Differentially Private Optimal Transport: Application to Domain Adaptation

Nam LeTien[1] , **Amaury Habrard**[1] and **Marc Sebban**[1]
[1]Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d'Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France
{tien.le, amaury.habrard, marc.sebban}@univ-st-etienne.fr

Abstract

Optimal transport has received much attention during the past few years to deal with domain adaptation tasks. The goal is to transfer knowledge from a source domain to a target domain by finding a transportation of minimal cost moving the source distribution to the target one. In this paper, we address the challenging task of privacy preserving domain adaptation by optimal transport. Using the Johnson-Lindenstrauss transform together with some noise, we present the first differentially private optimal transport model and show how it can be directly applied on both unsupervised and semi-supervised domain adaptation scenarios. Our theoretically grounded method allows the optimization of the transportation plan and the Wasserstein distance between the two distributions while protecting the data of both domains. We perform an extensive series of experiments on various benchmarks (VisDA, Office-Home and Office-Caltech datasets) that demonstrates the efficiency of our method compared to non-private strategies.

1 Introduction

Optimal Transport (OT) [Villani, 2008] is a geometric theory that allows us to put a distance (e.g. Wasserstein distance, earth mover distance) on the space of probability measures. As probability measures occur in many scenarios in machine learning, OT has received a lot of attention during the past few years in this community [Cuturi, 2013; Frogner *et al.*, 2015; Arjovsky *et al.*, 2017]. For instance, probability measures take the form of *color histograms* when comparing images, *bags of words* in document analysis, *activation maps* in brain imaging, etc. Perhaps, the most exciting recent application of OT in machine learning comes from the wide use of generative models and the need of some notion of distance to measure the divergence between generated data and the actual generative model. Very recently, OT has been shown to be a useful and intuitive tool to address domain adaptation tasks [Courty *et al.*, 2017b; Courty *et al.*, 2017a; Shen *et al.*, 2018], where the goal is to transfer knowledge from a source domain to a target domain. In such a scenario, regularized OT aims at finding a transportation of minimal cost aligning labeled source data on unlabeled target examples. Then a classifier can be learned from the source examples and deployed on the target domain. OT for domain adaptation has also received recent attention from a theoretical perspective. In [Redko *et al.*, 2017], the authors show how the Wasserstein distance can be used to derive generalization bounds on the target error. All things considered, OT can be seen nowadays as a theoretically grounded competitive framework to perform transfer learning.

Optimizing the transportation that aligns the source and target distributions boils down to finding a coupling matrix that moves the source data to the target ones and allows us to induce the Wasserstein distance. Therefore, this procedure requires to share data from both sources at some point of the learning process that may appear to be an unacceptable solution when source and/or target data contain sensitive information (e.g. about the medical history of patients). The contributions of this paper lie in the setting of differentially private OT and its application to Domain Adaptation. Differential privacy [Dwork *et al.*, 2006] provides a strong theoretical-grounded guarantee for the privacy of individuals against attackers and has become the standard for formal privacy in machine learning. The basic idea of differential privacy is to introduce randomness in the communication that preserves privacy even against an adversary possessing arbitrary side information and having access to the communication. It turns out that there has been a rich amount of work on differentially private machine learning, such as in logistic regression [Chaudhuri and Monteleoni, 2009], principal component analysis [Hardt and Roth, 2013], boosting [Dwork *et al.*, 2010], support vector machines [Rubinstein *et al.*, 2009], or more recently on deep learning [Abadi *et al.*, 2016; Shokri and Shmatikov, 2015] and semi-supervised deep learning [Papernot *et al.*, 2017]. However, it is worth noting that privacy preserving domain adaptation is surprisingly under-developed. Recently, [Wang *et al.*, 2018] and [Guo *et al.*, 2018] proposed two similar methods, combining hypothesis transfer learning with private logistic regression. However, the first model needs to have access to a publicly available auxiliary dataset as a bridge to transfer knowledge from the source to the target, while the second trains on a fully labeled target data; both are strong constraints that do not hold in the standard (both unsupervised and semi-supervised) domain adaptation setting.

The objective of this paper is to fill the gap mentioned above and address the challenging privacy preserving problem when OT is used to transfer knowledge from a source to a target. To do so, we present two new algorithms. The first one, called *Differentially Private Optimal Transport* (DPOT), makes use of the Johnson-Lindenstrauss transform [Johnson and Lindenstrauss, 1984] which linearly projects a set of examples onto a small feature space by a random matrix that preserves the pairwise distances. Adapted to the context of OT, DPOT allows us to jointly compute the coupling matrix and the Wasserstein distance between two domains under the differential privacy constraint, by adding noise on the random projection. As far as we know, DPOT is the first differentially private optimal transport algorithm. Experimental results show that DPOT preserves the Wasserstein distance on the space of probability measures compared to a non private OT approach. We then build upon DPOT the first complete *differentially private domain adaptation* model (DPDA) where the learner can benefit from labeled source data to improve a different but related target task while still ensuring the privacy of each source of data. We provide theoretical guaranty of our method and demonstrate its efficiency empirically on various benchmarks.

The rest of this paper is organized as follows: Section 2 presents the required background in Optimal Transport, Domain Adaptation and Differential Privacy. In Section 3, based on the Johnson-Lindenstrauss transform, we introduce DPOT. Section 4 is devoted to the presentation of DPDA, the first Differentially Private Domain Adaptation algorithm based on Optimal Transport for which differential privacy guarantees are derived. In Section 5, we first show that DPOT allows us to efficiently approximate the true Wasserstein distance that would be obtained without privacy constraint. Then, we report the results of an extensive experimental study performed on various benchmarks (VisDA, Office-Home and Office-Caltech datasets) that demonstrate the efficiency of DPDA compared to non-private strategies.

2 Background and Related Work

2.1 Optimal Transport

Let μ_1 and μ_2 be two probability measures. Given some cost function c, optimal transport (OT) seeks a transportation plan of minimal cost that moves μ_1 to μ_2. Let Π be the space of joint probability distributions with marginals μ_1 and μ_2. In the relaxed formulation of Kantorovitch [1942], the optimal transportation takes the form of a distribution (or *coupling*) $\gamma_0 \in \Pi$ that minimizes the following quantity:

$$W(\mu_1, \mu_2) = \min_{\gamma \in \Pi} \int_{\Omega_s \times \Omega_t} c(x^s x^t) d\gamma(x^s, x^t),$$

where $c(x^s, x^t)$ is this paper the Euclidean distance between $x^s \in \Omega_s$ and $x^t \in \Omega_t$. The quantity $W(\mu_1, \mu_2)$ is called the *Wasserstein distance* between the two distributions μ_1 and μ_2. When μ_1, μ_2 are defined as empirical measures between two datasets X_s and X_t, then

$$\gamma_0 = \underset{\gamma \in \Pi}{\operatorname{argmin}} \langle \gamma, C \rangle_F, \tag{1}$$

where $\langle ., . \rangle_F$ is the Frobenius dot product and C is the cost matrix between X_s and X_t. Then, the Wasserstein distance between X_s and X_t is $W(X_s, X_t) = \langle \gamma_0, C \rangle_F$.

2.2 Domain Adaptation

Domain adaptation [Pan and Yang, 2010] aims at using available labeled data from a *source domain* (*e.g.* synthetic data) to facilitate the learning process in a *target domain* (*e.g.* real data) with a different underlying distribution. In *unsupervised* domain adaptation, we assume to have access to only unlabeled target data, while in *semi-supervised* tasks, a small amount of supervision is available, but not large enough to learn well only from the labeled target data. Most recent state-of-the-art methods either learn a common latent feature space between the two domains, *e.g.* DANN [Ganin *et al.*, 2016], or map the samples of one domain to the other, *e.g.* ADDA [Tzeng *et al.*, 2017].

Optimal transport was recently applied successfully to domain adaptation, where the Wasserstein distance is used as a divergence measure between the two domains that a domain adaptation algorithm aims at minimizing. In [Courty *et al.*, 2017b], the source samples are transported to the target domain using the coupling matrix γ_0. In [Courty *et al.*, 2017a], the authors optimize both the coupling matrix and a target prediction function, while [Shen *et al.*, 2018] minimize the domain divergence in the latent space using the Wasserstein-GAN [Arjovsky *et al.*, 2017].

2.3 Differential Privacy

Differential Privacy was introduced by Dwork *et. al.* [2006] and constitutes nowadays a standard for privacy guarantees.

Definition 1 (Differential Privacy, see eg. [Abadi *et al.*, 2016])**.** *A randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathbb{R}^d$ satisfies* (ε, δ)-differential privacy *if for any two datasets $X, X' \in \mathcal{X}^n$ differing by a single element and for any set of possible output $\mathcal{O} \subseteq Range(\mathcal{M})$:*

$$\mathbb{P}(\mathcal{M}(X) \in \mathcal{O}) \leq e^{\varepsilon} \mathbb{P}(\mathcal{M}(X') \in \mathcal{O}) + \delta.$$

There has been a rich amount of work on differentially private machine learning, such as in logistic regression [Chaudhuri and Monteleoni, 2009], principal component analysis [Hardt and Roth, 2013], boosting [Dwork *et al.*, 2010] and support vector machine [Rubinstein *et al.*, 2009]. Private deep learning methods were also introduced recently; [Abadi *et al.*, 2016] provided a differentially private SGD method for training a deep neural network, where at each step, the gradient is clipped and some noise is added to protect privacy. On the other hand, [Shokri and Shmatikov, 2015] proposed a distributed selective SGD: several models are trained independently using SGD and they selectively upload the gradients to a global one. Similarly, [Papernot *et al.*, 2017] train several teachers with private data, then predict pseudo-labels via a majority vote for unlabeled public data that are used by a student to train.

It is worth noticing that privacy preserving domain adaptation has not received much attention. Deep learning methods such as [Abadi *et al.*, 2016; Papernot *et al.*, 2017] are

designed for standard classification tasks with the assumption that all data are drawn from the same underlying distribution, which is not the case in domain adaption (although technically, these algorithms can be deployed out of the box for domain adaptation through training a source-only model) Private multiparty learning methods, such as [Pathak *et al.*, 2010; Hamm *et al.*, 2016] could also be applied to semi-supervised domain adaptation, but they face two drawbacks: (1) the amount of target labeled data is usually very limited for a meaningful aggregation, and (2) these models assume that the domain distributions are identical and so ignore the domain shift. As already mentioned, [Wang *et al.*, 2018] and [Guo *et al.*, 2018] proposed two methods that combine hypothesis transfer learning with private logistic regression. However, the first model needs to have access to a publicly available auxiliary dataset as a bridge to transfer knowledge from the source to the target, while the second trains on a fully labeled target data; both are strong constraints that do not hold in the standard (both unsupervised and semi-supervised) domain adaptation setting.

2.4 Johnson-Lindenstrauss Transform

Let $\mathcal{N}(0,\sigma)^{k\times\ell}$ be a $k\times\ell$ matrix where each entry is drawn i.i.d. from $\mathcal{N}(0,\sigma)$. A variant of the famous Johnson-Lindenstrauss Lemma [Johnson and Lindenstrauss, 1984] states that a linear transformation of a set of n data points to a much smaller subspace by such a random matrix has a high probability to retain the pairwise distances within $(1\pm\eta)$-factor for some $\eta\in[0,0.5]$ (see eg. [Blocki *et al.*, 2012]).

Theorem 1. Let $v_1,...,v_n$ be a set of n points in $\mathbb{R}^k$. For any $\eta,\ell>0$ and an $\mathcal{N}(0,\frac{1}{\ell})^{k\times\ell}$ matrix M, with probability at least $1-\frac{2}{\exp(\ell\eta^2/8)}$ the following holds for every i,j:

$$1-\eta\leq\frac{\left\|v_i^TM-v_j^TM\right\|_2^2}{\|v_i-v_j\|_2^2}\leq 1+\eta.$$

The Johnson-Lindenstrauss transform is widely used across many areas of computer science such as computational speedups, machine learning, information retrieval. The interested reader may find an extensive overview of the topic in the monograph written by Vempala [2005]. Due to its randomness, the Johnson-Lindenstrauss transform was also used in differential privacy. [Blocki *et al.*, 2012] showed that it preserves the differential privacy, as long as all singular values of the database are sufficiently large. [Kenthapadi *et al.*, 2013] showed that by simply applying the Johnson-Lindenstrauss transform together with some noise, one can publish the pairwise distances of elements while protecting the privacy of the original data.

As OT resorts to a pairwise distance matrix to find the optimal coupling and the Wasserstein distance, we suggest in the next sections to benefit from the Johnson-Lindenstrauss transform and the idea introduced in [Kenthapadi *et al.*, 2013] to build the first differentially private OT method (DPOT); then, we use it to design a new privacy-preserving domain adaptation algorithm (DPDA).

Algorithm 1 Differentially Private Optimal Transport

Input: X_s, X_t, and $\sigma,\ell>0$

1: Source generates an $\mathcal{N}(0,\frac{1}{\ell})^{k\times\ell}$ matrix M and an $\mathcal{N}(0,\sigma)^{k\times\ell}$ noise matrix Δ.
2: Source sends $\{M,\tilde{X}_s+\Delta\}$, where $\tilde{X}_s=X_sM$
3: Target computes $\tilde{C}=c(\tilde{X}_s+\Delta,\tilde{X}_t)-\ell\sigma^2$ where $\tilde{X}_t=X_tM$
4: Solve Problem (1) with cost matrix $\tilde{C}$ and return $\tilde{\gamma_0}$ and $\tilde{W}(X_s,X_t)$.

3 Differentially Private Optimal Transport

Let X_s and X_t be a source and a target dataset of size $n_s\times k$ and $n_t\times k$ (where k is the number of features) respectively. A standard OT procedure aims at finding the Wasserstein distance W and the empirical transportation plan γ_0 between X_s and X_t.

Usually, one has access to both X_s and X_t to calculate the cost matrix C, typically as the matrix of pairwise Euclidean distances $c(X_s,X_t)$ between samples of the source and target sets. Then we solve the optimization Problem (1) to get γ_0 and W. However, under privacy constraints, source and target parties may not be willing to release their data, making optimal transport a challenging task. To overcome this problem, let us use the result proved in Kenthapadi *et al.* [2013] stating that by applying the Johnson-Lindenstrauss transform with some additional Gaussian noise, one can publish the pairwise distances of elements privately (see Kenthapadi *et al.* [2013] for more details).

We exploit this idea in the context of OT where the objective is to align X_s on X_t. Therefore, the data we want to publish safely here is the source set X_s. Algorithm 1 describes the pseudo-code of our Differentially Private Optimal Transport (DPOT) algorithm. The main steps of DPOT are the following: In Step 1, we generate a random matrix M according to [Johnson and Lindenstrauss, 1984] and a noise matrix Δ according to [Kenthapadi *et al.*, 2013]. Step 2 boils down to publishing both M and the source data $\tilde{X}_s+\Delta$ after a random projection of X_s by M in a subspace of size ℓ and the addition of the noise Δ. To benefit from the privacy guarantees of [Kenthapadi *et al.*, 2013] over the pairwise (source/target) distances, one compute the distance matrix $\tilde{C}$ in Step 3 between the published source data $\tilde{X}_s+\Delta$ and the projected target examples $\tilde{X}_t=X_tM$. Note that $\ell\sigma^2$ is subtracted from each entry to cancel the bias caused by Δ (indeed, σ^2 is known by the target party but not Δ). Step 4 is devoted to the resolution of Problem (1).

Kenthapadi *et al.* [2013], Theorem 1, showed that the mechanism of publishing M and the noisy data $\tilde{X}_s+\Delta$ satisfies the (ϵ,δ)-differential privacy of Definition 1, for any $\varepsilon,\delta>0$ and $\sigma\geq w\frac{\sqrt{2(\ln(\frac{1}{2\delta})+\varepsilon)}}{\varepsilon}$ where $w=\max_{1\leq i\leq k}\left(\sum_{j=1}^{\ell}M_{ij}^2\right)^{\frac{1}{2}}$, that is the ℓ_2-norm sensitivity of M, which is tightly concentrated around 1 [Kenthapadi *et al.*, 2013]. Therefore, Algorithm 1 comes directly with the following privacy guarantee.

Theorem 2. *Algorithm 1 is (ε,δ)-differentially private for any $\varepsilon,\delta>0$ and $\sigma\geq w\frac{\sqrt{2(\ln(\frac{1}{2\delta})+\varepsilon)}}{\varepsilon}$.*

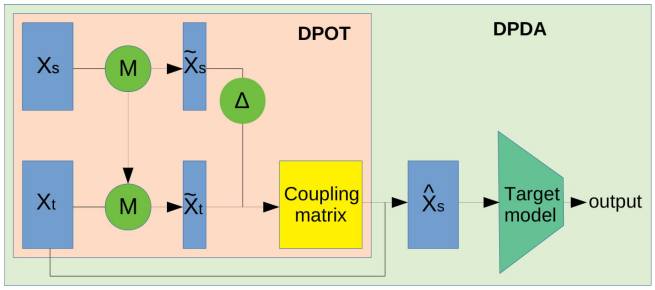

Figure 1: The architecture of DPOT and DPDA.

In Section 5, we will report experimental results showing that by satisfying the previous constraint on the noise σ, the Wasserstein distance $\tilde{W}(X_s, X_t)$ computed by Algorithm 1 remains very close to the true $W(X_s, X_t)$, while preserving the privacy of the data. Therefore, we can now benefit from $\tilde{W}(X_s, X_t)$ and $\tilde{\gamma}_0$ to transfer knowledge from a source domain to a target one to deal with domain adaptation in a differential privacy setting.

4 Differentially Private Domain Adaptation

Extending DPOT for domain adaptation requires (i) to transport the source data to the target domain thanks to a private barycentric mapping, (ii) to transfer the source labels while ensuring privacy, (iii) to add some appropriate domain adaptation regularizers for computing the coupling matrix.

Barycentric mapping. Once the coupling matrix $\tilde{\gamma}_0$ is computed, the source samples can be moved to the target domain using the geodesics of the Wasserstein metric. The *barycentric mapping* [Reich, 2013] from each source sample $x_i^s \in X_s$ to its corresponding image $\widehat{x}_i^s$ in the target domain is as follows

$$\widehat{x}_i^s = \underset{x \in \mathbb{R}^k}{\operatorname{argmin}} \sum_j \tilde{\gamma}_0[i,j] c(x, x_j^t). \tag{2}$$

The *barycentric image* $\widehat{x}_i^s$ can be understood as the empirical optimal location on the target domain to transport x_i^s w.r.t. the coupling matrix $\tilde{\gamma}_0$. Let $\widehat{X}_s$ be the set of barycentric images of the source data. It is now possible to train a model from the transported labeled source data $(\widehat{X}_s, Y_s)$ and use it in the target domain [Courty *et al.*, 2017b].

Private barycentric mapping. Generally in domain adaptation, the target agent must have access to the labeled data from the source. However, under privacy constraints, this is not possible anymore. Fortunately, by combining DPOT algorithm with the barycentric mapping, we can privately transport the source data in the target domain. Indeed, as shown in Equation 2, only $\tilde{\gamma}_0$ and X_t are involved in the calculation of the barycentric mapping which does not depend on the source data. The private barycentric images can be conveniently rewritten as $\widehat{X}_s = n_s \tilde{\gamma}_0 X_t$ as shown in [Perrot *et al.*, 2016] in a non private scenario.

Transmitting source labels. To perform domain adaptation in both unsupervised and semi-supervised settings, the target party needs to have access to the source labels Y_s. To enhance the privacy of our domain adaptation algorithm, we also aim at adding some noise on the labels. To do so, we follow the principle of *Histogram Queries* mentioned in [Dwork *et al.*, 2014]: The source first reorders its data (X_s, Y_s) such that all samples of label 1 appear first, then samples of label 2 and so on. Hence Y_s is now equivalent to a vector $v(Y_s)$ of length q counting the number of samples for each label, where q is the number of labels. The source party can now safely publish a noisy version of $v(Y_s)$ as a histogram query by adding a Laplacian noise $Lap(\frac{1}{\varepsilon'})^q$, which is $(\varepsilon', 0)$-differentially private [Dwork *et al.*, 2014].

Regularization techniques. There are various techniques in the literature to regularize the coupling matrix for domain adaptation tasks. In our method, we utilize two of them. The first one is the *entropic regularization* [Cuturi, 2013] that allows us to transform Problem 1 into a strictly convex problem and is defined as follows:

$$R_e(\gamma) = -\sum_{i,j} \gamma_{ij}(\log \gamma_{ij} - 1).$$

The second one is the *group Lasso regularization* [Courty *et al.*, 2017b],

$$R_g(\gamma) = \sum_j \sum_s \|\gamma(I_s, j)\|_2,$$

where $\|\cdot\|_2$ is the ℓ_2-norm, I_s contains the indices of rows in γ related to source samples associated to label s, $\gamma(I_s, j)$ is then a vector containing the coefficients of the j^{th} column of γ associated to label s. This regularization encourages a coupling where a given target sample receives masses from source samples having the same labels. Thus, the coupling matrix is now computed by the following equation:

$$\tilde{\gamma}_0 = \underset{\gamma \in \Pi}{\operatorname{argmin}} \langle \gamma, \tilde{C} \rangle_F + \lambda_e R_e(\gamma) + \lambda_g R_g(\gamma) \tag{3}$$

where λ_s, λ_c are hyper-parameters. In the semi-supervised setting, we can further exploit the available target label data by prohibiting masses from being transported between samples with different labels, which does not require any additional hyper-parameter [Courty *et al.*, 2017b].

Algorithm 2 Differentially Private Domain Adaptation

Input: (X_s, Y_s), X_t and $\sigma, \varepsilon', \ell > 0$

1: Source reorders (X_s, Y_s) and sends noisy $v(Y_s) + Lap(\frac{1}{\varepsilon'})^q$
2: Source runs Steps 1,2 of Algorithm 1 on the ordered X_s
3: Target follows Steps 3,4 of Algorithm 1 solving Problem (3) and returns $\tilde{\gamma}_0$ and $\tilde{W}$
4: Target computes the barycentric images $\widehat{X}_s = n_s \tilde{\gamma}_0 X_t$
5: Target trains his own model using $(\widehat{X}_s, \tilde{Y}_s)$, where $\tilde{Y}_s = v(Y_s) + Lap(\frac{1}{\varepsilon'})^q$, together with labeled data in X_t (if any)

Differentially Private Domain Adaptation Algorithm. We now have all the necessary blocks for our differentially private domain adaptation algorithm DPDA, see the pseudo-code in Algorithm 2, where the two parties collaboratively compute a regularized coupling matrix using Eq. (3), which

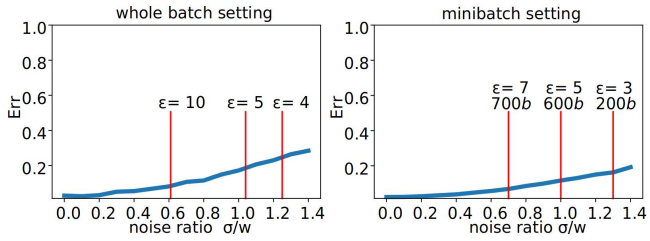

Figure 2: Evolution between noise-ratio and the accuracy of DPOT. On the right, 200b means "200 minibatches".

allows the target to obtain the barycentric images of the source data for his own training. We have shown that both mechanisms of transmitting $\{M, \tilde{X}_s + \Delta\}$ and the noisy vector $v(Y_s)$ from the source are differentially private. According to the Composition Theorem ([Dwork *et al.*, 2014] Theorem 3.16), the composition of differentially private mechanisms is also differentially private. This means that Algorithm 2 is also differentially private.

Theorem 3. *Algorithm 2 is* $(\varepsilon + \varepsilon', \delta)$*-differentially private for any* $\varepsilon, \varepsilon', \delta > 0$ *and* $\sigma \geq w\frac{\sqrt{2(\ln(\frac{1}{2\delta})+\varepsilon)}}{\varepsilon}$.

The whole workflow of our differentially private method is shown is Figure 2.

5 Experiments

5.1 Private Optimal Transport with DPOT

In this section, we empirically validate our private optimal transport method DPOT by analyzing the relation between the privacy budget (ε, δ) and the accuracy of the Wasserstein distance $\tilde{W}$ computed by DPOT. The error between $\tilde{W}$ and the true Wasserstein distance W is measured here by $Err = \left|\frac{\tilde{W}-W}{W}\right|$. We perform an experiment on the Office-Home dataset [Venkateswara *et al.*, 2017], using two domains with a moderate number of samples, Clipart (4365 samples) and Product (4439 samples). We follow the experimental protocol used in [Abadi *et al.*, 2016] and [Papernot *et al.*, 2017] by setting the privacy budget to $\delta = \frac{1}{1.2n_s}$, where n_s is the size of the source domain, and using a some moderate ε with 1-digit.

Besides, it is widely known that minibatch subsampling can enhance privacy [Abadi *et al.*, 2016; Balle *et al.*, 2018]. Recently, [Abadi *et al.*, 2016] introduced the *privacy accountant* theorem to get a tight bound on the total *privacy budget* (ε, δ) (i.e. the allowed leakage amount of privacy during the whole training process) when training with multibatches. In the following, we will perform experiments on two settings: *whole batch* and *minibatch* to demonstrate the effect of using minibatches on the privacy budget.

Whole batch setting. In the whole batch setting, the *noise-ratio* $\frac{\sigma}{w}$ can be calculated directly from the privacy budget (ε, δ) by Theorem 2. For example, when $\varepsilon = 4, 5, 10$, we get $\frac{\sigma}{w} = 1.25, 1.04, 0.61$, respectively. In Figure 2 (left), we plot the evolution of the error Err between $\tilde{W}$ and W given a noise-ratio $\frac{\sigma}{w}$ on the two whole datasets (we calculate Err over 20 runs and report the average). Keeping in mind that we want ε as small as possible, the figure shows that $Err = 0.08$ at $\varepsilon = 10$ (maximum acceptable value according to the literature), while when the privacy budget is reduced to $\varepsilon = 4$, the error reaches 22%. In order to reduce the approximation error, the next experiment shows that using minibatches allows us to drastically reduce the privacy budget.

Minibatch setting. In the minibatch setting, the relationship between the noise-ratio $\frac{\sigma}{w}$ and the privacy budget (ε, δ) can be calculated via the privacy accountant [Abadi *et al.*, 2016]. For $\delta = \frac{1}{1.2n_s}$ and a minibatch of size 128, the noise-ratio $\frac{\sigma}{w}$ depends on both ε and the number of minibatches we wish to run. For example, if $\frac{\sigma}{w} = 1$ then for $\varepsilon = 3, 5, 7$ we can run $200, 600, 1200$ minibatches respectively, while when $\frac{\sigma}{w} = 0.7$, we can run only $4, 60, 200$ batches for $\varepsilon = 3, 5, 7$. We can note in Figure 2 (right) that the error Err of DPOT (average over 1000 random minibatches), for $\frac{\sigma}{w} = 0.7, 1, 1.3$, is $7\%, 10\%$ and 15%, respectively. This is substantially lower than running on the whole batch. Empirically, we found that the effect of the noise-ratio $\frac{\sigma}{w}$ on the accuracy of DPOT in the minibatch setting is the same across datasets, but the privacy budget (or the number of allowed minibatches for a given budget) would be different since the size of the dataset plays an important role. For example, as the Synthetic domain of VisDA dataset [Peng *et al.*, 2017] has 150k samples, we can run 10k batches for $\varepsilon = 1$, $\frac{\sigma}{w} = 1$ or 30k batches for $\varepsilon = 2$, $\frac{\sigma}{w} = 0.8$.

5.2 Private Domain Adaptation with DPDA

The previous experimental study showed that minibatches allow us to improve the quality of the outputs of DPOT with a smaller privacy budget. For this reason, the next experiments are performed only in this setting.

Benchmarks. We evaluate our method on three domain adaptation benchmarks from the classical **Office-Caltech** dataset [Saenko *et al.*, 2010] to the more recent and challenging **VisDA** [Peng *et al.*, 2017] and **Office-Home** [Venkateswara *et al.*, 2017] datasets. The three datasets offer a wide range of scenarios: Office-Caltech is a very small dataset with 150-1000 samples per domain; VisDA contains a large collection (200k samples in total) of high quality images, while Office-Home has a moderate size (4000 samples/domain) but many (65) classes. In Office-Caltech and Office-Home, there are four different domains each, coming from different sources, while VisDA is a specific dataset for simulation-to-real adaptation, where the source domain contains 150k synthetic images rendering of 3D models from different angles and the target domain contains 50k samples of natural images. We conduct the experiments on both *unsupervised* and *semi-supervised* settings.

Baselines. Our baselines include the state-of-the-art private semi-supervised method **PATE** [Papernot *et al.*, 2017], where the teachers hold source data and the student holds target data, the non-private optimal transport domain adapation method **OTDA** [Courty *et al.*, 2017b] which we build our model DPDA upon, and another state-of-the-art non-private domain adaptation method **ADDA** [Tzeng *et al.*, 2017] using a domain-adversarial technique. We do not add [Wang *et al.*, 2018] and [Guo *et al.*, 2018] as baselines since both only

	PATE	DPDA	OTDA	ADDA
A→ C	81.4	87.6	88.3	86.9
A→ D	88.5	91.0	94.2	91.7
A→ W	77.9	96.4	95.5	96.8
C→ A	91.0	91.9	92.5	92.1
C→ D	85.9	92.0	92.5	91.7
C→ W	82.0	93.9	95.3	95.5
D→ A	67.2	89.1	92.5	89.8
D→ C	58.5	79.0	87.0	85.8
D→ W	81.3	96.1	98.7	96.9
W→ A	79.2	93.1	93.2	93.0
W→ C	70.9	86.0	87.6	87.6
W→ D	98.7	98.0	98.7	99.2
Average	80.2	91.2	93.0	92.3

Table 1: Performance (accuracy %) on Office-Caltech dataset in the unsupervised setting.

	PATE	DPDA	OTDA	ADDA
plane	87.7	88.5	89.2	96.2
bicycle	31.3	66.2	64.8	71.4
bus	76.4	75.8	75.7	76.2
car	69.6	59.1	58.6	44.3
horse	86.9	86.5	87.0	65.8
knife	53.4	70.4	62.1	83.0
motorcycle	80.8	69.9	74.6	87.7
person	60.1	71.6	70.8	44.8
plant	64.7	75.9	76.3	81.8
skateboard	22.9	49.3	50.5	68.7
train	55.4	86.8	87.7	91.4
truck	7.0	39.5	40.9	40.4
All	60.6	68.8	69.1	67.3

Table 2: Classwise performance (accuracy %) of VisDA dataset in the unsupervised setting.

work on binary-classification and the first requires an auxiliary public dataset while the second requires fully labeled target data.

For the Office-Caltech dataset, we use the common DeCAF features as input, while for VisDA and Office-Home, we use the NASNet model [Zoph *et al.*, 2018] with weights pre-trained on ImageNet as a base model to extract features from the images. All methods are written in Keras [Chollet, 2015] with the same target model architecture (a 3-layer neural network) for fair comparison. The coupling matrices are computed using the POT library [Flamary and Courty, 2017]. As already said, we train all the models using minibatches.

Hyper-parameters. For OTDA and our method DPDA, we set the hyper-parameters λ_e and λ_g of Eq. (3) to 0.01 and 0.1, respectively. In all benchmarks, we set the dimension of the subspace of our method $\ell = \frac{k}{10}$ and the noise-ratio $\frac{\sigma}{w} = 1.1$. For the privacy budget, we again follow the standard of [Abadi *et al.*, 2016; Papernot *et al.*, 2017] by setting $\delta = \frac{1}{1.2n_s}$, $\varepsilon = 2$ for VisDA and $\varepsilon = 8$ for the other datasets, except $\varepsilon = 20$ if the source is DSLR or Webcam in Office-Caltech since they have too few samples (150-200 in total). For PATE and DPDA, we use the privacy accountant tool [Abadi *et al.*, 2016] to keep track the privacy budget after each step as explained in Section 5.1. We run each test 3 times and report the average.

	PATE	DPDA	OTDA	ADDA	t-only	DPDA	OTDA
AC	37.4	39.2	44.2	41.3	23.3	43.4	46.5
AP	52.1	54.4	58.5	56.5	56.8	64.2	65.2
AR	59.7	61.4	67.2	67.1	51.3	67.7	68.6
CA	48.7	50.3	55.3	49.6	41.9	55.8	57.2
CP	54.2	57.8	61.7	61.2	56.8	67.6	69.1
CR	56.7	59.9	64.0	63.2	51.3	63.5	66.4
PA	51.5	53.5	53.9	50.4	41.9	57.6	58.0
PC	36.8	40.7	42.9	45.4	23.3	45.8	46.1
PR	65.8	67.9	68.9	69.5	51.3	70.8	71.3
RA	57.6	59.4	61.6	58.8	41.9	63.1	63.6
RC	39.4	43.8	46.2	49.3	23.3	47.9	49.1
RP	66.2	68.7	70.1	70.9	56.8	72.8	73.5
Avr	52.1	54.8	57.9	56.9	43.3	60.0	61.2

Table 3: Performance (accuracy %) on Office-Home dataset. Left: unsupervised, and right: semi-supervised.

Results. The results are presented in Tables 1, 2 and 3. In both Office-Caltech and VisDA benchmarks, our model DPDA performs at the same level as non-private state-of-the-art methods and even gets 1 point higher than ADDA in the VisDA benchmark, while PATE significantly lags behind. If we reduce the privacy budget for the source domains DSLR, Webcam in the Office-Caltech benchmark to $\varepsilon = 10$, then the performances of DPDA and PATE drop by 3-5 points. In the Office-Home dataset, DPDA is 2-3 points lower than non-private baselines but still safely outperforms PATE by 3 points. On the other hand, in the semi-supervised scenario on the Office-Home benchmark , when only 1 labeled target sample per class is allowed, DPDA manages to reduce the gap with OTDA significantly from 3 points (55 vs 58) to 1 point (60 vs 61). This means that DPDA benefits more from some labeled target data than OTDA. Behind the good behavior of DPDA to deal with domain adaptation tasks, the results reported in these tables confirm that DPOT is a very performing differentially private optimal transport method.

6 Conclusion

In this paper, we have proposed the first differentially private approach for optimal transport. The proposed algorithm DPOT is able to preserve the Wasserstein distance despite the noisy perturbations introduced used to ensure differential privacy. We have then designed the first complete differentially private domain adaptation model based upon DPOT. Our methods are justified by strong theoretical guarantees and the experimental evaluations illustrated that our approach allows one to obtain good results while ensuring a high privacy level of the data. Future work includes extensions of our framework to other applications of optimal transport, for instance, in differentially private distributed learning.

Acknowledgements

Work supported by the ACADEMICS grant of the IDEX-LYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005

References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS*, pages 308–318. ACM, 2016.

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv:1701.07875*, 2017.

[Balle *et al.*, 2018] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *NIPS*, pages 6280–6290, 2018.

[Blocki *et al.*, 2012] Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The johnson-lindenstrauss transform itself preserves differential privacy. In *FOCS*, pages 410–419. IEEE, 2012.

[Chaudhuri and Monteleoni, 2009] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In *NIPS*, pages 289–296, 2009.

[Chollet, 2015] François Chollet. Keras. https://keras.io, 2015.

[Courty *et al.*, 2017a] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *NIPS*, pages 3730–3739, 2017.

[Courty *et al.*, 2017b] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE TPAMI*, 39(9):1853–1865, 2017.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, pages 2292–2300, 2013.

[Dwork *et al.*, 2006] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284. Springer, 2006.

[Dwork *et al.*, 2010] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *FOCS*, pages 51–60. IEEE, 2010.

[Dwork *et al.*, 2014] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[Flamary and Courty, 2017] R'emi Flamary and Nicolas Courty. Pot python optimal transport library, 2017.

[Frogner *et al.*, 2015] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In *NIPS*, pages 2053–2061, 2015.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016.

[Guo *et al.*, 2018] Xiawei Guo, Quanming Yao, WeiWei Tu, Yuqiang Chen, Wenyuan Dai, and Qiang Yang. Privacy-preserving transfer learning for knowledge sharing. *arXiv preprint arXiv:1811.09491*, 2018.

[Hamm *et al.*, 2016] Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In *ICML*, pages 555–563, 2016.

[Hardt and Roth, 2013] Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *STOC*, pages 331–340. ACM, 2013.

[Johnson and Lindenstrauss, 1984] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[Kantorovich, 1942] Leonid Vitalievich Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

[Kenthapadi *et al.*, 2013] Krishnaram Kenthapadi, Aleksandra Korolova, Ilya Mironov, and Nina Mishra. Privacy via the johnson-lindenstrauss transform. *Journal of Privacy and Confidentiality*, 5(1), 2013.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[Papernot *et al.*, 2017] Nicolas Papernot, Martín , Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*, 2017.

[Pathak *et al.*, 2010] Manas Pathak, Shantanu Rane, and Bhiksha Raj. Multiparty differential privacy via aggregation of locally trained classifiers. In *NIPS*, pages 1876–1884, 2010.

[Peng *et al.*, 2017] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv:1710.06924*, 2017.

[Perrot *et al.*, 2016] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *NIPS*, pages 4197–4205, 2016.

[Redko *et al.*, 2017] Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *ECML-PKDD*, pages 737–753. Springer, 2017.

[Reich, 2013] Sebastian Reich. A nonparametric ensemble transform method for bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

[Rubinstein *et al.*, 2009] Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *arXiv preprint arXiv:0911.5708*, 2009.

[Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226. Springer, 2010.

[Shen *et al.*, 2018] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *AAAI*, 2018.

[Shokri and Shmatikov, 2015] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *CCS*, pages 1310–1321. ACM, 2015.

[Tzeng *et al.*, 2017] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017.

[Vempala, 2005] Santosh S Vempala. *The random projection method*, volume 65. American Mathematical Soc., 2005.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017.

[Villani, 2008] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[Wang *et al.*, 2018] Yang Wang, Quanquan Gu, and Donald Brown. Differentially private hypothesis transfer learning. In *ECML-PKDD*, pages 811–826. Springer, 2018.

[Zoph *et al.*, 2018] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *CVPR*, pages 8697–8710, 2018.