7th ICML Workshop on Automated Machine Learning (2020)

Geometric Dataset Distances via Optimal Transport

David Alvarez-Melis ALVAREZ.MELIS@MICROSOFT.COM
Microsoft Research

Nicolò Fusi FUSI@MICROSOFT.COM
Microsoft Research

Abstract

The notion of task similarity is at the core of various machine learning paradigms, such as domain adaptation and meta-learning. Current methods to quantify it are often heuristic, make strong assumptions on the label sets across the tasks, and many are architecture-dependent, relying on task-specific optimal parameters (*e. g.,* require training a model on each dataset). In this work we propose an alternative notion of distance between datasets that (i) is model-agnostic, (ii) does not involve training, (iii) can compare datasets even if their label sets are completely disjoint and (iv) has solid theoretical footing. This distance relies on optimal transport, which provides it with rich geometry awareness, interpretable correspondences and well-understood properties. Our results show that this novel distance provides meaningful comparison of datasets, and correlates well with transfer learning hardness across various experimental settings and datasets.

1. Introduction

A key hallmark of machine learning practice is that labeled data from the application of interest is usually scarce. For this reason, there is vast interest in methods that can combine, adapt and transfer knowledge across datasets and domains. Entire research areas are devoted to these goals, such as domain adaptation, transfer-learning and meta-learning. A fundamental concept underlying all these paradigms is the notion of *distance* (or more generally, *similarity*) between datasets. For instance, transferring knowledge across similar domains should intuitively be easier than across distant ones. Likewise, given a choice of various datasets to pretrain a model on, it would seem natural to choose the one that is closest to the task of interest.

Despite its evident usefulness and apparent simpleness, the notion of distance between datasets is elusive, and quantifying it efficiently and in a principled manner remains largely an open problem. Doing so involves various challenges that commonly arise precisely in the settings for which this notion would be most useful, such as the ones mentioned above. For example, in supervised learning settings the datasets consist of features and labels, and while defining a distance between the former is often —though not always— trivial, doing so for the labels is far from it, particularly if the label-sets across the two tasks are not identical (as is often the case for off-the-shelf pretrained models).

Current approaches to transfer learning that seek to quantify dataset similarity circumvent these challenges in various ingenious, albeit often heuristic, ways. A common approach is to compare the dataset via proxies, such as the learning curves of a pre-specified model (Leite and Brazdil, 2005) or its optimal parameters (Achille et al., 2019; Khodak et al., 2019) on a given task, or by making strong assumptions on the similarity or co-occurrence of labels across the two datasets (Tran et al., 2019). Most of these approaches lack guarantees, are highly dependent on the probe model used, and require

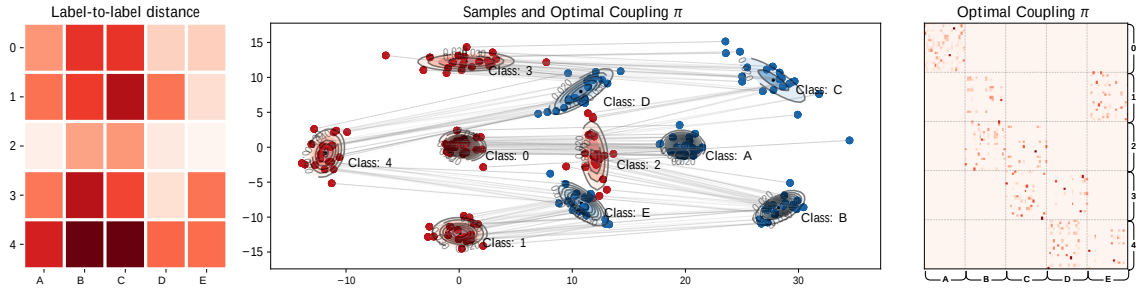

Figure 1: Our approach represents labels as distributions over features and computes Wasserstein distances between them (left). Combined with the usual metric between features, this yields a transportation cost between datasets. The optimal transport problem then characterizes the distance between them as the minimal possible cost of coupling them (optimal coupling * shown on the right).

training a model to completion (*e. g.,* to find optimal parameters) on each dataset being compared. On the opposite side of the spectrum are principled notions of discrepancy between domains Ben-David et al. (2007); Mansour et al. (2009), which nevertheless are often not computable in practice, or do not scale to the type of datasets used in machine learning practice.

In this work, we address some of these limitations by proposing an alternative notion of distance between datasets. At the heart of this approach is the use of optimal transport (OT) distances (Villani, 2008) to compare distributions over feature-label pairs in a geometrically-meaningful and principled way. In particular, we propose a hybrid Euclidean-Wasserstein distance between feature-label pairs across domains, where labels themselves are modeled as distributions over features vectors. As a consequence of this technique, our framework allows for comparison of datasets *even if their label sets are completely unrelated or disjoint*, as long as a distance between their features can be defined. This notion of distance between labels, a by-product of our approach, has itself various potential uses, *e. g.,* to optimally sub-sample classes from large datasets for more efficient pretraining.

In summary, we make the following contributions:

We introduce a notion of distance between datasets that is principled, flexible and computable

We show how to scale up computation of this distance to very large datasets

We provide extensive empirical evidence that this distance is highly predictive of transfer learning success across various domains, tasks and data modalities

Related Work[1]

Notions of (dis)similarity between data distributions have been proposed before. In the context of domain adaptation (Ben-David et al., 2007; Mansour et al., 2009), they are almost always loss and function-dependent and are often used to obtain generalization bounds (Cortes and Mohri, 2011), which despite powerful, their approximation quality relies on quantities that are often incomputable. A different line of work seeks to characterize dataset distances via parameter sensitivity, e.g., via the Fisher Information metric (Achille et al., 2019) or notions from information theory (Achille et al., 2018). Also related are methods that represent complex objects via distributions and compare them via optimal transport distances (Muzellec and Cuturi, 2018; Frogner et al., 2019).

1. We provide a more detailed discussion of related work in Appendix A.

2. Background on Optimal Transport

Optimal transport (OT) is a powerful and principled approach to compare probability distributions (Villani, 2008; Peyré and Cuturi, 2019). It considers a complete and separable metric space X, along with probability measures $\mu \in P(X)$ and $\nu \in P(X)$, continuous or discrete. The Kantorovich formulation Kantorovitch (1942) of the transportation problem reads:

$$\mathrm{OT}(\mu, \nu) \triangleq \min_{\pi \in \Pi(\mu, \nu)} \int_{X \times X} c(x, y) \, d\pi(x, y), \tag{1}$$

where $c(\cdot, \cdot) : X \times X \to \mathbb{R}^+$ is a cost function (the ground cost), and the set of couplings $\Pi(\mu, \nu)$ consists of joint probability distributions over the product space $X \times X$ with marginals μ and ν:

$$\Pi(\mu, \nu) \triangleq \{\pi \in P(X \times X) \mid P_{1\#}\pi = \mu, P_{2\#}\pi = \nu\}. \tag{2}$$

Whenever X is equipped with a metric d_X, it is natural to use it as ground cost, e.g., $c(x, y) = d_X(x, y)^p$ for some $p \geq 1$. In such case, $W_p(\mu, \nu) \triangleq \mathrm{OT}(\mu, \nu)^{1/p}$ is called the p-Wasserstein distance. The case $p = 1$ is also known as the Earth Mover's Distance (Rubner et al., 2000).

The measures μ and ν are rarely known in practice. Instead, one has access to finite samples $\{x^{(i)}\}_{i=1}^n \subset X$, $\{y^{(j)}\}_{j=1}^m \subset X$, which implicitly define discrete measures $\mu = \sum_{i=1}^n a_i \delta_{x^{(i)}}$ and $\nu = \sum_{j=1}^m b_j \delta_{y^{(j)}}$, where a, b are vectors in the probability simplex, and the pairwise costs can be compactly represented as an $n \times m$ matrix C, i.e., $C_{ij} = c(x^{(i)}, y^{(j)})$. In this case, Eq. (1) becomes a linear program, whose cubic complexity is often prohibitive. The entropy-regularized problem

$$\mathrm{OT}_\varepsilon(\mu, \nu) \triangleq \min_{\pi \in \Pi(\mu, \nu)} \int_{X^2} c(x, y) \, d\pi(x, y) + \varepsilon H(\pi \mid \mu \otimes \nu) \tag{3}$$

where $H(\pi \mid \mu \otimes \nu) = \int \log(d\pi / d(\mu \otimes \nu)) \, d\pi$, can be solved much more efficiently and with better sample complexity (Genevay et al., 2019) by using the Sinkhorn algorithm (Cuturi, 2013).

3. Optimal Transport between Datasets

The definition of dataset is notoriously inconsistent across the machine learning literature. Here we are interested in supervised learning, so we define a dataset D as a set of feature-label pairs $(x, y) \in X \times Y$ over a certain feature space X and label set Y. We will use the shorthand notation $z = (x, y)$ and $Z = X \times Y$. Henceforth, we focus on classification, so Y shall be finite. We consider two datasets D_A and D_B, and assume, for simplicity, that their feature spaces have the same dimensionality, but will discuss how to relax this assumption later on. On the other hand, we make no assumptions on the label sets Y_A and Y_B whatsoever. In particular, the classes these encode could be partially overlapping or related (e.g., *imagenet* and *cifar-10*) or completely disjoint (e.g., *cifar-10* and *mnist*). Although not a formal assumption of our approach, it will be useful to think of the samples in these two datasets as being drawn from joint distributions $P_A(x, y)$ and $P_B(x, y)$, i.e., $D_A = \{(x_A^{(i)}, y_A^{(i)})\}_{i=1}^n \sim P_A(x, y)$ and $D_B = \{(x_B^{(j)}, y_B^{(j)})\}_{j=1}^m \sim P_B(x, y)$.

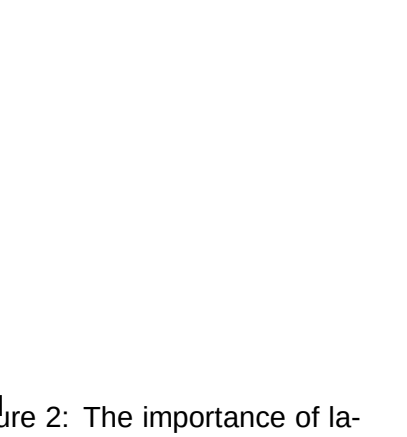

Figure 2: The importance of labels: the second pair of datasets are much closer than the first under the usual (label-agnostic) OT distance, while the opposite is true for our (label-aware) distance.

Our goal is to define a distance $d(D_A, D_B)$ without relying on external models or parameters. The interpretation above, viewed in light of Section 2, suggests comparing the datasets by computing an OT distance between their joint distributions. However, casting Problem (1) in this context requires a crucial component: a metric on Z, i.e., between pairs $(x, y), (x', y')$. If we had metrics on X and Y, we could define a metric on Z as $d_Z((x, y), (x', y')) = (d_X(x, x')^p + d_Y(y, y')^p)^{1/p}$, for $p \geq 1$. In most applications d_X is readily available, e.g., as the euclidean distance in the feature space. On the other hand d_Y will rarely be so, particularly between labels from unrelated label sets (e.g., between CARS in one image domain and DOGS in the other). If we had some prior knowledge of the label spaces, we could use it to define a notion of distance between pairs of labels. However, in the challenging but common case where no such knowledge is available, the only information we have about the labels is their occurrence in relation to the feature vectors. Thus, we can take advantage of the fact that we have a meaningful metric in X and use it to compare labels.

Formally, let $N_D(y) := \{x \in X \mid (x, y) \in D\}$ be the set of feature vectors with label y in dataset D, and let n_y be its cardinality. With this, a distance between two labels y and y' could be defined as the distance between the centroids of $N_D(y)$ and $N_D(y')$. But representing the collections $N_D(y)$ only through their mean is too simplistic for real datasets. Ideally, we would represent labels through the actual distribution over the feature space that they define, namely, by means of the map $y \mapsto \mu_y(X) := P(X \mid Y = y)$, of which $N_D(y)$ can be understood as a finite sample. If we use this representation, defining a distance between labels boils down to choosing a statistical divergence between their associated distributions. Here again we argue that OT is an ideal choice, since it: (i) yields a true metric, (ii) is computable from finite samples, which is crucial since the distributions are not available in analytic form, and (iii) is able to deal with sparsely-supported distributions.

The approach described so far grounds the comparison of the μ_y distributions to the feature space X, so we can simply use d_X^p as the optimal transport cost, leading to a p-Wasserstein distance between labels: $W_p^p(\mu_y, \mu_{y'})$, and in turn, to the following distance between feature-label pairs:

$$d_Z((x, y), (x', y')) := (d_X(x, x')^p + W_p^p(\mu_y, \mu_{y'}))^{1/p}. \tag{4}$$

With this notion of distance in Z at hand we can finally use optimal transport, which lifts this point-wise metric into a distance between measures (and therefore, between datasets):

$$d_{OT}(D_A, D_B) = \min_{\pi \in \Pi(\mu, \nu)} \int_{Z \times Z} d_Z(z, z')\, \pi(z, z'). \tag{5}$$

This defines a true metric (proof in Appendix B), the Optimal Transport Dataset Distance.

It remains to describe how the distributions μ_y are to be represented. We could treat the samples in $N_D(y)$ as support points of a uniform empirical measure, i.e., $\mu_y = \sum_{x^{(i)} \in N_D(y)} \frac{1}{n_y} \delta_{x^{(i)}}$, as described in Section 2. In this case, every evaluation of (4) would involve solving an OT problem, for a total worst-case $O(n^5 \log n)$ complexity, as we show in Appendix E.1. Instead, we propose an alternative representation of the μ_y as Gaussian distributions, which leads to a simple yet tractable realization of the general dataset distance (5). Formally, we model each μ_y as a Gaussian $\mathcal{N}(\hat{\mu}_y, \hat{\Sigma}_y)$ whose parameters are the sample mean and covariance of $N_D(y)$. The main advantage of this approach is that the 2-Wasserstein distance between Gaussians $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$ has an analytic form:

$$W_2^2(\mu_0, \mu_1) = \|\mu_0 - \mu_1\|_2^2 + \operatorname{tr}(\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{\frac{1}{2}} \Sigma_1 \Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}) \tag{6}$$

When using Eq. (6) in the point-wise distance (4), we denote the resulting distance as $d_{OT\text{-}LW}$.

Figure 4: OTDD vs. adaptation accuracy on mnist tasks (left) and mnist + augmentations (right).

4. Experiments

Dataset and model details are provided in the Appendix.

4.1 Dataset Selection for Transfer Learning

In this section, we test whether the OTDD can provide learning-free criterion on which to select a source dataset for transfer learning. We start with a simple domain adaptation setting, using usps, mnist and three of its extensions: fashion-mnist, kmnist and the emnist (letters). We first compute all pairwise OTDD distances (Fig 3). Despite both consisting of digits, mnist and usps are not the closest among these datasets according to the OTDD. The closest pair is instead mnist, (emnist), while fashion-mnist appears far from all others, particularly from mnist. Next, we compare these distances against the transferability between datasets, i.e., the gain in performance from using a model pertrained on the source domain and fine-tuning it on the target domain. To make these numbers comparable across dataset pairs, we report the relative drop in classification error brought by adaptation: $T(D_S \to D_T) = 100 \cdot \frac{\text{error}(D_S \to D_T) - \text{error}(D_T)}{\text{error}(D_T)}$. We run the adaptation task 10 times with different random seeds for each pair of datasets, and compare T against their distance. The strong significant correlation between these (Fig. 4) shows that the OTDD is highly predictive of transferability across these datasets. In particular, emnist led to the best adaptation to mnist, justifying the initially counter-intuitive value of the OTDD.

Figure 3: Pairwise OT Distances

4.2 Distance-Driven Data Augmentation

Data augmentation is another key aspect of transfer learning that has substantial empirical effect on the quality of the transferred model yet lacks principled guidelines. Here, we investigate if the OTDD could be used to compare and select among possible augmentations. For a fixed source-target dataset pair, we generate replicas of the source data with various transformations applied to it, compute their distance to the target dataset, and compare against the transferability as before.

Figure 5: Tiny-ImageNet $\to$ cifar-10.

We present results for small (mnist ! usps) and larger-scale (Tiny-ImageNet cifar-10) settings. The transformations we use on mnist consist of rotations by a fixed degree $\{30, \ldots, 180\}$, random rotations $(-180, 180)$, random affine transformations, center-crops and random crops. For Tiny-ImageNet we randomly vary brightness, contrast, hue and saturation. The models use are respectively the LeNet-5 and a ResNet-50. The results in both of these settings (Figs. 4 and 5) show, again, a strong significant correlation between these two. Note in particular that cropping mnist images substantialy improves performance on usps, while most rotations degrade transferability.

4.3 Transfer Learning for Text Classification

Natural Language Processing has seen a profound impact from large-scale transfer learning, partly due to the availability of off-the-shelf large language-models pretrained on massive amounts of the data (Devlin et al., 2019; Radford et al., 2019). While natural language inherently lacks the fixed-size continuous representation required by our framework to compute pointwise distances, we can take advantage of these pretrained models to embed sentences in a vector space with rich geometry. Here, we first embed the sentences of every dataset using the (base) bert model (Devlin et al., 2019) and then compute OTDD on these embedded datasets. We focus on the task of sentence classification, and consider the following datasets by Zhang et al. (2015): 15 news (ag), dbpedia (db), yelp reviews (5-way: yl_5, and binary: yl_+), amazon reviews (5-way: am_5, and binary: am_+), and yahoo answers (yh). We provide details for these datasets in the Appendix. As before, we simulate a challenging adaptation setting by keeping only 100 examples per target class. For every pair of datasets, we first fine-tune the bert model using the entirety of the source domain data, after which we fine-tune and evaluate on the target domain. Figure 6 shows that the OTDD is highly correlated with transferability in this setting too. Interestingly, adaptation sometimes substantially degrades performance, which suggests that off-the-shelf bert is powerful enough on its own to initialize many of these tasks, so that choosing the wrong domain for pretraining might be counterproductive.

Figure 6: Transfer in text classification.

5. Discussion

We have shown that the notion of distance between datasets proposed in this work is scalable and flexible enough to be used in realistic transfer learning scenarios, all the while offering appealing theoretical properties, interpretable comparisons and making minimal assumptions on the underlying datasets. There are many natural extensions of this framework. Here we assumed that the datasets were defined on feature spaces of the same dimension, but one could instead leverage a relational notion such as the Gromov-Wasserstein distance (Mémoli, 2011) to compute the distance between datasets whose features and not directly comparable. On the other hand, our efficient implementation relies on modeling groups of points with the same label as Gaussians, but this could be extended to more general distributions for which the Wasserstein distance has an analytic solution or at least can be computed efficiently, such as Gaussian mixture models (Delon and Desolneux, 2019).

References

Alessandro Achille, Glen Mbeng, and Stefano Soatto. Dynamics and reachability of learning tasks. arXiv e-prints, October 2018.

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task embedding for Meta-Learning. In Proceedings of the IEEE International Conference on Computer Vision, pages 6430 6439, 2019.

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, Advances in Neural Information Processing Systems 30, pages 1964 1974. Curran Associates, Inc., 2017.

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1881 1890, 2018.

David Alvarez-Melis, Tommi S Jaakkola, and Stefanie Jegelka. Structured optimal transport. In Amos Storkey And, editor, Proceedings of the Twenty-First International Conference on Arti cial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pages 1771 1780. PMLR, 2018.

Shun-Ichi Amari. Di erential-Geometrical Methods in Statistics, volume 28 of Lecture Notes in Statistics Springer New York, New York, NY, 1985. ISBN 9780387960562, 9781461250562.

Shun-Ichi Amari. Natural gradient works e ciently in learning. Neural Comput., 10(2):251 276, February 1998. ISSN 0899-7667.

Shun-Ichi Amari and Hiroshi Nagaoka. Methods of Information Geometry. Translations of Mathematical Monographs. American Mathematical Society, 2000. ISBN 9780821843024.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B Schölkopf, J C Platt, and T Ho man, editors, Advances in Neural Information Processing Systems 19, pages 137 144. MIT Press, 2007.

Tony F Chan, Gene H Golub, and Randall J Leveque. Algorithms for computing the sample variance: Analysis and recommendations. Am. Stat., 37(3):242 247, August 1983. ISSN 0003-1305.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. arXiv e-prints, December 2018.

G Cohen, S Afshar, J Tapson, and A van Schaik. EMNIST: Extending MNIST to handwritten letters. In 2017 International Joint Conference on Neural Networks (IJCNN), pages 2921 2926. IEEE, May 2017.

Corinna Cortes and Mehryar Mohri. Domain adaptation in regression. In Algorithmic Learning Theory, pages 308 323. Springer Berlin Heidelberg, 2011.

Nicolas Courty, Remi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1853–1865, September 2017. ISSN 0162-8828.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

Lieven De Lathauwer. Simultaneous matrix diagonalization: the overcomplete case. In *Proc. of the 4th International Symposium on ICA and Blind Signal Separation, Nara, Japan*, volume 8122, page 825. kecl.ntt.co.jp, 2003.

Julie Delon and Agnes Desolneux. A wasserstein-type distance in the space of gaussian mixture models. *arXiv e-prints*, July 2019.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, June 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Yonatan Dukler, Wuchen Li, Alex Lin, and Guido Montufar. Wasserstein of Wasserstein loss for learning generative models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1716–1725, Long Beach, California, USA, 2019. PMLR.

Charlie Frogner, Farzaneh Mirzazadeh, and Justin Solomon. Learning embeddings into entropic wasserstein spaces. In *International Conference on Learning Representations*, May 2019.

Matthias Gelbrich. On a formula for the L2 wasserstein metric between measures on euclidean and hilbert spaces. *Math. Nachr.*, 147(1):185–203, November 1990. ISSN 0025-584X.

Aude Genevay, Léna\ ic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583. PMLR, 2019.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, January 2008. ISBN 9780898717778.

J. J. Hull. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(5):550–554, May 1994. ISSN 0162-8828, 1939-3539.

L. Kantorovitch. On the translocation of masses. *Dokl. Akad. Nauk SSSR*, 37(7-8):227–229, 1942. ISSN 0002-3264.

Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive Gradient-Based Meta-Learning methods. In H Wallach, H Larochelle, A Beygelzimer, F d'Alché Buc, E Fox, and R Garnett, editors, Advances in Neural Information Processing Systems 32, pages 5915 5926. Curran Associates, Inc., 2019.

Alex Krizhevsky and Geo rey Hinton. Learning multiple layers of features from tiny images. 2009.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Sha eezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. arXiv e-prints, August 2019.

Yann LeCun, Corinna Cortes, and C J Burges. MNIST handwritten digit database. 2010.

Rui Leite and Pavel Brazdil. Predicting relative performance of classi ers from samples. In Proceedings of the 22nd international conference on Machine learning - ICML '05, pages 497 503. dl.acm.org, 2005.

Ling Li. Data Complexity in Machine Learning and Novel Classi cation Algorithms. PhD thesis, California Institute of Technology, 2006.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-Rao metric, geometry, and complexity of neural networks. In Proceedings of the 22nd International Conference on Arti cial Intelligence and Statistics. PMLR, 2019.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In The 22nd Conference on Learning Theory. arxiv.org, 2009.

Facundo Mémoli. Gromov Wasserstein distances and the metric approach to object matching. Found. Comput. Math., 11(4):417 487, August 2011. ISSN 1615-3375, 1615-3383.

Facundo Mémoli. Distances between datasets. In Laurent Najman and Pascal Romon, editors, Modern Approaches to Discrete Curvature, pages 115 132. Springer International Publishing, Cham, 2017. ISBN 9783319580029.

Boris Muzellec and Marco Cuturi. Generalizing point embeddings using the wasserstein space of elliptical distributions. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, Advances in Neural Information Processing Systems 31, pages 10237 10248. Curran Associates, Inc., 2018.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. Foundations and Trends® in Machine Learning, 11(5-6):355 607, 2019. ISSN 1935-8237.

Alec Radford, Je rey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. Better language models and their implications. OpenAI Blog https://openai.com/blog/better-language-models, 2019.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis., 40(2):99 121, November 2000. ISSN 0920-5691, 1573-1405.

Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In The IEEE International Conference on Computer Vision (ICCV), 2019.

Cédric Villani. Optimal transport, Old and New, volume 338. Springer Science & Business Media, 2008. ISBN 9783540710493.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. August 2017.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. Hierarchical optimal transport for document representation. In H Wallach, H Larochelle, A Beygelzimer, F d'Alché Buc, E Fox, and R Garnett, editors, Advances in Neural Information Processing Systems 32, pages 1599 1609. Curran Associates, Inc., 2019.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classi cation. In C Cortes, N D Lawrence, D D Lee, M Sugiyama, and R Garnett, editors, Advances in Neural Information Processing Systems 28, pages 649 657. Curran Associates, Inc., 2015.

Appendix A. Detailed Related Work

Discrepancy Distance Various notions of (dis)similarity between data distributions have been proposed in the context of domain adaptation, such as the d_A (Ben-David et al., 2007) and discrepancy distances (Mansour et al., 2009). These discrepancies depend on a loss function and hypothesis (i. e., predictor) class, and quantify dissimilarity through a supremum over this function class. The latter discrepancy in particular has proven remarkably useful for proving generalization bounds for adaptation (Cortes and Mohri, 2011), and while it can be estimated from samples, bounding the approximation quality relies on quantities like the VC-dimension of the hypothesis class, which might not be always known or easy to compute.

Dataset Distance via Parameter Sensitivity The Fisher information metric (FIM) is a classic notion from information geometry (Amari, 1985; Amari and Nagaoka, 2000) that characterizes a parametrized probability distribution locally through the sensitivity of its density to changes in the parameters. In machine learning, it has been used to analyze and improve optimization approaches (Amari, 1998) and to measure the capacity of neural networks (Liang et al., 2019). In recent work, Achille et al. (2019) use this notion to construct vector representations of tasks, which they then use to de ne a notion of similarity between these. They show that this notion recovers taxonomic similarities and is useful in meta-learning to predict whether a certain feature extractor will perform well in a new task. While this notion shares with ours its agnosticism of the number of classes and their semantics, it di ers in the fact that it relies on a probe network trained on a speci c dataset, so its geometry is heavily in uenced by the characteristics of this network. A related information-theoretic notion of complexity that can be used to characterize tasks is the Kolmogorov Structure Function (Li, 2006), which Achille et al. (2018) use to de ne a notion of *reachability* between tasks.

Optimal Transport-based distributional distances The general idea of representing complex objects via distributions, which are then compared through optimal transport distances, is an active area of research. Also driven by the appeal of their closed-form Wasserstein distance, Muzellec and Cuturi (2018) propose to embed objects as elliptical distributions, which requires di erentiating through these distances, and discuss various approximations to scale up these computations. Frogner et al. (2019) extend this idea but represent the embeddings as discrete measures (*i. e.*, point clouds) rather than Gaussian/Elliptical distributions. Both of these works focus on embedding and consider only within-dataset comparisons. Also within this line of work, Delon and Desolneux (2019) introduce a Wasserstein-type distance between Gaussian mixture models. Their approach restricts the admissible transportation couplings themselves to be Gaussian mixture models, and does not directly model label-to-label similarity. More generally, the Gromov-Wasserstein distance (Mémoli, 2011) has been proposed to compare collections across di erent domains (Mémoli, 2017; Alvarez-Melis and Jaakkola, 2018), albeit leveraging only features, not labels.

Hierarchical OT distances The distance we propose can be understood as a hierarchical OT distance, *i. e.*, one where the ground metric itself is de ned through an OT problem. This principle has been explored in other contexts before. For example, Yurochkin et al. (2019) use a hierarchical OT distance for document similarity, de ning a inner-level distance between topics and a outer-level distance between documents using OT. (Dukler et al., 2019) on the other hand use a nested Wasserstein distance as a loss for generative model training, motivated by the observation that the Wasserstein distance is better suited to comparing images than the usual pixel-wise L_2 metric used as ground metric. Both the goal, and the actual metric, used by these approaches di ers from ours.

Optimal Transport for Domain Adaptation Using label information to guide the optimal transport problem towards class-coherent matches has been explored before, e.g., by enforcing group-norm penalties (Courty et al., 2017) or through submodular cost functions (Alvarez-Melis et al., 2018). These works are focused on the unsupervised domain adaptation setting, so their proposed modifications to the OT objective use only label information from one of the two domains, and even then, do so without explicitly defining a metric between these. Furthermore, they do not lead to proper distances, and these works deal with a single static pair of tasks, so they lack analysis of the distance across multiple source and target datasets.

Appendix B. OTDD is a Proper Distance

Proposition 1 $d_{OT}(D_A, D_B)$ defines a valid metric on $\mathcal{P}(X \times \mathcal{P}(X))$ the space of measures over feature and label-distribution pairs.

Proof Whenever the cost function used is a metric in a given space X, the optimal transport problem itself defines a distance (the Wasserstein distance) on $\mathcal{P}(X)$ (Villani, 2008, Chapter 6). Therefore, it suffices to show that the cost function d_Z defined in Eq. (4) is indeed a distance. Clearly, it is symmetric because both d_X and W_p are. In addition, since both of these are distances:

$$d_Z(z, z') = 0 \;, \quad d_X(x, x') = 0 \wedge W_p(\alpha_y, \alpha_{y'}) = 0 \;, \quad x = x', \quad \alpha_y = \alpha_{y'}, \quad z = z'$$

Finally, we have that

$$\begin{aligned} d_Z(z_1, z_3) &= \left(d_X(x_1, x_3)^p + W_p(\alpha_{y_1}, \alpha_{y_3})^p \right)^{\frac{1}{p}} \\ &\leq \left(d_X(x_1, x_2)^p + d_X(x_2, x_3)^p + W_p(\alpha_{y_1}, \alpha_{y_2})^p + W_p(\alpha_{y_2}, \alpha_{y_3})^p \right)^{\frac{1}{p}} \\ &= \left(d_Z(z_1, z_2)^p + d_Z(z_2, z_3)^p \right)^{\frac{1}{p}} \leq d_Z(z_1, z_2) + d_Z(z_2, z_3) \end{aligned}$$

where the last step is an application of Minkowski's inequality. Hence, d_Z satisfies the triangle inequality, and therefore it is a metric on $Z = X \times \mathcal{P}(X)$. We therefore conclude that the value of the optimal transport (5) that uses this metric as a cost function is a distance itself. ■

Appendix C. A Gelbrich-Type Bound

Representing label-defined distributions as Gaussians might seem like a heuristic choice driven only by algebraic convenience. However, Proposition 3 shows that this approximation lower-bounds the distance that would be obtained had it been computed using the label distances on the true distributions (regardless of their form). This result is a direct extension of the following well-known bound for the 2-Wasserstein distance due to Gelbrich (1990):

Lemma 2 (Gelbrich bound) Suppose $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$ are any two measures with mean vectors $\mu_\alpha, \mu_\beta \in \mathbb{R}^d$ and covariance matrices $\Sigma_\alpha, \Sigma_\beta \in \mathbb{S}_+^d$ respectively. Then,

$$W_2^2\left(\mathcal{N}(\mu_\alpha, \Sigma_\alpha), \mathcal{N}(\mu_\beta, \Sigma_\beta)\right) \leq W_2^2(\alpha, \beta) \tag{7}$$

where $W_2^2\left(\mathcal{N}(\mu_\alpha, \Sigma_\alpha), \mathcal{N}(\mu_\beta, \Sigma_\beta)\right)$ is as in Eq. (6).

For our setting, we have:

Proposition 3 For any two datasets D_A, D_B, we have:

$$d_{\text{OT-N}}(D_A, D_B) \quad d_{\text{OT}}(D_A, D_B) \tag{8}$$

Furthermore, if the label distributions $_y$ are all Gaussian or elliptical, these quantities are equal, i.e., $d_{\text{OT-N}}$ is exact.

Proof In the notation of Section 3, Lemma 2 implies that for every feature-label pair (x, y) and $z' = (x', y')$, we have:

$$d_X(x, x') + W_2^2(\ N(\ _y,\ _y), N(\ _{y'},\ _{y'})) \quad d_X(x, x') + W_2^2(\ _y,\ _{y'}); \tag{9}$$

and therefore

$$\int d_Z(z, z')\, d \qquad \int d_Z(z, z')\, d \tag{10}$$

for every coupling $2\ (\ ,\ )$. In particular, for the minimizing , we obtain that

$$d_{\text{OT}}(D_A, D_B, N) \quad d_{\text{OT}}(D_A, D_B) \tag{11}$$

Clearly, Gelbrich's bound holds with equality when and are indeed Gaussian. More generally, equality is attained for elliptical distributions with the same density generator (Kuhn et al., 2019). This immediately implies equality of the two quantities in equation (11) in that case. ■

Appendix D. Computational Considerations

Since our goal in this work is to use the proposed dataset distance as a tool for tasks like transfer learning in realistic (i.e., large) machine learning datasets, scalability is crucial. Indeed, most compelling use cases of any notion of distance between datasets will involve computing it repeatedly on very large samples. While estimation of Wasserstein and more generally, optimal transport distances is known to be computationally expensive in general, in Section 2 we brie y discussed how entropy regularization can be used to trade-o accuracy for runtime. Recall that both the general and Gaussian versions of the dataset distance proposed in Section 3 involve solving optimal transport problems (though the latter, owing the closed form solution of subproblem (6), only requires optimization for the global problem). Therefore, both of bene t from approximate OT solvers.

But further speed-ups are possible. For $d_{\text{OT-N}}$, a simple and fast implementation can be obtained if (i) the metric in X coincides with the ground metric in the transport problem Y, and (ii) all covariance matrices commute. While (ii) will rarely occur in practice, one could use a diagonal approximation to the covariance or simultaneous matrix diagonalization (De Lathauwer, 2003). In either case, Eq. (6) further simpli es to $W_2^2(N(\ ,\ ), N(\ ,\ )) = \| \quad \|_2^2 + \| ^{\frac{1}{2}} \quad ^{\frac{1}{2}} \|_2^2$, so the pointwise distance $d(z, z')$ can be computed by creating augmented representations of each dataset, whereby each pair (x, y) is represented as a stacked vector $= [x;\ _y; \text{vec}(\ _y^{1=2})]$ for the corresponding label mean and covariance. Then $\| - \|_2^2 = d_Z(x, y, x', y')^2$ for d_Z as de ned in Eq. (4). Therefore, in this case the OTDD can be immediately computed using an o -the-shelf OT solver on these augmented datasets. While this approach is appealing computationally, here instead we focus on a exact version that does not require diagonal or commuting covariance approximations, and leave empirical evaluation of this approximate approach for future work.

What we propose next is motivated by the observation that, unlike usual OT distances for which the cost of computing pair-wise distance is negligible compared to the complexity of optimization, in our case the former dominates, since it involves computing multiple OT distances itself. In order to speed up computation, we first precompute and store in memory all label-to-label pairwise distances $d(\mu_y, \mu_{y'})$, and retrieve them on-demand during the optimization of the global OT problem.

For $d_{\text{OT-N}}$, computing the label-to-label distances $d(\mathcal{N}(\hat{\mu}_y, \hat{\Sigma}_y), \mathcal{N}(\hat{\mu}_{y'}, \hat{\Sigma}_{y'}))$ is dominated by the cost of matrix square roots, which if done exactly involves a full eigendecomposition. Instead, it can be computed approximately using the Newton-Schulz iterative method (Higham, 2008; Muzellec and Cuturi, 2018). Besides runtime, loading all examples of a given class to memory (to compute means and covariances) might be infeasible for large datasets (especially if running on GPU), so we instead use a two-pass stable online batch algorithm to compute these statistics (Chan et al., 1983).

The following result summarizes the time complexity of our two distances and sheds light on the trade-off between precision and efficiency they provide.

Theorem 4 For datasets of size n and m, with p and q classes, dimension d, and maximum class size n, both d_{OT} and $d_{\text{OT-N}}$ incur in a cost of $O(nm \log(\max\{n, m\}) \varepsilon^{-3})$ for solving the global OT problem ε-approximately, while the worst-case complexity for computing the label-to-label pairwise distances (4) is $O\left(nm(d + n^3 \log n + dn^2)\right)$ for d_{OT} and $O\left(nmd + pqd^3 + d^2n(p+q)\right)$ for $d_{\text{OT-N}}$.

In most practical applications, the cost of computing pairwise distances will dominate, making $d_{\text{OT-N}}$ superior. For example, if $n = m$ and the largest class size is $O(n)$, this step becomes $O(n^5 \log n)$ prohibitive for all but toy datasets for d_{OT} but only $O(n^2 d + d^3)$ for $d_{\text{OT-N}}$.

Appendix E. Time Complexity Analysis

For the analyses in this section, assume D_S and D_T respectively have n and m labeled examples in $\mathbb{R}^d$ and k_s, k_t classes. In addition, let $N_D^S(i) := \{x \in X \mid (x, y = i) \in D\}$ be the subset of examples in D_S with label i, and define analogously $N_D^T(j)$. The denote the cardinalities of these subsets as n_s^i, $|N_s^{(i)}|$ and analogously for n_t^j.

Direct computation of the distance (4) involves two main steps:

(i) computing pairwise pointwise distances (each requiring solution of a label-to-label OT sub-problem), and

(ii) a global OT problem between the two samples.

Step (ii) is identical for both the general distance d_{OT} and its Gaussian approximation counterpart $d_{\text{OT-N}}$, so we analyze it first. This is an OT problem between two discrete distributions of size n and m, which can be solved exactly in $O\left((n+m)nm \log(nm)\right)$ using interior point methods or Orlin's algorithm for the uncapacitated min cost flow problem (Peyré and Cuturi, 2019). Alternatively, it can be solved ε-approximately in $O(nm \log(\max\{n, m\}) \varepsilon^{-3})$ time using the Sinkhorn algorithm (Altschuler et al., 2017).

We next analyze step (i) individually for the two OTDD versions. Combined, they provide a proof of Theorem 4.

E.1 Pointwise distance computation for d_{OT}

Consider a single pair of points $(x, y = i) \in D_A$ and $(x', y' = j) \in D_B$. Evaluating $\|x - x'\|$ has $O(d)$ complexity, while $W(\cdot_y, \cdot_{y'})$ is an $n_s^i \times n_t^j$ OT problem which itself requires computing a distance matrix (at cost $O(n_s^i n_t^j d)$), and then solving the OT problem, which as discussed before, be done exactly in $O((n_s^i + n_t^j) n_s^i n_t^j \log(n_s^i + n_t^j))$ or ε-approximately in $O(n_s^i n_t^j \log(\max\{n_s^i, n_t^j\})\varepsilon^{-3})$.

For simplicity, let us denote $n_s = \max_i n_s^i$, and $n_t = \max_j n_t^j$ the size of the largest label cluster in each dataset, and $n = \max\{n_s, n_t\}$ the overall largest one. Using these, and combining all of the above, the overall worst case complexity for the computation of the nm pairwise distances can be expressed as

$$O\left(nm(d + n^3 \log n + dn^2)\right), \tag{12}$$

which is what we wanted to show.

E.2 Pointwise distance computation for $d_{OT\text{-}N}$

As before, consider a pair of points $(x, y = i) \in D_A$ and $(x', y' = j) \in D_B$ whose cluster sizes are n_s^i and n_t^j respectively. As mentioned in Section D, for $d_{OT\text{-}N}$ we first compute all the per-class means and covariance matrices. This step is clearly dominated by latter, which costs $O(d^2 n_s^i)$.[2] Considering all labels from both datasets, this amounts to a worst-case complexity of $O(d^2(k_s n_s + k_t n_t))$. Once the means and covariances have been computed, we precompute all the $k_s k_t$ pair-wise label-to-label distances $W_2(\cdot_y, \cdot_{y'})$ using Eq. (6). This computation is dominated by the matrix square roots. If done exactly, these involve a full eigendecomposition, at cost $O(d^3)$, so the total cost for this step is $O(k_s k_t d^3)$. Finally, while computing the pairwise distance, we will incur $O(nmd)$ to obtain $\|x - x'\|$. Putting all of these together, and replacing n_s, n_t by n, we obtain a total cost for precomputing all the point-wise distances of $O(nmd + k_s k_t d^3 + d^2 n(k_s + k_t))$.

Appendix F. Dataset Details

Dataset	Input Dimension	Number of Classes	Train Examples	Test Examples	Source
usps	16 × 16	10	7291	2007	(Hull, 1994)
mnist	28 × 28	10	60K	10K	(LeCun et al., 2010)
emnist (letters)	28 × 28	26	145K	10K	(Cohen et al., 2017)
kmnist	28 × 28	10	60K	10K	(Clanuwat et al., 2018)
fashion-mnist	28 × 28	10	60K	10K	(Xiao et al., 2017)
Tiny-ImageNet	64 × 64[z]	200	100K	10K	(Deng et al., 2009)
cifar-10	32 × 32	10	50K	10K	(Krizhevsky and Hinton, 2009)
ag-news	768[y]	4	120K	7.6K	(Zhang et al., 2015)
DBpedia	768[y]	14	560K	70K	(Zhang et al., 2015)
YelpReview (Polarity)	768[y]	2	560K	38K	(Zhang et al., 2015)
YelpReview (Full Scale)	768[y]	5	650K	50K	(Zhang et al., 2015)
AmazonReview(Polarity)	768[y]	2	3.6M	400K	(Zhang et al., 2015)
AmazonReview(Full Scale)	768[y]	5	3M	650K	(Zhang et al., 2015)
Yahoo Answers	768[y]	10	1.4M	60K	(Zhang et al., 2015)

Table 1: Summary of datasets used in this work. x: we rescale usps digits to 28 × 28 for comparison to the *nist datasets. z: we rescale Tiny-ImageNet to 32 × 32 for comparison to cifar-10. y: for text datasets, variable-length sentences are embedded to fixed-dimensional vectors by using

2. technically, this would be $O(d^\omega n_s^i)$ where ω is the coefficient of matrix multiplication, but we take $\omega = 3$ for simplicity.

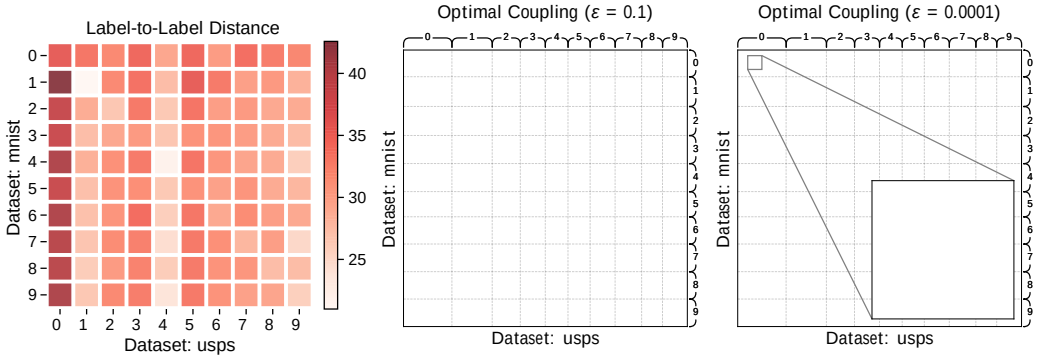

Figure 7: Dataset Distance between MNIST and USPS. **Left**: The label Wasserstein distances —computed without knowledge of the relation between labels across domains— recover expected relations between classes in the two domains. **Center/Right**: The optimal coupling * for different regularization levels exhibits a block-diagonal structure, indicating class-coherent matches across domains.

Appendix G. Optimization and Training Details

For the adaptation experiments on the *NIST datasets, we use a LeNet-5 architecture (two convolutional layers, three fully conntected ones) with ReLU activations trained for 20 epochs using ADAM with learning rate 1 10^{-3} and weight decay 1 10^{-6} It was fine-tuned for 10 epochs on the target domain(s) using the same optimization parameters. When transferring, we freeze the convolutional layers and fine-tune only the top three layers.

For the Tiny-ImageNet to CIFAR-10 adaptation results, we use a ResNet-50 trained for 300 epochs using SGD with learning rate 0.1 momentum 0.9 and weight decay 1 10^{-4} It was fine-tuned for 30 epochs on the target domain using SGD with same parameters except 0.01 learning rate. We discard pairs for which the variance on adaptation accuracy is beyond a certain threshold.

For the text classification experiments, we use a pretrained BERT architecture (the **`bert-base-uncased`** model of the **`transformers`**[3] library). We first embed all sentences using this model. Then, for each pair of source/target domains, we first fine-tune using ADAM with learning rate 2 10^{-5} for 10 epochs on the full source domain data, and the fine-tune on the restricted target domain data with the same optimization parameters for 2 epochs.

Our implementation of the OTDD relies on the **`pot`**[4] and **`geomloss`**[5] python packages.

3. **`huggingface.co/transformers/`**
4. **`pot.readthedocs.io/en/stable/`**
5. **`www.kernel-operations.io/geomloss/`**