PYRAMID MINI-BATCH OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligning distributions by minimizing optimal transport distances has been shown to be effective in a variety of machine learning settings, including generative modeling and domain adaptation. Computing optimal transport distances over large numbers of data points is intractable, so mini-batch-based optimal transport must be used, but it risks computing inaccurate distances between two distributions when the randomly sampled pairs of mini-batches are not optimal pairs. In this work, we propose a geometric mini-batch sampling scheme which orders the mini-batches using pyramid-based encodings. By building geometrically consistent batches, Pyramid Mini-Batching significantly improves the quality of optimal transport approximations and downstream alignments with minimal computational overhead. We perform experiments over the Discrete Optimal Transport benchmark to demonstrate the effectiveness of this strategy over multiple established optimal transport settings and see that our approach improves estimates of OT distances by nearly 30% for single pass estimation, and when attempting to minimize optimal transport distance is ten times more effective than with random mini-batch sampling.

1 INTRODUCTION

Optimal transport costs, especially Wasserstein distances, have proven to be a useful measure of the distance between distributions in a variety of settings. In addition to providing a geometrically meaningful distance measure, Wasserstein distances also produce stable measures for distributions that lack mutual support, as is often the case in domain adaptation or generative modeling settings (Arjovsky et al., 2017), which learn to minimize the distance measured on mini-batches to align or generate distributions. Given n data points, solving for exact Wasserstein distances requires $O(n^3 log(n))$ computations, which is not tractable for large n. Many approaches have been proposed to efficiently estimate Wasserstein distances, with mini-batch training being one of the most straightforward.

A general limitation of learning alignments with mini-batch optimal transport approaches is that the quality of learned alignment is strongly influenced by the ability of a mini-batch to fully capture the *meaningful* diversity of the sampled distributions. Even if the distance within mini-batches can be efficiently computed, the approximation is restricted by the mini-batch size. For example, when learning to align model features between two distributions that contain the 1K label space of ImageNet, a random batch of 256 samples from both distributions would inevitably assign couplings between samples that represent different labels. Aligning examples of mismatched classes can lead to detrimental performance on downstream classification tasks, and with a random mini-batching approach the only way to overcome this is to increase the size of the mini-batch, which results in a significant increase in the computation time.

Our approach overcomes this limitation by building batches that are *geometrically consistent* with their accompanying batch from the other distribution. Instead of computing optimal transport on random mini-batches, our approach constructs mini-batches considering the coupling over the whole set of data points before splitting them into mini-batches. We exploit the locality properties of tree-based hashing structures to efficiently sample near-corresponding pairs for each mini-batch set. We encode every data point into a multi-resolution pyramid structure and extract the implied coupling of data points from the structure. Mini-batches are constructed so that correspondences fall into the same mini-batch. Pyramid encoding and coupling inference adds little to the overall run-time, but significantly improves accuracy and reduces bias for small batch sizes making it particularly well-suited for large-scale learning. Our key insight is that estimated coupling from pyramid encodings are used to seed the mini-batches in near-linear time.

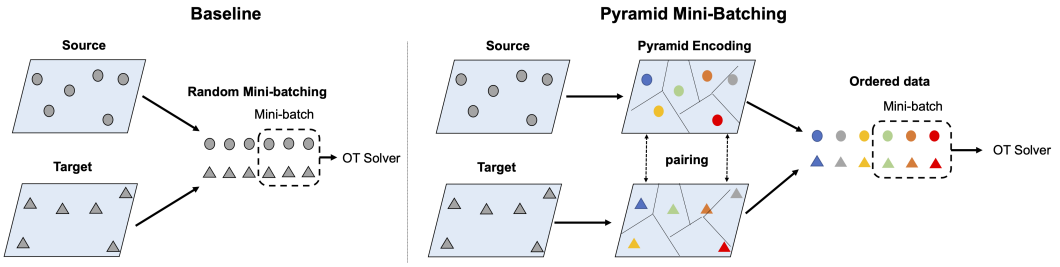

Figure 1: Unlike baseline optimal transport with randomly sampled mini-batches, our approach first jointly encodes the entire source and target datasets according to our Pyramid Matching structure and subsequently extracts mini-batches according to this structure. Exact Optimal Transport solvers can then be used on these mini-batches to either create a global transport plan or minimize an OT loss.

We note that our approach is complementary to and can be combined with existing efficient mini-batch optimal transport methods. While existing methods focus on how to approximate the distance within the sampled mini-batches, our approach improves the construction of mini-batches by considering global data structure.

Our results shows that by using multi-resolution encodings to guide the creation of mini-batches we can substantially improve the model's ability to align representations according to optimal transport distances. Over synthetic and real-world problem settings, we demonstrate the ability of our approach to produce more accurate estimates of optimal transport distance and drastically improve the ability to learn with respect to optimal transport distances.

To summarize, we (1) propose a geometric sampling approach for mini-batch optimal transport, (2) propose a multi-resolution encoding strategy that works well over disjoint high-dimensional domains, and (3) show our method outperforms popular alternatives on the challenging task of minimizing Wasserstein distance in complex domains.

2 Wasserstein Distance : Review

We begin with a discussion of the Wasserstein distance and common ways of computing it (or approximations of it). We consider the case of optimal transport on $\mathbb{R}^d$. Assume datasets X and Y are subsets of $\mathbb{R}^d$ and μ and υ are probability measures on X and Y, respectively. For simplicity, we assume that both datasets are of size N. If $\Pi(\mu, \upsilon) := \pi : \int \pi(x, y)dx = \upsilon(y), \int \pi(x, y)dy = \mu(x)$ is the transportation plan between μ and υ, the Wasserstein distance is defined as follows:

$$W_p(\mu, \upsilon) = \left(\min_{\pi \in \Pi(\mu,\upsilon)} \mathbb{E}_{\pi(x,y)} \|x - y\|^p) \right)^{1/p} \tag{1}$$

where π transports a point in distribution μ to another point in distribution υ . As our work is suited for the discrete optimal transport setting, we can further assume that we have discrete measures of the form $\mu = \Sigma_{i=1}^n \mu_i \delta_{x_i}$ and $\upsilon = \Sigma_{i=1}^n \upsilon_i \delta_{y_i}$ where δ_x and δ_y are the Dirac masses at point x and y, respectively. Letting $c_{ij} = ||x_i - y_j||^p$, c forms a rectangular grid of resolution $n \times n$ in $\mathbb{R}^2$. The amount of mass transported from x_i to y_j can be represented by π_{ij}. Thus the Wasserstein distance in the discrete optimal transport setting can be represented as follows:

$$\begin{aligned} \mathbb{W}_p = \min \left(\Sigma_{i=1}^n \Sigma_{j=1}^n c_{ij} \pi_{ij} \right)^{1/p} \\ \text{subject to } \Sigma_{j=1}^n \pi_{ij} = \mu_i \forall_i = 1, ..., n \\ \Sigma_{i=1}^n \pi_{ij} = \upsilon_j \forall_j = 1, ..., n \\ \pi_{ij} \geq 0. \end{aligned}$$

Network simplex and interior point methods are common efficient solvers for Wasserstein distance, both of which have a computational complexity of at least $O(N^3 log(N))$, which becomes intractable as N grows beyond a few thousand points.

Mini-Batch Optimal Transport As machine learning applications often deal with settings where N is greater than 100,000 data points, mini-batch optimal transport has become a popular approach

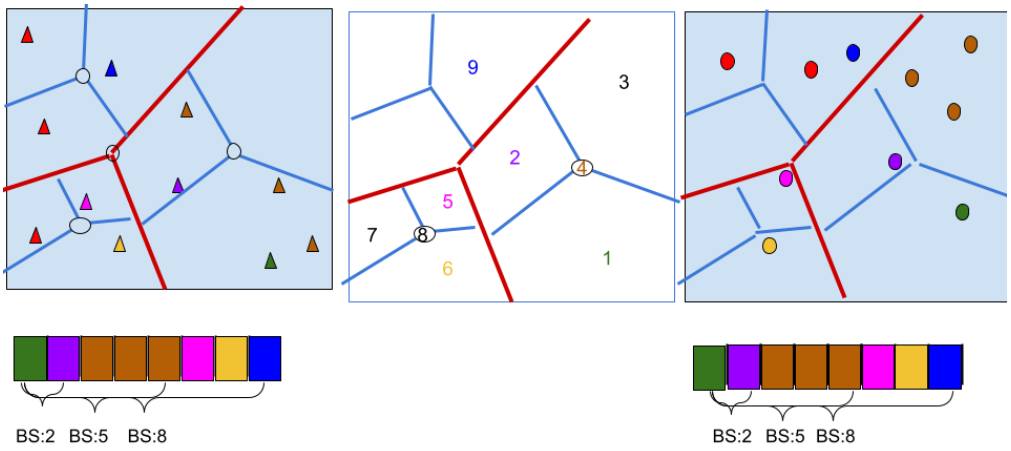

Figure 2: Visualization of how a batch of various sizes $(2, 5, 8)$ may be extracted with our Pyramid Mini-Batching (PMB) approach. The source distribution is mapped on the lefthand side of the figure, and the target distribution is mapped on the righthand side of the figure. The center panel illustrates the order in which nodes are visited when building a batch according to this structure, up until 8 data points are selected. Points selected at the same step of tree traversal share the same color and nodes which did not contribute any pairs are represented as black. We first select all available pairings from a leaf node, then all of its sibling nodes, and subsequently the parent node. After exhausting the parent node, we then transition to a leaf node of one its siblings and repeat the procedure until we've constructed a batch of the desired size.

for dealing with the intractability of this problem (Fatras et al., 2019; Sommerfeld et al., 2019; Feydy et al., 2019). We let x^m denote a batch of m samples from X and let X_k^m denote a set of k such batches. As such the mini-batch Wasserstein estimation can be defined as following:

$$\mathbb{W}_p^m = k^{-1} \sum_{x^m, y^m \in (X_k^m, Y_k^m)} \mathbb{W}_p(x^m, y^m)$$

The above stated mini-batch Wasserstein estimate is not a valid distance metric, as $\mathbb{W}_p^m(X, X) \neq 0$, and it also produces biased gradients which may not lead to the optimal minimum. Furthermore, the quality of the mini-batch Wasserstein estimate is dependent on the batch size, which is practically limited by the $O(N^3 logN)$ computational cost of evaluating large mini-batches. Our approach is primarily designed to address this limitation by allowing for more accurate estimation of Wasserstein distances over small batch sizes.

Pyramid Match Kernels Pyramid Match Kernels (Grauman & Darrell, 2005) have been shown to be an effective method for quickly estimating Wasserstein-1 (or equivalently "Earth mover's") distance. Pyramid Match Kernels operate by embedding a point set X into a high-dimensional multi-resolution histogram Ψ_X, which is used for performing implicit couplings between different point sets.

In order to build the histogram, Ψ_X, the feature space of R^d is hierarchically divided into bins and the histogram reflects the number of points from X which map into each bin. The histogram contains L hierarchical layers, where each bin at layer l is sub-divided into k bins at layer $l + 1$. Thus the multi-resolution histogram $\Psi(\cdot)$ can be represented as the concatenation of L histograms at each layer $\Psi(\cdot) = [H_0, H_1, ..., H_{L-1}]$. As each bin in layers $l > 0$ is a subdivision of a bin in layer $l - 1$, bin relationships between layers can be modeled as parent-child relationships. Pyramid Match Kernels rely on the parent-child relationship between bins and the fact that the maximum distance between points in parent bin will be greater than the maximum distance between points in a child bin to rapidly compute a Wasserstein distance approximation. The number of matches present in a particular bin is the minimum of the number of points present in this bin in either distribution and the number of new matches present in a bin is this number minus the number of matches present in all of the bins children. In this setting each bin has a fixed cost with parents having a greater cost than their children and the total cost can be computed in $O(nL)$ time.

Algorithm 1: PMBatchExtract

Data: $\Psi(A), \Psi(B), n$
Result: batchA, batchB, n
batchA $= \emptyset$, batchB $= \emptyset$;
for $\psi(A), \psi(B) \in (\Psi(A)$.children, $\Psi(B)$.children*)* **do**
  $b_a, b_b, n \leftarrow$ `PMBatchExtract`$(\psi(A), \psi(B), n)$;
  batchA.extend(b_a), batchB.extend(b_b) ;
  if $n == 0$ **then**
    return batchA, batchB, n
  end
end
if $n \neq 0$ **then**
  $t = \min(n, \Psi(A)$.cnt, $\Psi(B)$.cnt);
  $b_a = \Psi(A)$.indices.pop(t);
  $b_b = \Psi(B)$.indices.pop(t);
  batchA.extend(b_a), batchB.extend(b_b), $n = n - t$;
end

The fast approximations provided by Pyramid Match Kernels and related works (Indyk & Thaper, 2003; Grauman & Darrell, 2005; Lazebnik et al., 2006; Grauman & Darrell, 2006; Backurs et al., 2020) have been shown to be useful in a variety of settings including image retrieval and object classification. However, these approaches do not readily lend themselves to serving as a useful loss function and as such they have not been explored for for learning alignments, which is of growing interest in the machine learning community.

3 Pyramid Mini-Batching

A major limitation of learning alignments with mini-batch optimal transport approaches is that the quality of learned alignment is strongly influenced by the ability of a mini-batch of size n to fully capture the *meaningful* diversity of the sampled distributions. While *meaningful* diversity is best defined in relation to a downstream task, a reasonable example might be capturing all the classes present in a distribution. Optimal transport estimations which rely on random mini-batches are bound to produce sub-optimal couplings and the subsequent quality of the learned alignments is limited by the relationship of the diversity of the dataset and the size of the mini-batches. Our approach aims to overcome this limitation by ensuring that batch x^m captures the diversity of batch y^m and vice versa.

3.1 Approximate Couplings for Minibatch Sampling

Following this intuition, we present Pyramid Mini-Batching (PMB) as a means of building batches better suited for estimating and learning with Wasserstein distances. Using the tree structure from Pyramid Match Kernels, we can quickly construct paired batches from empirical distributions with large numbers of samples. Given n samples from two distributions that we wish to align, X, Y, we first encode each distribution according to our hierarchical pyramid structure Ψ_X, Ψ_Y. This can be done in $O(ndL)$ time and with $O(nL)$ memory, as we must store which index is mapped into each node.

Using Ψ_X and Ψ_Y, we can extract a batch of size n by a depth-first traversal of the tree until we have extracted n approximate pairings which share a local neighborhood. Starting at the root node, for each node we randomly select a non-empty child of the node. If no non-empty children nodes exist we up to n of the available approximate couplings up from the current node. If we have not yet filled a batch of size n, we move the siblings of the current node and finally parents of the current node. Couplings at a specific node can be attained by randomly sampling $\min(H_l^j(X), H_l^j(Y))$ indices from bin (i, j) in X and Y and adding these indices to the respective batches. Mini-batches can be constructed in this manner $O(nL)$ time. Algorithm 1 details how to construct a batch of size n from a Pyramid encoding of the two distributions $\Psi(A), \Psi(B)$ and Figure 2 illustrates this process. The Wasserstein distance of each batch (N) can be computed in $O(N^3 logN)$ with n/N batches required

Algorithm 2: PMBuildIntNode

Data: A', B', k
Result: Ψ
if $|A'| == 0$ **then**
  return $\emptyset$
children $= \emptyset$;
centers $=$ `GetClusters`(A', k);
assignments$_A$ $=$ `AssignClusters`$(A',$ centers$)$;
assignments$_B$ $=$ `AssignClusters`$(B',$ centers$)$;
for $i \in 0 \rightarrow k$ **do**
  $\psi =$ `PMBuildIntNode`$(B'[$assignments$_B[i]], A'[$assignments$_A[i]], k)$;
  `// We swap which distribution is used to build clusters`
  children.*insert*(ψ);
$\Psi =$ `node`(centers, children)

for a complete pass over the distributions the complexity of this step is $O(nN^2logN)$. We see that for most settings $O(nN^2logN) >> O(ndL)$ and our Pyramid mini-batching adds little to the overall run-time of our approach.

3.2 INTERLEAVING-PMB FOR DISJOINT DISTRIBUTIONS

If distributions X and Y are completely disjoint, existing pyramid matching binning schemes will offer no improvement over random batching. Figure 14 illustrates that for disjoint distributions the encoding tree may only contain matching points at the global/(most course) level. We introduce *Interleaving-PMB* as means of mitigating this issue. This approach generates two sets of sample points P_X and P_Y from X and Y respectively. At each level of the hierarchy, cluster centers are determined by the available points in either P_X or P_Y. Cluster assignments are then made for both point sets based on the determined centers, and the source of the cluster centers will alternate at the next level of the hierarchy. In the case where points from P_X and P_Y are mapped to different bin at the first layer, we derive no benefit from the tree structure because approximate couplings will not be available until we reach the root (global) node. We treat this case as a degenerate case because our Pyramid Mini-Batching would yield the same results as random batching. Our interleaving strategy ensures, that higher order nodes will contain smaller sets for potential matches by clustering based on either P_X or P_Y at that level and mapping points from both distributions to these clusters. An alternative approach is to simply concatenate the point sets P_X and P_Y into P_C, and perform hierarchical clustering over the new point set. We refer to this approach as *Joint-PMB*.

4 EXPERIMENTS

We conduct experiments to develop a better understanding of the effectiveness of Pyramid Mini-Batching approaches on solving discrete optimal transport alignment problems. PMB is designed to be effective in transport settings involving a large number of samples and high-dimensional domains. As such we evaluate our approach and various baselines in settings with varying levels of dimensionality and dataset sizes. For the largest of these problems, we are unable to solve the exact optimal transport problem and must rely on proxies such as visualization to evaluate how well our approximations are behaving. The datasets we use in our experiments are as follows:

- **DOTMark** (Schrieber et al., 2016): Discrete Optimal Transport benchmark of 1-channel images of various size. Unless otherwise stated our results are calculated over two images of the Classic Images subset of size 64×64.
- **Synth**: We create synthetic datasets by sampling from Guassian mixture models where we vary the number of centers to create more challenging distribution matches. Centers are selected uniformly at random from within the range $[-10, 10] \in \mathbb{R}^d$, with $d = 100$ unless otherwise stated.

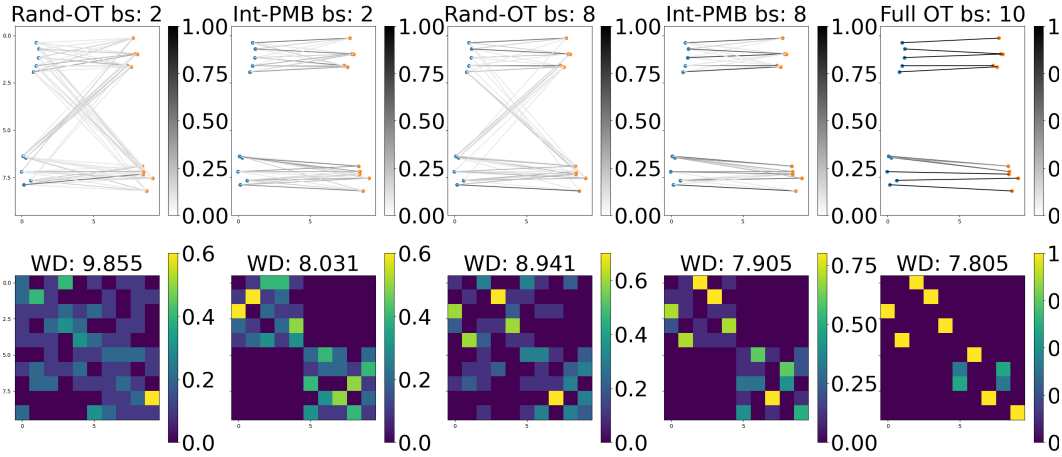

Figure 3: We display the resulting couplings and Optimal Transport maps from 10 full runs of Int-PMB-OT and Rand-OT for batch sizes 2, 8, and 10 over a 2D dataset with 10 points. For smaller batch sizes Int-PMB-OT does a much better job at approximating Wasserstein distance and learning OT maps that adhere to the underlying cluster structure. Both the accuracy of the estimate and sparsity of the learned maps indicate that the smoothing caused by mini-batches is greatly mitigated by PMB approaches.

- **ImageNet**: We align images from ImageNet to ImageNet-C. The latter provides several realistic synthetic perturbations to the ImageNet validation dataset which often present challenges for models trained without exposure to these perturbations.

In order to better understand the behavior of our Wasserstein approximations during optimization we use the non-parametric formulation of gradient flows as described in Feydy et al. (2019). We choose non-parametric gradient flows for these experiments because they represent the best-case of what function can be learned to minimize the associated losses; in later experiments we show the impact of our approach when learning parameters of a neural network. For a given target distribution B, gradient flows are used to model the distribution $A(t)$ which follows gradient descent at each iteration to minimize the loss $A(t) \mapsto L(A(t), B)$. Starting from the initial set at time $t = 0$, we simply integrate the ODE

$$\dot{A}(t) = -N\nabla_A L(A(t))$$

with a Euler scheme and evaluate the evolution of A(t) at time $t = 1, 5, 50$.

4.1 Estimation

PMB in 2D Similar to Fatras et al. (2019), we illustrate the OT matrix between two empirical distributions of 10 samples each in 2D in Figure 3. Our 2D distributions have cluster structure and the ordering of our samples is based on their cluster assignment. We see how the cluster structure is maintained by the PMB sampling and the overall error rate is considerably lower, indicating that with smaller batch sizes PMB is able to better model the optimal transport plan than with random mini-batching.

ImageNet to ImageNet-C We seek to align the features extracted from the penultimate layer of a Resnet-50, however in this instance we are unable to directly solve the optimal transport problem and thus rely of the predictive performance of the aligned distributions to assess how effective our method is at this task. In this setting, we observe the how well the optimal couplings at each batch size correspond to examples from the correct label space. $X, Y \in \mathbb{R}^{50k,1024}$. Results are shown in Table 2. We see that batches PMB approaches contain on between significantly more accurate label matches than random mini-batches. These results are most pronounced in the small batch regime, as when batch size is 10 our approach produce 30x more label matches.

Batch Sizes			
Algorithm	10	100	1000
Rand	25639.1	18636.8	17690.5
INT-PMB-OT	19076.4	17987.6	14654.8
JOINT-PMB-OT	19210.4	17969.4	14616.7

Table 1: Across all batch sizes we see that our approach provides better estimates of Wasserstein-2 distance on the Classic Images from the DOTMark dataset. As mini-batch approaches are an upper-bound to Wasserstein distance, our 20-30% reduction and estimate represents a tighter approximation. Subsequent experiments illustrate how the benefit of better estimates are carried downstream into alignments.

Batch Sizes							
Algorithm	10	50	200	500	1000	2000	5000
Source-Acc	0.68						
Target-Acc	0.53						
Rand-OT	0.006	0.02	0.06	0.11	0.16	0.22	0.32
INT-PMB-OT	0.18	0.29	0.40	0.48	0.54	0.60	0.69
JOINT-PMB-OT	0.21	0.31	0.40	0.47	0.52	0.59	0.66

Table 2: Pyramid Mini-Batching dramatically increases the number of validly classified couplings. This table shows what percentage of assigned couplings between the penultimate layer of a pretrained ResNet-50 on ImageNet-Val and its derived ImageNet-C-Glass-Blur contain the same ground truth labels for various batch sizes. Couplings between non-matching labels could have negative impacts on downstream tasks and while increasing batch size helps to mitigate this, it comes with extra computational costs. Our PMB methods of batch size 10 achieve similar correct classification percentages as random batches of size 2000. For batch sizes of 2000, a classifier based on our coupling approaches from the penultimate layer would out perform a pretrained model.

4.2 Alignment

DOTMark Extending the results from the DOTMark estimation task on Classic Images of size 64, we also learn alignments between the samples in this benchmark. Using a batch size of 100, we are able to reduce the Wasserstein distance between two images by 88% using random mini-batches. Int-PMB and Joint-PMB are both able to reduce the original Wasserstein distance by 99% or offer 10x reduction in comparison with random mini-batches. These drastic improvements in one-dimensional alignments are further exemplified when we examine high dimensional spaces.

High Dimensional Alignment Following the high-dimensional setting described in the previous section, we evaluate the ability of the various approaches to minimize Wasserstein distance by computing gradient flows over various time scales and reporting the residual Wasserstein distance between X and $\hat{Y}$. We see that our approach drastically outperforms random mini-batching approaches on this task. By examining the results at various timesteps, we gather a sense of how the approaches compare on a first pass $t = 1$ as well as how the approaches compare near the limits of their performance at $t = 50$. Gains in performance of our approach magnify as the model is allow to continue optimizing with results representing a $5x$ improvement in performance visible in Table 3

5 Related Work

We refer readers to Peyré et al. (2019) for a detailed overview of various approaches used for large-scale optimal transport. Here we focus our discussion on closely related approaches and alternatives commonly used for alignment. The first methods of interest are mini-batch estimates that compute exact optimal transport distance over a small subset of the data distribution, as used in (Courty et al., 2017; Bhushan Damodaran et al., 2018; Fatras et al., 2019; Sommerfeld et al., 2019). This approach has the benefit of being computable quickly for small N samples and the accuracy is independent of feature dimensionality. Additionally, not including the entire distributions in the matches can also be viewed as a form of regularization.

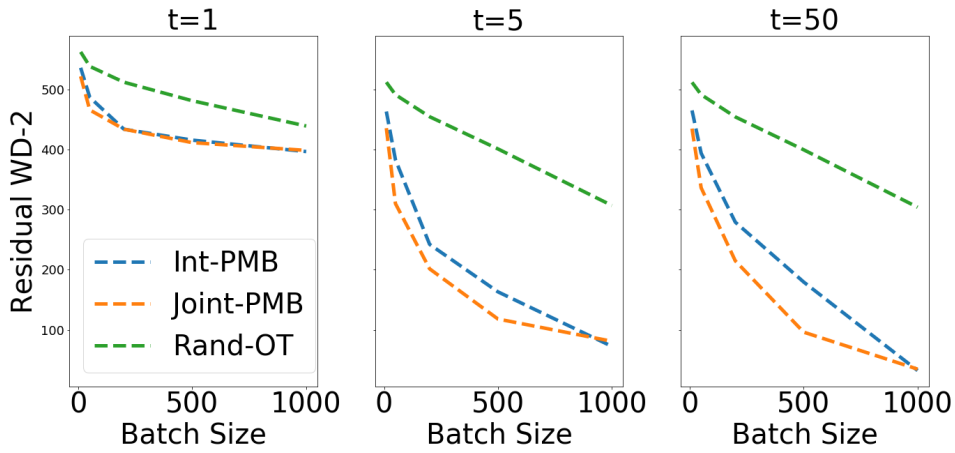

Figure 4: Pyramid Mini-Batching approaches allow us to substantially ($> \%70$) improve the alignments found via gradient flows. As we align distributions by intergrating over the gradient flows, we evaluate how well the aligned distribution has minimized Wasserstein distance at different steps, $t = 1, 5, 50$. We see that across all batch sizes and all stages of optimization our approach significantly reduces the residual Wasserstein distance and better aligns the two distributions. These experiments are carried out over our synthetic dataset where each distribution has 1000 cluster centers, 100-dimensional features, and 10k data points.

-	10 Centers			100 Centers			1k Centers		
Algorithm (t=50)	n=10	n=100	n=1000	n=10	n=100	n=1000	n=10	n=100	n=1000
WD-2	79.29			74.07			70.81		
Rand	35.81	21.62	16.24	50.36	34.60	20.875	53.64	48.24	31.86
INT-PMB	19.36	14.13	4.50	22.18	12.71	4.27	37.65	18.63	8.82
JOINT-PMB	15.96	10.58	2.79	18.54	9.16	3.57	32.86	15.29	4.98

Table 3: Across datasets with different numbers of modalities and various batch sizes, PMB approaches improve learned alignments between (2x-5x) when compared to a random mini-batch baseline. By aligning distributions according based on gradient flows at $t = 50$, we can get a better sense of the limitations an approach will be able to fully align discrete distributions. We see that the number of modes present in a dataset has an impact on how well Wasserstein-2 distance can be minimized and that PMB approaches significant increase the ability to minimize this OT distances. PMB batches of size 10 perform on par with random batches of size 1000 across all of these experiments on synthetic datasets.

However, this approach may suffer if batches are not large enough to sufficiently capture the diversity of the distributions. While batch size can be increased to account for this, eventually time and space limitations will hamper this approach.

Sinkhorn Distance (Cuturi, 2013) adds entropic regularization to the optimal transport distance, resulting in a formula that can be solved much more efficiently. However, its algorithmic complexity is still $O(n^2 \log n)$, and it adds a hyperparameter for the entropic regularization term which must be tuned for the problem at hand. The Kantorovich-Rubinstein duality of the optimal transport problem states that the maximum of the set of 1-Lipschitz continuous functions separating the two distributions is equivalent to the solution of the optimal transport problem. However searching the set of all 1-Lipschitz continuous functions is also intractable. In practice, this is estimated by clipping or normalizing the gradients of distribution discriminator (Arjovsky et al., 2017; Gulrajani et al., 2017).

Sliced Wasserstein distances (Bonneel et al., 2015) are also able to estimate Wasserstein distance in linear time with regards to sample size, making it an interesting competitor to our approach. This approach randomly projects multidimensional points down into 1-d spaces where Wasserstein distance can be computed in linear time. With sufficient random projections, this is an unbiased

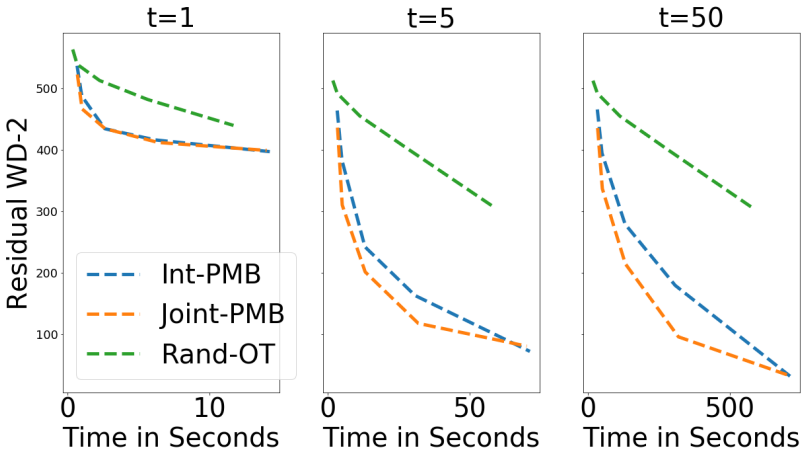

Figure 5: This figure illustrates the marginal impact of the added computational time of constructing mini-batches to the quality of the learned alignment. Using the same experimental setting as Figure 4, we show that the dramatic improvement in alignment quality caused by our approach add little to the overall run-time. We measure the residual Wasserstein distance after $t = 1, 5, 50$ steps on a synthetic dataset where each distribution has 1000 cluster centers, 100-dimensional features, and 10k data points.

estimate of optimal transport distance. However, it does not leverage the natural distribution of points in these high dimensional spaces, and may produce less accurate estimates of optimal transport cost than our approach. As illustrated in our experiments, our approach is able to more reliably minimize the Wasserstein metric.

The effectiveness of optimal transport in aligning distributions is shown in domain adaptation (Bhushan Damodaran et al., 2018; Lee et al., 2019) and generative modeling (Arjovsky et al., 2017; Gulrajani et al., 2017). We have empirically shown that our new approach can show comparable or superior results with respect to the scalability of sample size and input dimension. DeepJDOT (Bhushan Damodaran et al., 2018) and WGAN (Arjovsky et al., 2017) minimize mapping functions over the Wassertein objective like the experiments in this paper. Both of these approaches produced significant improvements over contemporary approaches, and we show how this approximation can equal or improve upon results produced by these methods, with the potential to scale better to on larger datasets/sample sizes.

As discussed above, a variety of multiresolution histogram intersection schemes have been proposed (Indyk & Thaper, 2003; Grauman & Darrell, 2005; Lazebnik et al., 2006) which inspired our work. While visualization and assessment of worst-case (per-bin) coupling schemes were reported in Grauman & Darrell (2006), it was not used for improving the approximate distance, and was not used in a differentiable scheme for the purpose of aligning distributions.

6 Conclusion

By providing a fast mechanism to sample geometrically consistent batches between two distributions, Pyramid Mini-Batching presents a compelling approach to learn distributional alignments. Pyramid Mini-Batching produces improved estimates of Wasserstein distances and drastically improves upon the quality of alignments. While this approach is presented in an Optimal Transport context, this approach has potential applications in a variety of alignment settings. We present one mechanism for using locality-sensitive tree structures to build compatible mini-batches, but several alternative approaches could be developed; we hope than the promising results presented here can inspire more work on sample selection for alignment and optimal transport problems. Future directions include incorporating the above mentioned approaches into generative modeling and domain adaptation pipelines.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Arturs Backurs, Yihe Dong, Piotr Indyk, Ilya Razenshteyn, and Tal Wagner. Scalable nearest neighbor search for optimal transport. In *International Conference on Machine Learning*, pp. 497–506. PMLR, 2020.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. 2018.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. 2013.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: asymptotic and gradient properties. *arXiv preprint arXiv:1910.04091*, 2019.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690. PMLR, 2019.

Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. 2005.

Kristen Grauman and Trevor Darrell. Approximate correspondences in high dimensions. 2006.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. 2017.

Piotr Indyk and Nitin Thaper. Fast image retrieval via embeddings. In *3rd international workshop on statistical and computational theories of vision*, volume 2, pp. 5, 2003.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. 2006.

Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. 2019.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Jörn Schrieber, Dominic Schuhmacher, and Carsten Gottschlich. Dotmark–a benchmark for discrete optimal transport. *IEEE Access*, 5:271–282, 2016.

Max Sommerfeld, Jörn Schrieber, Yoav Zemel, and Axel Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105):1–23, 2019.