Differentially Private Correlation Alignment for Domain Adaptation

Kaizhong Jin , **Xiang Cheng*** , **Jiaxi Yang** and **Kaiyuan Shen**
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China
{kzjin91,chengxiang,yangjiaxi,shenkaiyuan}@bupt.edu.cn

Abstract

Domain adaptation solves a learning problem in a target domain by utilizing the training data in a different but related source domain. As a simple and efficient method for domain adaptation, correlation alignment transforms the distribution of the source domain by utilizing the covariance matrix of the target domain, such that a model trained on the transformed source data can be applied to the target data. However, when source and target domains come from different institutes, exchanging information between the two domains might pose a potential privacy risk. In this paper, for the first time, we propose a differentially private correlation alignment approach for domain adaptation called PRIMA, which can provide privacy guarantees for both the source and target data. In PRIMA, to relieve the performance degradation caused by perturbing the covariance matrix in high dimensional setting, we present a random subspace ensemble based covariance estimation method which splits the feature spaces of source and target data into several low dimensional subspaces. Moreover, since perturbing the covariance matrix may destroy its positive semi-definiteness, we develop a shrinking based method for the recovery of positive semi-definiteness of the covariance matrix. Experimental results on standard benchmark datasets confirm the effectiveness of our approach.

1 Introduction

Supervised machine learning will encounter poor generalization performance with limited labeled data, while manual labeling of sufficient training data for emerging application domains is labor-intensive and time-consuming. This motivates the development of domain adaptation, a technique that aims to transfer a model from a source domain where sufficient training data are available to a target domain where few or no labeled data are available.

In many settings, the source and target domains do not want to share the raw data or statistics due to privacy concerns. For example, the source and target domains may come from two different medical institutions who hold medical records related to certain types of drugs or procedures. In order to perform domain adaptation without privacy leakage, we adopt the notion of differential privacy [Dwork *et al.*, 2006b], which provides strict and verifiable privacy guarantees, and makes almost no assumptions about the attacker's background knowledge. Recently, LeTien *et al.* [LeTien *et al.*, 2019] proposed a differentially private optimal transport model for privacy-preserving domain adaptation. However, optimal transport relies on two restrictive assumptions, which are: 1) the target samples are an affine transformation of the source samples; 2) the source and target domains have shared features.

In contrast to optimal transport which exploits only shared features, correlation alignment [Sun *et al.*, 2016] can use both shared and domain specific features by capturing the feature correlations of source and target domains, which is particularly crucial when the two domains contain few or even no shared features. What's more, compared with the recent deep adaptation approaches [Long *et al.*, 2015; Ganin *et al.*, 2016] which are quite complex and expensive, requiring tuning of many hyperparameters, correlation alignment is more efficient and can achieve better or comparable performance. Without requiring any target labels, correlation alignment first transforms the source distribution to as close as possible to the target one by utilizing the covariance matrix of the target domain. Then a model trained on the transformed source data is applied to the target data. Clearly, the covariance matrix of the target domain utilized by the source domain may leak sensitive information of the target data [Dwork *et al.*, 2014]. Moreover, the model applied by target domain may leak sensitive information of the source data [Abadi *et al.*, 2016].

In this paper, we start from a straightforward approach, which can provide privacy guarantee for correlation alignment based domain adaptation, and allow both the source and target domains to perform domain adaptation without trusting each other. Specifically, to obtain a private covariance matrix of the target domain, we perturb the real covariance matrix by Gaussian noise [Dwork *et al.*, 2014; Ge *et al.*, 2018]. In addition, to obtain a private model, we perturb gradients by Gaussian noise during the gradient based model training [Bassily *et al.*, 2014; Abadi *et al.*,

*Corresponding Author

2016]. However, such a straightforward approach has two limitations: 1) the upper error bound of the perturbed covariance matrix grows linearly with the dimension [Wang and Xu, 2019], which makes the straightforward approach suffer from poor performance in high-dimensional setting; 2) it is not clear whether the covariance matrix can keep positive semi-definiteness when it is perturbed, a property that is normally expected from a classifier. To address these limitations of the straightforward approach, we propose our differentially **PRI**vate correlation alignment approach for do**M**ain **A**daptation, which is referred to as PRIMA. In PRIMA, to reduce the large error caused by perturbing covariance matrix in high-dimensional setting, we present a random subspace ensemble based covariance estimation method which splits the feature spaces of source and target data into several low dimensional subspaces. In particular, we employ the forward selection strategy to determine the dimension of each subspace. In addition, to tackle the problem that the positive semi-definiteness of the covariance matrix is destroyed, we develop a shrinking based method for the positive semi-definite matrix recovery by computing the nearest positive semi-definite matrix to the perturbed covariance matrix.

Our contributions can be summarized as follows:

- We present a differentially private correlation alignment approach for domain adaptation, call PRIMA, to protect data privacy of both the source and target domains. To our knowledge, PRIMA is the first correlation alignment based approach for differentially private domain adaptation. We prove that PRIMA achieves (ϵ, δ)-differential privacy for the source and target data respectively.
- Based on random subspace ensemble, we propose a differentially private covariance estimation method for the high-dimensional setting, where the forward selection strategy is employed to determine subspace dimension.
- To recover the positive semi-definiteness of the covariance matrix, we develop a shrinking based method. Utility analysis shows that the recovered covariance matrix can provide better utility than the non-recovered one.
- We demonstrate the superiority of PRIMA through two concrete examples: logistic regression (LR) and deep neural network (DNN). Experimental results show that PRIMA outperforms the state-of-the-art methods.

2 Related Work

In the literature of differentially private domain adaptation, the most related work to ours is [LeTien *et al.*, 2019]. Using random projection, LeTien *et al.* [LeTien *et al.*, 2019] propose a differentially private domain adaptation approach based on optimal transport. However, optimal transport can only work if the target samples are an affine transformation of the source samples. In addition, optimal transport is sensitive to outliers of the source domain that have no correspondence in the target one, and can work only if there exist shared features in both domains. Wang *et al.* [Wang *et al.*, 2020] propose a differentially private deep domain adaptation approach which uses an adversarial-learning strategy to construct domain-invariant features for classifying the unlabeled target data. However, their approach supposes that the source and target data owners trust each other and is only applicable to deep models. Wang *et al.* [Wang *et al.*, 2018] propose a differentially private multiple-source hypothesis transfer learning approach. However, their approach needs to have access to a publicly available auxiliary dataset. Yao *et al.* [Yao *et al.*, 2019] first propose a privacy-preserving logistic regression approach by stacking, then combine the proposed approach with hypothesis transfer learning. However, their approach trains on a fully labeled target data, and cannot be applied to the scenario where no labeled data are available in the target domain.

3 Preliminaries

3.1 Domain Adaptation

Domain adaptation (DA) aims at adapting a model trained in a source domain for use in a target domain, where the source and target domains may be different but related. Let $X_s \in \mathbb{R}^d$ and $X_t \in \mathbb{R}^d$ be the feature spaces of source and target data respectively. We denote the source data as: $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{n_s}$, where $x_s^i \in X_s$, and y_s^i is the label of x_s^i. Similarly, we denote the target data as: $\mathcal{D}_t = \{(x_t^j)\}_{j=1}^{n_t}$, where $x_t^j \in X_t$. Let $\mathcal{P}(X_s)$ and $\mathcal{Q}(X_t)$ be the marginal distributions of X_s and X_t, respectively. The key assumption of domain adaptation is that $\mathcal{P}(X_s) \neq \mathcal{Q}(X_t)$, but $\mathcal{P}(Y_s|X_s) = \mathcal{Q}(Y_t|X_t)$.

As a simple yet efficient method for domain adaptation, correlation alignment (CORAL) aligns the marginal distributions of the source and target domains by aligning their covariance matrices. To minimize the distance between the covariance matrices of the source and target domains, correlation alignment applies a transformation matrix $\mathbf{H}$ to the original source feature space X_s and uses the Frobenius norm as the matrix distance metric:

$$\min_{\mathbf{H}} ||\mathbf{C}_{\vec{s}} - \mathbf{C}_t||_F^2 = \min_{\mathbf{H}} ||\mathbf{H}^\top \mathbf{C}_s \mathbf{H} - \mathbf{C}_t||_F^2, \qquad (1)$$

where $\mathbf{C}_s$ and $\mathbf{C}_t$ are the covariance matrices of X_s and X_t respectively, $\mathbf{C}_{\vec{s}}$ is the covariance matrix of the transformed source data $\{X_s\mathbf{H}, Y_s\}$. After CORAL transforms the source data, a model trained on the transformed source data can be directly applied to target domain.

3.2 Differential Privacy

Definition 1 (Differential privacy (DP) [Dwork *et al.*, 2006a; Dwork *et al.*, 2006b]). *A randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathbb{R}^d$ satisfies (ϵ, δ)-differential privacy if for any two datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$ differing by a single element and for any set of possible output $\mathcal{O} \subseteq Range(\mathcal{M})$:*

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{O}] \leq e^\epsilon \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{O}] + \delta, \qquad (2)$$

where ϵ is the privacy budget and δ is the failure probability.

Theorem 1 (Gaussian mechanism [Dwork *et al.*, 2006a]). *Given a function $f : \mathcal{X}^n \to \mathbb{R}^d$, the Gaussian Mechanism $\mathcal{M}_G(\mathcal{D}, f, \epsilon) = f(\mathcal{D}) + z_g$, where z_g is drawn from Gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$, satisfies (ϵ, δ)-DP for some $\delta > 0$, if $\sigma \geq \sqrt{2\log(1.25/\delta)}\Delta_2(f)/\epsilon$. Here $\Delta_2(f)$ is the ℓ_2-sensitivity of the function f, i.e. $\Delta_2(f) = sup_{\mathcal{D},\mathcal{D}'}||f(\mathcal{D}) - f(\mathcal{D}')||_2$.*

3.3 Random Subspace Ensemble

Random subspace ensemble (RSE) or simply Random Subspace (RS) [Skurichina and Duin, 2002] is an ensemble learning method which samples from the original feature space and builds one classifier on each subspace. Given training data $X = \{(x^i)\}_{i=1}^n$, whose rows represent the set of data samples and whose columns represent the set of features. Let $\mathcal{F}^{1:d} = \{\mathcal{F}^1, \mathcal{F}^2, ..., \mathcal{F}^d\}$ be the set of features, where $\mathcal{F}^i$ is the i-th feature of X. In RSE, one randomly selects p ($p \ll d$) features from $\mathcal{F}^{1:d}$, and obtains a p-dimensional subspace. Assume m is the ensemble size, the i-th subspace is $\hat{X}^i = (\mathcal{F}^{i,1}, \mathcal{F}^{i,2}, ..., \mathcal{F}^{i,p})^\top$ ($i \in [m]$), where $\mathcal{F}^{i,j}$ ($j \in [p]$) is randomly selected from $\mathcal{F}^{1:d}$. Then one constructs classifiers $\psi^i(\hat{X}^i)$ in the random subspaces $\hat{X}^i$ and combines m classifiers by simple majority voting.

4 Proposed Approach

4.1 Overview

Our approach PRIMA consists of three phases: RSE based covariance estimation, positive semi-definite matrix recovery and model training. In the first phase, target domain employs RSE to split its feature space into several low dimensional subspaces, and perturbs the covariance matrix of each subspace. In particular, to determine the dimension of each subspace, the target domain uses the forward selection strategy in generating subspaces. In the second phase, the target domain modifies each perturbed covariance matrix to make it positive semi-definite via shrinking, and sends the modified matrices to the source domain. In the last phase, the source domain first generates subspaces according to the splitting results of the target domain, and then performs correlation alignment to transform each subspace by exploiting the received covariance matrices. Finally, the source domain trains models on the transformed source subspaces, and perturbs gradients during the gradient-based training. Figure 1 illustrates an overview of our approach.

4.2 RSE based Covariance Estimation

To obtain a private estimation of the covariance matrix of the target domain $\mathbf{C}_t$, we add noise sampled from Gaussian distribution $\mathcal{N}(0, 2log(1.25/\delta)/\epsilon^2)$ to every element in the real covariance matrix $\mathbf{C}_t = \frac{1}{n_t} X_t X_t^\top$, where the sensitivity is at most one [Dwork *et al.*, 2014; Ge *et al.*, 2018; Jiang *et al.*, 2016; Wang and Xu, 2019]. As shown in [Dwork *et al.*, 2014; Amin *et al.*, 2019], the upper error bound of the perturbed covariance matrix is $d\sqrt{log(1/\delta)}/n_t\epsilon$, where d is the dimension of the target data. Here, the bound $d\sqrt{log(1/\delta)}/n_t\epsilon$ represents the Frobenius distance between the perturbed covariance matrix and the real covariance matrix, which is tight. We can see that the upper error bound is quite large in high dimensional setting.

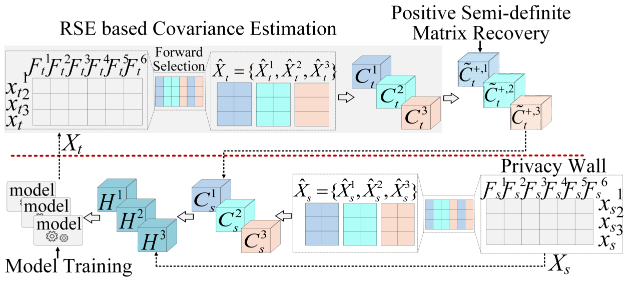

RSE based Covariance Estimation
Positive Semi-definite Matrix Recovery
Forward Selection
$\mathcal{F}_t^1 \mathcal{F}_t^2 \mathcal{F}_t^3 \mathcal{F}_t^4 \mathcal{F}_t^5 \mathcal{F}_t^6$
x_t^1 x_t^2 x_t^3 x_t
$\hat{X}_t = \{\hat{X}_t^1, \hat{X}_t^2, \hat{X}_t^3\}$
C_t^1 C_t^2 C_t^3
$\tilde{C}_t^{+,1}$ $\tilde{C}_t^{+,2}$ $\tilde{C}_t^{+,3}$
X_t
Privacy Wall
model
model
model
H^1 H^2 H^3
C_s^1 C_s^2 C_s^3
$\hat{X}_s = \{\hat{X}_s^1, \hat{X}_s^2, \hat{X}_s^3\}$
$\mathcal{F}_s^1 \mathcal{F}_s^2 \mathcal{F}_s^3 \mathcal{F}_s^4 \mathcal{F}_s^5 \mathcal{F}_s^6$
x_s^1 x_s^2 x_s^3 x_s
X_s
Model Training

Figure 1: An illustration of PRIMA.

To address the above limitation, we first employ RSE to split the original feature space X_t into m low dimensional subspaces. In particular, to avoid the allocation of privacy budget ϵ which results in much added noise, we leverage the composition property of differential privacy [McSherry, 2009] and split X_t into non-overlapping subspaces $\hat{X}_t^{1:m} = \{\hat{X}_t^1, \hat{X}_t^2, ..., \hat{X}_t^m\}$. We denote the dimension of $\hat{X}_t^i$ by p_t^i and have $\sum_{i=1}^m p_t^i = d$. Then, we add Gaussian noise to the corresponding covariance matrices $\hat{\mathbf{C}}_t^{1:m} = \{\hat{\mathbf{C}}_t^1, \hat{\mathbf{C}}_t^2, ..., \hat{\mathbf{C}}_t^m\}$ of $\hat{X}_t^{1:m}$. Finally, we obtain m perturbed covariance matrices $\tilde{\mathbf{C}}_t^{1:m} = \{\tilde{\mathbf{C}}_t^1, \tilde{\mathbf{C}}_t^2, ..., \tilde{\mathbf{C}}_t^m\}$.

Clearly, the choice of the dimension p_t^i of each subspace $\hat{X}_t^i$ ($i \in [m]$) is a serious dilemma: a small p_t^i leads to significant information loss of the subspace $\hat{X}_t^i$, while a large p_t^i leads to poor utility of the perturbed covariance matrix $\tilde{\mathbf{C}}_t^i$. To choose a proper dimension p_t^i of each subspace, we first define an criterion to measure the information loss of $\hat{X}_t^i$ and the utility of $\tilde{\mathbf{C}}_t^i$, then employ the forward selection strategy [Caruana and Freitag, 1994] to determine the dimension p_t^i of each subspace by minimizing such criterion.

The criterion contains two types of errors: reconstruction error and noise error. Specifically, we use the reconstruction error [Farahat *et al.*, 2013] to measure the information loss of the subspace $\hat{X}_t^i$, and use the noise error to measure the utility of the perturbed covariance matrix $\tilde{\mathbf{C}}_t^i$. The reconstruction error $R(\hat{X}_t^i)$ measures the sum of squared errors between the original data X_t and the reconstructed data based on $\hat{X}_t^i$. In particular, the reconstruction error is defined by: $R(\hat{X}_t^i) = ||X_t - \mathbf{P}_{\hat{X}_t^i} X_t||_F^2$, where $\mathbf{P}_{\hat{X}_t^i}$ is an $n_t \times n_t$ projection matrix that projects the columns of X_t onto the span of the subspace $\hat{X}_t^i$ of columns. The projection matrix $\mathbf{P}_{\hat{X}_t^i}$ can be calculated as: $\mathbf{P}_{\hat{X}_t^i} = \hat{X}_t^i((\hat{X}_t^i)^\top \hat{X}_t^i)^{-1}(\hat{X}_t^i)^\top$. The upper error bound of the perturbed covariance matrix $\tilde{\mathbf{C}}_t^i$ measures the utility of $\tilde{\mathbf{C}}_t^i$, which can be regarded as the noise error. In particular, the noise error can be defined by: $G(\hat{X}_t^i) = \frac{p_t^i\sqrt{log(1/\delta)}}{n_t\epsilon}$. As a result, the criterion is formulated as follows:

$$\arg\min \sum\nolimits_{i=1}^m ||X_t - \mathbf{P}_{\hat{X}_t^i} X_t||_F^2 + \frac{p_t^i\sqrt{log(1/\delta)}}{n_t\epsilon}$$
$$s.t. \sum\nolimits_{i=1}^m p_t^i = d \tag{3}$$
$$0 < p_t^i \le d, m \ge 1.$$

To minimize criterion (3), we adopt the forward selection strategy. Specifically, the forward selection strategy begins with an empty set $\mathcal{S}$, and adds features sampled from feature set $\mathcal{F}_t$ one by one; meanwhile, deletes the sampled features from $\mathcal{F}_t$. Once the errors (i.e., the sum of reconstruction and noise errors) of current $\mathcal{S}$ are larger than the previous one, the forward selection strategy stops the above process. The size of $\mathcal{S}$ is the dimension of the generated subspace $\mathcal{S}^{\top}$. To determine the dimensions of all the subspaces, we repeat the forward selection strategy until $\mathcal{F}_t$ is empty.

4.3 Positive Semi-definite Matrix Recovery

The positive semi-definiteness, which is a property of the covariance matrix $\mathbf{C}_t$, may be destroyed by the RSE based covariance estimation. Clearly, destroying the positive semi-definiteness will lead to poor performance of domain adaptation. To this end, we present a shrinking based method to recover the positive semi-definiteness of the perturbed covariance matrix $\tilde{\mathbf{C}}_t^i$ $(i \in [m])$.

Shrinking is widely used in statistical estimation [Ledoit and Wolf, 2004; Higham *et al.*, 2016]. Given an indefinite matrix $\mathbf{\Sigma}_0$, and a positive semi-definite (PSD) matrix $\mathbf{\Sigma}_1$, shrinking is to seek the elementwise minimal change to $\mathbf{\Sigma}_0$ in the direction $\mathbf{\Sigma}_1$ - $\mathbf{\Sigma}_0$ by forming a convex linear combination $\mathbf{\Omega}(\alpha) = \alpha\mathbf{\Sigma}_1 + (1-\alpha)\mathbf{\Sigma}_0$ of $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$, where $\alpha \in [0,1]$ is the shrinking parameter. Based on the idea of shrinking, our task is to seek the nearest PSD matrix to $\tilde{\mathbf{C}}_t^i$ by computing $\mathbf{\Omega}(\alpha^i) = \alpha^i\mathbf{\Sigma}_t^i + (1-\alpha^i)\tilde{\mathbf{C}}_t^i$, and ensure that the recovered covariance matrix $\mathbf{\Omega}(\alpha^i)$ provide better utility than the perturbed one $\tilde{\mathbf{C}}_t^i$.

To guarantee $\mathbf{\Omega}(\alpha^i)$ is PSD, the key is to choose an optimal shrinking parameter α_*^i to satisfy:

$$\alpha_*^i = min\{\alpha^i \in [0,1] : f(\alpha^i) \geq 0\}. \tag{4}$$

The function f is defined by $f(\alpha^i) = \lambda_{min}(\mathbf{\Omega}(\alpha^i))$, where λ_{min} denotes the smallest eigenvalue of a matrix. $f(\alpha^i) \geq 0$ is due to the fact that a matrix is PSD if and only if its smallest eigenvalue is non-negative. Since f is concave and continuous [Higham *et al.*, 2016], and $f(0) < 0, f(1) = \lambda_{min}(\mathbf{\Omega}(1)) = \lambda_{min}(\mathbf{\Sigma}_t^i)$, it follows that α_*^i is the unique zero of f in (0, 1) if the matrix $\mathbf{\Sigma}_t^i$ is PSD. Therefore, we use a bisection method to choose the optimal shrinking parameter, which is to find a zero of function f on a given interval $(0,1)$. Specifically, we first let $\alpha_\ell^i = 0$ and $\alpha_\gamma^i = 1$, then repeatedly bisect the interval $[\alpha_\ell^i, \alpha_\gamma^i]$ and select the subinterval in which the function f changes sign. The process is continued until the selected subinterval is sufficiently small. The final bisection point is the optimal shrinking parameter α_*^i. Therefore the nearest PSD matrix to $\tilde{\mathbf{C}}_t^i$ is $\mathbf{\Omega}(\alpha_*^i)$. For convenience, we denote $\mathbf{\Omega}(\alpha_*^i)$ as $\mathbf{C}_t^{i,+}$.

Moreover, to ensure that the utility of the recovered covariance matrix $\mathbf{\Omega}(\alpha^i)$ is better than $\tilde{\mathbf{C}}_t^i$, we should choose the PSD matrix $\mathbf{\Sigma}_t^i$ which satisfies $||\mathbf{\Sigma}_t^i||_F < ||\tilde{\mathbf{C}}_t^i||_F$. Since correlation alignment performs domain adaptation by minimizing the distance in the Frobenius norm between the covariance matrices of the source and target domains, the distance in the Frobenius norm between the recovered PSD matrix $\mathbf{C}_t^{i,+}$ and the real covariance matrix $\mathbf{C}_t^i$ determines the utility of $\mathbf{C}_t^{i,+}$. From Theorem 2, we can see $||\mathbf{\Sigma}_t^i||_F$ used in the procedure of the PSD matrix recovery guarantees the distance between $\mathbf{C}_t^{i,+}$ and $\mathbf{C}_t^i$ less than the distance between $\tilde{\mathbf{C}}_t^i$ and $\mathbf{C}_t^i$, which ensures the recovered covariance matrix $\mathbf{C}_t^{i,+}$ provides better utility than the perturbed matrix $\tilde{\mathbf{C}}_t^i$.

Theorem 2. *Let* $\mathbf{Z}$ *denote the noise matrix introduced by the RSE based Covariance Estimation. For any* $\alpha^i \in (0,1)$, $||\mathbf{C}_t^{i,+} - \mathbf{C}_t^i||_F^2 < ||\tilde{\mathbf{C}}_t^i - \mathbf{C}_t^i||_F^2$, *if* $\mathbf{\Sigma}_t^i$ *satisfies* $||\mathbf{\Sigma}_t^i||_F < ||\tilde{\mathbf{C}}_t^i||_F$.

Proof. If $||\mathbf{\Sigma}_t^i||_F < ||\tilde{\mathbf{C}}_t^i||_F$, we get

$$||\mathbf{\Sigma}_t^i||_F < ||\tilde{\mathbf{C}}_t^i||_F \leq ||\tilde{\mathbf{C}}_t^i||_F - \frac{1}{\alpha^i}||\mathbf{Z}||_F + \frac{1}{\alpha^i}||\mathbf{Z}||_F$$

$$\Leftrightarrow ||\mathbf{\Sigma}_t^i||_F < ||\tilde{\mathbf{C}}_t^i - \frac{\mathbf{Z}}{\alpha^i}||_F + \frac{1}{\alpha^i}||\mathbf{Z}||_F \tag{5}$$

$$\Leftrightarrow \alpha^i||\mathbf{\Sigma}_t^i||_F - \alpha^i||\tilde{\mathbf{C}}_t^i - \frac{\mathbf{Z}}{\alpha^i}||_F - ||\mathbf{Z}||_F < 0$$

$$\Leftrightarrow ||\alpha^i\mathbf{\Sigma}_t^i - \alpha^i\tilde{\mathbf{C}}_t^i + \mathbf{Z}||_F - ||\mathbf{Z}||_F < 0 \tag{6}$$

where (5) and (6) follow from the triangle inequality of the Frobenius norm.

It is clear that $||\mathbf{C}_t^{i,+} - \mathbf{C}_t^i||_F^2 - ||\tilde{\mathbf{C}}_t^i - \mathbf{C}_t^i||_F^2 = (||\mathbf{C}_t^{i,+} - \mathbf{C}_t^i||_F + ||\tilde{\mathbf{C}}_t^i - \mathbf{C}_t^i||_F)(||\mathbf{C}_t^{i,+} - \mathbf{C}_t^i||_F - ||\tilde{\mathbf{C}}_t^i - \mathbf{C}_t^i||_F)$, where $||\mathbf{C}_t^{i,+} - \mathbf{C}_t^i||_F + ||\tilde{\mathbf{C}}_t^i - \mathbf{C}_t^i||_F > 0$ follows from the definition of the Frobenius norm. Thus, to guarantee $||\mathbf{C}_t^{i,+} - \mathbf{C}_t^i||_F^2 < ||\tilde{\mathbf{C}}_t^i - \mathbf{C}_t^i||_F^2$, we need to ensure $||\mathbf{C}_t^{i,+} - \mathbf{C}_t^i||_F - ||\tilde{\mathbf{C}}_t^i - \mathbf{C}_t^i||_F < 0$. Since $\mathbf{C}_t^{i,+} = \alpha^i\mathbf{\Sigma}_t^i + (1-\alpha^i)\tilde{\mathbf{C}}_t^i$, we have

$$\begin{aligned}
&||\mathbf{C}_t^{i,+} - \mathbf{C}_t^i||_F - ||\tilde{\mathbf{C}}_t^i - \mathbf{C}_t^i||_F \\
&= ||\alpha^i\mathbf{\Sigma}_t^i + (1-\alpha^i)\tilde{\mathbf{C}}_t^i - \mathbf{C}_t^i||_F - ||\tilde{\mathbf{C}}_t^i - \mathbf{C}_t^i||_F \\
&= ||\alpha^i\mathbf{\Sigma}_t^i - \alpha^i\tilde{\mathbf{C}}_t^i + \mathbf{Z}||_F - ||\mathbf{Z}||_F
\end{aligned}$$

Based on (6), to ensure $||\mathbf{C}_t^{i,+} - \mathbf{C}_t^i||_F - ||\tilde{\mathbf{C}}_t^i - \mathbf{C}_t^i||_F < 0$, we require $||\mathbf{\Sigma}_t^i||_F < ||\tilde{\mathbf{C}}_t^i||_F$, which completes the proof. □

4.4 Model Training

After the PSD matrix recovery phase, the source domain first generates m subspaces $\hat{X}_s^{1:m} = \{\hat{X}_s^1, \hat{X}_s^2, ..., \hat{X}_s^m\}$ by exploiting the results of the forward selection strategy, and then performs correlation alignment on each subspace to get m transformed source subspaces $\{(\hat{X}_s^1\mathbf{H}^1, Y_s), (\hat{X}_s^2\mathbf{H}^2, Y_s), ..., (\hat{X}_s^m\mathbf{H}^m, Y_s)\}$. According to [Jiang *et al.*, 2013], the models trained on the transformed source subspaces may leak sensitive information of the source data. To obtain m private models trained on m transformed source subspaces, we use the differentially private mechanism proposed in [Abadi *et al.*, 2016] to perturb gradients using Gaussian noise during the stochastic gradient descent and use the moments accountant theorem [Abadi *et al.*, 2016] to get a tight bound on the total privacy budget. Specifically, at each iteration, one forms a batch b of $(\hat{X}_s^i\mathbf{H}^i, Y_s)$, and clips the gradient $\mathbf{g}$ in L_2 norm by dividing it by max$(1, \frac{||\mathbf{g}||_2}{c})$. This ensures that the L_2 sensitivity of $\mathbf{g}$ is bounded by c. Then one computes the perturbed gradient $\tilde{\mathbf{g}} = 1/|b|\sum_{x\in b}\mathbf{g} + N(0, \sigma^2c^2I_d)$ using the Gaussian mechanism with variance σ^2. Finally, we obtain m private models $\mathbf{W} = \{w^i\}_{i=1}^m$.

4.5 Privacy Analysis

Theorem 3. *PRIMA achieves (ϵ, δ)-DP for the source and target data respectively, for any $\epsilon, \delta > 0$ and $\sigma \geq \sqrt{2\log(1.25/\delta)}/\epsilon$.*

Proof. For the target domain, applying Gaussian mechanism to the covariance matrix $\hat{\mathbf{C}}_t^i$ $(i \in [m])$ can satisfy (ϵ, δ)-DP. Since the perturbed covariance matrices $\tilde{\mathbf{C}}_t^{1:m}$ are estimated on m disjoint subspaces, the parallel composition property of DP [McSherry, 2009] ensures that the RSE based covariance estimation phase satisfies (ϵ, δ)-DP. The positive semidefinite matrix recovery phase depends only on the perturbed covariance matrix $\tilde{\mathbf{C}}_t^i$, which also satisfies (ϵ, δ)-DP by the post-processing property of DP [Dwork *et al.*, 2006b]. In summary, PRIMA achieves (ϵ, δ)-DP for the target data.

For the source domain, when training each model, we perturb gradients in each iteration, and use the moments accountant theorem [Abadi *et al.*, 2016] to keep track of the privacy parameters ϵ, δ. According to [Abadi *et al.*, 2016] and Theorem 2, the process of training the model w^i satisfies (ϵ, δ)-DP. Similarly, the parallel composition property ensures that the model training phase satisfies (ϵ, δ)-DP. In summary, PRIMA achieves (ϵ, δ)-DP for the source data. □

5 Experiments

5.1 Benchmark Datasets

We evaluate our approach on two popular domain adaptation benchmark datasets. The first one is Office-Caltech10 dataset [Gong *et al.*, 2012], which contains 10 object categories from an office environment (e.g., keyboard, laptop, etc.) in 4 sources: Webcam (W), DSLR (D), Amazon (A), and Caltech256 (C) with 958, 295, 157 and 1,123 image samples respectively. We follow [LeTien *et al.*, 2019] to encode each source into 4096-dimensional feature vectors. Using each source as a domain, we get four domains leading to 12 domain adaptation tasks: A→D (train on A, test on D, the same below), A→C, and so on. The second one is Amazon review dataset [Blitzer *et al.*, 2006], which contains Amazon reviews on 4 domains, namely Book (BK), DVD (DV), Kitchen (KT) and Electronics (EL). There are 1000 positive and 1000 negative reviews on each domain. We follow [Wang *et al.*, 2020] to encode every review into a 5,000-dimensional feature vector by using the bag-of-word method, and perform domain adaptation under 4 pairs of source and target domains: KT → DV, DV → BK, BK → EL, EL → KT. For data preprocessing, we normalize each feature into the interval [0, 1].

5.2 Experimental Setup

We showcase the application of our approach in LR and DNN classifiers. For LR, since the Office-Caltech10 dataset is multiclass classification, we need to construct multiple binary models (one for each class), and split the privacy budget across sub-models. We use the simplest composition property [McSherry, 2009], and divide the privacy budget evenly. For DNN, we follow the standard of [LeTien *et al.*, 2019] that all methods are written with the same model architecture (a 3-layer neural network) for fair comparison.

There are 5 prime parameters in PRIMA. Among them, ϵ, δ, σ are privacy parameters, batch size b, clipping bound c are model training parameters. We follow the experimental protocol used in [Abadi *et al.*, 2016] by setting $\sigma = 4$, $\delta = 10^{-5}$, and compute the value of ϵ as a function of the training epochs E. We follow the experimental protocol of [Abadi *et al.*, 2016] again by setting c as the median of the unclipped gradients over the course of training. Empirically, batch size b is set to 25. For both datasets, our experiments are repeated for 20 times, and average accuracy is presented.

5.3 Experimental Results

Performance of Our Approach

In this part, we compare our approach **PRIMA** against two privacy-preserving domain adaptation approaches: **DPDA** [LeTien *et al.*, 2019], **GDPDA** [Wang *et al.*, 2020], as well as four non-private domain adaptation methods **CORAL** [Sun *et al.*, 2016], **SA** [Fernando *et al.*, 2013], **OTDA** [Courty *et al.*, 2016] and **DANN** [Ganin *et al.*, 2016]. The private approaches DPDA and GDPDA are implemented based on the non-private methods OTDA and DANN respectively. We do not compare with DPDA-Target which is also proposed by [Wang *et al.*, 2020], as it can only protect the privacy of the target data. Besides, since the covariance matrix used in PRIMA is widely used as a fundamental ingredient for subspace based domain adaptation [Fernando *et al.*, 2013; Cui *et al.*, 2014], PRIMA can be extended to subspace based domain adaptation. Therefore, we show the results on one PRIMA variant, which has the same settings as PRIMA, except that it replaces CORAL by SA [Fernando *et al.*, 2013]. We denote such variant as **PRISA**.

Tables 1, 2, 3 and 4 show results on both datasets with $\epsilon = 2$. Figure 2 gives the results of varying privacy budget ϵ on pair C → A. As can be seen, PRIMA and PRISA achieve better performance than the competitors DPDA and GDPDA in almost all cases. Specifically, when fixing $\epsilon = 2$, the accuracy of PRIMA and PRISA drop around 1% over non-private methods while DPDA and GDPDA drop by $3\% - 5\%$ over non-private methods. This is because, compared with DPDA and GDPDA, PRIMA and PRISA can guarantee the source and target domains have much closer distributions by improving the utility of the perturbed covariance matrix. From Figure 2, we can observe that PRIMA and PRISA achieve competitive accuracies on a wide range of values for ϵ. By contrast, DPDA and GDPDA show an unstable behavior. For example, when ϵ varies from 4 to 1, the accuracy drops from 0.90 to 0.79 for the LR classifier of DPDA.

Ablation Study

In order to analyze the effects of the RSE based covariance estimation and PSD matrix recovery methods, the following methods are compared: 1) **Basic**, which first perturbs the covariance matrix of the target domain, and then perturbs the gradients during model training; 2) **RSE+Basic**, which uses RSE to split the source and target feature spaces into several subspaces respectively, then applies the Basic method to each subspace; 3) **RSE+Basic+FS**, which applies the forward selection strategy to **RSE+Basic**; 4) **Basic+Shrinking**, which applies the shrinking based PSD matrix recovery method to

Method	A → C	A → D	A → W	C → A	C → D	C → W	D → A	D→ C	D → W	W → A	W → C	W → D	AVG
OTDA	85.4	85.1	83.2	92.3	87.5	84.3	92.1	84.2	95.9	90.5	81.7	96.3	88.2
SA	85.1	85.9	81.8	93.2	87.9	85.1	93.4	86.5	97.6	88.1	80.4	98.7	88.6
CORAL	86.2	87.2	83.5	94.7	90.4	87.8	94.7	88.3	98.3	89.6	83.8	99.8	90.4
DPDA	80.7	81.3	78.1	87.4	84.0	81.2	88.0	81.7	92.1	87.3	77.8	93.2	84.4
PRISA	83.7	84.3	80.4	91.7	85.3	84.3	91.8	84.9	96.1	87.2	78.9	97.9	87.2
PRIMA	85.6	85.4	82.7	92.8	87.6	87.0	93.2	86.1	97.6	88.3	81.9	98.7	88.9

Table 1: Accuracy (%) on Office-Caltech10 dataset for LR

Method	A → C	A → D	A → W	C → A	C → D	C → W	D → A	D→ C	D → W	W → A	W → C	W → D	AVG
DANN	87.9	82.5	77.8	93.3	91.2	89.6	84.7	82.1	98.9	82.9	81.3	99.8	87.7
OTDA	88.3	93.1	94.9	92.5	92.5	94.1	92.5	87.0	98.7	93.2	87.6	98.7	92.7
SA	88.5	92.9	93.4	92.9	93.8	93.7	91.1	88.2	98.1	93.5	89.2	99.8	92.9
CORAL	89.2	94.1	94.3	93.0	94.5	93.2	92.1	90.7	98.9	94.4	89.6	99.8	93.7
GDPDA	82.5	79.1	74.2	90.2	87.9	86.7	81.8	78.9	95.3	78.2	77.2	96.1	84.0
DPDA	84.1	90.3	91.8	89.3	88.7	91.8	87.6	79.1	93.9	90.0	83.5	94.2	88.6
PRISA	86.9	91.8	92.1	91.5	92.6	91.9	89.7	86.8	97.3	91.8	87.5	98.4	91.5
PRIMA	87.6	93.2	93.7	92.1	92.3	91.4	91.6	88.1	97.6	92.5	87.9	98.1	92.2

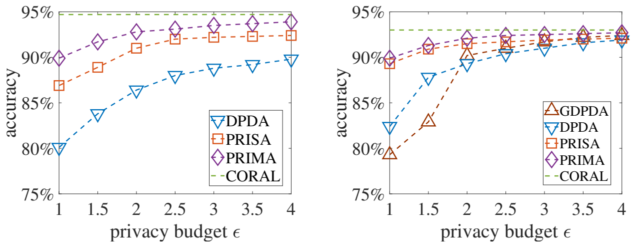
Table 2: Accuracy (%) on Office-Caltech10 dataset for DNN

Method	KT → DV	DV → BK	BK → EL	EL → KT	AVG
OTDA	75.3	72.1	73.8	78.1	74.8
SA	78.4	74.7	75.6	79.3	77.0
CORAL	73.9	78.3	76.3	83.6	78.0
DPDA	72.5	69.9	71.4	75.2	72.3
PRISA	76.8	73.6	74.9	78.4	75.9
PRIMA	72.1	76.9	75.6	82.8	76.8

Table 3: Accuracy (%) on Amazon review dataset for LR

Method	KT → DV	DV → BK	BK → EL	EL → KT	AVG
DANN	72.1	75.3	74.5	83.8	76.4
OTDA	76.4	75.2	78.8	81.2	77.9
SA	80.4	79.6	78.5	82.3	80.2
CORAL	76.9	81.3	81.7	87.4	81.8
GDPDA	68.3	72.2	70.1	79.6	72.5
DPDA	75.5	74.1	77.3	80.3	76.8
PRISA	79.8	79.0	77.8	81.7	79.5
PRIMA	76.1	80.4	80.9	86.6	81.0

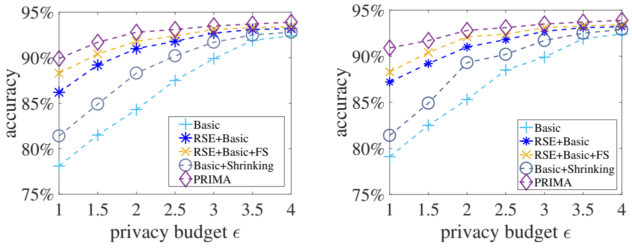
Table 4: Accuracy (%) on Amazon review dataset for DNN



(a) LR

(b) DNN

Figure 2: Accuracy (%) on pair C → A with differential ϵ



(a) LR

(b) DNN

Figure 3: Ablation study results on pair C → A

Basic. We compare PRIMA with these four methods by performing LR on pair C → A.

Figures 3(a) and 3(b) give the results of varying privacy budget ϵ. From the results, we observe that: 1) RSE+Basic achieves a higher accuracy than Basic, which indicates that the RSE based covariance estimation can significantly improve the utility of the perturbed covariance matrix by decreasing dimension; 2) RSE+Basic+FS outperforms RSE+Basic, from which we can conclude that the forward selection strategy can effectively generate subspaces; 3) Basic+Shrinking achieves better performance comparing with Basic, which shows that the PSD matrix recovery method can further improve the utility of the perturbed covariance matrix.

6 Conclusion

In this paper, we propose the first differentially private approach for correlation alignment based domain adaptation. We show that our approach guarantees (ϵ, δ)-differential privacy. Experimental results on real world data demonstrate the superiority of our approach.

Acknowledgments

We sincerely thank the anonymous reviewers for their helpful comments. This work was supported by the National Natural Science Foundation of China (Grant No. 61872045), the Innovation Research Group Project of NSFC (61921003).

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, 2016.

[Amin *et al.*, 2019] Kareem Amin, Travis Dick, Alex Kulesza, Andres Munoz, and Sergei Vassilvitskii. Differentially private covariance estimation. In *NeurIPS*, pages 14213–14222, 2019.

[Bassily *et al.*, 2014] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS*, pages 464–473. IEEE, 2014.

[Blitzer *et al.*, 2006] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, pages 120–128, 2006.

[Caruana and Freitag, 1994] Rich Caruana and Dayne Freitag. Greedy attribute selection. In *Machine Learning Proceedings 1994*, pages 28–36. Elsevier, 1994.

[Courty *et al.*, 2016] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *TPAMI*, 39(9):1853–1865, 2016.

[Cui *et al.*, 2014] Zhen Cui, Wen Li, Dong Xu, Shiguang Shan, Xilin Chen, and Xuelong Li. Flowing on riemannian manifold: Domain adaptation by shifting covariance. *IEEE transactions on cybernetics*, 44(12):2264–2273, 2014.

[Dwork *et al.*, 2006a] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Eurocrypt*, pages 486–503. Springer, 2006.

[Dwork *et al.*, 2006b] 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.*, 2014] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *STOC*, pages 11–20, 2014.

[Farahat *et al.*, 2013] Ahmed K Farahat, Ali Ghodsi, and Mohamed S Kamel. Efficient greedy feature selection for unsupervised learning. *Knowledge and Information Systems*, 35(2):285–310, 2013.

[Fernando *et al.*, 2013] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, pages 2960–2967, 2013.

[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.

[Ge *et al.*, 2018] Jason Ge, Zhaoran Wang, Mengdi Wang, and Han Liu. Minimax-optimal privacy-preserving sparse pca in distributed systems. In *AISTATS*, 2018.

[Gong *et al.*, 2012] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073. IEEE, 2012.

[Higham *et al.*, 2016] Nicholas J Higham, Natasa Strabic, and Vedran Sego. Restoring definiteness via shrinking, with an application to correlation matrices with a fixed block. *SIAM Review*, 58(2):245–263, 2016.

[Jiang *et al.*, 2013] Xiaoqian Jiang, Zhanglong Ji, Shuang Wang, Noman Mohammed, Samuel Cheng, and Lucila Ohno-Machado. Differential-private data publishing through component analysis. *Transactions on data privacy*, 6(1):19, 2013.

[Jiang *et al.*, 2016] Wuxuan Jiang, Cong Xie, and Zhihua Zhang. Wishart mechanism for differentially private principal components analysis. In *AAAI*, pages 1730–1736, 2016.

[Ledoit and Wolf, 2004] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

[LeTien *et al.*, 2019] Nam LeTien, Amaury Habrard, and Marc Sebban. Differentially private optimal transport: Application to domain adaptation. In *IJCAI*, pages 2852–2858, 2019.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105. PMLR, 2015.

[McSherry, 2009] Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *SIGMOD*, pages 19–30, 2009.

[Skurichina and Duin, 2002] Marina Skurichina and Robert PW Duin. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, 5(2):121–135, 2002.

[Sun *et al.*, 2016] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, pages 2058–2065, 2016.

[Wang and Xu, 2019] Di Wang and Jinhui Xu. Differentially private high dimensional sparse covariance matrix estimation. *arXiv preprint arXiv:1901.06413*, 2019.

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

[Wang *et al.*, 2020] Qian Wang, Zixi Li, Qin Zou, Lingchen Zhao, and Song Wang. Deep domain adaptation with differential privacy. *TIFS*, 15:3093–3106, 2020.

[Yao *et al.*, 2019] Quanming Yao, Xiawei Guo, James T Kwok, Wei-Wei Tu, Yuqiang Chen, Wenyuan Dai, and Qiang Yang. Privacy-preserving stacking with application to cross-organizational diabetes prediction. In *IJCAI*, pages 4114–4120, 2019.