Optimal transport for novelty and out-of-distribution detection

Internship

Expected starting: March/April 2024

Key-words. Machine learning, robust optimal transport, out-of-distribution samples, statistical testing.

Context. For a decision-making system trained on data to be reliable, it must possess the ability to adjust its decisions based on differences between the distribution $p_{train}(X_{train}, Y_{train})$ of training samples and that of test samples $p_{test}(X_{test}, Y_{test})$. In case of distribution shift, deep-based-approaches may be overconfident and tend to treat the given inputs as one of the previously seen situations leading to mislabelling. This underscores the challenges in detecting out-of-distribution (OOD) samples, where the test point x_0 is marginally sampled from $p_{test}(x_0) \neq p_{train}(x_0)$, or recognizing that point x_0 belongs to an unseen class (involving a new type of object in the scenes for instance) [1]. Additionally, given the multimodal nature of inputs and variations in sensor availability, samples may not be embedded into the same space, posing further challenges related to *incomparable spaces*. Our approach envisions employing optimal transport theory [2] to develop algorithms addressing out-of-distribution detection, aiming for a *robust* optimal transport framework.

Optimal transport (OT) [2] has become a potent tool for computing distances (a.k.a. Wasserstein or earth mover's distances) between data distributions, facilitated by new computational schemes that make transport computations tractable [3]. Its applications span computer vision, statistics, imaging, and it has been integrated into machine learning for efficient problem-solving in classification or transfer learning [4]. OT's advantage lies in its ability to compare high-dimensional empirical probability measures, considering the geometry of underlying metric spaces and accommodating discrete measures. It also offers tools, like the Gromov-Wasserstein distance, for comparing distributions not residing in the same ground space.

The classical optimal transport problem seeks a transportation map preserving total mass between two probability distributions, requiring their masses to be equal. This might be overly restrictive in applications such as color or shape matching, where distributions have arbitrary masses or only a fraction needs transporting. Similar challenges arise when datasets X_{train} and/or X_{test} contain outliers that should be excluded from the transportation plan. These scenarios are addressed by unbalanced [5] or partial OT formulations [6], allowing removal of mass from distributions. Various algorithms have been devised to solve the problem, with [7] solving the exact partial problem when given the total mass to be transported between two empirical distributions, and [8] devising algorithms for the unbalanced problem, offering a *regularization path* for unbalanced OT when formulated as a penalized regression problem.

Scientific objectives and expected achievements. The primary goal of the internship is to investigate the behavior of optimal transport (OT) in scenarios where distributions are tainted by outliers or out-of-distribution (OOD) samples and to formulate a robust OT framework. Existing studies, such as those by Mukherjee et al. [9] and Balaji et al. [10], have utilized OT in such contexts, employing a straightforward rule that identifies points significantly distant from the other distribution as outliers. While approaches

like the regularization path [8] or OT profiles [11] have been effective in selecting optimal regularization parameters, particularly using techniques like the elbow rule, they may fall short when dealing with points that are OOD but situated "between" the two distributions.

Conversely, Monge-Kantorovich (MK) quantiles and ranks, introduced by Chernozhukov et al. [12] with a comprehensive review available in [13], present an alternative. This method replaces the traditional "left-to-right" ordering of samples with a "center-outward" approach applicable in $\mathbb{R}^d$. MK quantiles have proven successful in devising statistical tests, as demonstrated in studies such as [14] and [15], particularly for testing independence.

The internship's specific objectives include: i) examining how the placement of outliers influences the OT solution, ii) developing a robust OT formulation with statistical guarantees, leveraging MK quantiles, and iii) implementing the solution in the POT toolbox [16], a tool developed by team members. The intern will benefit from the expertise gained through ongoing collaborations with academic partners specializing in this domain.

Furthermore, the internship will explore the integration of partial-OT-based loss in deep learning approaches as a means to evaluate the proposed methods. Ensuring scalability will be a crucial aspect of the method's development. Additionally, investigations into adapting the approach for incomparable spaces will be undertaken.

Research environnement/Location The research will take place either within the LITIS laboratory (`https://www.insa-rouen.fr/recherche/laboratoires/litis`) located at INSA Rouen, France or either in the new team from IRISA (MALT) which deals with machine learning and IA in structure environments. The internship will be jointly supervised by Gilles Gasso (LITIS) and Laetitia Chapel (IRISA).

Candidate profile Applicants are expected to be graduated in applied mathematics/statistics and/or machine learning and show an excellent academic profile. Beyond, good programming skills are expected.

Application procedure Send a resume to Gilles Gasso (gilles.gasso@insa-rouen.fr) and Laetitia Chapel (laetitia.chapel@irisa.fr). Potential candidates will be contacted for interview. Feel free to contact us for any question.

References

[1] A. Shafaei, M. Schmidt, and J. J. Little, "Does your model know the digit 6 is not a cat? a less biased evaluation of" outlier" detectors," 2018.

[2] C. Villani, *Optimal transport: old and new*. Springer Science, 2008.

[3] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*, 2013, pp. 2292–2300.

[4] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[5] J.-D. Benamou, "Numerical resolution of an "unbalanced" mass transport problem," *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, vol. 37, no. 5, pp. 851–868, 2003.

[6] A. Figalli, "The optimal partial transport problem," *Archive for rational mechanics and analysis*, vol. 195, 2010.

[7] L. Chapel, M. Z. Alaya, and G. Gasso, "Partial optimal transport with applications on positive-unlabeled learning," in *NeurIPS*, 2020.

[8] L. Chapel, R. Flamary, H. Wu, C. Févotte, and G. Gasso, "Unbalanced optimal transport through non-negative penalized linear regression," *arXiv preprint arXiv:2106.04145*, 2021.

[9] D. Mukherjee, A. Guha, J. M. Solomon, Y. Sun, and M. Yurochkin, "Outlier-robust optimal transport," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7850–7860.

[10] Y. Balaji, R. Chellappa, and S. Feizi, "Robust optimal transport with applications in generative modeling and domain adaptation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 934–12 944, 2020.

[11] A. Phatak, S. Raghvendra, C. Tripathy, and K. Zhang, "Computing all optimal partial transports," in *The Eleventh International Conference on Learning Representations*, 2022.

[12] V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry, "Monge–kantorovich depth, quantiles, ranks and signs," *The Annals of Statistics*, 2017.

[13] M. Hallin, "Measure transportation and statistical decision theory," *Annual Review of Statistics and Its Application*, vol. 9, pp. 401–424, 2022.

[14] P. Ghosal and B. Sen, "Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing," *The Annals of Statistics*, vol. 50, no. 2, pp. 1012–1037, 2022.

[15] H. Shi, M. Drton, M. Hallin, and F. Han, "Center-outward sign-and rank-based quadrant, spearman, and kendall tests for multivariate independence," *arXiv preprint arXiv:2111.15567*, 2021.

[16] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier *et al.*, "Pot: Python optimal transport," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 3571–3578, 2021.