not differentiable. But any solution $\boldsymbol{\alpha}^*$ of the dual problem (2.20) is a sub-gradient of $W_{\mathbf{C}}(\mathbf{a},\mathbf{b})$ with respect to the source distribution weights $\mathbf{a}$. It can then be used for instance to estimate a Wasserstein Barycenter as in [Cuturi 2014a]. The gradient *w.r.t.* the position of the diracs when working with empirical distributions requires to differentiate through c (See [Cuturi 2014a, sec. 4.3]).

Convergence of Wasserstein distance One very interesting property of the Wasserstein distance is its ability to measure a meaningful distance between distributions that do not have a shared support. One particular case of this is the Wasserstein distance between a distribution ν and its empirical counterpart

$\nu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ where x_i are IID realizations from ν. In this case the Wasserstein distance has been proven to converge with the following speed [Fournier 2015]

$$E[W_1(\nu, \nu_n)] = O\left(n^{-\frac{1}{d}}\right), \tag{2.24}$$

where we recall that d is the dimensionality of Ω. This convergence speed is particularly slow for high dimensional distributions and seems to pale in comparison to kernel MMD [Gretton 2012] that is known to be $O\left(n^{-\frac{1}{2}}\right)$, independent from d. Sharpest bounds of $O\left(n^{-\frac{1}{s}}\right)$ with $s \leq d$ depending on geometrical properties of the distribution ν have been obtained in [Weed 2017].

2.2.2 Entropic regularization

A major recent result of optimal transport that allowed its use in numerous ML applications has been the seminal works of Cuturi [Cuturi 2013] about entropic regularized optimal transport.

Entropic regularized OT in the primal Cuturi proposed to solve the following optimization problem

$$\min_{\mathbf{T}\in\mathcal{P}(\mathbf{a},\mathbf{b})} \quad \langle \mathbf{T}, \mathbf{C}\rangle_F + \lambda\Omega_e(\mathbf{T}) \tag{2.25}$$

with $\lambda \geq 0$ the regularization parameter and $\Omega_e(\mathbf{T}) = \sum_{i,j} T_{i,j}(\log(T_{i,j}) - 1)$ the entropic regularization term. The second term promotes a smooth joint distribution $\mathbf{T}$. Interestingly, problem 2.25 is equivalent to regularizing with the Kullback-Leibler (KL) divergence between $\mathbf{T}$ and a matrix of ones. Note that in this document we will use the regularization defined above but other equivalent formulations use a KL between $\mathbf{T}$ and the independent distribution $\mathbf{a}\mathbf{b}^\top$ instead of a uniform matrix.

This regularization makes the problem strongly convex and less sensitive to small changes in the distribution. It will make the optimization problem easier to solve as discussed in the following. But it will spread the mass in the OT matrix, making it non-sparse when $\lambda > 0$ which can make the interpretation of the OT matrix more difficult.

Sinkhorn-Knopp and Bregman projections The problem (2.25) is strictly convex and can be solved with the well known Sinkhorn-Knopp Matrix scaling Algorithm [Sinkhorn 1967]. Indeed the solution can be expressed as

$$\mathbf{T} = \mathrm{Diag}(\mathbf{u})\mathbf{K}\,\mathrm{Diag}(\mathbf{v}), \quad \text{with} \quad \mathbf{K} = \exp(-\mathbf{C}/\lambda) \tag{2.26}$$

where the exponential function is taken element-wise.

The Sinkhorn-Knopp Algorithm (see Alg.2.1) updates iteratively $\mathbf{u}$ and $\mathbf{v}$ by projecting on the left and right marginals until convergence. It is a very simple algorithm that relies only on matrix multiplications and point-wise operations [Cuturi 2013]. It can be easily accelerated on CPU and GPU and can solve in parallel several OT problems using the same cost matrix with different marginals ($\mathbf{u}$ and $\mathbf{v}$ become matrices but the algorithm is similar). This algorithm has been shown to have a linear convergence [Altschuler 2017] and is known to be very fast especially for a large regularization since the convergence coefficient depends directly on λ. Note that one can also perform Greedy update by updating the components of $\mathbf{u}$ and $\mathbf{v}$ that are farthest from convergence and proposed in the Greenkhorn Algorithm of [Altschuler 2017].

The SK algorithm is in fact a special case of Bregman Projections as discussed in details in [Benamou 2015]. Problem (2.25) can indeed be reformulated as

$$\min_{\mathbf{T}\in\mathcal{P}(\mathbf{a},\mathbf{b})} \quad KL(\mathbf{T}|\mathbf{K}) \tag{2.27}$$

where $KL(\mathbf{T}|\mathbf{K}) = \sum_{i,j} T_{i,j}\log(\frac{T_{i,j}}{K_{i,j}}) - T_{i,j} + K_{i,j}$ is the Kullback-Leibler divergence between the matrices $\mathbf{T}$ and $\mathbf{K}$. The Bregman Projection algorithm can then be applied by projecting alternatively the current

Algorithm 2.1 Sinkhorn-Knopp Algorithm (SK). $\oslash$ denotes the point-wise division.

Require: $\mathbf{a}, \mathbf{b}, \mathbf{C}, \lambda$
$\mathbf{u}^{(0)} = \mathbf{1}, \mathbf{K} = \exp(-\mathbf{C}/\lambda)$
for i in $1, \dots, n_{it}$ **do**
  $\mathbf{v}^{(i)} = \mathbf{b} \oslash \mathbf{K}^\top \mathbf{u}^{(i-1)}$ // Update right scaling
  $\mathbf{u}^{(i)} = \mathbf{a} \oslash \mathbf{K}\mathbf{v}^{(i)}$ // Update left scaling
end for
return $\mathbf{T} = \text{Diag}(\mathbf{u}^{(n_{it})})\mathbf{K}\,\text{Diag}(\mathbf{v}^{(n_{it})})$

$\mathbf{T}$ matrix in the KL sense onto the left and right marginals. Numerous generalizations of this very simple algorithm, to barycenter or multi-marginal OT are presented in [Benamou 2015].

It is also interesting to note that a relaxed formulation of OT where the constraints are relaxed as a weighted additive term measuring the KL divergence between the marginals of $\mathbf{T}$ and the objective marginals $\mathbf{a}, \mathbf{b}$ [Frogner 2015, Liero 2018, Chizat 2018]. This very elegant formulation can also be solved by Bregman projections and allows to compute unbalanced OT when the total masses of $\mathbf{a}$ and $\mathbf{b}$ are different [Benamou 2003].

Entropic regularized OT in the dual Problem (2.25) can also be reformulated in the dual as

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^{n^s}, \boldsymbol{\beta} \in \mathbb{R}^{n^t}} \quad \boldsymbol{\alpha}^T \mathbf{a} + \boldsymbol{\beta}^T \mathbf{b} - \lambda \sum_{i,j} \exp((\alpha_i + \beta_j - C_{i,j})/\lambda) \tag{2.28}$$

where the optimal dual variables have a direct relation with the scaling vectors in Sinkhorn $(\mathbf{u}^\star, \mathbf{v}^\star) = (\exp(\boldsymbol{\alpha}^\star/\lambda), \exp(\boldsymbol{\beta}^\star/\lambda))$. Note that thanks to the regularization, the dual problem is unconstrained which means that the problem can be solved by any gradient based algorithm such as L-BFGS [Cuturi 2016, Blondel 2018]. Similarly to unregularized OT, one can express a *c-transform* for the regularized OT problem above and find the regularized $\boldsymbol{\alpha}^c_\lambda$ as

$$(\boldsymbol{\alpha}^c_\lambda)_j = \lambda \log(b_j) - \lambda \log \left(\sum_i \exp((\alpha_i - C_{i,j})/\lambda) \right) \tag{2.29}$$

which can be seen as a soft-min that converges to the true minimum (2.21) when $\lambda \to \infty$. The semi dual can also be obtained and the problem solved only *w.r.t.* $\boldsymbol{\alpha}$ [Genevay 2016].

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^{n^s}} \quad \boldsymbol{\alpha}^T \mathbf{a} - \lambda \sum_j b_j \log \left(\sum_i \exp((\alpha_i - C_{i,j})/\lambda) \right) \tag{2.30}$$

The formulation above is a sum of smooth loss functions and can be solved using stochastic optimization [Genevay 2016] as discussed in section 2.3.2.

Sinkhorn distance and Auto-differentiation Now that we have defined the regularized OT problem we discuss how to compute the Wasserstein distance in this case. A first approximation would be to approximate the Wasserstein distance as the optimal value in problem (2.25) such that $W^\lambda_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \langle \mathbf{T}_\lambda, \mathbf{C} \rangle_F + \lambda \Omega_e(\mathbf{T}_\lambda)$. Since the optimization problem (2.25) is strictly convex and differentiable it has a unique gradient *w.r.t.* $\mathbf{a}$ that can be shown to be the dual variable $\boldsymbol{\alpha}^\star = \lambda \log(\mathbf{u})$ [Cuturi 2014a]. But the regularization term will introduce an important smoothing of the gradient [Luise 2018]. A more sensible option is to use the following expression without the entropic term

$$OT^\lambda_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \langle \mathbf{T}_\lambda, \mathbf{C} \rangle_F \tag{2.31}$$

where $\mathbf{T}_\lambda$ is the OT matrix in Equation (2.25). This term has been sometimes called the Sinkhorn loss or sharp Sinkhorn and has been used to estimate barycenters [Luise 2018] and auto-encoders [Patrini 2018].

But it is not a divergence because the regularization $\lambda > 0$ prevent $OT_{\mathbf{C}}^{\lambda}(\mathbf{a}, \mathbf{a})$ to be 0. In order to address this problem [Genevay 2017b] has proposed the following formulation coined Sinkhorn Divergence

$$\widetilde{W}_{\mathbf{C}}^{\lambda}(\mathbf{a}, \mathbf{b}) = W_{\mathbf{C}}^{\lambda}(\mathbf{a}, \mathbf{b}) - \frac{1}{2} W_{\mathbf{C}}^{\lambda}(\mathbf{a}, \mathbf{a}) - \frac{1}{2} W_{\mathbf{C}}^{\lambda}(\mathbf{b}, \mathbf{b}) \tag{2.32}$$

that requires three times the computational complexity of (2.31).

Also note that the OT loss described above in (2.31) is sightly more difficult to optimize since the problem is now a bi-level optimization problem that requires in theory the use of the implicit function theorem to compute a gradient [Luise 2018]. An elegant approach that can be used to optimize over (2.31) and (2.31) is auto-differentiation of the Sinkhorn algorithm which has been proposed in [Genevay 2017b, Flamary 2018]. The individual operations in Algorithm 2.1 are indeed all differentiable and modern auto-differentiation tools can compute the gradient along the Sinkhorn iterations with a small overhead. This last approach has been used to estimate Generative Adversarial Networks [Genevay 2017a].

Regularized Wasserstein Barycenters When using entropic regularization [Benamou 2015] has proposed a very elegant Bregman projection algorithm to estimate efficiently a Wasserstein barycenter. Note that as proposed in [Solomon 2015] when the histograms have a separable structure, support on a regular grid for instance, the operators can be performed with a convolution which can greatly decrease the complexity of the problem. This allowed for applications on 3D images for computing population averages of fMRI images that take into account the geometry of the data [Gramfort 2015, Wang 2018]. A very detailed discussion about Wasserstein barycenters and their numerical estimation is provided in [Peyré 2017, Section 9.2]. Note that [Bigot 2018b, Bigot 2018a] have studied the statistical properties and convergence of the Wasserstein barycenters with entropic regularization.

Convergence speed of Sinkhorn distance The convergence speed of the Sinkhorn Divergence has been investigated first numerically in [Genevay 2017b] that suggested that it was more robust than Wasserstein with a convergence speed independent from d. The theoretical proof was provided in [Genevay 2018] that showed that the Sinkhorn divergence converges better than Wasserstein with speed $O(n^{-\frac{1}{2}})$ albeit depending on the regularization term λ interpolating between Wasserstein and MMD.

2.3 General regularization and stochastic optimization

In this section we discuss a more general framework of regularized OT and the recent development of stochastic optimization to solve the OT problem.

2.3.1 General regularized OT problems

Regularized Optimal Transport Entropy has been up to now the preferred regularization for optimal transport mostly due to its efficient optimization algorithm. But depending on prior knowledge about the distribution other regularization can be used instead. The general optimization problem is of the form

$$\min_{\mathbf{T} \in \mathcal{P}(\mathbf{a}, \mathbf{b})} \quad \langle \mathbf{T}, \mathbf{C} \rangle_F + \lambda \Omega(\mathbf{T}) \tag{2.33}$$

where Ω is a general regularization term. A number of possible regularizations has been investigated in [Dessein 2016] where the resulting distance is called ROT Mover's distance. They showed that the alternative projection used for entropic regularization can be adapted to a large family of regularization.

Quadratic regularization One particular regularization is the squared Frobenius norm of the transport matrix of the form

$$\Omega_F(\mathbf{T}) = \sum_{i,j} T_{i,j}^2 \tag{2.34}$$

This regularization has the advantage that while it also promotes a spreading of the mass, the resulting matrix stays sparse along the regularization path [Blondel 2018] as opposed to entropic regularization that looses sparsity for $\lambda > 0$. This regularization has been used for solving transport over a graph in [Essid 2018] where a Newton-type solver has been proposed to solve the problem. Also note that the dual formulation for this regularization is of the form

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^{n^s},\boldsymbol{\beta}\in\mathbb{R}^{n^t}} \quad \boldsymbol{\alpha}^T\mathbf{a} + \boldsymbol{\beta}^T\mathbf{b} - \frac{1}{4\lambda}\sum_{i,j}\max(0,\alpha_i+\beta_j-C_{i,j})^2 \tag{2.35}$$

which is a smooth function that resembles the SVM Squared Hinge of SVM-Rank loss [Laporte 2014a] and can be solved efficiently with L-BFGS [Blondel 2018]. Finally the quadratic regularization can be generalized to a Mahalanobis regularization, for instance [Ferradans 2014, Flamary 2014d] proposed to regularize the OT matrix using a graph of neighborhood between empirical samples which results in a quadratic regularization *w.r.t.* $\mathbf{T}$.

Group lasso and structured regularization Regularization can make the optimization problem more robust and efficient to solve, but one can also use it to encode additional information available about the data. For instance if we know that the bins in the histogram from the source distribution are somewhat grouped together it becomes sensible to promote the sharing of mass only among those groups. To this end, we proposed in [Courty 2014, Courty 2016a] to use the following regularization

$$\Omega_{p,q}(\mathbf{T}) = \sum_{j,k}\left(\sum_{i\in\mathcal{G}_k} T_{i,j}^p\right)^{\frac{q}{p}} + \lambda_e\sum_{i,j} T_{i,j}(\log(T_{i,j})-1) \tag{2.36}$$

where $\mathcal{G}_k$ contains the non-overlapping index of the bins in group k. This regularization will promote group sparsity for $q \leq 1$ and $p \geq 1$ as discussed more in details in [Courty 2016a]. Note that we keep the entropic regularization term because the non-regularized OT is already sparse and we need to spread the mass in order to make groups appear. In terms of optimization we proposed in [Courty 2014] to use a Majoration-Minimization approach for the non convex case $p = 1, q = \frac{1}{2}$ that consists in solving at each iteration of the algorithm a linearization of the concave group lasso with Sinkhorn. The classical convex group lasso $p = 2, q = 1$ regularization is a little more complex but can be solved using a generalized conjugate gradient as discussed in the following.

The group lasso discussed above supposes that the groups of bins are known a-priori, which is a strong assumption. In [Alvarez-Melis 2017] the authors propose to use the very general framework of submodularity to encode structural information in OT. In practice their formulation allows for the joint estimation of the groups and the OT matrix.

Projections, Conjugate and Generalized Conjugate gradients In this paragraph we discuss the different optimization algorithms that can be used to solve the general regularized OT problem.

First, alternating projection approaches are known to work particularly well on entropic regularization [Benamou 2015]. These alternating projection approaches can be generalized to a wide class of regularization [Dessein 2016] terms. But in the general case, one also needs to project onto the positive orthant (which was not necessary with entropic regularization) so a more general Dikstra Algorithm [Bauschke 2011, Chapter 29] has to be used. Also note that the efficiency of those alternative projection methods depends on the existence of a fast close form projection operator or else it will need to use an iterative solver for each projection [Dessein 2016].

There exists a very simple approach that was proposed in [Ferradans 2014] to solve regularized OT: Conditional Gradient (CG). This approach relies on optimizing at each iteration a linearization of the optimized loss which can be done with a solver for non-regularized OT (See Algorithm 2.2.a). This approach is very general and will even converge for non-convex regularization term [Lacoste-Julien 2016] but solving non-regularized OT problem at each iteration can be computationally intensive.

Algorithm 2.2 Conjugate Gradient (CG) and Generalized Conjugate Gradient (GCG)

(a) Conjugate Gradient (CG)

Require: $\mathbf{a}, \mathbf{b}, \mathbf{C}, \lambda, \Omega$

$\mathbf{T}^{(0)}$ = Solve OT (2.2) with cost matrix $\mathbf{C}$

for i in $1, \dots, n_{it}$ **do**

  $\mathbf{G} = \mathbf{C} + \lambda \nabla \Omega(\mathbf{T}^{(i-1)})$

  $\mathbf{S}$ = Solve OT (2.2) with cost matrix $\mathbf{G}$

  Select step α with $\Delta = \mathbf{S} - \mathbf{T}^{(i-1)}$

$$\min_{\alpha} \left\langle \mathbf{T}^{(i-1)} + \alpha\Delta, \mathbf{C} \right\rangle_F + \lambda\Omega(\mathbf{T}^{(i-1)} + \alpha\Delta)$$

  $\mathbf{T}^{(i)} = \mathbf{T}^{(i-1)} + \alpha\Delta$

end for

return $\mathbf{T}^{(n_{it})}$

(b) Generalized Conjugate Gradient (GCG)

Require: $\mathbf{a}, \mathbf{b}, \mathbf{C}, \lambda, \Omega_{gcg} = \Omega_0 + \lambda_e \Omega_e$ (Eq. (2.37))

$\mathbf{T}^{(0)}$ = Solve Sinkhorn with cost matrix $\mathbf{C}$

for i in $1, \dots, n_{it}$ **do**

  $\mathbf{G} = \mathbf{C} + \lambda \nabla \Omega_0(\mathbf{T}^{(i-1)})$

  $\mathbf{S}$ = Solve Sinkhorn with cost matrix $\mathbf{G}$

  Select step α with $\Delta = \mathbf{S} - \mathbf{T}^{(i-1)}$

$$\min_{\alpha} \left\langle \mathbf{T}^{(i-1)} + \alpha\Delta, \mathbf{C} \right\rangle_F + \lambda\Omega_{gcg}(\mathbf{T}^{(i-1)} + \alpha\Delta)$$

  $\mathbf{T}^{(i)} = \mathbf{T}^{(i-1)} + \alpha\Delta$

end for

return $\mathbf{T}^{(n_{it})}$

This is the reason why we proposed to use a Generalized conditional gradient (GCG) [Bredies 2009] to solve the optimization problem that regularized by a general term plus an entropic term in [Courty 2016a, Rakotomamonjy 2015] of the form

$$\Omega_{gcg}(\mathbf{T}) = \Omega_0(\mathbf{T}) + \lambda_e \Omega_e(\mathbf{T}) \tag{2.37}$$

where $\Omega_e(\mathbf{T}) = \sum_{i,j} T_{i,j}(\log(T_{i,j}) - 1)$ is the entropic regularization of (2.25). GCG allows to linearize at each iteration of the algorithm only part of the objective function, which allows us to linearize Ω_0 but keep the entropic term when solving the inner problem. In practice it means that each iteration of the algorithm can use the Sinkhorn algorithm that can be much quicker on some problems. It has been successfully used for group lasso regularization [Courty 2016a, Rakotomamonjy 2015] and has similar complexity to the MM algorithm used for non-convex group lasso but still need to perform a line search which can incur computational cost.

2.3.2 The rise of stochastic optimization

Stochastic optimization has been used with tremendous results for large scale optimization in ML, in particular to train neural networks on large scale datasets [Bottou 2010]. The main idea is that when optimizing a complex function that is a sum of numerous functions, such as empirical loss on a dataset for instance, one do not need to compute an exact gradient on the whole dataset at each update of the parameter but use a sensible (and fast to compute) approximation. It is natural to try and apply those approaches to solve the OT problem in order to speedup computation for large problems.

Stochastic optimization in the semi-dual The first stochastic optimization algorithm applied to OT has been proposed by [Genevay 2016]. In this work they express the dual (2.28) and semi-dual (2.30) problem for entropic regularized OT and show that it has a structure amenable to stochastic optimization. They propose to use Stochastic Average Gradient (SAG) in the semi-dual which is proven to converge linearly to the solution of the strongly convex optimization problem.

Stochastic optimization in the dual The seminal works of [Genevay 2016] also proposed to solve the entropic OT problem between continuous distributions in the dual. To this end they propose to estimate dual potentials in a Reproducing Kernel Hilbert Space (RKHS). This approach is very elegant, and can handle both continuous and semi-discrete formulations, but since it relies on kernel machines it does not scale well on large datasets.

A large scale approach using neural networks to solve the non-regularized dual problem has been proposed in [Arjovsky 2017]. In a nutshell, they proposed to solve the non-regularized OT problem (2.7) with $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ directly in the dual and use a neural network to estimate the dual potential ϕ (the

second potential is a closed form $\phi^c = -\phi$ in this case). This elegant formulation is hard to implement in practice since the constraint on the Lipschitz constant $\phi \in \text{Lip}^1$ cannot be enforced on a neural network so the authors had to use a coarse approximation based on parameter thresholding. This meant that one cannot recover the Wasserstein distance at the end of the optimization due to an unknown constant. This approach and its numerous extensions are discussed more in detail in Chapter 5.

Finally the approach discussed above is limited to the Euclidean cost function and only solves an approximate non-regularized OT. We proposed in [Seguy 2018] to solve the regularized dual problem similarly to [Genevay 2016] but using neural networks to estimate the dual potential instead of kernel machines. For both entropic and Frobenius regularization, this dual problem is indeed unconstrained and can be solved with stochastic gradient on the neural network parameters. Note that this approach also works for any ground metric and the estimated transport matrix converges weakly to the non-regularized OT when the regularization goes to 0 [Seguy 2018, Theorem 1].

Stochastic Wasserstein barycenters Several recent works also address the problem of estimating Wasserstein barycenter on large datasets with stochastic optimization. In particular [Staib 2017] proposed a stochastic projected gradient ascent that can be implemented in parallel and can handle streaming data. The recent works of [Claici 2018] also propose a stochastic optimization of Wasserstein barycenters but also allow the estimation of the support of the barycenters. These approaches allow the estimation of un-regularized barycenters, which can be interesting of large datasets when the distributions are "peaky" since as discussed above the regularization can have a "low pass" effect that can lead to loss of information.

Chapter 3

Mapping with Optimal Transport

Contents

This chapter discusses the problem of estimating an OT mapping from empirical distributions. It will also introduce one of our contribution to domain adaptation that uses the estimated mapping.

3.1 Optimal transport mapping estimation

While mathematicians have investigated in depth the properties and existence of optimal transport mapping between continuous distributions [Brenier 1991], they rarely asked the question of estimating a mapping from empirical distributions since as discussed in the previous chapter, this mapping might not even exist.

But in machine learning, we often suppose that we have access to a finite sampling from an unknown continuous distribution. Which means that even though there is no mapping between two given samples, there might exist a smooth mapping between the underlying distributions. The question of estimating a reasonable continuous mapping from the empirical distribution is then a classical learning problem and this mapping can be used in numerous applications as discussed in the next sections.

3.1.1 Barycentric mapping

A very elegant approach for mapping discrete samples has been proposed and used in [Reich 2013] for sampling in Sequential Monte Carlo Methods. It has also been used in [Ferradans 2014] to adapt the colors of the pixels between images. They proposed to use regularized OT to map the RGB color values between two images. The main idea is that since the transport matrix $\mathbf{T}$ contains the information of mass displacement, it can be used to find a reasonable position for a given displaced sample. For a given source sample $\mathbf{x}_i^s$ and the OT matrix $\mathbf{T}$ its transported position can be estimated by solving the following optimization problem

$$\widehat{m}_{\mathbf{T}}(\mathbf{x}_i^s) = \arg\min_{\mathbf{x}} \quad \sum_j T_{i,j} c(\mathbf{x}, \mathbf{x}_j^t). \tag{3.1}$$

The approximate mapping above gives the barycenter, *w.r.t.* the cost c, of the target samples weighted by the OT matrix.

Barycentric mapping with $c(x,y) = \|x-y\|^2/2$ One particular case of the formulation 3.1 is when using squared Euclidean loss as proposed in [Ferradans 2014]. In this case the barycenter is a weighted sum of the target samples:

$$\widehat{m}_{\mathbf{T}}(\mathbf{x}_i^s) = \frac{1}{\sum_j T_{i,j}} \sum_j T_{i,j}\mathbf{x}_j^t = \arg\min_{\mathbf{x}} \quad \sum_j T_{i,j}\|\mathbf{x}-\mathbf{x}_j^t\|^2. \tag{3.2}$$

It has nice geometric properties, for instance the displaced samples will always be inside the convex hull of the samples in the target distribution. It is also very efficient to compute since it can be done with a matrix product between $\mathbf{T}$ and the target sample matrix $\mathbf{X}_t$ with the following expression:

$$\widehat{\mathbf{X}}_s = \text{Diag}\left(\frac{1}{\mathbf{T}\mathbf{1}_{n_t}}\right)\mathbf{T}\mathbf{X}_t \tag{3.3}$$

where the division in the diagonal is done component-wise. The application is made even faster when the OT matrix $\mathbf{T}$ is sparse. Note that a similar operation is performed when updating the position of the Wasserstein barycenter in [Cuturi 2014a].

Barycentric effect and limits The approach proposed above is very elegant and fast when the OT matrix $\mathbf{T}$ is already estimated. It also converges to the true Monge mapping for the squared Euclidean loss when the number of samples goes to infinity as we proved in [Seguy 2018, Theorem 2]. But it has a few limits that will be discussed here.

First, the use of regularization has some important effect on the barycentric mapping. Indeed regularization, has the effect of spreading the mass between all target samples which means that all samples will impact the position of the displaced samples. It can be seen in practice as a shrinkage effect where all displaced samples will converge to the center of mass for a large regularization. While this can be seen as denoising (averaging target samples), it also makes the choice of the regularization term very sensitive.

Second this approach allows to find a displacement only for samples already in the distributions when $\mathbf{T}$ was estimated. In this sense it does not allow "out of sample" prediction because if new samples are available in the source and target distributions, the OT problem has to be solved again. It can also be very complex to solve the OT problem for a large number of samples. These problem have been handled elegantly in [Ferradans 2014] using k-means clustering as a subsampling of the data. It greatly decreases the number of samples of the distributions (quantification in color space) but some information is obviously lost. In order to have a better mapping, they proposed to store the displacement for each pixel to its cluster and apply this displacement after transport. This solution is limited since it does not provide a continuous mapping and might have artifacts in large dimension.

3.1.2 Continuous mapping estimation

The problem of estimating a continuous mapping that provides out of sample prediction has been surprisingly seldom investigated in the literature. In this section we introduce what is to our knowledge the first practical results in this field and discuss two contributions we proposed recently. It is also a difficult problem as suggested by the theoretical results in [Hütter 2019] that exhibit a minimax bound in $O(n^{-1/d})$ that is similar to the estimation error of Wasserstein distance.

Regression on the barycentric mapping In their works, [Stavropoulou 2015] proposed a very elegant two-step approach for estimating a continuous mapping from discrete distributions. The first step consists in estimating the optimal transport matrix $\mathbf{T}$ between the discrete distributions. The second step consists in estimating a mapping m on a polynomial basis that approximates the barycentric mapping using $\mathbf{T}$

$$\min_m \quad \sum_i \|m(\mathbf{x}_i^s) - \hat{m}_{\mathbf{T}}(\mathbf{x}_i^s)\|^2 \tag{3.4}$$

where $\hat{m}_{\mathbf{T}}$ is the barycentric mapping (3.1) applied to all samples and m is a linear weighting of polynomial basis. The problem can be obviously solved using a least square solver. One limit of their approach is that they did not perform regularization of m which can lead to over-fitting when the number of samples in the distribution is small.

Joint OT and mapping estimation In order to address the problem of estimating an OT mapping when only a small number of samples is available, we proposed to use regularization [Perrot 2016]. The main idea is to estimate simultaneously the OT matrix and the mapping approximating its barycentric mapping with

$$\min_{m\in\mathcal{H},\mathbf{T}\in\mathcal{P}} \quad \langle \mathbf{T},\mathbf{C}\rangle_F + \sum_i \|m(\mathbf{x}_i^s) - \hat{m}_{\mathbf{T}}(\mathbf{x}_i^s)\|^2 + \lambda\|m\|_{\mathcal{H}}^2 \tag{3.5}$$

where $\mathcal{H}$ can be the set of linear functions, a Reproducible Kernel Hilbert Space (RKHS) and $\lambda > 0$ is a regularization parameter. The joint estimation is interesting because the mapping estimation will have an effect on the transport and vice-versa. While the problem is clearly a least square estimation controlled by the OT matrix, you can also see problem (3.5) as a transport problem regularized by a mapping model. Note that we did not use regularized OT here since the mapping data fitting term can be seen as a regularization for $\mathbf{T}$. Another interest of the regularization is that it can help providing bounds for the transport map approximation (See [Perrot 2016, Section 3.3]).

The problem (3.5) can be solved using an alternate optimization algorithm (block coordinate descent). When updating the transport matrix with a fixed mapping, the problem is a quadratic regularized OT. When updating the mapping one can use any least square solver. The main limit of this approach is that when we work in a RKHS, our mapping is a kernel machine, which is known to scale poorly to a large number of training samples. While we also proposed to use a neural network for m in the paper, the problem of solving the large scale OT in the alternative optimization remains and relies on OT solvers of complexity $O(n^3)$.

Large scale OT and mapping estimation The limited scalability of the previous approach is a problem when trying to learn from very large datasets. So we proposed an efficient stochastic gradient approach in [Seguy 2018] that builds on neural network estimation of the dual potentials. When those optimal dual potentials are estimated, one can use the following close form primal/dual formula to recover the optimal OT matrix :

$$T^\star_{i,j} = \exp((\alpha^\star_i + \beta^\star_j - C_{i,j})/\lambda) \tag{3.6}$$

$$T^\star_{i,j} = \frac{1}{2\lambda}\max(0, (\alpha^\star_i + \beta^\star_j - C_{i,j})) \tag{3.7}$$

where (3.6) is the solution with entropic regularization and (3.7) for quadratic regularization. It is then easy to see that one can directly estimate a mapping m minimizing the loss

$$\min_m \quad \sum_{i,j} T^*_{i,j}\|x^t_j - m(x^s_i)\|^2 \tag{3.8}$$

which again can be estimated using stochastic gradient descent when m is a neural net. This loss is clearly equivalent to minimizing problem (3.4) but stays separable with respect to both i, j which make it more amenable than the close form solution (3.3). Note that it can be easily extended to general losses c which made the approach in [Seguy 2018] the first large scale mapping estimation that can accommodate any ground cost.

Linear Monge mapping When the OT mapping between two distributions is linear, there exists a close form solution relying only on first and second order moments [Flamary 2019, Proposition 1]. This close form is the same as the Monge mapping between Gaussian distributions defined in Equation (2.9) with $m(\mathbf{x}) = \mathbf{A}(\mathbf{x} - \mathbf{m}_s) + \mathbf{m}_t$ where $\mathbf{A}$ is defined in Equation (2.10). We have investigated the quality of estimation of the linear Monge mapping when using empirical estimates for the moments and have proven an expected mapping error of $O(n^{-1/2})$.

3.2 Optimal Transport for Domain Adaptation (OTDA)

Now that we discussed how to approximate a Monge mapping from empirical distributions, we show how those mappings can be used in machine learning. The first application of these mappings to the best of our knowledge has been done in [Ferradans 2014]. They treat images as empirical distributions of the pixels in 3D (color space) and propose to map those pixels between distributions/images. This process also called color grading will transport the colors of one image onto the other. This very elegant use of the Monge map can be seen as Domain Adaptation where a dataset is processed through a Monge map to fit another dataset.

We discuss in the next sections our contribution that we called Optimal Transport for Domain Adaptation (OTDA) where we use a similar approach to perform domain adaptation for classification problem.

3.2.1 Principle of OTDA

Domain Adaptation (DA) is a problem in machine learning that is part of the larger Transfer Learning family. The main problem that is addressed in unsupervised DA is how to predict classes on a new (target) dataset that is different from the available (source) training dataset. In order to train a classifier f that works well on the target data, we have access to samples (x_i^s, y_i^s) drawn from the source joint feature/target distribution $\mathcal{P}_s$ (whose feature marginal is defined as μ_s) and only feature samples x_j^t from the marginal μ_t of the joint target distribution $\mathcal{P}_t$.

Domain adaptation in the literature There exist numerous approaches for domain adaptation and providing a full state of the art is beyond the objective of this document but we will discuss shortly the main approaches before introducing our contribution.

First there has been a lot of work in DA based on re-weighting schemes [Sugiyama 2008]. The main idea is that if the distribution of the data is different between source and target, one can re-weight the samples in the source to compensate for this discrepancy. A classifier can then be estimated by minimizing the weighted source distribution. These approaches have been shown to work very well in numerous case but a classic failure scenario would be when the distributions do not overlap.

Another type of DA approaches can be categorized as subspaces methods. When the datasets are high dimensional, one can suppose that there exists a subspace that is discriminant in both source and target domains. This subspace can then be used to train a robust classifier. A standard approach is to minimize the divergences between the two projected distributions [Si 2010]. Note that since source labels are often available, one can also include this knowledge in the subspace estimation (to conserve discrimination after projection) [Long 2014].

Finally the last approaches aim at aligning the source and target distribution through a more complex representation than linear subspace. [R. Gopalan 2014] propose to align the distributions by following the geodesic between source and target distributions. Geodesic flow kernel [Gong 2012] aim at adapting distribution using a projection in a Grassmann manifold and computing kernels using the geodesic flow in this manifold. Note that both methods above aim at finding a way to compensate for the change in distribution between the source and target domain. The same philosophy has been investigated for neural networks where the feature extraction aims at being indistinguishable between the domain using Domain Adversarial Neural Network (DANN) [Ganin 2016]. Other approaches have aimed at minimizing a divergence between the distributions in the NN embedded space using covariance matrix alignment (CORAL) [Sun 2016], minimization of the Maximum Mean Discrepancy (MMD) [Tzeng 2014] or Wasserstein in the embedded space [Shen 2018]. A recent survey in the domain of visual adaptation that has been very active recently is given in [Csurka 2017].

Three-step adaptation with OT Our contribution to domain adaptation supposes that there exists a transformation between the distributions. So we try and estimate this transformation in order to adapt the source domain to the target before learning the classifier. While there is in theory a very

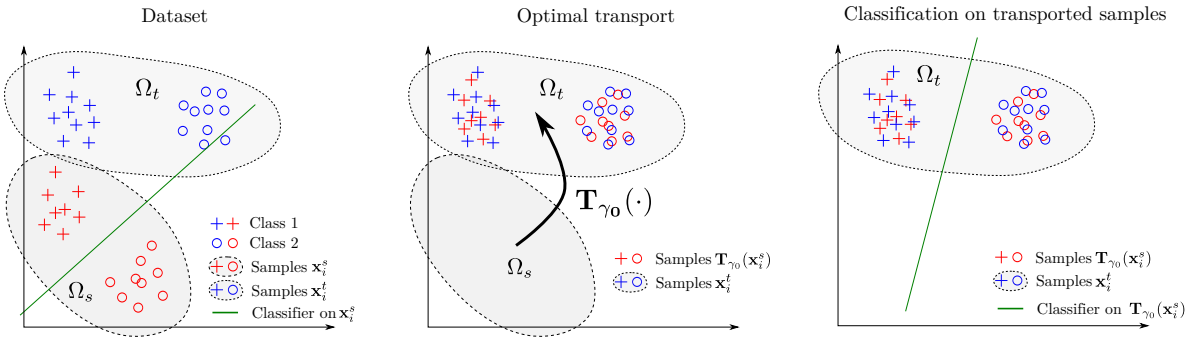


Figure 3.1: Illustration of the principle of OTDA. (left) an Optimal Transport mapping is estimated between the source (red) and target (blue) distributions (center) the mapping is applied to the source samples. (right) a classifier is estimated on the displaced source samples.

large number of mappings of the data that can explain the transformation, we choose to use the optimal transport Monge mapping since it is the one corresponding to the least effort (a sound approach often found in nature). Our Optimal Transport for Domain Adaptation (OTDA) [Courty 2016a] approach consists in three steps:

1. Estimate a mapping $\hat{m}$ between source and target feature distribution.
2. Apply the mapping $\hat{m}$ to the source samples (they keep their labels).
3. Train a classifier on the displaced source samples $(\hat{m}(x_i^s), y_i^s)$.

Those three steps are illustrated in Figure 3.1. This approach has been received well in the DA community because it is very simple and seems to work well in practice. In the next section we will discuss the theoretical assumptions and limits and extensions of the approach.

Assumptions and theory We have defined several assumptions required for OTDA to perform well in [Courty 2016a]. First we suppose that the difference between the distributions is due to a push-forward mapping in the feature space such that $\mu_t = m\#\mu_s$. In addition to this assumption we suppose that the labels are preserved through the mapping such that $\mathcal{P}_s(x, y) = \mathcal{P}_t(m(x), y)$. These two assumptions are reasonable and correspond to a number of real life situations such as a change in the acquisition conditions, sensor drifts, thermal noise in signal processing. It also implies that the expected loss of a given classifier f on the target domain can be computed from the source domain with

$$R_t(f_t) := \mathbb{E}_{(x,y)\sim\mathcal{P}_t}[L(y, f_t(x))] = \mathbb{E}_{(x,y)\sim\mathcal{P}_s}[L(y, f_t(m(x))] =: R_s(f_t \circ m) \tag{3.9}$$

where R_s (resp. R_t) is the expected risk in source (resp target) domains and L is a loss of Lipschitz constant M_L *w.r.t.* the x variable ($M_L = 1$ for SVM Hinge loss for instance independently of the class y). If in addition we suppose that the classifier f_t has a Lipschitz constant M_f then we can bound the expected generalization error on the target distribution with

$$\begin{aligned} R_t(f_t) &= \mathbb{E}_{(x,y)\sim\mathcal{P}_s}[L(y, f_t(\hat{m}(x) - (\hat{m}(x) - m(x)))] \\ &\leq \mathbb{E}_{(x,y)\sim\mathcal{P}_s}[L(y, f_t(\hat{m}(x))] + M_f M_l \mathbb{E}_{x\sim\mu_s}[\|\hat{m}(x) - m(x)\|] \end{aligned} \tag{3.10}$$

which clearly justifies step 3 in our approach that consists in estimating a classifier on the transported samples (and hopefully minimize the error in target). Note that the left term above encodes the mapping estimator $\hat{m}$ and is hard to bound in practice. The term on the right measures the convergence of the estimated Monge mapping toward the true one. The theoretical works of [Hütter 2019] provide a convergence speed of $O(n^{-1/d})$. Also note that [Seguy 2018, Theorem 2] proves a weak convergence of

the barycentric mapping to the true mapping which suggests that the second term can be very small with a large number of samples.

When one has access to a continuous mapping it might also be a good idea to train a good classifier f_s in the source domain and use m^{-1} to bring the new samples from target providing a target classifier $f_t = f_s \circ m^{-1}$. In this case we expressed in [Flamary 2019] the generalization error as

$$R_t(f_s \circ \hat{m}^{-1}) \leq R_s(f_s) + M_f M_L \mathbb{E}_{x\sim\mu_s}\left[\|\hat{m}^{-1}(m(x)) - \hat{m}^{-1}(\hat{m}(x))\|\right] \tag{3.11}$$

If $\hat{m}(\mathbf{x}) = \hat{\mathbf{A}}(\mathbf{x} - \hat{\mathbf{m}}_s) + \hat{\mathbf{m}}_t$ is the linear Monge mapping as defined in 2.9 using empirical estimates then we have

$$R_t(f_s \circ \hat{m}^{-1}) \leq R_s(f_s) + M_f M_L \|\widehat{\mathbf{A}}^{-1}\| \, \mathbb{E}_{x\sim\mu_s}\left[\|\mathbf{A} - \hat{\mathbf{A}}\|\right] \tag{3.12}$$

The bounds above are particularly interesting because they suggest that if we can estimate convergent estimators of classifiers in source domain and mapping between distributions independently, we can reach the performance of the Bayes classifier on target data. A generalization bound depending on the number of samples used to train f_s and the number of samples used to estimate $\hat{m}^{-1}$ is provided in [Flamary 2019, Theorem 3].

One limit of the assumptions above is that our approach cannot handle the case of target shift domain adaptation where the ratio of classes change between source and target distributions. Numerical experiments in the domain of remote sensing where the proportion of classes vary between images in [Tuia 2015c] show indeed a dramatic loss of performance for large deviations.

The devil in the regularization As discussed in the previous chapter, exact OT is often very sensitive to the samples and complex to solve. This is why we proposed in our first paper [Courty 2014] to use entropic regularization. It leads to a clear gain in performance when using the barycentric mapping but the OT matrix becomes dense. This is a problem because one might want to keep some sparsity at least between classes in order to avoid a collapsing of the samples.

This is why we proposed to use non convex group lasso introduced in Eq. (2.36) with $p = 1, q = \frac{1}{2}$ in [Courty 2014]. Since we have access only to the labels in the source domain we use those labels to provide groups for each column in the OT matrix. This means that for every target samples of unknown class, we promote group sparsity *w.r.t.* the classes of the source samples, which forces every target sample to "choose" a class among the groups. As discussed in the previous chapter, the non convex problem can be solved using a linearization of the concave part with DCA which consists in performing several Sinkhorn solver until convergence.

Other kind of regularization were later proposed in [Flamary 2014d] and [Courty 2016a]. Similarly to the graph based regularization in [Ferradans 2014], we proposed to regularize the OT matrix with respect to the graph of neighbors of the samples in [Flamary 2014d]. It was particularly interesting when adapting surfaces modeled as a 3D distribution of nodes in order to maintain a reasonable manifolds during the mapping. We also proposed in [Courty 2016a] the GCG algorithm 2.2.b to solve the group lasso with $p = 2, q = 1$ that has the advantage to be convex. Note that all those regularizations provided better performances in the numerical experiments but the group lasso (both convex and non-convex) usually yield the best gain.

Finally we note the generalization of [Alvarez-Melis 2017] that proposed to use a sub-modular regularization term that promotes structures in the OT matrix while simultaneously estimating those structures. This very elegant extension is also proposed with an algorithm to solve the problem but remains computationally intensive.

How to map ? In [Courty 2014,Courty 2016a], we originally proposed to perform barycentric mapping in order to map the source samples. This is obviously limited due to the lack of out of sample mapping. [Perrot 2016] proposed a first step in estimating a continuous mapping that can displace new samples but was limited to small to medium sized datasets due to computational complexity. Our more recent work in [Seguy 2018] proposes a scalable estimation that can be estimated in two steps using stochastic gradient

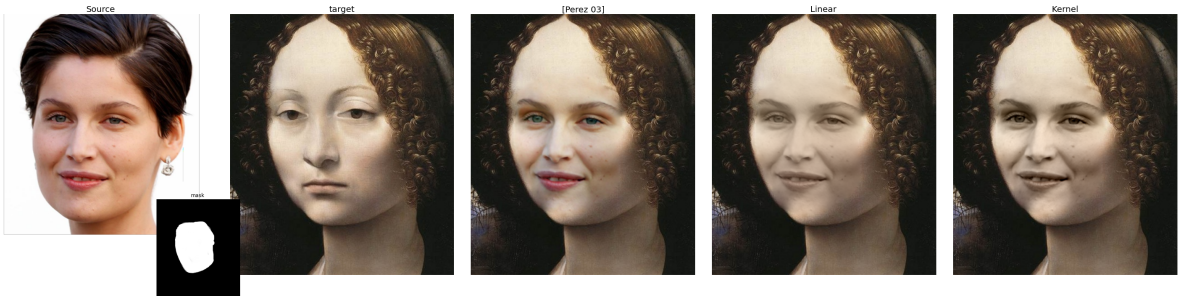


Figure 3.2: Illustration of gradient adaptation for seamless copy in image between images having different color gradient distributions. (two-left) the two images and the mask for the copy. (middle) Seamless copy of [Pérez 2003] that creates false colors (two-right) gradient adaptation with linear and non-linear mapping that produce more realistic colors.

descent and deep learning. [Seguy 2018] also proved weak convergence of the barycentric mapping to the true Monge mapping when the number of samples becomes large. But the community still lacks, to the best of our knowledge, proper convergence speed and concentration inequalities concerning the quality of such approximations.

3.2.2 Applications and extensions

OTDA in practice We compared OTDA to several approaches of the state of the art in [Courty 2016a] on visual adaptation problems. It was often better than state of the art methods but has been shown to be sensitive to the ground metric. For instance it worked surprisingly well when using Euclidean ground loss on raw images which is known to be a particularly bad representation. But when using Euclidean distance on features estimated from deep learning framework [Donahue 2014] the adaptation performance was unsurprisingly even better.

OTDA has also been used in several biomedical applications. [Gayraud 2017] has applied OTDA to the problem of P300 detection in Brain Computer Interfaces. It allowed to adapt between different subjects and can potentially decrease the time needed for calibration. OTDA was also applied to the problem of Computer Aided Diagnostic of Prostate cancer from MRI in [Gautheron 2017] to adapt between patients. Finally [Chambon 2018] showed that OTDA can improve the performance of sleep stage classification from ElectroEncephaloGrams.

Also note that while we proposed OTDA in the context of classification, learning a mapping between distributions and applying it as proposed originally by [Ferradans 2014] also has numerous applications. We already discussed color gradients that adapt the color space between images but we proposed an extension that focuses not on the colors but on the gradients in the image in [Perrot 2016]. Basically we proposed to perform a better seamless copy between images using Poisson image editing [Pérez 2003] by adapting the gradient prior to image reconstruction. An example of our adaptation is available in Figure 3.2 showing that adapting the gradients leads to better integration in the target image (both colors and dynamic properties). Finally the mapping can also be used to perform registration between two point clouds and can be used in remote sensing for registering LiDAR acquisitions and detect erosion of a coastal cliff [Courty 2016b].

Multi-domain adaptation and target shift OTDA can be extended to multi-domain when several source datasets are available. It can for instance be applied in parallel as proposed in [Gayraud 2017]. Some theoretical insight on the quality of the adaptation was first presented in [Redko 2016] where a generalization bond is exhibited both for single and multiple domains. Those bounds are very general but they measure the discrepancy between the domains and use this discrepancy to bound the target error. We believe that under the OTDA assumptions discussed above the generalization can be independent

from this discrepancy and depends only on the number of samples (see Eq. (3.10)).

The problem of target shift (ratio of classes changing) can be handled in the multi-domain case by estimating the ratio of classes in the target domain simultaneously with the OT matrices [Redko 2019]. In this case, the estimated ratio is used to re-weight the source samples and can be applied even when sources and the target have very different proportions. Another way to look at this proportion estimation is to estimate a barycenter of the proportions from the source domains optimal *w.r.t.* the OT that better fits the target data. We show in the paper that our objective function is indeed minimal when the correct class ratio is estimated.

Semi supervised Domain Adaptation OT for semi supervised learning has been proposed in the seminal work of [Solomon 2014a] where they proposed to use the OT matrix to propagate labels between samples or on a graph. OTDA can also be extended to the semi-supervised DA problem when a few target samples have known labels. The problem is easier in this case since more information is available and can help guide the transport. [Rousselle 2015] proposed to perform post-processing on the OT matrix by removing the mass that goes between samples that have different classes. This works in practice but the resulting matrix is not a transport matrix since some mass is discarded in the process. We proposed an alternative in [Courty 2016a] that relies on a very simple linear regularization term of the form

$$\Omega_{ss}(\mathbf{T}) = \langle \mathbf{T}, \mathbf{M} \rangle_F \qquad \text{where } M_{i,j} = \begin{cases} 0 & \text{if class } y_j^t \text{ unknown or } y_j^t = y_i^s \\ \infty & \text{if } y_j^t \neq y_i^s \end{cases} \tag{3.13}$$

This term forbids the OT solver to put mass between two samples that are known to have different classes and in practice guides the whole OT problem. We achieved important performance gain *w.r.t.* the unsupervised case in our experiments using this very simple regularization.

Chapter 4

Optimal Transport between histograms

Contents

In this chapter I discuss the use of OT for learning from histogram data. I first provide a short state of the art of machine learning on histogram data with OT and then present two of our contributions, namely Optimal Spectral Transportation (OST) and Deep Wasserstein Embeddings (DWE).

Data as histograms In this chapter I focus on histogram data, which can be described as discrete distributions $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ with a fixed support $\{\mathbf{x}_i\}_i$. In this case a dataset contains many histograms that have the same support $\mathbf{x}_i$ (for instance a regular grid) but have different weights a_i in their bins. A natural divergence used on this kind of data is the Kullback–Leibler divergence but it treats all the components of the histograms independently. The strength of the Wasserstein distance on this kind of data is that it can encode the relation between the bins through their positions.

Note that a lot of datasets can be seen as histograms. For instance images (color or black and white) are acquired by the physical process of counting photons on a regular-grid charge-coupled device (CCD). They are then natural histograms after normalization. One can also see a power spectrum as a histogram since it basically describes how the power is spread among different frequency bands. Finally a coarse but efficient modeling of text can be used as word counts, yielding very sparse histograms where the bins are all the words in a dictionary.

4.1 State of the art in Machine Learning

The ML community has only recently applied OT and the corresponding Wasserstein distance to ML problems. In this section we present some works that have used OT for both unsupervised and supervised machine learning.

4.1.1 Unsupervised learning with OT

Dictionary learning In dictionary learning, one wants to find simultaneously a dictionary describing a dataset along with the representation of the examples of the dataset on this dictionary. The problem is often formulated as seeking for matrices $\mathbf{D} \in \mathbb{R}^{n\times p}$ and $\mathbf{H} \in \mathbb{R}^{p\times n}$ that approximate the matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ with $\mathbf{A} \approx \mathbf{DH}$, hence the name of matrix factorization often used un practice. Special cases of Dictionary learning are the Principal Component Analysis and its positive counterpart Non-negative Matrix Factorization (NMF) [Lee 2001]. When the data $\mathbf{a}_k \in \Delta_n$ and $\mathbf{h} \in \Delta_p$ are both histograms and the dictionary elements (columns) of $\mathbf{D}$ are also histograms $\mathbf{d}_i \in \Delta \forall i$ divergences specific to distributions such as Kullback-Leibler [Lee 2001] or Itakura Saito [Févotte 2009] are often used. The Wasserstein distance on the columns of $\mathbf{A}$ can also be used as a data fitting term, leading to the following optimization problem [Sandler 2011] :

$$\min_{\mathbf{h}_k\in\Delta_p \forall k, \mathbf{D}\in\Delta_n^p} \quad \sum_k W_{\mathbf{C}}(\mathbf{a}_k, \mathbf{D}\mathbf{h}_k) \tag{4.1}$$

where the ground cost matrix $\mathbf{C}$ encodes geometrical relationship between the components of the histograms. Dictionary learning has been first extended to the Wasserstein distance data fitting in [Sandler 2011] but relied on slow linear programming solvers which limited its application on large histograms/datasets. The use of regularized OT as proposed in [Rolet 2016], promotes smoothness in the estimated dictionary elements and provides a significant speedup in the numerical computation for large datasets. Finally note that since it can be difficult to design a meaningful metric for matrix $\mathbf{C}$, [Zen 2014] proposed to estimate this matrix simultaneously with the dictionary similarly to what was proposed in [Cuturi 2014b] for Wasserstein distance.

Nonlinear unmixing with Wasserstein simplex A natural but non-obvious extension of linear unmixing is to perform non-linear unmixing using Wasserstein Barycenters. Indeed while a linear mixing model $\mathbf{Dh}$ can be seen as a Euclidean barycenter (since $\mathbf{h} \in \Delta$, the linear model is a weighted mean), one can use a more complex barycenter such as the Wasserstein barycenter [Benamou 2015] to fit the data. When using a Wasserstein data fitting term the estimation problem becomes

$$\min_{\mathbf{h}_k\in\Delta\forall k, \mathbf{D}\in\Delta^p} \quad \sum_k W_{\mathbf{C}}(\mathbf{a}_k, WB(\mathbf{h}_k, \mathbf{D})), \quad \text{with} \quad WB(\mathbf{h}, \mathbf{D}) = \arg\min_{\mathbf{b}\in\Delta} \sum_i h_i W_{\mathbf{C}}(\mathbf{b}, \mathbf{d}_i) \tag{4.2}$$

where WB is the Wasserstein barycenter of the dictionary elements in $\mathbf{D}$ weighted by $\mathbf{h}$. This elegant model has been proposed in [Schmitz 2017] and the authors propose a very interesting comparison between the Euclidean (linear) and Wasserstein barycenter models and both the Euclidean and Wasserstein data fitting terms. This extension of nonlinear mixing on the Wasserstein simplex suggests a lot of extensions of non-supervised methods among which the Principal Component Analysis discussed next.

Wasserstein Principal Geodesic Analysis Principal Component Analysis (PCA) seeks for directions that maximize the variance in the data. The natural divergence in space when computing variances is the Euclidean distance but PCA can be extended to more complex spaces such as geodesics. Principal Geodesic Analysis (PGA), has first been introduced by Fletcher et al. [Fletcher 2004]. It can be seen as a generalization of PCA on general Riemannian manifolds. Its goal is to find a set of directions, called geodesic directions or principal geodesics, that best encode the statistical variability of the data on the manifold.

Fletcher gives a generalization of this problem for complete geodesic spaces by extending three important concepts: *variance* as the expected value of the squared Riemannian distance from mean, *Geodesic subspaces* as a portion of the manifold generated by principal directions, and a *projection* operator onto that geodesic sub-manifold. The space of probability distribution equipped with the Wasserstein metric defines a geodesic space with a Riemannian structure [Santambrogio 2014], and an application of PGA is then an appealing tool for analyzing distributional data. However, as noted in [Seguy 2015, Bigot 2017], a direct application of Fletcher's original algorithm is intractable because space of probability distribution is infinite dimensional and there is no analytical expression for the exponential or logarithmic maps

allowing to travel to and from the corresponding Wasserstein tangent space. An efficient algorithm to perform PGA with entropic regularization was proposed in [Seguy 2015] with application to image data. Recent advances have proposed to solve both a Geodesic PCA and Log-PCA in the Wasserstein space using Forward-Backward splitting [Cazelles 2018]. Note that we will discuss an efficient embedding in section 4.3 that can greatly speedup the estimation of Wasserstein PGA.

Training Restricted Boltzman Machine with Wasserstein A Restricted Boltzmann machine (RBM) is a generative stochastic artificial neural network that can learn a probability distribution. It is often optimized by maximizing its likelihood over a dataset. When the data consists in histograms, the Kullback–Leibler divergence is often used as a reconstruction loss. [Montavon 2016] proposed to use the Wasserstein distance instead and has shown that it can help better estimate sensible models since it can encode the geometry of the histograms during learning. For instance they estimated RBM on small images and successfully used it for image completion and denoising.

4.1.2 Supervised learning and classification with OT

Learning with a Wasserstein Loss Similarly to the unsupervised case, the Wasserstein distance can be used as a data fitting term for learning a probabilistic predictor since its output is an histogram. The use of OT is particularly interesting in the multi-label scenario where a given sample can have multiple labels and a separable loss such as KL will treat them all independently. In this case the labels of a given sample can be seen as a histogram where non-zero bins denote the presence of a class. [Frogner 2015] proposed to use the Wasserstein loss as a data fitting in multi-label classification, leading to the following learning problem for a training set $\{\mathbf{x}_i, \mathbf{y}_i\}_i$

$$\min_f \quad \sum_{k=1}^{N} W_{\mathbf{C}}(f(\mathbf{x}_i), \mathbf{y}_i) \tag{4.3}$$

where f is a model with an output in the probability simplex, such as a neural network with softmax output. One major contribution of [Frogner 2015] was to propose a very elegant loss matrix $\mathbf{C}$ that would have been impossible to design manually for a large number of classes. They proposed to use a semantic representation of the classes by computing the Euclidean distance between the class representation in the `word2vec` embedding [Mikolov 2013]. This embedding is known to provide semantic euclidean distance and will encode in $\mathbf{C}$ the semantic relations between the classes. In other words, mistakes between classes that are similar will have a smaller loss (penalizing less an error) than between classes that are very dissimilar. They proposed to use regularized OT to solve the problem and proposed an extension relaxing the marginal constraints of the OT problem similarly to [Chizat 2018] to handle a different number of classes in the samples.

Word Mover's Distance (WMD) Wasserstein distance in a `word2vec` embedding can also be used to measure similarity between text seen as histograms of word occurrences. In [Kusner 2015], the authors propose this application called Word Mover's Distance (WMD) as a reference to Rubner's EMD. In this case, a document is represented as a very sparse histogram and each word can be represented in a semantic `word2vec` space. A simple example why this works well in practice is the use of synonyms in a text. When using WMD two synonyms will be very close in the embedding and the OT will find a similar correspondence between the words, which cannot be obtained with independent divergences.

The `word2vec` embedding can encode semantic representation between words but it has not been estimated with a specific text classification objective. It means that one can estimate a metric in this space that can maximize accuracy of a classifier. An extension of WMD that estimates a Mahalanobis metric to be used for WMD has been proposed in [Huang 2016]. It relies on regularized OT for optimizing a linear operator $\mathbf{A}$ applied to the data before computing WMD. Note that this approach is strongly related to Ground Metric Learning as proposed for un-regularized OT in [Cuturi 2014b].

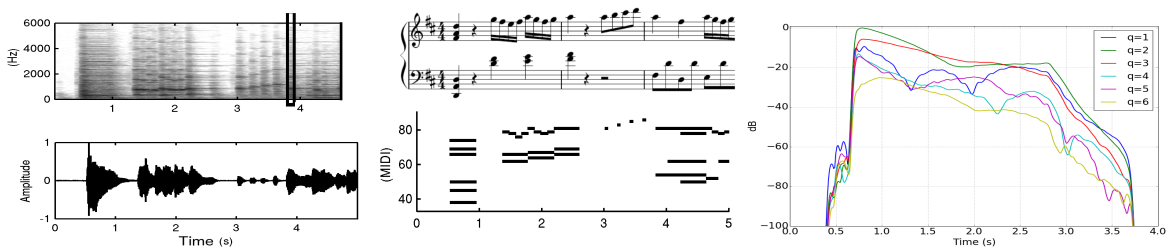


Figure 4.1: (left) Audio sequence with its time and time-frequency representation, (center) its corresponding sheet music and MIDI, (right) temporal evolution of the power in the harmonics when one note is played.

4.2 Optimal Spectral Transportation (OST)

In this section we present Optimal Spectral Transportation (OST) that is an adaptation of OT to musical unmixing. To this end, we first introduce the problem of musical spectral unmixing and discuss the use of the Wasserstein divergence for unmixing [Flamary 2016]. Next we discuss the resulting optimization problem and regularized variants of OST.

4.2.1 Musical spectral unmixing

Musical spectral unmixing with linear model Being able to separate different sources in an audio sequence is a difficult problem but can be very useful in applications such as audio restoration [Févotte 2009]. In the particular case of musical data, one unmixing of particular interest consists in determining automatically the sequence of musical notes that were played from the raw audio or time frequency representation of an audio sequence (Figure 4.1(left)). The main objective is to be able to reconstruct a sheet music or its discrete MIDI counterpart (Figure 4.1(center)) from a musical sequence. A state of the art approach is to estimate a robust dictionary $\mathbf{D}$ modeling each notes and to perform linear unmixing with KL divergence [Smaragdis 2006] on each spectrum in a time frequency representation. In this case one want to estimate on a given spectrum $\mathbf{a}$ the mixing coefficients $\mathbf{h}$ that minimize the following linear unmixing optimization problem

$$\min_{\mathbf{h}\in\Delta} \quad KL(\mathbf{a}, \mathbf{D}\mathbf{h}) \tag{4.4}$$

where KL is the Kullback-Leibler divergence.

While the approach above works very well in practice it still has a few limits that we will discuss now. First we can see in the time frequency representation in (Figure 4.1(left)) that musical data has an harmonic structure. This means that a given note played on an instrument will have in the Fourier domain power on its fundamental frequency but also its harmonics (integer multiples of the fundamental). This harmonic structure can be embedded in the dictionary elements of $\mathbf{D}$ but it forces the spectrum to have fixed magnitudes at each harmonics. This can be a problem since changing instruments will change those magnitudes and make the unmixing less robust. Another problem occurs if the instrument is not well tuned, its fundamental frequency and corresponding harmonics will be slightly changed which leads to a large discrepancy in a separable divergence such as KL. Finally, as illustrated in (Figure 4.1(right)), the magnitude of the harmonics do change along time when a note is played. When using fixed dictionary elements for the notes this might lead to a variability of the unmixing along time for a whole note. In the following we show that Wasserstein distance can be designed to be robust to the variabilities discussed above.

Linear unmixing with Wasserstein Linear unmixing can be seen as a special case of dictionary learning where the dictionary elements are known *a priori*. It consists in estimating a representation of an observation $\mathbf{a} \in \Delta$ as a weighted sum $\mathbf{D}\mathbf{h} = \sum_k h_k \mathbf{d}_k$ of dictionary elements from a known dictionary

$\mathbf{D} \in \mathbb{R}_+^{n \times p}$ where the dictionary elements (columns) of $\mathbf{D}$ are also histograms $\mathbf{d}_i \in \Delta \forall i$. It is a classical problem in remote sensing, signal and image processing and the representation $\mathbf{h}$ is often estimated by minimizing a divergence (Euclidean, Kullback-Leibler [Lee 2001], Itakura-Saito [Févotte 2009]) between the observation $\mathbf{a}$ and the reconstructed estimate $\mathbf{Dh}$ with respect to both $\mathbf{D}$ and $\mathbf{h}$. One can also use the Wasserstein distance to perform linear unmixing with

$$\min_{\mathbf{h} \in \Delta} \quad W_{\mathbf{C}}(\mathbf{a}, \mathbf{Dh}) \tag{4.5}$$

This very simple extension to linear unmixing can encode the relationship between the bins of the histograms in the $\mathbf{C}$ matrix, during the estimation. For instance we applied this formulation for linear unmixing of planetary hyperspectral observations in [Nakhostin 2016] where the Wasserstein distance is robust to small displacement along the frequencies. We will see in the following how this problem can be adapted to musical data.

4.2.2 Optimal spectral transportation and optimization

Optimal spectral transportation We proposed to use the Wasserstein distance for linear unmixing of audio spectrum in [Flamary 2016]. The main question that has to be answered when using Wasserstein distance in this case is the design of the matrix $\mathbf{C}$. A first tempting option is to use the quadratic loss between the frequencies of the bins in the spectrum. This corresponds to the W_2^2 distance and will be robust to small change along the frequency (badly tuned instrument) but a change in the magnitude of the harmonics will be very costly, meaning the Wasserstein distance will be very sensitive to this problem.

In order to be robust to both small frequency change and harmonic magnitude, we proposed to use a loss that takes into account the harmonic structure in the data and is quasi-invariant to it. We proposed to use a ground cost matrix of term

$$C_{ij} = \min_{q=1,\dots,\left\lceil \frac{f_i}{f_j} \right\rceil} (f_i - q f_j)^2 + \varepsilon \, \delta_{q \neq 1}, \tag{4.6}$$

where the f_i are the frequency of bin i in the spectrum. The main idea here is that mass from frequency f_j in the model $\mathbf{Dh}$ can move to all its harmonics in the observed sample $\mathbf{a}$ for a small cost (ε) making the corresponding Wasserstein loss invariant to changes in harmonics magnitudes. The small $\varepsilon > 0$ is here to penalize the case where all the mass of a fundamental goes to the harmonics since in this case it means that the note is in the next octave (it has no mass on the fundamental frequency). Finally note that this loss is locally quadratic which means that it will allow small movement of mass along the spectrum and will be robust to mistuned instruments.

The optimization problem for OT can be computationally expensive, but since a lot of information is encoded in the cost matrix $\mathbf{C}$ one can wonder if the dictionary elements are still necessary. Indeed if we take one dictionary element for a given note as a Dirac vector where the mass is 1 at the fundamental frequency and 0 everywhere else there is no need for designing or estimating dictionary elements anymore. Note that the invariant properties of (4.6) means that this simple Dirac dictionary element will be close in the Wasserstein sense to all its harmonic variants with different magnitudes.

Optimization problem The resulting unmixing optimization problem can be expressed as

$$\min_{\mathbf{h} \in \Delta, \mathbf{T} \geq \mathbf{0}} \quad \langle \mathbf{T}, \mathbf{C} \rangle_F \quad \text{s.t.} \quad \mathbf{T}^\top \mathbf{1} = \mathbf{Dh}, \mathbf{T}\mathbf{1} = \mathbf{a} \tag{4.7}$$

When the dictionary elements are Diracs on the fundamental frequencies, the $\mathbf{D}$ is a selection operator and the reconstructed model $\mathbf{Dh}$ can only have mass on p frequency positions (the fundamentals). The problem can then be reformulated as

$$\min_{\tilde{\mathbf{T}} \geq \mathbf{0}} \quad \langle \tilde{\mathbf{T}}, \tilde{\mathbf{C}} \rangle_F \quad \text{s.t.} \quad \tilde{\mathbf{T}}\mathbf{1} = \mathbf{a} \tag{4.8}$$

where $\tilde{\mathbf{C}}, \tilde{\mathbf{T}}$ are sub-matrices of size $n \times p$ and the solution of the unmixing can be computed with $\mathbf{h} = \tilde{\mathbf{T}}^\top \mathbf{1}$. The problem above can be solved independently for each line of $\mathbf{T}$ with an argmax on $\mathbf{C}$ that can be pre-computed in $O(np)$, leading to a final complexity on one sample of only $O(n)$.

One obvious limit is that the mass from observation $\mathbf{a}$ at frequency i will always be associated to the same dictionary element having the minimal loss $C_{i,j}$ in line i. This means that by construction this unmixing is not sparse which can be a problem since detecting notes will require post processing or sparsity promoting regularization. In the next section we discuss two regularization terms that we proposed that can help in having more robust unmixing.

4.2.3 Regularization and applications

Entropic regularization A first regularization that can lead to more robust estimations is the entropic regularization. When adding this regularization term, problem (4.8) can be reformulated as a Bregman projection [Benamou 2015] that can be solved in closed form using a matrix multiplication of complexity $O(np)$. It is interesting to note that the entropic regularization has the effect of replacing the argmax along the lines of $\mathbf{C}$ by a softmax whose smoothness is controlled by λ. This regularization works well in practice but comes with a significant caveat: the estimation is even less sparse than the original OT problem (4.8) which makes the identification of the active musical notes more difficult.

Sparsity promoting regularization In order to promote a sparse unmixing and an automatic detection of musical notes, we proposed to use the following regularization term

$$\Omega_{cspare}(\mathbf{T}) = \sum_j \left(\sum_i T_{i,j} \right)^{\frac{1}{2}} \tag{4.9}$$

that will promote column sparsity hence sparsity in the unmixing since $\mathbf{h} = \mathbf{T}^\top \mathbf{1}$. This regularization term is very similar to (2.36) and the resulting optimization problem can be solved similarly with DC algorithm. At each iteration the concave term (4.9) can be approximated by its linear majorization resulting in the iterative solving of problem (4.8) where the loss matrix $\mathbf{C}$ is updated at each iteration. While this iterative approach is slower than the original problem, it has been observed to converge in a few iterations (≤ 10) and can still be computed efficiently.

OST in action We applied in [Flamary 2016] OST on a toy generated dataset where the exact dictionary elements are known perfectly and compared it with KL unmixing of [Smaragdis 2006]. We generated data with slightly shifted fundamental frequencies and varying harmonic magnitudes and observed that indeed the proposed approach is robust to those complex and non-linear variabilities while being very fast ($\sim$ms per frame). OST seems to beneficiate from both regularization on the toy data regularization.

Next we applied our approach on the well known MAPS Dataset [Emiya 2010] that consists in several piano sequences from classical music ($m = 60$ notes). We can compare to a ground truth on this dataset since the musical notes are available as MIDI files. On this dataset, OST performed similarly as KL+dictionary but was more computationally efficient (≥ 70 times quicker).

Finally since OST is a very quick approach, we implemented a demonstration program in Python using the Pygame library were we can show the spectrogram and the estimated musical notes in real time. The code is open source and available on GitHub [1].

4.3 Learning Deep Wasserstein Embeddings (DWE)

In this section we address the problem of speeding up multiple computations of OT distance and how to perform efficient data mining in the Wasserstein space. I present to this end a paper published at ICLR 2018 [Courty 2018] that relies on the supervised learning of a Wasserstein embedding with deep learning.

[1] https://github.com/rflamary/OST

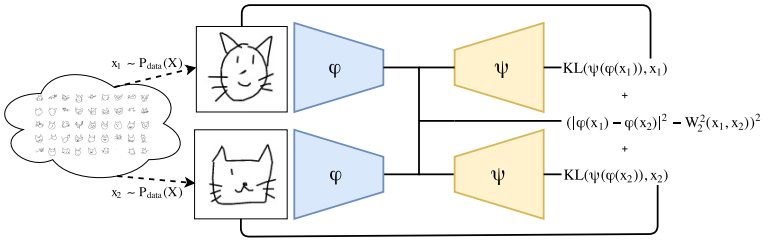


Figure 4.2: Architecture of the Deep Wasserstein Embedding: two samples are drawn from the data distribution and set as input of the same network (ϕ) that computes the embedding. The embedding is learnt such that the squared Euclidean distance in the embedding mimics the Wasserstein distance. The embedded representation of the data is then decoded with a different network (ψ), trained with a Kullback-Leibler divergence loss.

4.3.1 Deep Wasserstein Embedding

I discuss here how our method, coined DWE for Deep Wasserstein Embedding, learns, in a supervised way, a new representation of the data that can speedup Wasserstein distance computation. Our objective is to find a deep embedding that takes histograms as input but provides a Euclidean distance in the embedding that mimics the Wasserstein distance between the original histograms. We proposed to learn those embeddings in [Courty 2018] in a supervised way. To this end we need a pre-computed dataset that consists of pairs of histograms $\{\mathbf{a}_i^1, \mathbf{a}_i^2\}_{i\in 1,\dots,n}$ of dimensionality d and their corresponding W_2^2 Wasserstein distance $\{d_i = W_2^2(\mathbf{a}_i^1, \mathbf{a}_i^2)\}_{i\in 1,\dots,n}$. This dataset can be pre-computed offline in parallel but will be used in the following to learn an embedding emulating the Wasserstein space.

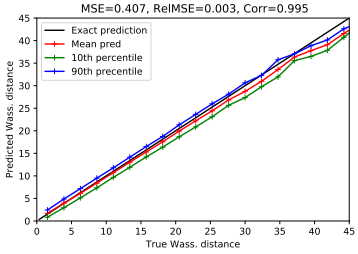
Siamese networks A way to estimate a meaningful embedding that can be used more broadly is to use a siamese neural network [Bromley 1994]. Originally designed for metric learning purpose and similarity learning (based on labels), this type of architecture is usually defined by replicating a network which takes as input two samples from the same learning set, and learns a mapping to new space with a contrastive loss. It has mainly been used in computer vision, with successful applications to face recognition [Chopra 2005] or one-shot learning for example [Koch 2015].

Deep Wasserstein Embedding We proposed in [Courty 2018] to learn an embedding network φ that takes as input a histogram and projects it in a given Euclidean space of $\mathbb{R}^p$ as illustrated in Figure 4.2 on small images. In practice, this embedding should mirror the geometrical properties of the Wasserstein space. We also propose to regularize the computation of this embedding by adding a reconstruction loss based on a decoding network ψ. The reconstruction has two important implications: first we observed empirically that it eases the learning of the embedding and improves the generalization performance of the network (see experimental results in appendix of [Courty 2018]) by forcing the embedded representation to catch sufficient information of the input data to allow a good reconstruction. This type of auto-encoder regularization loss has been discussed in [Yu 2013] in the different context of embedding learning. Second, using a decoder network allows the reconstruction from embedding, which is of prime importance in several data-mining tasks (discussed in the next subsection).

An overall picture depicting the whole process is given in Figure 4.2. The global objective function reads

$$\min_{\phi,\psi} \quad \sum_i \left\| \|\varphi(\mathbf{a}_i^1) - \varphi(\mathbf{a}_i^2)\|^2 - d_i \right\|^2 + \lambda \sum_i \left[\mathrm{KL}(\psi(\varphi(\mathbf{a}_i^1)), \mathbf{a}_i^1) + \mathrm{KL}(\psi(\varphi(\mathbf{a}_i^2)), \mathbf{a}_i^2) \right] \tag{4.10}$$

where $\lambda > 0$ weights the two data fitting terms and $\mathrm{KL}(,)$ is the Kullback-Leibler divergence. This choice is motivated by the fact that the Wasserstein metric operates on probability distributions and KL can be efficiently computed and derived.



Method	W_2^2/sec
LP network simplex (1 CPU)	192
DWE Indep. (1 CPU)	3 633
DWE Pairwise (1 CPU)	213 384
DWE Indep. (GPU)	233 981
DWE Pairwise (GPU)	10 477 901

Figure 4.3: Prediction performance on the MNIST dataset. (Figure) The test performance are as follows: MSE=0.41, Relative MSE=0.003 and Correlation=0.995. (Table) Computational performance of W_2^2 and DWE given as average number of W_2^2 computation per seconds for different configurations.

4.3.2 Fast data mining in the Wasserstein space

Once the functions φ and ψ have been learned, several data mining tasks can be operated in the Wasserstein space. We discuss here the potential applications of our computational scheme and its wide range of applications on problems where the Wasserstein distance plays an important role. Though our method is not an exact Wasserstein estimator, we empirically show in the numerical experiments that it performs very well and competes favorably with other classical computation strategies.

Wasserstein barycenters Barycenters in Wasserstein space were first discussed by [Agueh 2011]. Designed through an analogy with barycenters in a Euclidean space, the Wasserstein barycenters of a family of measures are defined as minimizers of a weighted sum of squared Wasserstein distances. In our framework, approximate barycenters can be obtained with

$$\bar{\mathbf{a}} = \arg\min_{\mathbf{a}\in\Delta} \sum_i \alpha_i W(\mathbf{a}, \mathbf{a}_i) \approx \psi(\sum_i \lambda \varphi(\mathbf{a}_i)), \tag{4.11}$$

where $\mathbf{a}_i$ are the data samples and the weights α_i obeys the following constraints: $\sum_i \lambda_i = 1$ and $\lambda_i > 0$. When the weights are uniform and the whole data collection is considered, the barycenter is the Wasserstein population mean, also known as Fréchet mean [Bigot 2017].

Principal Geodesic Analysis in Wasserstein space We propose a novel PGA approximation as the following procedure: *i)* find $\overline{x}$ the approximate Fréchet mean of the data as $\overline{x} = \frac{1}{N}\sum_i^N \varphi(x_i)$ and subtract it to all the samples *ii)* build recursively a linear subspace $V_k = \texttt{span}(v_1, \cdots, v_k)$ in the embedding space (v_i being of the dimension of the embedded space) by solving the following maximization problem:

$$v_1 = \text{argmax}_{|v|=1} \sum_{i=1}^{n} (v.\varphi(x_i))^2, \quad v_k = \text{argmax}_{|v|=1} \sum_{i=1}^{n} \left((v.\varphi(x_i))^2 + \sum_{j=1}^{k-1} (v_j.\varphi(x_i))^2 \right). \tag{4.12}$$

which is strictly equivalent to performing PCA in the embedded space. Any reconstruction from the corresponding subspace to the original space is conducted through ψ.

4.3.3 Applications of DWE

Numerical precision and computational performance The true and predicted values for the Wasserstein distances on never seen MNIST images are given in Fig. 4.3. We can see that we reach a good precision with a test MSE of 0.4 and a relative MSE of $2e-3$. The correlation is of 0.995 and the quantiles show that we have a very small uncertainty with only a slight bias for large values where only a small number of samples is available for training. These results show that a good approximation of the W_2^2 can be performed by our approach (≈1e-3 relative error).

We also investigated the ability of our approach to compute W_2^2 efficiently. To this end we computed the average speed of Wasserstein distance computation on test dataset to estimate the number of W_2^2

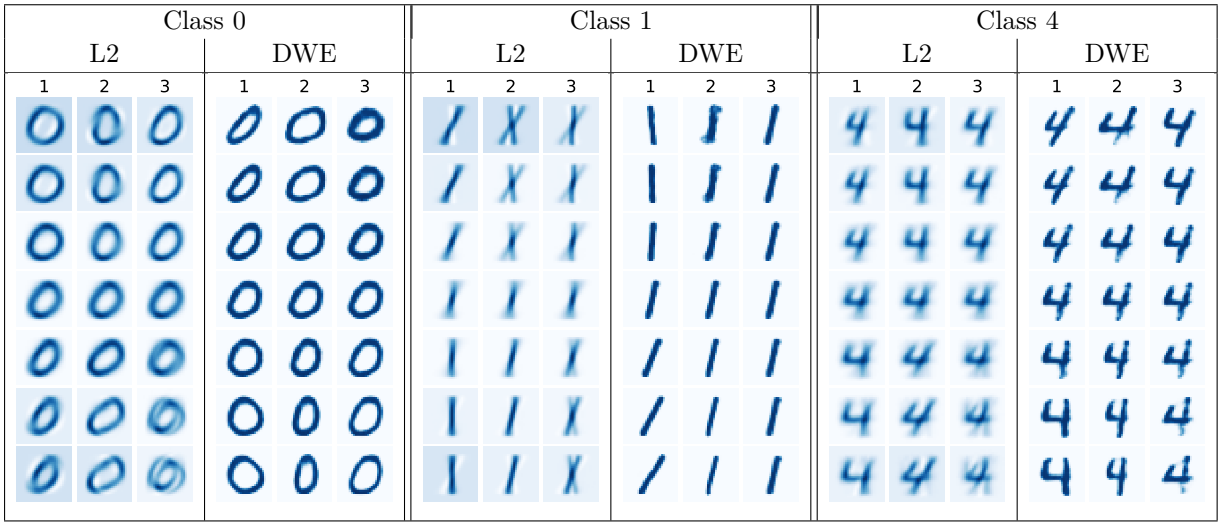


Figure 4.4: Principal Geodesic Analysis for classes 0,1 and 4 from the MNIST dataset for squared Euclidean distance (L2) and Deep Wasserstein Embedding (DWE). For each class and method we show the variation from the barycenter along one of the first 3 principal modes of variation.

computations per second in the Table of Fig. 4.3. Note that there are 2 ways to compute the W_2^2 with our approach denoted as Indep. and Pairwise. This comes from the fact that our W_2^2 computation is basically a squared Euclidean norm in the embedding space. The first computation measures the time to compute the W_2^2 between independent samples by projecting both in the embedding and computing their distance. The second computation aims at computing all the pairwise W_2^2 between two sets of samples and this time one only needs to project the samples once and compute all the pairwise distances, making it more efficient. Note that the second approach would be the one used in a retrieval problem where one would just embed the query and then compute the distance to all or a selection of the dataset to find a Wasserstein nearest neighbor for instance. The speedup achieved by our method is very impressive even on CPU with speedup of x18 and x1000 respectively for Indep. and Pairwise. But the GPU allows an even larger speedup of respectively x1000 and x500 000 with respect to a state-of-the-art C compiled Network Simplex LP solver of the POT Toolbox [Flamary 2017b, Bonneel 2011]. Of course this speed-up comes at the price of a time-consuming learning phase, which makes our method better suited for mining large scale datasets.

Principal Geodesic Analysis We report in Figure 4.4 the Principal Component Analysis (L2) and Principal Geodesic Analysis (DWE) for 3 classes of the MNIST dataset. We can see that using Wasserstein to encode the displacement of mass leads to more semantic and nonlinear subspaces such as rotation/width of the stroke and global sizes of the digits. This is well known and has been illustrated in [Seguy 2015]. Nevertheless our method allows for estimating the principal component even in large scale datasets and our reconstruction seems to be more detailed compared to [Seguy 2015] maybe because our approach can use a very large number of samples for subspace estimation.

Chapter 5

Optimal Transport between empirical distributions

Contents

This chapter focuses on the use of OT on empirical distributions. Those distributions are very common in machine learning since all the datasets used for training have a finite number of examples that are supposed to be drawn independently from an unknown data distribution μ_d. One particular property of those empirical distributions is that they nearly never overlap which means that they are particularly difficult to compare for a learning purpose.

One very efficient approach to compare those empirical distributions have been to use classic divergences such as ℓ_2 after the distributions have been convolved by a kernel (Parzen windows). This approach known as Maximum Mean Discrepancy (MMD) [Fortet 1953] has been used with success for two-sample tests [Gretton 2012] and generative modeling [Dziugaite 2015, Sutherland 2016]. Still the kernel has the effect of spreading the mass around the diracs of the samples which might lose part of the information encoded in the empirical distributions.

In the remaining of the chapter I first provide a short state of the art of the use of OT on empirical distributions in machine learning and discuss more in details two of our contributions in this domain. The first is the Wasserstein Discriminant Analysis that can be seen as a generalization of Fisher Discriminant Analysis for non-linearly separable data. The second is a domain adaptation approach based on minimizing the Wasserstein distance between joint feature/label distributions.

5.1 State of the art in Machine Learning

This short state of the art discusses the different approaches proposed in recent years that use Wasserstein distance on empirical distributions.

5.1.1 Unsupervised learning and Generative Adversarial Network

Generative Adversarial Networks (GAN) A major problem of unsupervised learning is to estimate a realistic data generator from a finite dataset. Early approaches aimed at learning simple Gaussian generative representations such as PCA as illustrated by the well known EigenFaces application [Sirovich 1987] or Gaussian mixtures [Ghahramani 1996]. More complex generative models such as Restricted Boltzman Machines (RBM) [Hinton 2012] have also been proposed for highly nonlinear distributions. While those approaches managed to generate realistic samples they all had in common the fact that on complex images those samples seemed to be smooth and lack details.

This problem lead to the proposition of Generative Adversarial Networks in [Goodfellow 2014a] which objective is to generate samples that are indistinguishable from real data. They propose to find a generator G that takes random IID samples drawn for instance from the Gaussian distribution μ_g as input and transform them into realistic samples of the data distribution μ_d. The proposed optimization problem consists in estimating simultaneously a generator G and a discriminator D optimizing the following problem

$$\min_G \max_D \quad E_{\mathbf{x}\sim\mu_d}[\log D(\mathbf{x})] + E_{\mathbf{z}\sim\mu_g}[\log(1 - D(G(\mathbf{z})))]. \tag{5.1}$$

Note that in the problem above each models G and D are adversaries in the sense that they try to respectively minimize and maximize the objective value. We want the generator G to generate realistic image, it means that the objective is to have a discriminator D failing to classify between generated and real images. [Zhao 2016] has shown that at convergence those models reach a Nash equilibrium and the generated distribution is almost equal to the data distribution. The seminal work of Goodfellow had a tremendous impact in the machine learning community because of the quality of the generated images and the apparent simplicity of the optimization problem. Extension to convolutional models have also shown that the space at the input of the generator has semantic meaning [Radford 2015], meaning that some simple arithmetic in the generator space allows for instance to add glasses to a person in the image space. Still the training of GAN, *i.e.* the optimization of (5.1), is notoriously hard to do in practice and requires numerous tricks [Salimans 2016] due to the fact that the small magnitude of the gradients when updating G on a discriminator D that manages to separate the classes.

Wasserstein Generative Adversarial Networks (WGAN) The difficulty to estimate a GAN led to the proposal of [Arjovsky 2017]. Wasserstein GAN aims at finding a generator G minimizing the W_1 Wasserstein distance between the data distribution μ_d and the push-forward $G\#\mu_g$ with the following optimization problem

$$\min_G \quad W_1^1(G\#\mu_g, \mu_d) \tag{5.2}$$

The problem above has several nice properties, for instance the gradient of the Wasserstein distance never vanish until true equality between the distributions which makes the problem easier to solve that classical GAN (5.1). Also note the Wasserstein GAN fits into the family of minimum Wasserstein estimators [Bassetti 2006]. The problem can be expressed in the dual of the OT problem with

$$\min_G \sup_{\phi\in\mathrm{Lip}^1} \quad \mathbb{E}_{\mathbf{x}\sim\mu_d}[\phi(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim\mu_g}[\phi(G(\mathbf{z}))] \tag{5.3}$$

that is separable *w.r.t.* the data distribution and the generator. The problem above bears a strong resemblance to original GAN since the dual potential ϕ can be seen as an adversarial discriminator. It also illustrates the fact that the Wasserstein GAN is a special case of f-GAN that minimizes an f-divergence [Nowozin 2016].

On the implementation side, [Arjovsky 2017] proposed to use a neural network to estimate the dual potential ϕ. This brings several problems since it is very difficult to force a neural network to have a given Lipschitz constant. The main idea proposed in [Arjovsky 2017] was to perform weight clipping so as to limit the magnitude of the linear operators in the neural network, hence limiting the Lipschitz constant.

The WGAN objective value is then a lower bound on the actual scaled Wasserstein distance as shown below

$$\max_{f \in \text{NN class}} L_{WGAN}(f, G) \leq \sup_{\|\phi\|_L \leq K} L_{WGAN}(\phi, G) = K \cdot W_1^1(G\#\mu_g, \mu_d) \tag{5.4}$$

where the Wasserstein distance is multiplied by an unknown coefficient K (the true Lipschitz constant of ϕ). Since the neural networks has a specific architecture, it might never reach KW_1, which means that the optimized value is not exactly the Wasserstein distance but it can be shown that the gradients are aligned. Alternatively, [Gulrajani 2017] recently proposed to promote the gradient norm to be one, *i.e.*

$$\min_G \sup_{f \in \text{NN class}} \mathbb{E}_{\mathbf{x}\sim\mu_d}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim\mu_g}[f(G(\mathbf{z}))] + \lambda \mathbb{E}_{\mathbf{x}\sim\mu_{st}}[(||\nabla f(\mathbf{x})||_2 - 1)^2] \tag{5.5}$$

where μ_{st} is a distribution of samples drawn on straight lines between samples from the source and target distribution so as to promote the gradient constraints in the simplex of the whole source + target data. This is a relaxation of the Lipschitz constraint but makes sense since the gradient of the dual potential of Wasserstein is of norm 1 almost everywhere. A variant of Variational Auto-Encoders called Wasserstein Auto-encoders has been proposed in [Tolstikhin 2017]. Wasserstein GAN and its improved version above have had a strong impact in the GAN community due to their easier to learn optimization problem.

Note that other approaches relying on Wasserstein distance have been proposed to train Generative Networks. [Genevay 2017b, Genevay 2017a] proposed to use entropic regularized Sinkhorn divergence as fitting term for a model. Sliced Radon Wasserstein has also been proposed as an efficient alternative [Deshpande 2018].

5.1.2 Supervised learning and domain adaptation

The use of the Wasserstein distance on datasets seen as empirical distributions for supervised learning is quite recent and can be mostly grouped in two major families: robust optimization and domain adaptation/transfer learning.

Distributionally robust optimization Distributionally robust optimization is a very elegant approach that has been used with success in machine learning for providing better generalization. When the training data is an empirical joint feature/label distribution $\hat{\mathcal{P}}$, robust optimization consists in optimizing the following problem

$$\min_f \max_{\mathcal{P}, \mathcal{D}(\hat{\mathcal{P}}, \mathcal{P}) \leq \varepsilon} \mathbb{E}_{\mathbf{x}, y \sim \mathcal{P}}[L(y, f(\mathbf{x}))] \tag{5.6}$$

where $\{\mathcal{P} | \mathcal{D}(\hat{\mathcal{P}}, \mathcal{P}) \leq \varepsilon\}$ is called the uncertainty set in the language of robust optimization. The problem above aims at finding a predictor f that works well even in the worst case distribution $\mathcal{P}$ in the uncertainty set hence bringing robustness. Since $\hat{\mathcal{P}}$ is an empirical distribution, it seems natural to use the Wasserstein distance to define the uncertainty set. In this case it gives a nice geometrical intuition about the authorized displacements of the samples. For instance when using W_1, if only one sample moves in $\hat{\mathcal{P}}$, then it can move anywhere in the Euclidean ball of radius ε.

Using the Wasserstein distance in robust optimization has been studied simultaneously by several groups. An application to robust logistic regression has been proposed in [Shafieezadeh-Abadeh 2015], and was extended to several losses in [Shafieezadeh-Abadeh 2017] which also provided concentration inequalities. The authors in [Blanchet 2016] proposed to use this formulation to estimate a linear Lasso estimator using the Wasserstein distance for defining the uncertainty set. [Esfahani 2018] has shown that the optimization problem can in certain cases be reformulated as a Linear program and provided concentration inequalities. [Gao 2016] provided a generalization to non-empirical distributions and do not require the distributions to have a compact support. Also note that there is a strong relation between robust optimization and adversarial training *i.e.* training classifiers robust to adversarial examples [Madry 2017]. Robust optimization has also been used to investigate the generalization of adversarial training in [Sinha 2018].

Finally [Lee 2017] provided a generalization bound and showed that such optimization can be used in the case of Domain Adaptation. The main idea is that if you know how far away the source and target distributions are with $W_1(\mathcal{P}_s, \mathcal{P}_t)$, then you can perform domain adaptation by solving problem (5.6) with $\varepsilon = W_1(\mathcal{P}_s, \mathcal{P}_t)$ which will ensure that the classifier will work well on the target distribution since it is included in the uncertainty set. This approach is actually quite elegant but not applicable when the distributions are very different for instance when transformations occur on the feature space as discussed for OTDA in Chapter 3.

Domain adaptation and transfer learning Domain adaptation and transfer learning for deep neural network bring several challenges. In particular the problem of unsupervised domain adaptation try to estimate a classifier on new data $\hat{\mu}_t$ with no labels available but with access to another dataset $\hat{\mathcal{P}}_s$ of data marginals μ_s where labels are available. One major family of approaches aims at learning a classifier $f \circ g$ that works well on the source distributions but that will also have a similar representation across datasets in the embedding g. The global optimization problem can often be expressed as the following

$$\min_{f,g} \quad \mathbb{E}_{\mathbf{x},y\sim\hat{\mathcal{P}}_s}[L(y, f(g(\mathbf{x}))] + \lambda \mathcal{D}(g\#\hat{\mu}_s, g\#\hat{\mu}_t) \tag{5.7}$$

where $\mathcal{D}$ is a divergence between the data distributions of the samples in the feature space obtained by g. One of the first approaches proposed to use Maximum Mean Discrepancy to measure the dissimilarity in the feature space [Tzeng 2014]. DeepCORAL proposed to minimize the squared Frobenius norm between the covariance matrices of the two distributions [Sun 2016] hence focussing only on the second order moments of the distributions. Domain Adversarial Neural Network use as the name suggests an adversarial approach to promote similarity between the distributions [Ganin 2016].

Finally the Wasserstein distance has also been proposed as a measure of similarity between the distributions in the feature space in [Shen 2018]. In their work that bears resemblance to [Ganin 2016] they propose to use the dual formulation as in [Arjovsky 2017] with the gradient penalty of [Gulrajani 2017]. This leads to the following optimization problem

$$\min_{f,g} \quad \mathbb{E}_{\mathbf{x},y\sim\hat{\mathcal{P}}_s}[L(y, f(g(\mathbf{x}))] + \lambda \max_{\phi} \left\{\mathbb{E}_{\mathbf{x}\in\mu_s}[\phi(g(\mathbf{x}))] - \mathbb{E}_{\mathbf{x}\in\mu_t}[\phi(g(\mathbf{x}))] - \gamma \mathbb{E}_{\mathbf{x}\in\mu_{st}}[\|\nabla_{\mathbf{x}}\phi(g(\mathbf{x})) - 1\|^2]\right\} \tag{5.8}$$

where the part on the right is an approximation of the W_1 Wasserstein distance. This method seems to work very well in practice but shares a limit with all the formulations (5.7) above: they treat the label information in the source and the feature information separately through two objectives. While this makes sense when no labels are available in the target domain, we believe a more general formulation that works on joint feature/label distribution can help enhance the performance. This approach will be discussed in section 5.3.

5.2 Wasserstein Discriminant Analysis (WDA)

In this section, we present our contribution to the problem of estimating a discriminant linear subspace from empirical distributions [Flamary 2018]. Estimating a linear subspace of the data is one major family of dimensionality reduction methods [Van Der Maaten 2009, Burges 2010]. Although very simple, linear subspaces have many advantages. They are easy to interpret, and can be inverted, at least in a least-squares way. This latter property has been used for instance in PCA denoising [Zhang 2010]. Linear projection is also a key component in random projection methods [Fern 2003] or compressed sensing and is often used as a first pre-processing step, such as the linear part in a neural network layer. Finally, linear projections only imply matrix products and stream therefore particularly well on any type of hardware (CPU, GPU, DSP).

The objective of supervised linear projection approach is to find a projector $\mathbf{P} : \mathbb{R}^d \to \mathbb{R}^p$, $p \ll d$, such that the embeddings of the projected points $\mathbf{P}\mathbf{x}_i$ of $\mu = \frac{1}{n}\sum_i \delta_{\mathbf{x}_i}$ separate the classes in $\{y_i\}_i$ with respect to a criterion. Note that the projector $\mathbf{P}$ can be seen as a push-forward operator, and the distribution of

projected samples can be expressed as in $\mathbf{P}\#\mu$. Also note that in the following we denote the empirical distribution for class c as $\mu^c = \frac{1}{n_c}\sum_{i,y_i=c}\delta_{\mathbf{x}_i}$ where n_c is the number of samples from class c. The data distribution μ is a weighted sum of the class distributions $\mu = \sum_c \frac{n_c}{n}\mu^c$.

5.2.1 Fisher ratio and Wasserstein discriminant

Fisher Discriminant Analysis (FDA) and local approaches FDA is one of the most common approach to perform supervised dimensionality reduction. Given an empirical training dataset μ and its corresponding classes $\{y_i\}_i$, the goal of FDA is to learn a linear map that can discriminate classes using linear classifiers. FDA attempts to maximize w.r.t. $\mathbb{P}$ the sum of *all* squared distances $\|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_{j'}\|^2$ between pairs of samples from different classes c, c' while minimizing the sum of *all* distances $\|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_j\|^2$ between pairs of samples within the same class c [Friedman 2001, **S**4.3]. Because of this, it is well documented that the performance of FDA degrades when class distributions are multi-modal.

Several variants of FDA have been proposed to tackle non-linearly separable problems [Friedman 2001, **S**12.4]. For instance, a localized version of FDA was proposed by [Sugiyama 2007], which boils down to discarding the computation for all pairs of points that are not neighbors. On the other hand, originally designed to operate with a k-nearest neighbor classifier, the first techniques that were proposed to learn metrics [Xing 2003] used a *global* criterion, namely a sum on all pairs of points. Later on, variations that focused instead exclusively on *local* interactions, such as LMNN [Weinberger 2009], were shown to be far more efficient in practice.

Wasserstein Discriminant Analysis We introduced in [Flamary 2018] a novel approach called Wasserstein Discriminant Analysis (WDA). WDA is built on the regularized Wasserstein loss (2.31) to compute similarity between the class distributions μ_c. The criterion we proposed to optimize is the following:

$$\max_{\mathbf{P}\in\Delta} \quad \frac{\sum_{c,c'>c} W_\lambda(\mathbf{P}\#\mu^c, \mathbf{P}\#\mu^{c'})}{\sum_c W_\lambda(\mathbf{P}\#\mu^c, \mathbf{P}\#\mu^c)} \tag{5.9}$$

where $\Delta = \{\mathbf{P} = [\mathbf{p_1}, \ldots, \mathbf{p}_p] \quad | \quad \mathbf{p}_i \in \mathbb{R}^d, \|\mathbf{p}_i\|_2 = 1 \quad \text{and } \mathbf{p}_i^\top \mathbf{p}_j = 0 \text{ for } i \neq j\}$ is the Stiefel manifold [Absil 2009], the set of orthogonal $d \times p$ matrices; $\mathbf{P}\#\mu^c$ is the distribution of projected samples from class c. W_λ is the regularized Wasserstein loss defined in (2.31). The ratio in equation (5.9) is very similar to the ratio of variances in FDA, we want to maximize the ratio of the regularized Wasserstein distances between inter class populations and between the intra-class population with itself, when these points are considered *in their projected space*. This is a major difference with other local approaches that rely on relations between samples estimated in the original space.

Regularized OT and covariance matrices The entropic-regularized Wasserstein loss measures the dissimilarity between empirical distributions by considering pairwise distances between samples. For a given value of the regularization parameter λ the and the optimal OT matrix is $\mathbf{T}$ and the regularized OT loss becomes

$$W_\lambda(\mathbf{P}\#\mu^c, \mathbf{P}\#\mu^{c'}) = \frac{1}{n_c n_{c'}}\sum_{i,j} T_{i,j}\|\mathbf{P}\mathbf{x}_i^s - \mathbf{P}\mathbf{x}_j^t\|^2 = \langle \mathbf{P}^T\mathbf{P}, \mathbf{C}_\lambda^{c,c'} \rangle \tag{5.10}$$

with

$$\mathbf{C}_\lambda^{c,c'} = \frac{1}{n_c n_{c'}}\sum_{i,j} T_{i,j}(\mathbf{P}\mathbf{x}_i^c - \mathbf{P}\mathbf{x}_j^{c'})(\mathbf{P}\mathbf{x}_i^c - \mathbf{P}\mathbf{x}_j^{c'})^T \tag{5.11}$$

where $\mathbf{C}_\lambda^{c,c'}$ can be seen as a covariance matrix weighted by the relation between samples in the OT matrix $\mathbf{T}$. When the regularization parameter is small (5.10) will be similar to Wasserstein distance. When the regularization increases, the entropic regularization will spread the mass across more samples, hence enlarging the neighborhood between the samples. Finally note that the fact that the value of (5.10) is always > 0 for any $\lambda > 0$ is necessary since it leads to a well posed optimization problem in (5.9) (the denominator cannot be 0).

5.2.2 Optimization problem and applications

Optimization problem Using the definition (5.10) of regularized Wasserstein distance, we can write the Wasserstein Discriminant Analysis optimization problem as

$$\max_{\mathbf{P}\in\Delta} \quad \left\{ J(\mathbf{P},\mathbf{T}(\mathbf{P})) = \frac{\sum_{c,c'>c}\langle \mathbf{P}^T\mathbf{P}, \mathbf{C}_\lambda^{c,c'}\rangle}{\sum_c \langle \mathbf{P}^T\mathbf{P}, \mathbf{C}_\lambda^{c,c}\rangle} = \frac{\langle \mathbf{P}^T\mathbf{P}, \mathbf{C}_{b,\lambda}\rangle}{\langle \mathbf{P}^T\mathbf{P}, \mathbf{C}_{w,\lambda}\rangle} \right\} \tag{5.12}$$

where $\mathbf{C}_{b,\lambda} = \sum_{c,c'>c} \mathbf{C}_\lambda^{c,c'}$ and $\mathbf{C}_{w,\lambda} = \sum_c \mathbf{C}_\lambda^{c,c}$ are the between and within cross-covariance matrices that depend on the optimal matrices $\mathbf{T}^{c,c'}$ that depend on $\mathbf{P}$. Problem (5.12) can be reformulated as the following bi-level optimization problem

$$\max_{\mathbf{P}\in\Delta} \quad J(\mathbf{P},\mathbf{T}(\mathbf{P})) \tag{5.13}$$

$$\text{s.t. } \mathbf{T}(\mathbf{P}) = \underset{\mathbf{T}\in U_{n_c n_{c'}}}{\arg\min} \quad E(\mathbf{T},\mathbf{P}) \tag{5.14}$$

where $\mathbf{T} = \{\mathbf{T}^{c,c'}\}_{c,c'}$ contains all the transport matrices between classes and the inner problem function E is defined as the sum of regularized OT problems of the form (2.31) for all pairs $c, c' \geq c$ yield a solution $\mathbf{T}(\mathbf{P})$ as the concatenation of all the regularized OT matrices between the pairs. Optimization problem (5.13)-(5.14) is a bi-level optimization problem, which can be solved using gradient descent [Colson 2007]. Indeed, J is differentiable with respect to $\mathbf{P}$. This comes from the fact that optimization problems in Equation (5.14) are all strictly convex, making solutions of the problems unique, hence $\mathbf{T}(\mathbf{P})$ is smooth and differentiable [Bonnans 1998].

Optimization algorithm One can compute the gradient of J directly *w.r.t.* $\mathbb{P}$ using the chain rule as follows

$$\nabla_{\mathbb{P}} J(\mathbf{P},\mathbf{T}(\mathbf{P})) = \frac{\partial J(\mathbf{P},\mathbf{T})}{\partial \mathbf{P}} + \sum_{c,c'\geq c} \frac{\partial J(\mathbf{P},\mathbf{T})}{\partial \mathbf{T}^{c,c'}} \frac{\partial \mathbf{T}^{c,c'}}{\partial \mathbf{P}} \tag{5.15}$$

The first term in gradient (5.15) supposes that $\mathbf{T}$ is constant and can be computed (Eq. 94-95 [Petersen 2008]) as

$$\frac{\partial J(\mathbb{P},\mathbf{T})}{\partial \mathbb{P}} = \mathbb{P}\left(\frac{2}{\sigma_w^2}\mathbf{C}_b - \frac{2\sigma_b^2}{\sigma_w^4}\mathbf{C}_w\right) \tag{5.16}$$

with $\sigma_w^2 = \langle \mathbf{P}^T\mathbf{P}, \mathbf{C}_w\rangle$ and $\sigma_b^2 = \langle \mathbf{P}^T\mathbf{P}, \mathbf{C}_b\rangle$. The second term in (5.15) is much more difficult to compute because of the Jacobian $\partial \mathbf{T}^{c,c'}/\partial \mathbf{P}$. A possible way to compute the Jacobian $\partial \mathbf{T}^{c,c'}/\partial \mathbb{P}$ is to use the implicit function theorem as in hyperparameter estimation in ML [Bengio 2000, Chapelle 2002]. Sadly in our case it requires inverting a very large matrix, which does not scale in practice. It also assumes that the exact optimal transport $\mathbf{T}_\lambda$ is obtained at each iteration, which is clearly an approximation since we only have the computational budget for a finite, and usually small, number of Sinkhorn iterations. In the paper we proposed to differentiate the transportation matrices obtained after running exactly L Sinkhorn iterations, with a predefined L using auto-differentiation similarly to [Bonneel 2016, Genevay 2017b] and as discussed in section 2.2.2.

In addition to automatic differentiation of the Sinkhorn algorithm, we proposed to use classical manifold optimization tools such as projected gradient of [Schmidt 2008] or a trust region algorithm as implemented in Manopt/Pymanopt [Boumal 2014, Koep 2016]. The latter toolbox includes tools to optimize over the Stiefel manifold, notably automatic conversions from Euclidean to Riemannian gradients.

Connection between WDA and FDA Equation (5.12) exhibits a classical ratio between the inter- and intra-class variance but with covariance matrices that depend on the projector $\mathbf{P}$. When the regularization parameter λ grows to infinity, the transport matrices between $\mathbf{P}\#\mu^c$ and $\mathbf{P}\#\mu^{c'}$ converge to

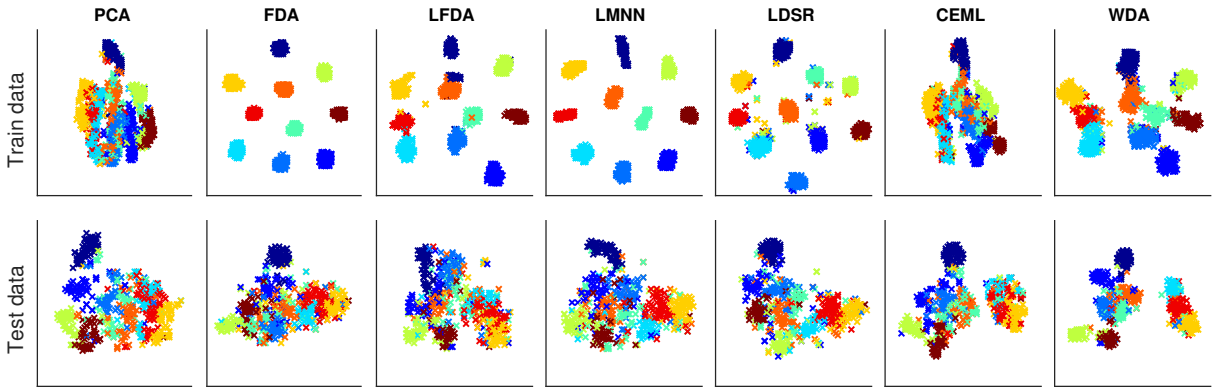


Figure 5.1: 2D tSNE of the MNIST samples linearly projected on $p = 10$ for different approaches. (first line) training set (second line) test set. The method providing the best subspace estimation is the one separating the classes on test data, in this case WDA.

$\mathbf{T} = \frac{1}{n_c n_{c'}} \mathbf{1}_{n_c, n_{c'}}$ and do not depend on the data anymore. The covariance matrices become

$$\mathbf{C}^{c,c'} = \frac{1}{n_c n_{c'}} \sum_{i,j} (\mathbf{P}\mathbf{x}_i^c - \mathbf{P}\mathbf{x}_j^{c'})(\mathbf{P}\mathbf{x}_i^c - \mathbf{P}\mathbf{x}_j^{c'})^T$$

and the matrices $\mathbf{C}_{w,\infty}$ and $\mathbf{C}_{b,\infty}$ correspond then to intra- and inter-class covariance matrices as used in FDA. Since these matrices do not depend on $\mathbb{P}$, the optimization problem (5.9) boils down to the usual Rayleigh quotient which can be solved using a generalized eigen-decomposition of $\mathbf{C}_w^{-1}\mathbf{C}_b$ as in FDA. Note that infinitely regularized WDA is equivalent to FDA when the classes are balanced (in the unbalanced case one needs to weight the covariance matrices with the class ratios).

Applications on real data We investigated the performance of WDA on real datasets such as Caltech and several UCI datasets in [Flamary 2018]. The proposed approach ranked the best across datasets. We now discuss more in details the application to the well known MNIST dataset.

In a first experiment, we wanted to measure how robust our approach is with only few training samples despite high-dimensionality of the problem. To this end, we draw $n = 1000$ samples for training and report the KNN prediction error for the different subspace methods when projecting onto $p = 10$ and $p = 20$ dimensions (detailed results in [Flamary 2018, Fig. 5]). For both p, WDA found a better subspace than the original space which suggests that most of the discriminant information available in the training dataset has been correctly extracted. Conversely, the other approaches struggle to find a relevant subspace in this configuration. In addition to better prediction performance, we want to emphasize that in this configuration, WDA leads to a dramatic compression of the data from 784 to 10 or 20 features while preserving most of the discriminative information.

To gain a better understanding of the corresponding embedding, we have further projected the data from the 10-dimensional space to a 2-dimensional one using t-SNE [Van der Maaten 2008]. In order to make the embeddings comparable, we have used the same initializations of t-SNE for all methods. The resulting 2D projections on the test samples are shown in Figure 5.1. We can clearly see the overfitting behavior of FDA, LFDA [Sugiyama 2007], LMNN [Weinberger 2009] and LDSR [Suzuki 2013] that separate accurately the training samples but fail to separate the test samples. Instead, WDA is able to disentangle classes in the training set while preserving generalization abilities.

5.3 Joint Distribution Domain Adaptation (JDOT)

We introduced a first Domain Adaptation approach based on mapping between distributions in Chapter 3. While this approach has very nice properties and works very well in practice it requires the strong

assumption that the labels are transported along the samples which clearly limits its application in practice. Another limit of the mapping approach is that it is a two-step method that requires first a good approximation for a mapping and then a classifier estimation. In the following we discuss another approach proposed originally in [Courty 2017] that relaxes the "label conservation through mapping" assumption and estimates a classifier simultaneously with the OT between source and target distributions.

5.3.1 Model and theory

The main idea behind Joint Distribution Domain Adaptation (JDOT) [Courty 2017] is to work in the joint feature/label space and align the source and target distributions.

Distributions and proxy distribution The source and target joint distributions are denoted respectively $\mathcal{P}_s$ and $\mathcal{P}_t$. In practice we have only access to a finite source dataset with labels $\hat{\mathcal{P}}_s$ and a finite number of the target examples with no labels $\hat{\mu}_t$. Since we do not have access to the full joint distribution in the target space, it is impossible to compute an OT between those. This is why we proposed in JDOT to use the following target proxy distribution

$$\hat{\mathcal{P}}_t^{\,f} = (\mathbf{x}, f(\mathbf{x}))_{\mathbf{x}\sim\hat{\mu}_t} \tag{5.17}$$

where $f : \Omega \to \mathcal{C}$ is a classifier we want to estimate. This distribution will not be equal to the true $\mathcal{P}_t$ but is a reasonable approximation if we have a good classifier f.

JDOT formulation Now that we have defined the proxy distribution $\hat{\mathcal{P}}_t^{\,f}$ we can use it to express the proposed JDOT optimization problem:

$$\min_{f\in\mathcal{H}} \quad W_{\mathcal{D}}(\hat{\mathcal{P}}_s, \hat{\mathcal{P}}_t^{\,f}) \tag{5.18}$$

where $\mathcal{H}$ is the functional space of f that can be a Reproducing Kernel Hilbert Space (RKHS) or defined through a neural network architecture and the ground loss for the Wasserstein distance is expressed as

$$\mathcal{D}(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2) = \alpha\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \mathcal{L}(y_1, y_2) \tag{5.19}$$

where $\mathcal{L}$ is a Lipschitz continuous loss (regression or classification) and $\alpha > 0$ is a parameter that weights the impact of the features *w.r.t.* the impact of the label loss. In other words we want to estimate the classifier f that best aligns the proxy distribution with the source joint distribution. Note that with this formulation we choose to use a loss that is separable between features and labels since it is much easier to optimize as discussed in the next section. Also note that in oder to avoid overfitting we have added a regularization term on f during the numerical experiments.

Problem (5.18) can be expressed with the primal formulation of OT as

$$\min_{f\in\mathcal{H},\mathbf{T}\in\mathcal{P}} \quad \sum_{i,j} T_{i,j}\left(\alpha\|\mathbf{x}_i^s - \mathbf{x}_j^t\|^2 + \mathcal{L}(y_i^s, f(\mathbf{x}_j^t))\right) \tag{5.20}$$

We can see from this formulation that the OT matrix will be optimal *w.r.t.* the joint feature/label loss $\mathcal{D}$ and will have an effect of label propagation during the training of f. Indeed if some mass is transferred through $\mathbf{T}$ between source sample i and target sample j, minimizing (5.20) *w.r.t.* f will minimize the discrepancy between $f(x_i^s)$ and label y_j^t hence treating sample x_i^s as if he was from class y_j^t. This will be discussed in more details in the next section when we discuss the proposed optimization algorithm.

Generalization bound We provided in [Courty 2017] a generalization bound for JDOT. This bound is of interest since it shows that the generalization error in the target domain can be bounded by the proposed JDOT loss, which explains why the minimization problem (5.18) works well in practice. In

order to achieve this generalization bound we extended the notion of probabilistic Lipschitzness introduced in [Urner 2011,Ben-David 2012]. The extension that we called Probabilistic Transfer Lipschitzness (PTL), assumes that a labeling function must comply with two close instances of each domain w.r.t. a coupling Π between the source and target feature distributions.

Definition 1. *(**Probabilistic Transfer Lipschitzness**) Let μ_s and μ_t be respectively the source and target distributions. Let $\phi : \mathbb{R} \to [0, 1]$. A labeling function $f : \Omega \to \mathbb{R}$ and a joint distribution $\Pi(\mu_s, \mu_t)$ over μ_s and μ_t are ϕ-Lipschitz transferable if for all $\lambda > 0$:*

$$Pr_{(\mathbf{x}_1,\mathbf{x}_2)\sim\Pi(\mu_s,\mu_t)} \left[|f(\mathbf{x}_1) - f(\mathbf{x}_2)| > \lambda d(\mathbf{x}_1, \mathbf{x}_2)\right] \leq \phi(\lambda).$$

Intuitively, given a deterministic labeling functions f and a coupling Π, it bounds the probability of finding pairs of source-target instances labelled differently in a $(1/\lambda)$-ball with respect to Π.

We recall the definition the expected loss in the target domain $R_t(f)$ as $R_t \stackrel{\text{def}}{=} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}_t}\mathcal{L}(y, f(\mathbf{x}))$, the expected error on the source domain is defined similarly as $R_s(f)$. We assume the loss function $\mathcal{L}$ to be bounded, symmetric, k-lipschitz and satisfying the triangle inequality. We can now give our main result (simplified version):

Theorem 5.1. *Let f be any labeling function of $\in \mathcal{H}$. Let $\Pi^* = \arg\min_{\Pi\in\Pi(\mathcal{P}_s,\mathcal{P}_t{}^f)} \int_{(\Omega\times\mathcal{C})^2} \alpha\|\mathbf{x}_i^s - \mathbf{x}_j^t\|^2 + \mathcal{L}(y_s, y_t)d\Pi(\mathbf{x}_s, y_s; \mathbf{x}_t, y_t)$ and $W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}_t{}^f})$ the associated 1-Wasserstein distance. Let $f^* \in \mathcal{H}$ be a Lipschitz labeling function that verifies the ϕ-probabilistic transfer Lipschitzness (PTL) assumption w.r.t. Π^* and that minimizes the joint error $R_s(f^*) + R_t(f^*)$ w.r.t all PTL functions compatible with Π^*. We assume the input instances are bounded s.t. $|f^*(\mathbf{x}_1) - f^*(\mathbf{x}_2)| \leq M$ for all $\mathbf{x}_1, \mathbf{x}_2$. Let $\mathcal{L}$ be any symmetric loss function, k-Lipschitz and satisfying the triangle inequality. Consider a sample of N_s labeled source instances drawn from $\mathcal{P}_s$ and N_t unlabeled instances drawn from μ_t, and then for all $\lambda > 0$, with $\alpha = k\lambda$, we have with probability at least $1 - \delta$ that:*

$$R_t(f) \leq W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}_t{}^f}) + \sqrt{\frac{2}{c'}\log(\frac{2}{\delta})}\left(\frac{1}{\sqrt{N_S}} + \frac{1}{\sqrt{N_T}}\right) + R_s(f^*) + R_t(f^*) + kM\phi(\lambda).$$

The detailed proof of Theorem 5.1 is given in the supplementary material of [Courty 2017]. The bound on the target error above is interesting to interpret. The first two terms correspond to the objective function (5.18) we propose to minimize accompanied with a sampling bound. The last term $\phi(\lambda)$ assesses the probability under which the probabilistic Lipschitzness does not hold. The remaining two terms involving f^* correspond to the joint error minimizer illustrating that domain adaptation can work only if we can predict well in both domains, similarly to existing results in the literature [Mansour 2009, Ben-David 2010]. If the last terms are small enough, adaptation is possible if we are able to align well $\mathcal{P}_s$ and $\mathcal{P}_t{}^f$, provided that f^* and Π^* verify the PTL. Finally, note that $\alpha = k\lambda$ and tuning this parameter is thus actually related to finding the Lipschitz constants of the problem.

5.3.2 Optimization and application

Learning with JDOT According to the hypotheses on f and $\mathcal{L}$, Problem (5.20) is smooth and the constraints are separable according to f and $\mathbf{T}$. Hence, a natural way to solve the problem (5.20) is to rely on alternate optimization w.r.t. both parameters $\mathbf{T}$ and f. This algorithm is also known as Block Coordinate Descent (BCD) or Gauss-Seidel method (the pseudo code of the algorithm is given in the appendix of [Courty 2017]). Block optimization steps are discussed with further details below.

1. Solving with fixed f boils down to a classical OT problem with a loss matrix $\mathbf{C}$ such that $C_{i,j} = \alpha d(\mathbf{x}_i^s, \mathbf{x}_j^t) + \mathcal{L}(y_i^s, f(\mathbf{x}_j^t))$. We can use classical OT solvers such as the network simplex algorithm, but other strategies can be considered, such as regularized OT [Cuturi 2013] or stochastic versions [Genevay 2016].

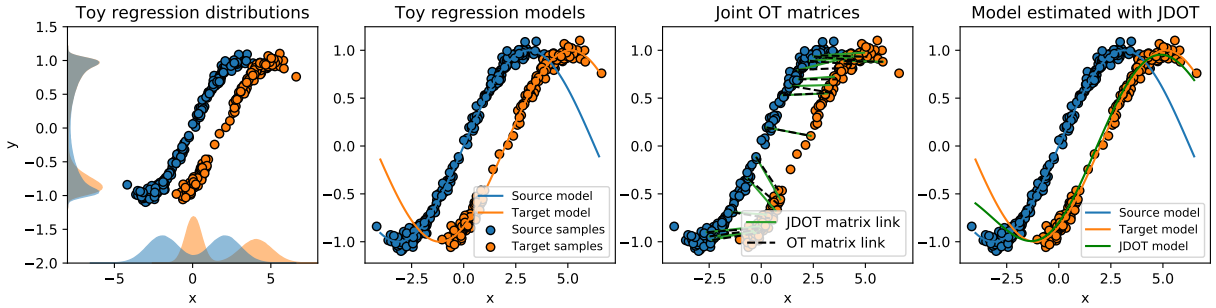


Figure 5.2: Illustration of JDOT on a 1D regression problem. (left) Source and target empirical distributions and marginals (middle left) Source and target models (middle right) OT matrix on empirical joint distributions and with JDOT proxy joint distribution (right) estimated prediction function f.

2. The optimization problem with fixed $\mathbf{T}$ leads to a new learning problem expressed as

$$\min_{f \in \mathcal{H}} \quad \sum_{i,j} T_{i,j} \mathcal{L}(y_i^s, f(\mathbf{x}_j^t)) + \lambda \Omega(f) \tag{5.21}$$

Note how the data fitting term elegantly and naturally encodes the transfer of source labels y_i^s through estimated labels of test samples with a weighting $T_{i,j}$ in the optimal transport matrix.

Let us now discuss briefly the convergence of the proposed algorithm. Owing to the 2-block coordinate descent structure, to the differentiability of the objective function in Problem (5.20) and constraints on f (or its kernel trick parameters) and $\mathbf{T}$ are closed, non-empty and convex, convergence result of Grippo et al. [Grippo 2000] on 2-block Gauss-Seidel methods directly applies. It states that if the sequence $\{\mathbf{T}^k, f^k\}$ produced by the algorithm has limit points then every limit point of the sequence is a critical point of Problem (5.20).

Estimating f for least square regression problems We detail the use of JDOT for transfer least-square regression problem i.e when $\mathcal{L}$ is the squared-loss. In this context, when the optimal transport matrix $\mathbf{T}$ is fixed the learning problem boils down to

$$\min_{f \in \mathcal{H}} \quad \sum_{j} \frac{1}{n_t} \|\hat{y}_j - f(\mathbf{x}_j^t)\|^2 + \lambda \|f\|^2 \tag{5.22}$$

where the $\hat{y}_j = n_t \sum_j \mathbf{T}_{i,j} y_i^s$ is a weighted average of the source target values. Note that this simplification results from the properties of the quadratic loss and that it may not occur for a more complex regression loss. An illustration of JDOT for a simple 1D regression problem can be seen in Figure 5.2. From this Figure we can see that not only the estimated function f performs very well on the target data, but also that the estimated OT matrix with the proxy distribution is actually very similar to the one on the true joint distribution.

Estimating f for hinge loss classification problems We now aim at estimating a multiclass classifier with a one-against-all strategy. We suppose that the data fitting is the binary squared hinge loss of the form $\mathcal{L}(y, f(\mathbf{x})) = \max(0, 1 - yf(\mathbf{x}))^2$. In a One-Against-All strategy we often use the binary matrices $\mathbf{P}$ such that $P_{i,k}^s = 1$ if sample i is of class k else $P_{i,k}^s = 0$. Denote as $f_k \in \mathcal{H}$ the decision function related to the k-vs-all problem. The learning problem (5.21) can now be expressed as

$$\min_{f_k \in \mathcal{H}} \quad \sum_{j,k} \hat{P}_{j,k} \mathcal{L}(1, f_k(\mathbf{x}_j^t)) + (1 - \hat{P}_{j,k}) \mathcal{L}(-1, f_k(\mathbf{x}_j^t)) + \lambda \sum_k \|f_k\|^2 \tag{5.23}$$

where $\hat{\mathbf{P}}$ is the transported class proportion matrix $\hat{\mathbf{P}} = \frac{1}{N_t} \mathbf{T}^\top \mathbf{P}^s$. Interestingly this formulation illustrates that for each target sample, the data fitting term is a convex sum of hinge loss for a negative and positive label with weights estimated from $\mathbf{T}$.

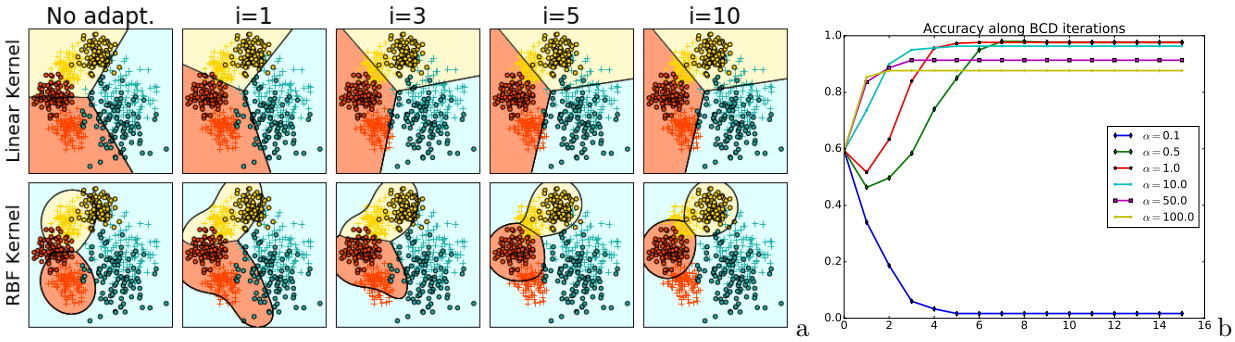


Figure 5.3: Illustration of JDOT for a classification problem. (a): Decision boundaries for linear and RBF kernels on selected iterations. The source domain is depicted with crosses, while the target domain samples are class-colored circles. (b): Evolution of the accuracy along 15 iterations of the method for different values of the α parameter;

We illustrate the behavior of our method on a 3-class toy example (Figure 5.3). We consider a classification problem using the hinge loss and $\mathcal{H}$ is a Reproducing Kernel Hilbert Space. Source domain samples are drawn from three different 2D Gaussian distributions with different centers and standard deviations. The target domain is obtained rotating the source distribution by $\pi/4$ radian. Two types of kernel are considered: linear and RBF. In Figure 5.3.a, one can observe on the first column of images that using directly a classifier learned on the source domain leads to bad performances because of the rotation. We then show the iterations of the block coordinate descent which allows one to recover the true labels of the target domain. It is also interesting to examine the impact of the α parameter on the success of the method. In Figure 5.3.b, we show the evolution of classification accuracy for six different α in the case of RBF kernel. Relying mostly on the label cost ($\alpha = \{0.1\}$) leads to a deterioration of the final accuracy. Using only the input space distance ($\alpha = \{50, 100\}$) allows a performance gain. But it is clear that using both losses with $\alpha = \{0.5, 1, 10\}$ leads to the best performance. Also note the small number of iterations required (< 10) for achieving a steady state.

JDOT in practice We applied JDOT on two different transfer tasks of classification and regression on real datasets. We provided an Open Source Python implementation of JDOT with examples for reproducing the figures on GitHub [1].

JDOT was evaluated on classification tasks on the Caltech Office [Saenko 2010] and Amazon review datasets [Blitzer 2006]. For Caltech office JDOT was compared to all the methods evaluated in [Courty 2016a] and performed better in average than all other approaches. On Amazon review, JDOT was compared to the state of the art approach on this dataset namely Domain Adversarial Neural Network [Ganin 2015] which was outperformed by JDOT using the same neural network architecture.

For the regression task, we used the cross-domain indoor Wifi localization dataset that was proposed by Zhang and co-authors [Zhang 2013], and recently studied in [Gong 2016]. Two cases of adaptation were considered: transfer across periods, for which three time periods t1, t2 and t3 are considered, and transfer across devices, where three different devices are used to collect the signals in the same straight-line hallways (hallway1-3). JDOT had encouraging performance (best out of 3) on the adaptation between periods. On the adaptation between devices, it had tremendous performances with 99% accuracy in the localization when best competitors were around 87%.

5.3.3 DeepJDOT and extensions

JDOT had very encouraging results on real data but the experiments were limited to small scale datasets due to the global OT optimization problem in (5.18). Another question raised by JDOT is the space in

[1] JDOT code: https://github.com/rflamary/JDOT

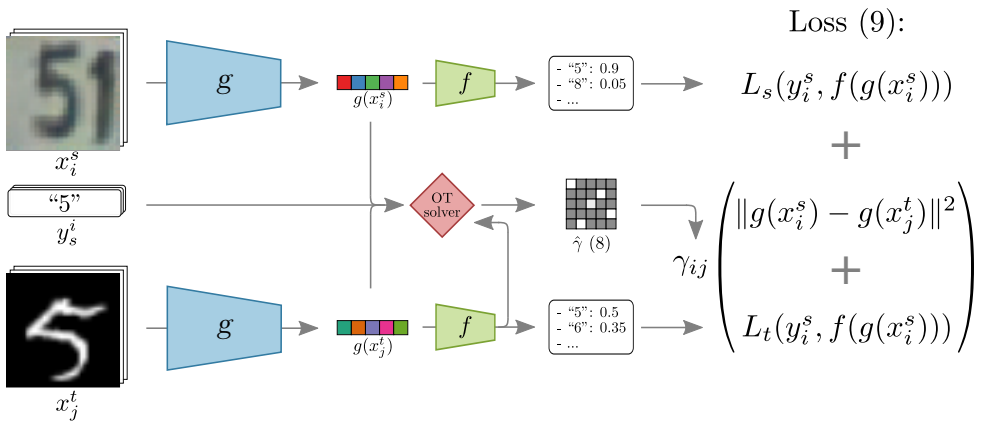


Figure 5.4: Overview of the DeepJDOT method. While the structure of the feature extractor g and the classifier f are shared by both domains, they are represented twice to distinguish between the two domains. Both the latent representations and labels are used to compute per batch a coupling matrix γ that is used in the global loss function.

which the similarity between features should be computed. In [Courty 2017] we used the raw data space which is known to perform rather poorly on images.

Learning feature extraction and adaptation jointly The main idea behind DeepJDOT is to learn a representation of the data g simultaneously with the classifier f on this representation. The final classifier is $f \circ g$ and the JDOT computes the distance in the feature space using the embedding of the data through g instead of the raw original space. The proposed optimization problem can be expressed as

$$\min_{\mathbf{T}\in\mathcal{P},f,g} \quad \frac{1}{n^s}\sum_i L_s\left(y_i^s, f(g(x_i^s))\right) + \sum_{i,j} T_{ij}\left(\alpha\|g(x_i^s) - g(x_j^t)\|^2 + \lambda_t \mathcal{L}\left(y_i^s, f(g(x_j^t))\right)\right). \tag{5.24}$$

We can see from the equation above that it is an extension of JDOT where the feature extraction g is learned simultaneously and used to compute the transport between the joint distributions. But learning a representation working on unlabeled target data is prone to overfitting so we also add a classification loss on the original data that will help in avoiding catastrophic forgetting of the source domain. An illustration of the loss and architecture of the network is given in Figure 5.4.

The devil in the approximation The problem described in (5.24) is very complex to optimize for large datasets due to the global OT problem. We proposed in [Damodaran 2018b] to solve the problem with a stochastic approximation using minibatches from both the source and target domains [Genevay 2017b]. This approach has two major advantages: it is scalable to large datasets and can be easily integrated in modern deep learning frameworks. The objective function (5.24) is approximated by sampling a mini-batch of size m, leading to the following optimization problem:

$$\min_{f,g} \quad \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\left(y_i^s, f(g(x_i^s))\right) + \min_{\gamma\in\mathcal{P}}\sum_{i,j}^{m}\gamma_{ij}\left(\alpha\|g(x_i^s) - g(x_j^t)\|^2 + \lambda_t\mathcal{L}\left(y_i^s, f(g(x_j^t))\right)\right)\right] \tag{5.25}$$

where $\mathbb{E}$ is the expected value with respect to the randomly sampled minibatches of size m drawn from both source and target domains. The classification loss functions for the source and target domains ($\mathcal{L}$) can be any general class of loss functions that are twice differentiable. We opted for a traditional cross-entropy loss in both cases. Note that, as discussed in [Genevay 2017b], the expected value over the minibatches does not converge to the true OT coupling between every pair of samples, which might lead to the appearance of connections between samples that would not have been connected in the full coupling. However, this can also be seen as a regularization that will promote sharing of the mass between

neighboring samples. Finally note that we did not use the regularized version of OT as in [Genevay 2017b], since it introduces an additional regularization parameter that should be cross-validated, which can make the model calibration more complex. Still, the extension of DeepJDOT to regularized OT is straightforward and could be beneficial for high-dimensional embeddings g.

DeepJDOT in practice We applied DeepJDOT to three visual adaptation datasets: Digits classification, Home-Office, and VisDA-2017 dataset. For digit classification we considered four data sources (domains) from the digits classification field: MNIST [Lecun 1998], USPS [Hull 1994], MNIST-M, and the Street View House Numbers (SVHN) [Netzer 2011] dataset. The Office-Home dataset [Venkateswara 2017] contains around 15′500 images in 65 categories from four different domains: artistic paintings, clipart, product and real-world images. The Visual Domain Adaptation classification challenge of 2017 (VisDA-2017; [Peng 2017]) requires training a model on renderings of 3D models for each of the 12 classes and adapting to natural images sampled from MS-COCO [Lin 2014] (validation set) and YouTube BoundingBoxes [Real 2017] (test set), respectively. The test set performances were evaluated on the official competition server.

For all datasets, we compared to a number of state of the art approaches such as DeepCORAL [Sun 2016] and DANN [Ganin 2016]. Details of the numerical experiments are given in [Damodaran 2018b], but DeepJDOT had consistently better performances compared to all the other methods. We also did an ablation study where we removed some terms in the objective function and it illustrated the importance of each term in (5.25).

For an interpretation of the performance of DeepJDOT we visualized the quality of the embeddings for the source and target domain learnt by DeepJDOT and DANN using t-SNE embedding on the MNIST→MNIST-M adaptation task (Figure 5.5). As expected, on the model trained on source data the samples from the source domain are well clustered and target samples are more scattered. The t-SNE embeddings with the `DANN` are better but not able to align the distributions well. DeepJDOT perfectly aligns the source domain samples and target domain samples per class, which explains the good numerical performances reported above. The "tentacle"-shaped and near-perfect separation of the classes in the embedding illustrate the fact that DeepJDOT finds an embedding that both aligns the source/target distribution, but also maximizes the margin between the classes.

Label denoising with DeepJDOT The optimization problem from DeepJDOT can be adapted to other learning problems. Indeed the fact that the labels are propagated through the OT matrix suggests that when using regularization, multiple labels will be propagated on the samples performing some kind of label smoothing on the samples. For this reason we investigated the use of regularized OT in the DeepJDOT framework in the presence of label noise in [Damodaran 2018a]. In this case we proposed to optimize the following problem

$$\min_{f\in\mathcal{H}} \quad \sum_{i,j} T_{i,j}\left(\alpha\|\mathbf{x}_i-\mathbf{x}_j\|^2+\mathcal{L}(y_i,f(\mathbf{x}_j))\right) \tag{5.26}$$

where $\mathbf{T}$ is the solution of entropic regularized OT using ground loss 5.19 between the noisy training data $(\mathbf{x}_i, y_i)$ and the proxy distribution $(\mathbf{x}_i, f(\mathbf{x}_i))$. The regularization will perform a smoothing of the labels around its neighbors in the joint feature/label space which will provide robustness. The method has been evaluated on real life remote sensing applications and performed very well for training deep neural networks compared to other robust losses.

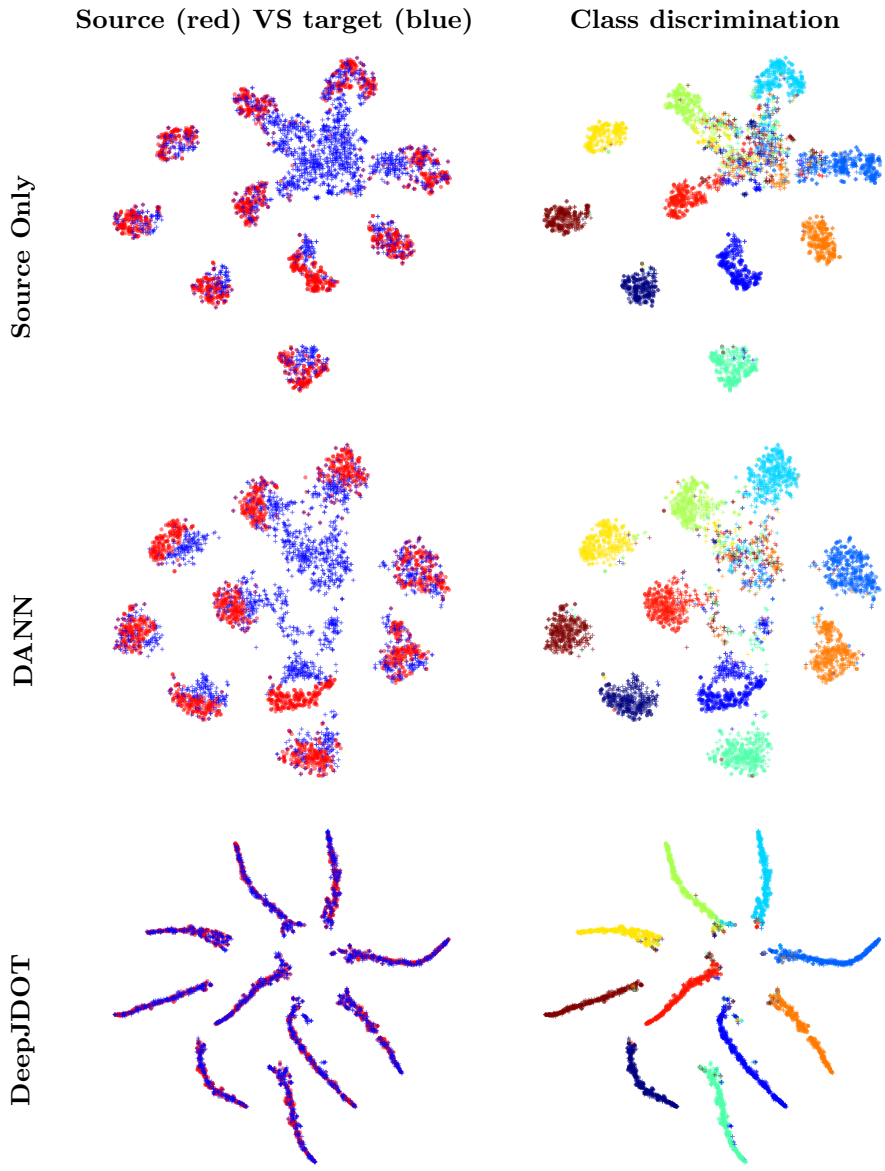


Figure 5.5: t-SNE embeddings of 2'000 test samples for MNIST (source) and MNIST-M (target) for `Source only` classifier, `DANN` and `DeepJDOT`. The left column shows domain comparisons, where colors represent the domain. The right column shows the ability of the methods to discriminate classes (samples are colored w.r.t. their classes).

Chapter 6

Optimal transport for structured data

Contents

In this short chapter, I discuss the application of OT when the distributions do not live in the same ambient space. In this case it is obvious that one cannot compute a ground metric c between samples in source and target domains. I will present first an extension of OT called the Gromov-Wasserstein distance (GW) proposed in [Mémoli 2011]. Next I introduce the Fused Gromov-Wasserstein distance (FGW) that we proposed recently to measure similarity between labeled graphs. The last section describes some applications of FGW on classical graph processing problems.

In the following I will define GW and FGW on discrete distributions for readability reasons but most of the discussed properties have been proven for general distributions (see [Mémoli 2011, Vayer 2018]).

6.1 Gromov-Wasserstein distance

The Gromov-Wasserstein (GW) distance has been introduced in [Mémoli 2011] to provide a similarity measure and a matching between objects. The main idea behind GW is to look at (and align) pairwise relations between samples in their respective domains and to seek for a transport between these pairwise relations.

6.1.1 Definition and properties

Gromov-Wasserstein distance Let $\mu_X = \sum_i h_i \delta_{\mathbf{x}_i}$ be a source distribution having a support in $\mathcal{X}$, and $\mu_Y = \sum_j g_i \delta_{\mathbf{y}_j}$ be the target distribution with a support in $\mathcal{Y}$. The GW distance is defined as

$$\mathcal{GW}_p(\mathbf{D}, \mathbf{D}', \mu_X, \mu_Y) = \left(\min_{\mathbf{T} \in \Pi(\mu_X, \mu_Y)} \sum_{i,j,k,l} |D_{i,k} - D'_{j,l}|^p T_{i,j}\, T_{k,l} \right)^{\frac{1}{p}} \tag{6.1}$$

where $D_{i,k} = \|\mathbf{x}_i - \mathbf{x}_k\|, D'_{j,l} = \|\mathbf{y}_j - \mathbf{y}_l\|$ are the pairwise distances between samples in source and target respectively. The GW is a metric for $p \geq 1$ and the optimal matrix $\mathbf{T}$ in the problem above gives a correspondence between source and target samples that aligns the pairwise relationship between the samples across different metric spaces.

GW properties This distance has interesting properties such as being null if and only if there exists an isometry between the two distributions. It is also invariant to rotations and translations of the distributions. GW has been used to measure similarity between graphs where the pairwise distance between nodes can be computed for instance with shortest path. It is a sensible measure of similarity between surfaces in computer graphics where the rotation invariance is of particular interest. [Mémoli 2011]. It can also be used for the estimation of GW barycenters between distributions lying in different spaces with very nice interpolation between weighted surfaces in [Peyré 2016]. Finally note that the GW distance has been used as a data fitting term to train a generative model across different spaces in [Bunne 2019].

6.1.2 Solving Gromov-Wasserstein

Solving the optimization problem Solving the optimization problem in 6.1 is very difficult. It is a high dimensional non-convex constrained Quadratic Program (QP) that is much more complex and memory intensive than the Linear Program of classical OT. Since it is non convex one can only hope to find local stationary points in practice.

Interestingly solving the GW problem 6.1 is equivalent to solve a general regularized OT of the form 2.33 where $\mathbf{C} = \mathbf{0}$ and $\Omega(\mathbf{T})$ is the non-convex quadratic term. One simple approach that we implemented in the POT toolbox [Flamary 2017b] consists in using Conditional Gradient (CG) to find a local solution. CG relies on solving at each iteration a linearization of the objective function which in our case boils down to solving a classical OT of complexity $O(n^3 \log(n))$.

Another approach proposed in [Peyré 2016] was to regularize the problem with entropy. In this case the regularization has a smoothing (and convexification effect) and similarly to entropic OT the transport matrix is not sparse anymore. But the problem can be solved with a simple projected gradient descent where each projection can be done with the efficient Sinkhorn algorithm.

Gromov-Wasserstein in 1D GW has been introduced only recently in the mathematical and ML community and some of its behavior in special cases have not yet been investigated. But we looked at the problem in 1D and have proven that there exists a closed form for GW in 1D in [Vayer 2019b]. Interestingly 1D GW has also a closed form relying on quantile functions and similar to the solution for 1D Wasserstein. In a nutshell, for uniform weights and the same number of sample, the solution when the samples on the left and right have been sorted is either the identity matrix on the anti-identity matrix (invariance to rotations). This means that one can find the optimal solution by computing the GW loss on both (Can be efficiently computed in $O(n)$. The GW in 1D is a specific structure of Quadratic Assignment Problem that can be solved efficiently using a $O(n \log(n))$ sorting followed by a $O(n)$ computation (same complexity as 1D Wasserstein). We find it very interesting that in 1D one can solve and compute this problem in $O(n \log(n))$ when the computation of the objective value itself is $O(n^3)$ [Peyré 2016] in the general case.

This efficient solver in 1D opens the door to Sliced Radon Gromov Wasserstein which can be used on a very large number of samples for generative model training or Sliced GW barycenter estimation [Vayer 2019b]. For example, one can solve Gromov Wasserstein in 1D (so one projection in sliced radon configuration) on 1 million samples in about 10^{-2} seconds.

6.2 Fused Gromov-Wasserstein distance

In the previous section, we introduced GW that can measure similarity between graphs. But one limit of GW is that when the graphs in source and target are labeled it cannot encode this label. Before introducing the Fused Gromov-Wasserstein distance we first need to express the labeled graph as a distribution.

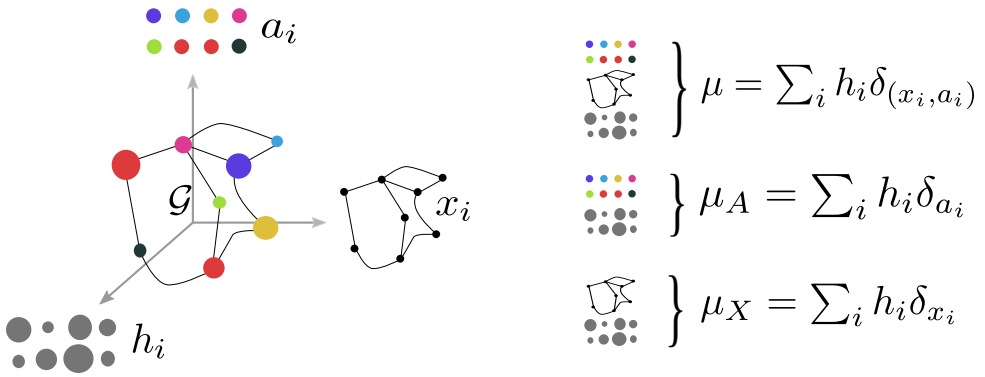


Figure 6.1: Illustration of a labeled graph modeled as a distribution. (Left) Labeled graph with $(a_i)_i$ its feature information, $(x_i)_i$ its structure information and histogram $(h_i)_i$ that measures the relative importance of the vertices. (Right) Associated structured data which is entirely described by a fully supported probability measure μ over the product space of feature and structure, with marginals μ_X and μ_A on the structure and the features respectively.

6.2.1 Structured object as a distribution

We proposed in [Vayer 2019a] to model a labeled graph as the following distribution (see Figure 6.1):

$$\mu = \sum_i h_i \delta_{(\mathbf{x}_i, \mathbf{a}_i)} \tag{6.2}$$

where the Diracs have a position in the joint structure/label. $\mathbf{x}_i$ encodes a position in the structure space, the structure relation between the nodes is encoded through the pairwise relations $\|\mathbf{x}_i - \mathbf{x}_j\|$ that can be computed implicitly for instance with a shortest path between nodes. The label in each node is denoted as $\mathbf{a}_i$ and it can be compared across graphs so it lies in the same space for all objects. h_i is the weight of each node in the distribution, by default one can use uniform weights. The joint structure/label space allows to encode simultaneously the graph structure and the label information.

6.2.2 Fused Gromov-Wasserstein

We proposed in [Vayer 2019a] the FGW between $\mu_s = \sum_i h_i \delta_{(\mathbf{x}_i, \mathbf{a}_i)}$ and $\mu_t = \sum_j g_j \delta_{(\mathbf{y}_j, \mathbf{b}_j)}$:

$$\mathcal{FGW}_{p,q,\alpha}(\mathbf{D}, \mathbf{D}', \mu_s, \mu_t) = \left(\min_{\pi \in \Pi(\mu_s, \mu_t)} \sum_{i,j,k,l} \left((1-\alpha) M_{i,j}^q + \alpha |D_{i,k} - D'_{j,l}|^q\right)^p T_{i,j}\, T_{k,l} \right)^{\frac{1}{p}} \tag{6.3}$$

where $M_{i,j} = c(\mathbf{a}_i, \mathbf{b}_j)$ measures the divergence between labels and $0 \leq \alpha \leq 1$ is a parameter that weights the relative importance of the labels and structures. The optimization problem above is very similar to the one of GW but comes with an additional linear term. It is clearly an interpolation between GW on structure alone ($\alpha = 1$) and Wasserstein on nodes labels $\alpha = 0$. In order to solve the optimization problem, we proposed in [Vayer 2019a] to use a Conditional Gradient that relies on iteratively solving linear OT problems. It can also be solved using entropic regularization with a slight change in the projected gradient algorithm of [Peyré 2016].

Theoretical properties The FGW distance described above has several nice theoretical properties that were proven in [Vayer 2018]. First FGW is a metric over structured data that is invariant to measure and feature preserving isometries. It is a true metric when $q = 1$ and a semi-metric when $q > 1$ far any $p \geq 1$. FGW also have nice geometrical properties for continuous distributions such as uniquely defined constant speed geodesics which open the door to Fréchet means (or barycenters). Finally it is an upper bound for the GW and Wasserstein distances and we have shown in [Vayer 2018] that FGW in the same space has the same sample complexity as Wasserstein distance of $O(n^{-1/d})$.

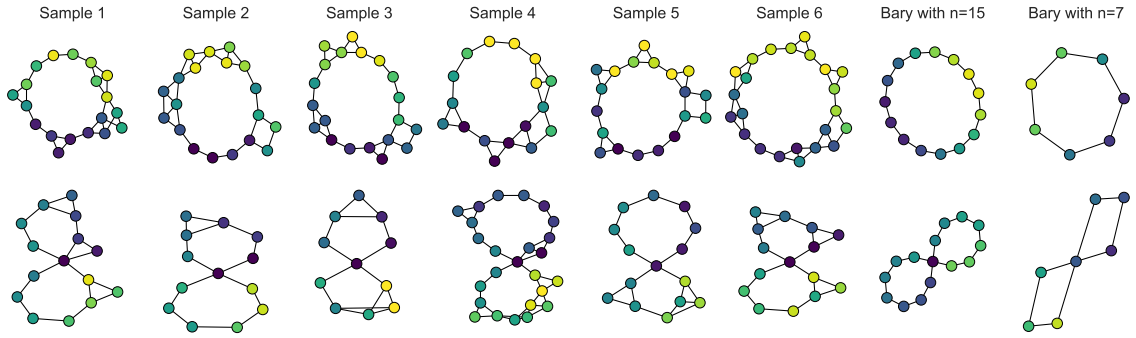


Figure 6.2: Illustration of *FGW* graph barycenter. Columns 1 to 6 are noisy samples that constitute the datasets. Columns 6 and 7 show the barycenters for each setting, with different number of nodes. Blue nodes indicates a feature value close to -1, yellow nodes close to 1.

6.2.3 Fused Gromov-Wasserstein barycenters

Since FGW is a meaningful similarity measure between graphs we proposed to estimate FGW barycenters corresponding to the following optimization problem:

$$\min_{\mathbf{C},\mu} \quad \sum_k \lambda_k \mathcal{FGW}_{p,q,\alpha}(\mathbf{D}, \mathbf{D}_k, \mu, \mu_k) \tag{6.4}$$

where $(\mathbf{D}_k, \mu_k)$ describes object k and λ_k are the relative weights of the objects. Note that optimizing $\mathbf{D}$ corresponds to the estimation of the structure of the barycenter. For instance when $\mathbf{D}_k$ corresponds to shortest path on the graph, one can recover the graph structure of the barycenter using a threshold on $\mathbf{D}$. The features of the graph (its labels) are encoded in μ through the matrix of labels $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\top$. The optimization problem above is hard to solve but one can relatively easily get a stationary point by using a block coordinate descent algorithm that corresponds to updating alternatively the FGW transport matrices and the structure/feature matrices of the barycenter similarly to what was proposed in [Peyré 2016].

6.3 Applications on graphs

FGW is a novel distance between labeled graphs and can be applied to numerous applications. We present in the following several illustrative applications of FGW.

6.3.1 Graph classification

FGW is a principled metric between labeled graph that can encode both the structure of the graph and the labels on the nodes and can be used to train template based classifiers. In [Vayer 2019a], we designed a very simple non-positive kernel of the form $k(\cdot,\cdot) = \exp(-FGW(\cdot,\cdot))$. This kernel was used to train support vector machines on several graph classification datasets. We achieved very competitive performance even when compared to deep learning methods such as Patchy-SAN Graph Convolutional Networks [Niepert 2016]. This result is particularly impressive since we compare our FGW on simple label/graph representation to a much more complex supervised feature extraction network but still can outperform it.

Note that an interesting development would be to train an embedding that can represent the label/structure spaces on each graph for prototype based classifier. Note that one can also directly train a linear classifier on prototypes with FGW distance, that would be an extension of the Dissimilarity Measure Machines framework proposed in [Rakotomamonjy 2018].

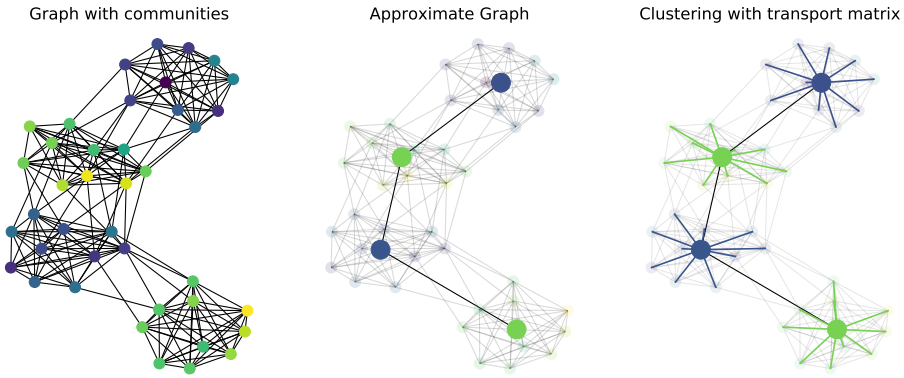


Figure 6.3: Example of community clustering on graphs using FGW. Community clustering with 4 communities and uniform features per cluster.

6.3.2 Graph barycenters and community clustering

Graph barycenters, compression and clustering Another interesting application of FGW is to find a barycenter for a family of graph. We illustrate barycenters for two families of "noisy" graphs (one with a circle structure and one with an ∞ structure) in Figure 6.2. The barycenters with $n = 15$ and $n = 7$ nodes are reported in the last two columns of the figure. One can see that the barycenters recover both the overall structure of the family and the labels (colors). Note that by selecting the number of nodes in the barycenter one can compress the graph or estimate a "high resolution" representation from all the samples. To the best of our knowledge, no other method can compute such graph barycenters. Finally since we can now compute Fréchet means of graphs, we can extend classical algorithms sur as K-means to estimate graph centroids from a dataset of clustered graphs. We show the evolution of the cluster centers and final clusters in [Vayer 2019a, Fig. 5].

Graph community clustering As discussed above, one can estimate a "compressed" version of a given graph with a smaller number of nodes. In practice this can be used to perform community clustering in graphs. We generate a community graph illustrated in the left column of Figure 6.3. We can see that the relation between the blocks is sparse and the features also follow the graph clusters (noisy but similar in each block). The graph approximation shown in the center column of Figure 6.3 is done with 4 nodes and we can recover both the blocks in the graph and the average feature on each blocks. Note that in addition to finding 4 nodes corresponding to the communities, we also have access as illustrated in the right column of Figure 6.3 to a transport matrix that gives the relation between the nodes in the original graph and the cluster nodes in the compressed graph.

Chapter 7

Concluding remarks

Contents

This chapter provides concluding remarks about this document. It first describes some current and future works that we plan to investigate. Finally I discuss some more fundamental and general questions that will be in my opinion key to the future of OT for ML.

7.1 Current and future work

7.1.1 Optimal Transport on graphs

The Fused Gromov-Wasserstein distance described in Chapter 6 has been illustrated on several examples of data mining on graphs. But a lot of work remains to be done in order to make it a general tool for use in Graph Processing. First I believe that it would be very interesting to investigate the link between FGW and Signal Processing on graphs [Shuman 2012] that is a very active community. One first direction would be to study at what happens on the eigenvectors and eigenvalues when computing the barycenter of two distributions that share the same graph.

Another direction we want to investigate is to propose some structure in the ground distance matrix of the graph barycenter such as sparsity and group structure. Promoting such a structure will help getting better estimators since we can encode prior knowledge directly in the estimation. Having an additive structure (with a dictionary) for the distance matrix might even lead to novel applications such as graph dictionary learning where a manifold with FGW metric can be estimated.

7.1.2 Estimating the Monge mapping

The problem of estimating the continuous Monge mapping from empirical distributions has only been investigated very recently. In my opinion, despite recent results [Flamary 2019, Hütter 2019, Paty 2019b] the community is still searching for a general estimator that has known statistical properties (in term of convergence) and can be computed in practice. The question of the use of regularized OT is still open since the effect of the regularization is to split the mass, which seems counterproductive for mapping estimation.

One important research direction in my opinion is to investigate how to find good estimator for nonlinear Monge mapping. [Paty 2019b] proposed in a recent paper to include regularity directly in the Brenier potentials, which leads to a very elegant (but not efficient) Quadratically Constrained QP (QCQP). They can also provide out of sample mapping by solving a smaller QCQP. Another approach also relying on smoothness of the mapping would be to estimate Monge mapping on finite order polynomial basis (linear, quadratic, ...). Interestingly, in this case the positivity constraint on the Jacobian of the mapping, corresponds to semi algebraic constraints on the polynomials. We plan on studying the effect of those constraints on the polynomial parameters to evaluate the quality of estimation of a linear and higher order polynomials mappings.

7.1.3 Wasserstein on minibatches

One major problem of OT is its computational complexity. It has been a problem for large scale implementation especially when used to train neural networks with stochastic gradient updates. Wasserstein GAN proposed to solve the problem in the dual and to estimate the dual variable with a NN [Arjovsky 2017] but some approximations had to be done since the constraints of the dual variable are impossible to encode in practice. Using regularized OT makes the problem scalable but still requires to solve large problems. In order to use it for training neural networks with SGD updates, [Genevay 2017a] proposed to compute the OT on minibatches. Since then, OT on minibatches has been used in practice in several recent works from domain adaptation [Damodaran 2018b] to GAN [Deshpande 2018, Kolouri 2018].

One important question is then: what happens when we minimize the expectation of OT on minibatches? It clearly leads to a biased solution *w.r.t.* the solution of Wasserstein distance. We plan on investigating why optimizing on minibatches works in practice and look at the statistical estimators and results of stochastic optimization. Another question is the effect of the size of the minibtach. In practice one can see this as a regularization parameter, so the effect of this regularization when combined with entropic regularization as proposed in [Genevay 2017a] is still an open problem.

7.1.4 Adversarial regularization with Wasserstein

One major recent development in deep learning was the introduction of adversarial training [Goodfellow 2014b, Miyato 2016], that aims at training neural networks that are robust to adversarial examples. This can be seen as a regularization of the neural network that will forbid quick change in classification output (*w.r.t.* the input) and promote robustness. Recent approaches such as virtual adversarial training (VAT) [Miyato 2018] create virtual adversarial samples that correspond to the closest samples that lead to the largest change in the classifier output.

Those methods have reached state of the art performances on semi supervised learning. But they treat all classes uniformly (the regularization is isotropic) despite the fact that on can often have access to some additional information (such as similarity between classes). We proposed in our preliminary works in [Damodaran 2019] to encode these relations in the estimation of adversarial examples using Wasserstein distance on the output of the classifier. This will allow us to promote smoothness between classes that are known to be similar/noisy but keep a complex border when necessary.

7.2 The big OT for ML questions

7.2.1 Statistical properties of Wasserstein distance

One well known bottleneck for the use of Wasserstein distance in ML from measuring similarities between empirical description is its very slow convergence of $O(n^{-1/d})$ in terms of number of samples n that depends on the dimension d. Similar convergence has also been proven for the Monge mapping estimation [Hütter 2019]. This leads to the following question: do we really want to use the Wasserstein distance or Monge mapping in practice?

Researchers have partially answered this question with the negative. In [Genevay 2018], the authors have studied the statistical properties of the Sinkhorn divergence and shown a much better convergence speed of $O(n^{-1/2})$ (for a fixed regularization parameter). Since exact OT can be computationally expensive, Sinkhorn divergence seems to be an elegant and efficient alternative. But the question of how to select the regularization term is still open and application dependent. We can also note the burgeoning Sliced Wasserstein distance that can in some way avoid the curse of dimensionality and has been used with success for training generative models [Kolouri 2018, Deshpande 2018, Liutkus 2018]. All the trending numerical approaches in ML seem to go away from the original Wasserstein, and they all correspond to some kind of regularization that attenuate the large variance of OT. The question of how to keep the nice geometric properties of OT and avoid its statistical shortcoming will be key to the future of OT for ML.

7.2.2 Large scale optimization

The slow convergence of Wasserstein discussed above suggests to use it on massive datasets. One major problem in this case is the complexity of solving the OT problem. It is known to be $O(n^3 \log(n))$ for exact OT and nearly $O(n^2)$ for entropic OT which still cannot scale well to very large datasets. Stochastic optimization procedures as proposed in [Genevay 2016] and [Seguy 2018] do scale but can still be very slow on large datasets for small regularization terms. This limits their application when the Sinkhorn divergence is the objective value of the optimization problem.

In order to tackle this complexity, we can observe as discussed in the previous section a trend that consists in optimizing the expectation of the Wasserstein distance over minibatches [Genevay 2017b, Damodaran 2018b]. This is interesting in practice since it allows the use of classical SGD with IID minibatches but the objective value is very different from the Wasserstein distance or its entropic counterpart. Other approaches tend to replace the Wasserstein with more efficient loss such as Sliced-Radon Wasserstein but still use minibatches [Kolouri 2018]. The question of how to scale OT solvers is strongly related to the discussion above about statistical properties and will also be central in a near future.

7.2.3 Learning the ground metric

Finally another important question in OT is the choice of the ground metric [Cuturi 2014b]. In practice, the Euclidean distance is often used but this is known to be a poor measure of similarity for instance when data lie in a manifold. For specific problem, one can design some ground metric using prior knowledge similarly to the design of a kernel for SVM as we did for OST [Flamary 2016]. We also proposed to estimate the equivalent of a Mahalanobis distance in [Flamary 2018] that will optimize a separability measure between classes. A similar approach estimating a robust subspace has been proposed in [Paty 2019a].

Some works such as [Genevay 2017b, Bellemare 2017] have proposed to learn an adversarial deep embedding (maximize *w.r.t.* the embedding) and it seems to work well for generative training. Still the question of finding a ground metric optimal with respect to a given objective without overfitting is important. While simple linear embeddings as discussed above are one way. The question of learning a regular ground metric so that it conserves smoothness (and why not convexity) remains an open problem that will have to be solved if we want to apply OT in a wide range of ML applications.

Bibliography

[Absil 2009] P-A Absil, Robert Mahony et Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. Princeton University Press, 2009. (Cited on page 41.)

[Agueh 2011] Martial Agueh et Guillaume Carlier. *Barycenters in the Wasserstein space*. SIAM Journal on Mathematical Analysis, vol. 43, no. 2, pages 904–924, 2011. (Cited on pages 10 and 34.)

[Altschuler 2017] Jason Altschuler, Jonathan Weed et Philippe Rigollet. *Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration*. In Advances in Neural Information Processing Systems, pages 1964–1974, 2017. (Cited on page 13.)

[Alvarez-Melis 2017] David Alvarez-Melis, Tommi S Jaakkola et Stefanie Jegelka. *Structured Optimal Transport*. arXiv preprint arXiv:1712.06199, 2017. (Cited on pages 16 and 24.)

[Ammanouil 2017] Rita Ammanouil, Andre Ferrari, Remi Flamary, Chiara Ferrari et David Mary. *Multi-frequency image reconstruction for radio-interferometry with self-tuned regularization parameters*. In European Conference on Signal Processing (EUSIPCO), 2017. (Cited on pages vii and viii.)

[Appell 1887] Paul Appell. *Mémoire sur les déblais et les remblais des systemes continus ou discontinus*. Mémoires présentes par divers Savants à l'Académie des Sciences de l'Institut de France, vol. 29, pages 1–208, 1887. (Cited on page 8.)

[Arjovsky 2017] Martin Arjovsky, Soumith Chintala et Léon Bottou. *Wasserstein gan*. arXiv preprint arXiv:1701.07875, 2017. (Cited on pages iii, 2, 17, 38, 40 and 58.)

[Bassetti 2006] Federico Bassetti, Antonella Bodini et Eugenio Regazzini. *On minimum Kantorovich distance estimators*. Statistics & probability letters, vol. 76, no. 12, pages 1298–1302, 2006. (Cited on page 38.)

[Bauschke 2011] Heinz H Bauschke, Patrick L Combettes*et al.* Convex analysis and monotone operator theory in hilbert spaces, volume 408. Springer, 2011. (Cited on page 16.)

[Bellemare 2017] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer et Rémi Munos. *The cramer distance as a solution to biased wasserstein gradients*. arXiv preprint arXiv:1705.10743, 2017. (Cited on page 59.)

[Ben-David 2010] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira et J. Wortman Vaughan. *A theory of learning from different domains*. Machine Learning, vol. 79, no. 1-2, pages 151–175, Mai 2010. (Cited on page 45.)

[Ben-David 2012] S. Ben-David, S. Shalev-Shwartz et R. Urner. *Domain Adaptation–Can Quantity compensate for Quality?* In Proc of ISAIM, 2012. (Cited on page 45.)

[Benamou 2003] Jean-David Benamou. *Numerical resolution of an "unbalanced" mass transport problem*. ESAIM: Mathematical Modelling and Numerical Analysis, vol. 37, no. 5, pages 851–868, 2003. (Cited on page 14.)

[Benamou 2015] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna et Gabriel Peyré. *Iterative bregman projections for regularized transportation problems*. SIAM Journal on Scientific Computing, vol. 37, no. 2, pages A1111–A1138, 2015. (Cited on pages 2, 13, 14, 15, 16, 28 and 32.)

[Bengio 2000] Yoshua Bengio. *Gradient-based optimization of hyperparameters.* Neural computation, vol. 12, no. 8, pages 1889–1900, 2000. (Cited on page 42.)

[Bertsekas 1981] Dimitri P Bertsekas. *A new algorithm for the assignment problem.* Mathematical Programming, vol. 21, no. 1, pages 152–171, 1981. (Cited on page 12.)

[Bhatia 2018] Rajendra Bhatia, Tanvi Jain et Yongdo Lim. *On the Bures–Wasserstein distance between positive definite matrices.* Expositiones Mathematicae, 2018. (Cited on page 9.)

[Bigot 2017] Jérémie Bigot, Raúl Gouet, Thierry Klein, Alfredo López*et al. Geodesic PCA in the Wasserstein space by convex PCA.* In Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, volume 53, pages 1–26. Institut Henri Poincaré, 2017. (Cited on pages 28 and 34.)

[Bigot 2018a] Jérémie Bigot, Elsa Cazelles et Nicolas Papadakis. *Data-driven regularization of wasserstein barycenters with an application to multivariate density registration.* arXiv preprint arXiv:1804.08962, 2018. (Cited on page 15.)

[Bigot 2018b] Jérémie Bigot, Elsa Cazelles et Nicolas Papadakis. *Penalization of barycenters in the Wasserstein space.* 2018. (Cited on page 15.)

[Blanchet 2016] Jose Blanchet, Yang Kang et Karthyek Murthy. *Robust Wasserstein profile inference and applications to machine learning.* arXiv preprint arXiv:1610.05627, 2016. (Cited on page 39.)

[Blitzer 2006] J. Blitzer, R. McDonald et F. Pereira. *Domain adaptation with structural correspondence learning.* In Proc. of the 2006 conference on empirical methods in natural language processing, pages 120–128, 2006. (Cited on page 47.)

[Blondel 2018] Mathieu Blondel, Vivien Seguy et Antoine Rolet. *Smooth and sparse optimal transport.* In Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (AISTATS), 2018. (Cited on pages 14 and 16.)

[Boisbunon 2014] A. Boisbunon, R. Flamary et A. Rakotomamonjy. *Active set strategy for high-dimensional non-convex sparse optimization problems.* In International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2014. (Cited on pages vii and viii.)

[Bonnans 1998] J Frédéric Bonnans et Alexander Shapiro. *Optimization problems with perturbations: A guided tour.* SIAM review, vol. 40, no. 2, pages 228–264, 1998. (Cited on page 42.)

[Bonneel 2011] N. Bonneel, M. van de Panne, S. Paris et W. Heidrich. *Displacement Interpolation Using Lagrangian Mass Transport.* ACM Transaction on Graphics, vol. 30, no. 6, pages 158:1–158:12, Décembre 2011. (Cited on pages 2, 12 and 35.)

[Bonneel 2014] N. Bonneel, J. Rabin, G. Peyr'e et H. Pfister. *Sliced and Radon Wasserstein Barycenters of Measures.* Journal of Mathematical Imaging and Vision, vol. 51, no. to appear, pages 22–45, to appear 2014. (Cited on page 10.)

[Bonneel 2016] Nicolas Bonneel, Gabriel Peyré et Marco Cuturi. *Wasserstein Barycentric Coordinates: Histogram Regression Using Optimal Transport.* ACM Transactions on Graphics, vol. 35, no. 4, 2016. (Cited on page 42.)

[Bordin 1884] Prix Bordin. *Géometrie: Prix Bordin.* Comptes rendus hebdomadaires des seances de l'Academie des sciences., vol. 281, no. 23, page 1165, 1884. (Cited on page 8.)

[Bottou 2010] Léon Bottou. *Large-scale machine learning with stochastic gradient descent.* In Proceedings of COMPSTAT'2010, pages 177–186. Springer, 2010. (Cited on page 17.)

[Boumal 2014] Nicolas Boumal, Bamdev Mishra, P-A Absil et Rodolphe Sepulchre. *Manopt, a Matlab toolbox for optimization on manifolds.* The Journal of Machine Learning Research, vol. 15, no. 1, pages 1455–1459, 2014. (Cited on page 42.)

[Bredies 2009] K. Bredies, D. A.A Lorenz et P. Maass. *A generalized conditional gradient method and its connection to an iterative shrinkage method.* Computational Optimization and Applications, vol. 42, no. 2, pages 173–193, 2009. (Cited on page 17.)

[Brenier 1991] Yann Brenier. *Polar factorization and monotone rearrangement of vector-valued functions.* Communications on pure and applied mathematics, vol. 44, no. 4, pages 375–417, 1991. (Cited on pages 8, 9 and 19.)

[Bromley 1994] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger et Roopak Shah. *Signature verification using a" siamese" time delay neural network.* In Advances in Neural Information Processing Systems, pages 737–744, 1994. (Cited on page 33.)

[Bunne 2019] Charlotte Bunne, David Alvarez-Melis, Andreas Krause et Stefanie Jegelka. *Learning Generative Models across Incomparable Spaces.* arXiv preprint arXiv:1905.05461, 2019. (Cited on page 52.)

[Burges 2010] Christopher JC Burges. Dimension reduction: A guided tour. Now Publishers, 2010. (Cited on page 40.)

[Cazelles 2018] Elsa Cazelles, Vivien Seguy, Jérémie Bigot, Marco Cuturi et Nicolas Papadakis. *Geodesic PCA versus Log-PCA of Histograms in the Wasserstein Space.* SIAM Journal on Scientific Computing, vol. 40, no. 2, pages B429–B456, 2018. (Cited on pages 2 and 29.)

[Chambon 2018] Stanislas Chambon, Mathieu N Galtier et Alexandre Gramfort. *Domain adaptation with optimal transport improves EEG sleep stage classifiers.* In Pattern Recognition in Neuroimaging, 2018. (Cited on page 25.)

[Chapelle 2002] Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet et Sayan Mukherjee. *Choosing multiple parameters for support vector machines.* Machine learning, vol. 46, no. 1-3, pages 131–159, 2002. (Cited on page 42.)

[Chizat 2018] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer et François-Xavier Vialard. *Unbalanced optimal transport: Dynamic and Kantorovich formulations.* Journal of Functional Analysis, vol. 274, no. 11, pages 3090–3123, 2018. (Cited on pages 14 and 29.)

[Chopra 2005] S. Chopra, R. Hadsell et Y. LeCun. *Learning a similarity metric discriminatively, with application to face verification.* In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 539–546. IEEE, 2005. (Cited on page 33.)

[Claici 2018] Sebastian Claici, Edward Chien et Justin Solomon. *Stochastic wasserstein barycenters.* arXiv preprint arXiv:1802.05757, 2018. (Cited on page 18.)

[Colson 2007] Benoît Colson, Patrice Marcotte et Gilles Savard. *An overview of bilevel optimization.* Annals of operations research, vol. 153, no. 1, pages 235–256, 2007. (Cited on page 42.)

[Courty 2014] N. Courty, R. Flamary et D. Tuia. *Domain adaptation with regularized optimal transport.* In ECML PKDD, Septembre 2014. (Cited on pages vii, 4, 16 and 24.)

[Courty 2016a] N. Courty, R. Flamary, D. Tuia et A. Rakotomamonjy. *Optimal transport for domain adaptation.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2016. (Cited on pages vii, 4, 16, 17, 23, 24, 25, 26 and 47.)

[Courty 2016b] Nicolas Courty, Rémi Flamary, Devis Tuia et Thomas Corpetti. *Optimal transport for data fusion in remote sensing.* In Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International, pages 3571–3574. IEEE, 2016. (Cited on page 25.)

[Courty 2017] Nicolas Courty, Remi Flamary, Amaury Habrard et Alain Rakotomamonjy. *Joint Distribution Optimal Transportation for Domain Adaptation.* In Neural Information Processing Systems (NIPS), 2017. (Cited on pages vii, 44, 45 and 48.)

[Courty 2018] Nicolas Courty, Rémi Flamary et Mélanie Ducoffe. *Learning Wasserstein Embeddings.* In International Conference on Learning Representation (ICMR), 2018. (Cited on pages vii, 4, 32 and 33.)

[Csurka 2017] Gabriela Csurka. *Domain adaptation for visual applications: A comprehensive survey.* arXiv preprint arXiv:1702.05374, 2017. (Cited on page 22.)

[Cuturi 2013] Marco Cuturi. *Sinkhorn Distances: Lightspeed Computation of Optimal Transportation.* In Neural Information Processing Systems (NIPS), pages 2292–2300. 2013. (Cited on pages iii, 2, 13 and 45.)

[Cuturi 2014a] M. Cuturi et A. Doucet. *Fast Computation of Wasserstein Barycenters.* In International Conference on Machine Learning (ICML), jun 2014. (Cited on pages 12, 14 and 20.)

[Cuturi 2014b] Marco Cuturi et David Avis. *Ground metric learning.* The Journal of Machine Learning Research, vol. 15, no. 1, pages 533–564, 2014. (Cited on pages 28, 29 and 59.)

[Cuturi 2016] Marco Cuturi et Gabriel Peyré. *A smoothed dual approach for variational Wasserstein problems.* SIAM Journal on Imaging Sciences, vol. 9, no. 1, pages 320–343, 2016. (Cited on page 14.)

[Damodaran 2018a] Bharath Damodaran, Rémi Flamary, Viven Seguy et Nicolas Courty. *An Entropic Optimal Transport Loss for Learning Deep Neural Networks under Label Noise in Remote Sensing Images.* 2018. (Cited on pages vii, viii and 49.)

[Damodaran 2018b] Bharath Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia et Nicolas Courty. *DeepJDOT: Deep Joint distribution optimal transport for unsupervised domain adaptation.* In European Conference in Computer Vision (ECCV), 2018. (Cited on pages vii, 48, 49, 58 and 59.)

[Damodaran 2019] Bharath Damodaran, Kilian Fatras, Sylvain Lobry, Rémi Flamary, Devis Tuia et Nicolas Courty. *Pushing the right boundaries matters! Wasserstein Adversarial Training for Label Noise.* 2019. (Cited on pages vii, viii and 58.)

[Deshpande 2018] Ishan Deshpande, Ziyu Zhang et Alexander G Schwing. *Generative modeling using the sliced wasserstein distance.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3483–3491, 2018. (Cited on pages 10, 39, 58 and 59.)

[Dessein 2016] Arnaud Dessein, Nicolas Papadakis et Jean-Luc Rouas. *Regularized Optimal Transport and the Rot Mover's Distance.* arXiv preprint arXiv:1610.06447, 2016. (Cited on pages 15 and 16.)

[Donahue 2014] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng et T. Darrell. *DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.* In International Conference on Machine Learning (ICML), pages 647–655, 2014. (Cited on page 25.)

[Dziugaite 2015] Gintare Karolina Dziugaite, Daniel M Roy et Zoubin Ghahramani. *Training generative neural networks via maximum mean discrepancy optimization.* arXiv preprint arXiv:1505.03906, 2015. (Cited on page 37.)

[Emiya 2010] V. Emiya, R. Badeau et B. David. *Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle.* IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 6, pages 1643–1654, 2010. (Cited on page 32.)

[Esfahani 2018] Peyman Mohajerin Esfahani et Daniel Kuhn. *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations.* Mathematical Programming, vol. 171, no. 1-2, pages 115–166, 2018. (Cited on page 39.)

[Essid 2018] Montacer Essid et Justin Solomon. *Quadratically Regularized Optimal Transport on Graphs.* SIAM Journal on Scientific Computing, vol. 40, no. 4, pages A1961–A1986, 2018. (Cited on page 16.)

[Fern 2003] Xiaoli Zhang Fern et Carla E Brodley. *Random projection for high dimensional data clustering: A cluster ensemble approach.* In ICML, volume 3, pages 186–193, 2003. (Cited on page 40.)

[Ferradans 2014] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré et Jean-François Aujol. *Regularized Discrete Optimal Transport.* SIAM Journal on Imaging Sciences, vol. 7, no. 3, 2014. (Cited on pages 16, 19, 20, 22, 24 and 25.)

[Ferrari 2014] A. Ferrari, D. Mary, R. Flamary et C. Richard. *Distributed image reconstruction for very large arrays in radio astronomy.* In IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM), 2014. (Cited on pages vii and viii.)

[Févotte 2009] Cédric Févotte, Nancy Bertin et Jean-Louis Durrieu. *Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis.* Neural computation, vol. 21, no. 3, pages 793–830, 2009. (Cited on pages 28, 30 and 31.)

[Flamary 2011] R. Flamary, F. Yger et A. Rakotomamonjy. *Selecting from an infinite set of features in SVM.* In European Symposium on Artificial Neural Networks, 2011. (Cited on pages vii and viii.)

[Flamary 2012a] R. Flamary et A. Rakotomamonjy. *Decoding finger movements from ECoG signals using switching linear models.* Frontiers in Neuroscience, vol. 6, no. 29, 2012. (Cited on page viii.)

[Flamary 2012b] R. Flamary, D. Tuia, B. Labbé, G. Camps-Valls et A. Rakotomamonjy. *Large Margin Filtering.* IEEE Transactions Signal Processing, vol. 60, no. 2, pages 648–659, 2012. (Cited on page viii.)

[Flamary 2014a] R. Flamary, N. Jrad, R. Phlypo, M. Congedo et A. Rakotomamonjy. *Mixed-Norm Regularization for Brain Decoding.* Computational and Mathematical Methods in Medicine, vol. 2014, no. 1, pages 1–13, April 2014. (Cited on page viii.)

[Flamary 2014b] R. Flamary, A. Rakotomamonjy et G. Gasso. *Learning Constrained Task Similarities in Graph-Regularized Multi-Task Learning.* In Argyriou A. Suykens J. A.K. Signoretto M., editeur, Regularization, Optimization, Kernels, and Support Vector Machines. 2014. (Cited on page vii.)

[Flamary 2014c] Remi Flamary et Claude Aime. *Optimization of starshades: focal plane versus pupil plane.* Astronomy and Astrophysics, vol. 569, no. A28, page 10, Sep 2014. (Cited on page viii.)

[Flamary 2014d] Rémi Flamary, Nicolas Courty, Devis Tuia et Alain Rakotomamonjy. *Optimal transport with Laplacian regularization: Applications to domain adaptation and shape matching.* In Neural Information Processing Systems (NIPS) Workshop on Optimal Transport and Machine Learning OTML, 2014. (Cited on pages 16 and 24.)

[Flamary 2015] R. Flamary, A. Rakotomamonjy et G. Gasso. *Importance Sampling Strategy for Non-Convex Randomized Block-Coordinate Descent.* In IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015. (Cited on page vii.)

[Flamary 2016] Remi Flamary, Cedric Fevotte, N. Courty et Valentin Emyia. *Optimal spectral transportation with application to music transcription.* In Neural Information Processing Systems (NIPS), 2016. (Cited on pages vii, 4, 30, 31, 32 and 59.)

[Flamary 2017a] Remi Flamary. *Astronomical image reconstruction with convolutional neural networks.* In European Conference on Signal Processing (EUSIPCO), 2017. (Cited on page viii.)

[Flamary 2017b] Rémi Flamary et Nicolas Courty. *POT Python Optimal Transport library.* 2017. (Cited on pages 35 and 52.)

[Flamary 2018] Rémi Flamary, Marco Cuturi, Nicolas Courty et Alain Rakotomamonjy. *Wasserstein discriminant analysis.* Machine Learning, 2018. (Cited on pages vii, 15, 40, 41, 43 and 59.)

[Flamary 2019] Rémi Flamary, Karim Lounici et André Ferrari. *Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation.* 2019. (Cited on pages vii, 4, 9, 21, 24 and 57.)

[Fletcher 2004] P. T. Fletcher, C. Lu, S. M. Pizer et S. Joshi. *Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape.* IEEE Trans. Medical Imaging, vol. 23, no. 8, pages 995–1005, 2004. (Cited on page 28.)

[Fortet 1953] Robert Fortet et Edith Mourier. *Convergence de la répartition empirique vers la répartition théorique.* In Annales scientifiques de l'École Normale Supérieure, volume 70, pages 267–285. Elsevier, 1953. (Cited on page 37.)

[Fournier 2015] Nicolas Fournier et Arnaud Guillin. *On the rate of convergence in Wasserstein distance of the empirical measure.* Probability Theory and Related Fields, vol. 162, no. 3-4, pages 707–738, 2015. (Cited on page 13.)

[Friedman 2001] Jerome Friedman, Trevor Hastie et Robert Tibshirani. The elements of statistical learning. Springer series in statistics Springer, Berlin, 2001. (Cited on page 41.)

[Frisch 2002] Uriel Frisch, Sabino Matarrese, Roya Mohayaee et Andrei Sobolevski. *A reconstruction of the initial conditions of the universe by optimal mass transportation.* Nature, vol. 417, no. 6886, page 260, 2002. (Cited on page 2.)

[Frogner 2015] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya et Tomaso A Poggio. *Learning with a Wasserstein loss.* In Advances in Neural Information Processing Systems, pages 2053–2061, 2015. (Cited on pages 14 and 29.)

[Ganin 2015] Y. Ganin et V. Lempitsky. *Unsupervised Domain Adaptation by Backpropagation.* In ICML, pages 1180–1189, 2015. (Cited on page 47.)

[Ganin 2016] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand et V. Lempitsky. *Domain-adversarial training of neural networks.* Journal of Machine Learning Research, vol. 17, no. 59, pages 1–35, 2016. (Cited on pages 22, 40 and 49.)

[Gao 2016] Rui Gao et Anton J Kleywegt. *Distributionally robust stochastic optimization with Wasserstein distance.* arXiv preprint arXiv:1604.02199, 2016. (Cited on page 39.)

[Gautheron 2017] Léo Gautheron, Carole Lartizien et Ievgen Redko. *Domain adaptation using Optimal Transport: application to prostate cancer mapping.* 2017. (Cited on page 25.)

[Gayraud 2017] Nathalie TH Gayraud, Alain Rakotomamonjy et Maureen Clerc. *Optimal Transport Applied to Transfer Learning For P300 Detection.* In 7th Graz Brain-Computer Interface Conference 2017, 2017. (Cited on page 25.)

[Genevay 2016] Aude Genevay, Marco Cuturi, Gabriel Peyré et Francis Bach. *Stochastic optimization for large-scale optimal transport.* In Advances in Neural Information Processing Systems, pages 3440–3448, 2016. (Cited on pages 2, 14, 17, 18, 45 and 59.)

[Genevay 2017a] Aude Genevay, Gabriel Peyré et Marco Cuturi. *GAN and VAE from an optimal transport point of view.* arXiv preprint arXiv:1706.01807, 2017. (Cited on pages 15, 39 and 58.)

[Genevay 2017b] Aude Genevay, Gabriel Peyré et Marco Cuturi. *Sinkhorn-AutoDiff: Tractable Wasserstein Learning of Generative Models.* arXiv preprint arXiv:1706.00292, 2017. (Cited on pages 15, 39, 42, 48, 49 and 59.)

[Genevay 2018] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi et Gabriel Peyré. *Sample Complexity of Sinkhorn divergences.* arXiv preprint arXiv:1810.02733, 2018. (Cited on pages 2, 15 and 59.)

[Ghahramani 1996] Zoubin Ghahramani, Geoffrey E Hinton*et al.* *The EM algorithm for mixtures of factor analyzers.* Rapport technique, Technical Report CRG-TR-96-1, University of Toronto, 1996. (Cited on page 38.)

[Givens 1984] Clark R Givens, Rae Michael Shortt*et al.* *A class of Wasserstein metrics for probability distributions.* The Michigan Mathematical Journal, vol. 31, no. 2, pages 231–240, 1984. (Cited on page 9.)

[Gong 2012] B. Gong, Y. Shi, F. Sha et K. Grauman. *Geodesic flow kernel for unsupervised domain adaptation.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2066–2073, 2012. (Cited on page 22.)

[Gong 2016] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour et B. Schölkopf. *Domain Adaptation with Conditional Transferable Components.* In ICML, volume 48, pages 2839–2848, 2016. (Cited on page 47.)

[Goodfellow 2014a] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio. *Generative adversarial nets.* In Advances in neural information processing systems, pages 2672–2680, 2014. (Cited on page 38.)

[Goodfellow 2014b] Ian J Goodfellow, Jonathon Shlens et Christian Szegedy. *Explaining and harnessing adversarial examples.* arXiv preprint arXiv:1412.6572, 2014. (Cited on page 58.)

[Gramfort 2015] Alexandre Gramfort, Gabriel Peyré et Marco Cuturi. *Fast optimal transport averaging of neuroimaging data.* In International Conference on Information Processing in Medical Imaging, pages 261–272. Springer, 2015. (Cited on page 15.)

[Gretton 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf et Alexander Smola. *A kernel two-sample test.* Journal of Machine Learning Research, vol. 13, no. Mar, pages 723–773, 2012. (Cited on pages 3, 13 and 37.)

[Grippo 2000] Luigi Grippo et Marco Sciandrone. *On the convergence of the block nonlinear Gauss–Seidel method under convex constraints.* Operations research letters, vol. 26, no. 3, pages 127–136, 2000. (Cited on page 46.)

[Gulrajani 2017] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin et A. Courville. *Improved training of wasserstein gans.* NIPS, 2017. (Cited on pages 39 and 40.)

[Harrane 2016] I. Harrane, R. Flamary et C. Richard. *Toward privacy-preserving diffusion strategies for adaptation and learning over networks.* In European Conference on Signal Processing (EUSIPCO), 2016. (Cited on page vii.)

[Harrane 2018a] Ibrahim Harrane, R. Flamary et C. Richard. *On reducing the communication cost of the diffusion LMS algorithm.* IEEE Transactions on Signal and Information Processing over Networks (SIPN), 2018. (Cited on page vii.)

[Harrane 2018b] Ibrahim Harrane, R. Flamary, C. Richard et R. Couillet. *Random matrix theory for diffusion LMS analysis.* In Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 2018. (Cited on page vii.)

[Hartley 2017] Philippa Hartley, Remi Flamary, Neal Jackson, A. S. Tagore et R. B. Metcalf. *Support Vector Machine classification of strong gravitational lenses.* Monthly Notices of the Royal Astronomical Society (MNRAS), 2017. (Cited on page viii.)

[Hinton 2012] Geoffrey E Hinton. *A practical guide to training restricted Boltzmann machines.* In Neural networks: Tricks of the trade, pages 599–619. Springer, 2012. (Cited on page 38.)

[Huang 2016] Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha et Kilian Q Weinberger. *Supervised Word Mover's Distance.* In Advances in Neural Information Processing Systems, pages 4862–4870, 2016. (Cited on page 29.)

[Hull 1994] J. J. Hull. *A database for handwritten text recognition research.* IEEE TPAMI, vol. 16, no. 5, pages 550–554, May 1994. (Cited on page 49.)

[Hütter 2019] Jan-Christian Hütter et Philippe Rigollet. *Minimax rates of estimation for smooth optimal transport maps.* arXiv preprint arXiv:1905.05828, 2019. (Cited on pages 20, 23, 57 and 58.)

[Jrad 2011] N. Jrad, M. Congedo, R. Phlypo, S. Rousseau, R. Flamary, F. Yger et A. Rakotomamonjy. *sw-SVM: sensor weighting support vector machines for EEG-based brain–computer interfaces.* Journal of Neural Engineering, vol. 8, no. 5, page 056004, 2011. (Cited on page viii.)

[Kantorovich 1942] L. Kantorovich. *On the translocation of masses.* C.R. (Doklady) Acad. Sci. URSS (N.S.), vol. 37, pages 199–201, 1942. (Cited on page 8.)

[Koch 2015] G. Koch, R. Zemel et R Salakhutdinov. *Siamese neural networks for one-shot image recognition.* In ICML Deep Learning Workshop, volume 2, 2015. (Cited on page 33.)

[Koep 2016] Niklas Koep et Sebastian Weichwald. *Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation.* Journal of Machine Learning Research, vol. 17, pages 1–5, 2016. (Cited on page 42.)

[Kolouri 2016] Soheil Kolouri, Serim Park, Matthew Thorpe, Dejan Slep**v**cev et Gustavo K Rohde. *Transport-based analysis, modeling, and learning from signal and data distributions.* arXiv preprint arXiv:1609.04767, 2016. (Cited on page 2.)

[Kolouri 2018] Soheil Kolouri, Phillip E Pope, Charles E Martin et Gustavo K Rohde. *Sliced-Wasserstein autoencoder: an embarrassingly simple generative model.* arXiv preprint arXiv:1804.01947, 2018. (Cited on pages 10, 58 and 59.)

[Kusner 2015] Matt Kusner, Yu Sun, Nicholas Kolkin et Kilian Weinberger. *From word embeddings to document distances.* In International Conference on Machine Learning, pages 957–966, 2015. (Cited on page 29.)

[Lacoste-Julien 2016] Simon Lacoste-Julien. *Convergence rate of Frank-Wolfe for non-convex objectives.* arXiv preprint arXiv:1607.00345, 2016. (Cited on page 16.)

[Laporte 2014a] L. Laporte, R. Flamary, S. Canu, S. Déjean et J. Mothe. *Nonconvex Regularizations for Feature Selection in Ranking With Sparse SVM.* Neural Networks and Learning Systems, IEEE Transactions on, vol. 25, no. 6, pages 1118–1130, 2014. (Cited on page 16.)

[Laporte 2014b] L. Laporte, R. Flamary, S. Canu, S. Déjean et J. Mothe. *Nonconvex Regularizations for Feature Selection in Ranking With Sparse SVM.* Neural Networks and Learning Systems, IEEE Transactions on, vol. 25, no. 6, pages 1118–1130, June 2014. (Cited on pages vii and viii.)

[Lecun 1998] Y. Lecun, L. Bottou, Y. Bengio et P. Haffner. *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, vol. 86, no. 11, pages 2278–2324, Nov 1998. (Cited on page 49.)

[Lee 2001] Daniel D Lee et H Sebastian Seung. *Algorithms for non-negative matrix factorization.* In Advances in neural information processing systems, pages 556–562, 2001. (Cited on pages 28 and 31.)

[Lee 2017] Jaeho Lee et Maxim Raginsky. *Minimax Statistical Learning and Domain Adaptation with Wasserstein Distances.* arXiv preprint arXiv:1705.07815, 2017. (Cited on page 40.)

[Lehaire 2014] J. Lehaire, R. Flamary, O. Rouvière et C. Lartizien. *Computer-aided diagnostic for prostate cancer detection and characterization combining learned dictionaries and supervised classification.* In IEEE International Conference on Image Processing (ICIP), 2014. (Cited on page viii.)

[Liero 2018] Matthias Liero, Alexander Mielke et Giuseppe Savaré. *Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures.* Inventiones mathematicae, vol. 211, no. 3, pages 969–1117, 2018. (Cited on page 14.)

[Lin 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár et C Lawrence Zitnick. *Microsoft coco: Common objects in context.* In European conference on computer vision, pages 740–755. Springer, 2014. (Cited on page 49.)

[Liutkus 2018] Antoine Liutkus, Umut Şimşekli, Szymon Majewski, Alain Durmus et Fabian-Robert Stöter. *Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions.* arXiv preprint arXiv:1806.08141, 2018. (Cited on pages 10 and 59.)

[Long 2014] M. Long, J. Wang, G. Ding, J. Sun et P. Yu. *Transfer Joint Matching for Unsupervised Domain Adaptation.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1410–1417, 2014. (Cited on page 22.)

[Luise 2018] Giulia Luise, Alessandro Rudi, Massimiliano Pontil et Carlo Ciliberto. *Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance.* arXiv preprint arXiv:1805.11897, 2018. (Cited on pages 14 and 15.)

[Madry 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras et Adrian Vladu. *Towards deep learning models resistant to adversarial attacks.* arXiv preprint arXiv:1706.06083, 2017. (Cited on page 39.)

[Malagò 2018] Luigi Malagò, Luigi Montrucchio et Giovanni Pistone. *Wasserstein Riemannian geometry of Gaussian densities.* Information Geometry, vol. 1, no. 2, pages 137–179, 2018. (Cited on page 9.)

[Mansour 2009] Y. Mansour, M. Mohri et A. Rostamizadeh. *Domain Adaptation: Learning Bounds and Algorithms.* In Conference on Learning Theory (COLT), pages 19–30, 2009. (Cited on page 45.)

[McCann 1997] R. J. McCann. *A convexity principle for interacting gases.* advances in mathematics, vol. 128, no. 1, pages 153–179, 1997. (Cited on pages 9 and 10.)

[Mémoli 2011] F. Mémoli. *Gromov-Wasserstein Distances and the Metric Approach to Object Matching.* Foundations of Computational Mathematics, pages 1–71, 2011. (Cited on pages 3, 51 and 52.)

[Metcalf 2019] R Benton Metcalf, M Meneghetti, Camille Avestruz, Fabio Bellagamba, Clécio R Bom, Emmanuel Bertin, Rémi Cabanac, Etienne Decencière, Rémi Flamary, Raphael Gavazzi*et al. The Strong Gravitational Lens Finding Challenge.* Astronomy and Astrophysics, vol. 625, page A119, 2019. (Cited on page viii.)

[Mikolov 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado et Jeff Dean. *Distributed representations of words and phrases and their compositionality.* In Advances in neural information processing systems, pages 3111–3119, 2013. (Cited on page 29.)

[Miyato 2016] Takeru Miyato, Andrew M Dai et Ian Goodfellow. *Adversarial training methods for semi-supervised text classification.* arXiv preprint arXiv:1605.07725, 2016. (Cited on page 58.)

[Miyato 2018] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama et Shin Ishii. *Virtual adversarial training: a regularization method for supervised and semi-supervised learning.* IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 8, pages 1979–1993, 2018. (Cited on page 58.)

[Monge 1781] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. De l'Imprimerie Royale, 1781. (Cited on pages 1 and 7.)

[Montavon 2016] Grégoire Montavon, Klaus-Robert Müller et Marco Cuturi. *Wasserstein training of restricted Boltzmann machines.* In Advances in Neural Information Processing Systems, pages 3718–3726, 2016. (Cited on page 29.)

[Mourya 2017a] Rahul Mourya, Andre Ferrari, Remi Flamary, Pascal Bianchi et Cedric Richard. *Distributed Approach for Deblurring Large Images with Shift-Variant Blur.* In European Conference on Signal Processing (EUSIPCO), 2017. (Cited on page vii.)

[Mourya 2017b] Rahul Mourya, Andre Ferrari, Remi Flamary, Pascal Bianchi et Cedric Richard. *Distributed Deblurring of Large Images of Wide Field-Of-View.* 2017. (Cited on page viii.)

[Nakhostin 2016] Sina Nakhostin, Nicolas Courty, Remi Flamary, D. Tuia et Thomas Corpetti. *Supervised planetary unmixing with optimal transport.* In Whorkshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS), 2016. (Cited on page 31.)

[Netzer 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu et Andrew Y. Ng. *Reading Digits in Natural Images with Unsupervised Feature Learning.* In NIPS worksophs, 2011. (Cited on page 49.)

[Niaf 2014] E. Niaf, R. Flamary, O. Rouvière, C. Lartizien et S. Canu. *Kernel-Based Learning From Both Qualitative and Quantitative Labels: Application to Prostate Cancer Diagnosis Based on Multiparametric MR Imaging.* Image Processing, IEEE Transactions on, vol. 23, no. 3, pages 979–991, March 2014. (Cited on page viii.)

[Niepert 2016] Mathias Niepert, Mohamed Ahmed et Konstantin Kutzkov. *Learning convolutional neural networks for graphs.* In International conference on machine learning, pages 2014–2023, 2016. (Cited on page 54.)

[Nowozin 2016] Sebastian Nowozin, Botond Cseke et Ryota Tomioka. *f-gan: Training generative neural samplers using variational divergence minimization.* In Advances in Neural Information Processing Systems, pages 271–279, 2016. (Cited on page 38.)

[Patrini 2018] Giorgio Patrini, Marcello Carioni, Patrick Forre, Samarth Bhargav, Max Welling, Rianne van den Berg, Tim Genewein et Frank Nielsen. *Sinkhorn AutoEncoders.* arXiv preprint arXiv:1810.01118, 2018. (Cited on page 14.)

[Paty 2019a] François-Pierre Paty et Marco Cuturi. *Subspace robust wasserstein distances*. arXiv preprint arXiv:1901.08949, 2019. (Cited on page 59.)

[Paty 2019b] François-Pierre Paty, Alexandre d'Aspremont et Marco Cuturi. *Regularity as Regularization: Smooth and Strongly Convex Brenier Potentials in Optimal Transport*. arXiv preprint arXiv:1905.10812, 2019. (Cited on pages 57 and 58.)

[Peleg 1989] Shmuel Peleg, Michael Werman et Hillel Rom. *A unified approach to the change of resolution: Space and gray-level*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pages 739–742, 1989. (Cited on page 2.)

[Peng 2017] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang et Kate Saenko. *VisDA: The Visual Domain Adaptation Challenge*, 2017. (Cited on page 49.)

[Pérez 2003] P. Pérez, M. Gangnet et A. Blake. *Poisson Image Editing*. ACM Trans. on Graphics, vol. 22, no. 3, 2003. (Cited on pages iii and 25.)

[Perrot 2016] M. Perrot, N. Courty, R. Flamary et A. Habrard. *Mapping estimation for discrete optimal transport*. In Neural Information Processing Systems (NIPS), 2016. (Cited on pages vii, 4, 21, 24 and 25.)

[Petersen 2008] Kaare Brandt Petersen, Michael Syskind Pedersen*et al.* *The matrix cookbook*. Technical University of Denmark, vol. 7, page 15, 2008. (Cited on page 42.)

[Peyré 2016] G. Peyré, M. Cuturi et J. Solomon. *Gromov-Wasserstein Averaging of Kernel and Distance Matrices*. In ICML 2016, Proc. 33rd International Conference on Machine Learning, New-York, United States, Juin 2016. (Cited on pages 52, 53 and 54.)

[Peyré 2017] Gabriel Peyré et Marco Cuturi. *Computational Optimal Transport*. Rapport technique, 2017. (Cited on pages 7, 10, 12 and 15.)

[R. Gopalan 2014] R. Li R. Gopalan et R. Chellappa. *Unsupervised Adaptation Across Domain Shifts By Generating Intermediate Data Representations*. IEEE Transactions on Pattern Analysis and Machine Intelligence, page To be published, 2014. (Cited on page 22.)

[Rabin 2011] Julien Rabin, Gabriel Peyré, Julie Delon et Marc Bernot. *Wasserstein barycenter and its application to texture mixing*. In International Conference on Scale Space and Variational Methods in Computer Vision, volume 6667 of *Lecture Notes in Computer Science*, pages 435–446. Springer, 2011. (Cited on page 2.)

[Radford 2015] Alec Radford, Luke Metz et Soumith Chintala. *Unsupervised representation learning with deep convolutional generative adversarial networks*. arXiv preprint arXiv:1511.06434, 2015. (Cited on page 38.)

[Rakotomamonjy 2011] A. Rakotomamonjy, R. Flamary, G. Gasso et S. Canu. *lp-lq penalty for sparse linear and sparse multiple kernel multi-task learning*. IEEE Transactions on Neural Networks, vol. 22, no. 8, pages 1307–1320, 2011. 13. (Cited on page vii.)

[Rakotomamonjy 2013] A. Rakotomamonjy, R. Flamary et F. Yger. *Learning with infinitely many features*. Machine Learning, vol. 91, no. 1, pages 43–66, 2013. (Cited on page vii.)

[Rakotomamonjy 2015] Alain Rakotomamonjy, Rémi Flamary et Nicolas Courty. *Generalized conditional gradient: analysis of convergence and applications*. arXiv preprint arXiv:1510.06567, 2015. (Cited on page 17.)

[Rakotomamonjy 2016] A. Rakotomamonjy, R. Flamary et G. Gasso. *DC Proximal Newton for Non-Convex Optimization Problems*. Neural Networks and Learning Systems, IEEE Transactions on, vol. 27, no. 3, pages 636–647, 2016. (Cited on page vii.)

[Rakotomamonjy 2018] Alain Rakotomamonjy, Abraham Traore, Maxime Berar, Remi Flamary et Nicolas Courty. *Distance Measure Machines*. 2018. (Cited on pages 10 and 54.)

[Real 2017] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan et Vincent Vanhoucke. *YouTube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video*. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 7464–7473. IEEE, 2017. (Cited on page 49.)

[Redko 2016] I. Redko, A. Habrard et M. Sebban. *Theoretical Analysis of Domain Adaptation with Optimal Transport*. ArXiv e-prints, Octobre 2016. (Cited on page 25.)

[Redko 2019] Ievgen Redko, Nicolas Courty, Rémi Flamary et Devis Tuia. *Optimal Transport for Multi-source Domain Adaptation under Target Shift*. In International Conference on Artificial Intelligence and Statistics (AISTATS), 2019. (Cited on pages vii and 26.)

[Reich 2013] Sebastian Reich. *A nonparametric ensemble transform method for Bayesian inference*. SIAM Journal on Scientific Computing, vol. 35, no. 4, pages A2013–A2024, 2013. (Cited on page 19.)

[Rolet 2016] Antoine Rolet, Marco Cuturi et Gabriel Peyré. *Fast Dictionary Learning with a Smoothed Wasserstein Loss*. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, pages 630–638, 2016. (Cited on page 28.)

[Rougeot 2017] Raphael Rougeot, Remi Flamary, Damien Galano et Claude Aime. *Performance of hybrid externally occulted Lyot solar coronagraph, Application to ASPIICS*. Astronomy and Astrophysics, 2017. (Cited on page viii.)

[Rougeot 2018] Raphaël Rougeot, Claude Aime, Cristian Baccani, Silvano Fineschi, Rémi Flamary, Damien Galano, Camille Galy, Volker Kirschner, Federico Landini, Marco Romoli*et al.* *Stray-light analysis for the externally occulted Lyot solar coronagraph ASPIICS*. In Space Telescopes and Instrumentation 2018: Optical, Infrared, and Millimeter Wave, volume 10698, page 106982T. International Society for Optics and Photonics, 2018. (Cited on page viii.)

[Rougeot 2019] Raphael Rougeot, Remi Flamary, David Mary et Claude Aime. *Influence of surface roughness on diffraction in the externally occulted Lyot solar coronagraph*. Astronomy and Astrophysics, 2019. (Cited on page viii.)

[Rousselle 2015] Denis Rousselle et Stéphane Canu. *Optimal transport for semi-supervised domain adaptation*. In Proceedings, page 373. Presses universitaires de Louvain, 2015. (Cited on page 26.)

[Rubner 1998] Y. Rubner, C. Tomasi et L.J. Guibas. *A metric for distributions with applications to image databases*. In ICCV, pages 59–66, Jan 1998. (Cited on pages 2 and 9.)

[Rubner 2000] Yossi Rubner, Carlo Tomasi et Leonidas J Guibas. *The earth mover's distance as a metric for image retrieval*. International journal of computer vision, vol. 40, no. 2, pages 99–121, 2000. (Cited on pages iii and 2.)

[Saenko 2010] K. Saenko, B. Kulis, M. Fritz et T. Darrell. *Adapting Visual Category Models to New Domains*. In European Conference on Computer Vision (ECCV), LNCS, pages 213–226, Berlin, Heidelberg, 2010. Springer-Verlag. (Cited on page 47.)

[Salimans 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford et Xi Chen. *Improved techniques for training gans*. In Advances in Neural Information Processing Systems, pages 2234–2242, 2016. (Cited on page 38.)

[Sandler 2011] Roman Sandler et Michael Lindenbaum. *Nonnegative matrix factorization with earth mover's distance metric for image analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, pages 1590–1602, 2011. (Cited on pages 2 and 28.)

[Santambrogio 2014] Filippo Santambrogio. *Introduction to optimal transport theory*. Notes, 2014. (Cited on pages 7 and 28.)

[Schmidt 2008] M Schmidt. *Minconf-projection methods for optimization with simple constraints in matlab*, 2008. (Cited on page 42.)

[Schmitz 2017] Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Maurice Ngolè Mboula, David Coeurjolly, Marco Cuturi, Gabriel Peyré et Jean-Luc Starck. *Wasserstein Dictionary Learning: Optimal Transport-based unsupervised non-linear dictionary learning*. arXiv preprint arXiv:1708.01955, 2017. (Cited on page 28.)

[Seguy 2015] Vivien Seguy et Marco Cuturi. *Principal geodesic analysis for probability measures under the optimal transport metric*. In Advances in Neural Information Processing Systems, pages 3312–3320, 2015. (Cited on pages 28, 29 and 35.)

[Seguy 2018] Vivien. Seguy, Bharath Bhushan Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet et Mathieu Blondel. *Large-Scale Optimal Transport and Mapping Estimation*. In International Conference on Leraning Representation (ICLR), 2018. (Cited on pages vii, 4, 18, 20, 21, 23, 24, 25 and 59.)

[Shafieezadeh-Abadeh 2015] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani et Daniel Kuhn. *Distributionally robust logistic regression*. In Advances in Neural Information Processing Systems, pages 1576–1584, 2015. (Cited on page 39.)

[Shafieezadeh-Abadeh 2017] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn et Peyman Mohajerin Esfahani. *Regularization via mass transportation*. arXiv preprint arXiv:1710.10016, 2017. (Cited on page 39.)

[Shen 2018] Jian Shen, Yanru Qu, Weinan Zhang et Yong Yu. *Wasserstein Distance Guided Representation Learning for Domain Adaptation*. In AAAI Conference on Artificial Intelligence, 2018. (Cited on pages 22 and 40.)

[Shuman 2012] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega et Pierre Vandergheynst. *The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains*. arXiv preprint arXiv:1211.0053, 2012. (Cited on page 57.)

[Si 2010] S. Si, D. Tao et B. Geng. *Bregman Divergence-Based Regularization for Transfer Subspace Learning*. IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 7, pages 929–942, July 2010. (Cited on page 22.)

[Sinha 2018] Aman Sinha, Hongseok Namkoong et John Duchi. *Certifying some distributional robustness with principled adversarial training*. 2018. (Cited on page 39.)

[Sinkhorn 1967] Richard Sinkhorn et Paul Knopp. *Concerning nonnegative matrices and doubly stochastic matrices*. Pacific Journal of Mathematics, vol. 21, no. 2, pages 343–348, 1967. (Cited on page 13.)

[Sirovich 1987] Lawrence Sirovich et Michael Kirby. *Low-dimensional procedure for the characterization of human faces*. Josa a, vol. 4, no. 3, pages 519–524, 1987. (Cited on page 38.)

[Smaragdis 2006] Paris Smaragdis, Bhiksha Raj et Madhusudana Shashanka. *A probabilistic latent variable model for acoustic modeling*. Advances in models for acoustic processing, NIPS, vol. 148, pages 8–1, 2006. (Cited on pages 30 and 32.)

[Solomon 2014a] Justin Solomon, Raif Rustamov, Leonidas Guibas et Adrian Butscher. *Wasserstein propagation for semi-supervised learning*. In International Conference on Machine Learning, pages 306–314, 2014. (Cited on page 26.)

[Solomon 2014b] Justin Solomon, Raif Rustamov, Guibas Leonidas et Adrian Butscher. *Wasserstein Propagation for Semi-Supervised Learning.* In International Conference on Machine Learning (ICML), pages 306–314, 2014. (Cited on page 2.)

[Solomon 2015] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du et Leonidas Guibas. *Convolutional wasserstein distances: Efficient optimal transportation on geometric domains.* ACM Transactions on Graphics (TOG), vol. 34, no. 4, page 66, 2015. (Cited on pages 2 and 15.)

[Staib 2017] Matthew Staib, Sebastian Claici, Justin M Solomon et Stefanie Jegelka. *Parallel streaming wasserstein barycenters.* In Advances in Neural Information Processing Systems, pages 2647–2658, 2017. (Cited on page 18.)

[Stavropoulou 2015] Faidra Stavropoulou et Johannes M uller. *Parametrization of random vectors in polynomial chaos expansions via optimal transportation.* SIAM Journal on Scientific Computing, vol. 37, no. 6, pages A2535–A2557, 2015. (Cited on page 20.)

[Sugiyama 2007] Masashi Sugiyama. *Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis.* The Journal of Machine Learning Research, vol. 8, pages 1027–1061, 2007. (Cited on pages 41 and 43.)

[Sugiyama 2008] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau et Motoaki Kawanabe. *Direct importance estimation for covariate shift adaptation.* Annals of the Institute of Statistical Mathematics, vol. 60, no. 4, pages 699–746, 2008. (Cited on page 22.)

[Sun 2016] Baochen Sun et Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation, pages 443–450. Springer International Publishing, Cham, 2016. (Cited on pages 22, 40 and 49.)

[Sutherland 2016] Dougal J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola et Arthur Gretton. *Generative models and model criticism via optimized maximum mean discrepancy.* arXiv preprint arXiv:1611.04488, 2016. (Cited on page 37.)

[Suzuki 2013] Taiji Suzuki et Masashi Sugiyama. *Sufficient dimension reduction via squared-loss mutual information estimation.* Neural computation, vol. 25, no. 3, pages 725–758, 2013. (Cited on page 43.)

[Takatsu 2011] Asuka Takatsu*et al.* *Wasserstein geometry of Gaussian measures.* Osaka Journal of Mathematics, vol. 48, no. 4, pages 1005–1026, 2011. (Cited on page 9.)

[Tolstikhin 2017] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly et Bernhard Schoelkopf. *Wasserstein auto-encoders.* arXiv preprint arXiv:1711.01558, 2017. (Cited on page 39.)

[Tuia 2010] D. Tuia, G. Camps-Valls, R. Flamary et A. Rakotomamonjy. *Learning spatial filters for multispectral image segmentation.* In IEEE Workshop in Machine Learning for Signal Processing (MLSP), 2010. (Cited on page viii.)

[Tuia 2014a] D. Tuia, N. Courty et R. Flamary. *A group-lasso active set strategy for multiclass hyperspectral image classification.* In Photogrammetric Computer Vision (PCV), 2014. (Cited on page viii.)

[Tuia 2014b] D. Tuia, M. Volpi, M. Dalla Mura, A. Rakotomamonjy et R. Flamary. *Automatic Feature Learning for Spatio-Spectral Image Classification With Sparse SVM.* Geoscience and Remote Sensing, IEEE Transactions on, vol. 52, no. 10, pages 6062–6074, Oct 2014. (Cited on page viii.)

[Tuia 2015a] D. Tuia, R. Flamary et M. Barlaud. *To be or not to be convex? A study on regularization in hyperspectral image classification.* In International Geoscience and Remote Sensing Symposium (IGARSS), 2015. (Cited on page viii.)

[Tuia 2015b] D. Tuia, R. Flamary et N. Courty. *Multiclass feature learning for hyperspectral image classification: sparse and hierarchical solutions.* ISPRS Journal of Photogrammetry and Remote Sensing, 2015. (Cited on page viii.)

[Tuia 2015c] D. Tuia, R. Flamary, A. Rakotomamonjy et N. Courty. *Multitemporal classification without new labels: a solution with optimal transport.* In 8th International Workshop on the Analysis of Multitemporal Remote Sensing Images, 2015. (Cited on page 24.)

[Tuia 2016] D. Tuia, R. Flamary et M. Barlaud. *Non-convex regularization in remote sensing.* Geoscience and Remote Sensing, IEEE Transactions on, 2016. (Cited on page viii.)

[Tzeng 2014] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko et Trevor Darrell. *Deep domain confusion: Maximizing for domain invariance.* arXiv preprint arXiv:1412.3474, 2014. (Cited on pages 22 and 40.)

[Urner 2011] R. Urner, S. Shalev-Shwartz et S. Ben-David. *Access to Unlabeled Data can Speed up Prediction Time.* In Proceedings of ICML, pages 641–648, 2011. (Cited on page 45.)

[Van der Maaten 2008] L. Van der Maaten et G. Hinton. *Visualizing data using t-SNE.* Journal of Machine Learning Research, vol. 9, no. 2579-2605, page 85, 2008. (Cited on page 43.)

[Van Der Maaten 2009] Laurens Van Der Maaten, Eric Postma et Jaap Van den Herik. *Dimensionality reduction: a comparative review.* Journal of Machine Learning Research, vol. 10, pages 66–71, 2009. (Cited on page 40.)

[Vayer 2018] Titouan Vayer, Laetita Chapel, Rémi Flamary, Romain Tavenard et Nicolas Courty. *Fused Gromov-Wasserstein distance for structured objects: theoretical foundations and mathematical properties.* 2018. (Cited on pages vii, 51 and 53.)

[Vayer 2019a] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard et Nicolas Courty. *Optimal Transport for structured data with application on graphs.* In International Conference on Machine Learning (ICML), 2019. (Cited on pages vii, 53, 54 and 55.)

[Vayer 2019b] Titouan Vayer, Rémi Flamary, Romain Tavenard, Laetitia Chapel et Nicolas Courty. *Sliced Gromov-Wasserstein.* 2019. (Cited on pages vii and 52.)

[Venkateswara 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty et Sethuraman Panchanathan. *Deep Hashing Network for Unsupervised Domain Adaptation.* In (IEEE) Conference on Computer Vision and Pattern Recognition (CVPR), 2017. (Cited on page 49.)

[Villani 2003] C. Villani. Topics in optimal transportation. Graduate Studies in Mathematics Series. American Mathematical Society, 2003. (Cited on page 7.)

[Villani 2009] C. Villani. Optimal transport: old and new, volume 338 of *Grundlehren der mathematischen Wissenschaften.* Springer, 2009. (Cited on page 7.)

[Wang 2018] Qi Wang, Ievgen Redko et Sylvain Takerkart. *Population Averaging of Neuroimaging Data Using L p Distance-based Optimal Transport.* In 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI), pages 1–4. IEEE, 2018. (Cited on page 15.)

[Weed 2017] Jonathan Weed et Francis Bach. *Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance.* arXiv preprint arXiv:1707.00087, 2017. (Cited on page 13.)

[Weinberger 2009] Kilian Q Weinberger et Lawrence K Saul. *Distance metric learning for large margin nearest neighbor classification.* The Journal of Machine Learning Research, vol. 10, pages 207–244, 2009. (Cited on pages 41 and 43.)

[Xing 2003] Eric P Xing, Andrew Y Ng, Michael I Jordan et Stuart Russell. *Distance metric learning with application to clustering with side-information.* Advances in neural information processing systems, vol. 15, pages 505–512, 2003. (Cited on page 41.)

[Yu 2013] W. Yu, G. Zeng, P. Luo, F. Zhuang, Q. He et Z. Shi. *Embedding with autoencoder regularization.* In ECML/PKDD, pages 208–223. Springer, 2013. (Cited on page 33.)

[Zen 2014] G. Zen, E. Ricci et N. Sebe. *Simultaneous Ground Metric Learning and Matrix Factorization with Earth Mover's Distance.* In Pattern Recognition (ICPR), 2014 22nd International Conference on, pages 3690–3695, Aug 2014. (Cited on page 28.)

[Zhang 2010] Lei Zhang, Weisheng Dong, David Zhang et Guangming Shi. *Two-stage image denoising by principal component analysis with local pixel grouping.* Pattern Recognition, vol. 43, no. 4, pages 1531–1549, 2010. (Cited on page 40.)

[Zhang 2013] K. Zhang, V. W. Zheng, Q. Wang, J. T. Kwok, Q. Yang et I. Marsic. *Covariate Shift in Hilbert Space: A Solution via Surrogate Kernels.* In ICML, 2013. (Cited on page 47.)

[Zhao 2016] Junbo Zhao, Michael Mathieu et Yann LeCun. *Energy-based generative adversarial network.* arXiv preprint arXiv:1609.03126, 2016. (Cited on page 38.)