Conditional Bures Metric for Domain Adaptation

You-Wei Luo[1] Chuan-Xian Ren[1,2*]
[1]School of Mathematics, Sun Yat-Sen University, China
[2]Pazhou Lab, Guangzhou, China
luoyw28@mail2.sysu.edu.cn, rchuanx@mail.sysu.edu.cn

Abstract

As a vital problem in classification-oriented transfer, unsupervised domain adaptation (UDA) has attracted widespread attention in recent years. Previous UDA methods assume the marginal distributions of different domains are shifted while ignoring the discriminant information in the label distributions. This leads to classification performance degeneration in real applications. In this work, we focus on the conditional distribution shift problem which is of great concern to current conditional invariant models. We aim to seek a kernel covariance embedding for conditional distribution which remains yet unexplored. Theoretically, we propose the Conditional Kernel Bures (CKB) metric for characterizing conditional distribution discrepancy, and derive an empirical estimation for the CKB metric without introducing the implicit kernel feature map. It provides an interpretable approach to understand the knowledge transfer mechanism. The established consistency theory of the empirical estimation provides a theoretical guarantee for convergence. A conditional distribution matching network is proposed to learn the conditional invariant and discriminative features for UDA. Extensive experiments and analysis show the superiority of our proposed model.

1. Introduction

Large-scale data with sufficient annotations are vital sources of machine learning. However, the data collected from the real-world scenarios are usually unlabeled and the manual annotations are expensive. Recent advances in transfer learning yields plenty of methods for dealing with the shortage of labeled data. These methods aim to transfer the knowledge on a labeled source domain to a target domain with few or no annotations, such setting is also known as domain adaptation [27].

The most common assumption in Unsupervised Domain Adaptation (UDA) is that the labeled source domain and unlabeled target domain have the same feature spaces, but different marginal distributions [27], *i.e.*, $\mathcal{X}^s = \mathcal{X}^t$, $P^s_X \neq P^t_X$. This assumption is also called covariate shift [29] and sample selection bias [35]. Ben-David *et al.* [2] give a theoretical insight into the domain adaptation problem, they show that the risk of the target domain is mainly bounded by the risk of the source domain and the discrepancy between distributions of two domains. Inspired by this theory, many methods are proposed to mitigate the discrepancy between feature distributions of the source and target domains, *e.g.*, explicit discrepancy minimization via Maximum Mean Discrepancy (MMD) [13, 21], domain invariant feature learning [26], Optimal Transport (OT) based feature matching [7, 20, 37], manifold based feature alignment [10], statistical moment matching [21, 32] and adversarial domain adaptation [9]. These methods are proved to be effective in minimizing the marginal discrepancy and alleviating the domain shift problem. However, this assumption may lead to the omission of discriminant information in the label distributions, which is described in Figure 1. Recent advancements [19, 22, 24] show that the adaptation models will be more discriminative on the target domain if the target label information (*e.g.*, pseudo labels) is explored carefully.

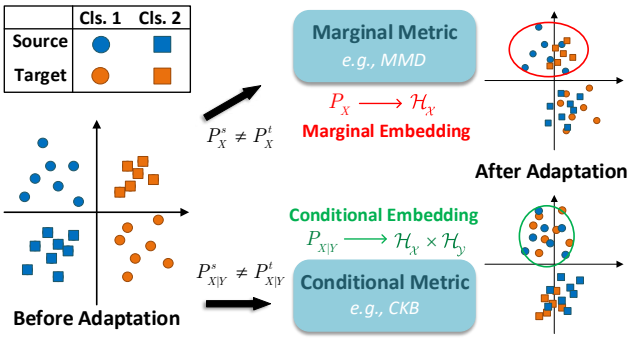


Figure 1. Illustration of the conditional shift problem. Previous metrics that only consider the marginal distribution discrepancy may lead to a misaligned conditional distribution, *i.e.*, the red circle region. On the bottom, the class-level alignment is achieved by exploiting the conditional distribution embedding metric.

Extended from the marginal shift assumption, the con-

*Corresponding Author.

ditional shift problem is studied to build a conditional invariant model [36], *i.e.*, $P^s_{X|Y} = P^t_{X|Y}$. The most critical problem is to construct a framework which can explicitly reflect the relation between different conditional distributions. Zhao *et al*. [38] prove a new generalization bound which quantitatively reflects the underlying structure of the conditional shift problem. Several works have also been made in the field of conditional/joint distribution matching for domain adaptation, *e.g*., multi-layer feature approximation [23], conditional variants of MMD [16, 19, 39], conditional invariant learning with causal interpretations [11, 28], OT based joint distribution models [4, 6].

In this paper, we aim to estimate the transport cost in Reproducing Kernel Hilbert Space (RKHS) for the continuous conditional distributions. Inspired by pioneering work [8], which employs the conditional covariance operator on the RKHS to characterize the independence, we define transport cost estimation on the set of conditional covariance operators called Conditional Kernel Bures (CKB) metric. By virtue of the conditional covariance operator and OT theory, we prove that the CKB metric reflects the discrepancy between two conditional distributions directly. This result can be taken as an extension of the marginal distribution embedding property in MMD [13] and kernel Bures metric [37]. An explicit empirical estimation of the CKB metric and its consistency theory are presented. Further, we apply it to the proposed conditional distribution matching network. Extensive experiment results show the effectiveness of the CKB metric and the superiority of the proposed model. Our contributions are summarized as follows.

- A novel CKB metric for characterizing conditional distribution discrepancy is proposed, and the kernel embedding property of the CKB metric is proved to show that it is well-defined on conditional distributions. This metric is also exactly the OT between conditional distributions, which provides an interpretable approach to understand the knowledge transfer mechanism.
- An explicit empirical estimation of the CKB metric is derived, which provides a computable measurement for conditional domain discrepancy. The asymptotic property of the estimation is proved which provides a rigorous theoretical guarantee for convergence.
- A conditional distribution matching network based on the CKB metric is proposed for discriminative domain alignment, and a joint distribution matching variant is further extended. The SOTA results in extensive experiments validate the model's effectiveness.

2. Related Work

Unsupervised Domain Adaptation. Based on the distribution shift assumption, the UDA methods can be roughly categorized as follows. Domain invariant feature learning methods like Transfer Component Analysis (TCA) [26] try to learn a set of transfer components that make the corresponding distribution robust to the change of domains. OT based methods mitigate the domain discrepancy by minimizing the cost of transporting the source samples to the target domain. It has been shown that OT alignment is equivalent to minimizing the KL divergence [7] or Wasserstein distance [37] between the distributions. Moment matching methods attempt to minimize the distribution discrepancy via statistical moments, *e.g*., Domain Adaptation Network (DAN) [21] for the first order matching and CORAL [32] the second order. Manifold alignment methods take the domains as the points on the manifold and align the domains under the manifold metric [10, 24]. Adversarial based methods [9, 33] alternatively optimize the feature generator and domain discriminator, which are respectively supposed to be domain-confusable and discriminative, to achieve domain confusion. Extended from the marginal distribution assumption, recent works [4, 6, 20, 22, 23] show that the models yield promising results by introducing the label information. Joint Adaptation Network (JAN) [23] builds a joint distribution alignment model via the features from different hidden layers. Conditional Domain Adversarial Network (CDAN) [22] extends the Domain Adversarial Neural Network (DANN) [9] by exploring a multilinear map to describe the conditional variables in adversarial training.

Optimal Transport. Recently, OT has been successively applied to the UDA problem [4, 6, 7, 20, 37]. Courty *et al*. [7] deal with UDA based on the Kantorovitch formulation of OT, which allows to define the well-known Wasserstein distance between the domain distributions. As a variant of Wasserstein distance, Bures metric has been of great interest to various research fields like quantum information, information theory and Riemannian geometry [3]. The original Bures metric is defined on the set of Positive Semi-Definite (PSD) matrices and cannot be used to measure the distribution discrepancy. In [37], Zhang *et al*. extend the OT problem to RKHS, and then define the kernel Wasserstein distance and kernel Bures metric. They show the covariance embedding in RKHS is injective which implies that the kernel Bures metric defines a metric on the distributions. However, these discrepancy measures mainly focus on the marginal distribution. To exploit the label information, joint distribution OT models [4, 6] seek an optimal joint transport map that minimizes the generalized cost associated to the joint space of features and labels $X \times Y$. Enhanced Transport Distance (ETD) [20] uses the prediction feedback from the classifier to reweigh the transport cost. Differing from the above OT based methods which are formulated on discrete joint distribution or marginal distribution, our work focuses on the explicit estimation of OT between conditional distributions under the continuous case.

3. OT for Conditional Distribution

In this section, we first review the definitions and properties of conditional covariance operator and Kantorovitch's OT in RKHS, which are the fundamentals of the proposed CKB metric. Then we present the theoretical definition and property of the CKB metric. Finally, we provide the empirical estimation and its asymptotic property.

3.1. Preliminary

Conditional Covariance Operators. Let $(\mathcal{X}, \mathcal{B})$ be a measure space with Borel σ-field $\mathcal{B}$. Denote $(\mathcal{H}_\mathcal{X}, k_\mathcal{X})$ as the RKHSs of $\mathcal{X}$, which is generated by the positive definite kernels $k_\mathcal{X}$. The mean element μ_X in $\mathcal{H}_\mathcal{X}$ with law P_X is given by $\mu_X = \mathbb{E}_X[\phi(X)]$, where ϕ is the nonlinear feature map of $\mathcal{H}_\mathcal{X}$. It is assumed that $\phi(x) = k_\mathcal{X}(x, \cdot)$ satisfies the reproducing properties $\langle \phi(x), \phi(x') \rangle_{\mathcal{H}_\mathcal{X}} = k_\mathcal{X}(x, x')$ and $\langle \phi(x), f \rangle_{\mathcal{H}_\mathcal{X}} = f(x), \forall f \in \mathcal{H}_\mathcal{X}$.

To explore the casual connection between $\mathcal{X}$ and $\mathcal{Y}$, we consider the pair $(X, Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$ with probability measure $P_{XY} \in \Pr(\mathcal{X}, \mathcal{Y})$, where $\Pr(\mathcal{X}, \mathcal{Y})$ is the set of Borel probability measures on $(\mathcal{X}, \mathcal{Y})$. Given a joint measure $(\mathcal{H}_\mathcal{X} \times \mathcal{H}_\mathcal{Y}, \mathcal{B}_\mathcal{X} \times \mathcal{B}_\mathcal{Y})$, its corresponding *cross-covariance operator* [1] $\mathbf{R}_{XY} : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{X}$ satisfies that $\forall\, f \in \mathcal{H}_\mathcal{X},\ g \in \mathcal{H}_\mathcal{Y}$,

$$\langle f, \mathbf{R}_{XY} g \rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}_{XY}[f(X)g(Y)] - \mathbb{E}_X[f(X)]\,\mathbb{E}_Y[g(Y)]$$

Formally, $\mathbf{R}_{XY}$ is defined as [30]

$$\mathbf{R}_{XY} = \mathbb{E}_{XY}[(\phi(X) - \mu_X) \otimes (\psi(Y) - \mu_Y)].$$

If Y equals to X, $\mathbf{R}_{XX}$ is just the covariance operator on $\mathcal{H}_\mathcal{X}$. Based on the *cross-covariance operator*, we further consider the conditional covariance of $\phi(X)$ w.r.t. the conditioning variable Y. The *conditional covariance operator* $\mathbf{R}_{XX|Y}$ is usually written as [8]

$$\mathbf{R}_{XX|Y} = \mathbf{R}_{XX} - \mathbf{R}_{XY}\mathbf{R}_{YY}^{-1}\mathbf{R}_{YX}.$$

Note that $\mathbf{R}_{YY}$ may be non-invertible, especially in the real-world applications with finite samples. When necessary conditions are fulfilled [8], the conditional covariance operator also satisfies that

$$\left\langle f, \mathbf{R}_{XX|Y} f \right\rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}_Y\left[\mathrm{Var}_{X|Y}[f(X)|Y]\right], \quad \forall\, f \in \mathcal{H}_\mathcal{X}.$$

Kantorovitch's OT in RKHS. For any two distributions $P_X^s, P_X^t \in \Pr(\mathcal{X})$, let $\Pi(P_X^s \times P_X^t)$ be the set of probabilistic couplings, the Kantorovitch formulation of OT is

$$\gamma^* = \inf_{\gamma \in \Pi(P_X^s \times P_X^t)} \int_{\mathcal{X} \times \mathcal{Y}} d^2(\mathbf{x}^s, \mathbf{x}^t) d\gamma(\mathbf{x}^s, \mathbf{x}^t). \qquad (1)$$

The Kantorovitch problem in Eq. (1) is also equivalent to the Wasserstein distance. Under the Gaussian measures, if the distributions P_X^s and P_X^t have the same expectations, the Wasserstein distance between them is equivalent to the Bures metric between their covariance matrices. Let $\mathbb{S}^+(d)$ be the set of $d \times d$ PSD matrices; for any PSD matrix $\mathbf{\Sigma}$, its unique square root $\sqrt{\mathbf{\Sigma}}$ is defined by $\mathbf{\Sigma} = \sqrt{\mathbf{\Sigma}}\sqrt{\mathbf{\Sigma}}$. The Bures metric is defined by

$$d_{\mathrm{B}}^2(\mathbf{\Sigma}_{XX}^s, \mathbf{\Sigma}_{XX}^t) = \mathrm{tr}\left(\mathbf{\Sigma}_{XX}^s + \mathbf{\Sigma}_{XX}^t - 2\mathbf{\Sigma}_{XX}^{st}\right),$$

where $\mathbf{\Sigma}_{XX}^{st} = \sqrt{\sqrt{\mathbf{\Sigma}_{XX}^s}\mathbf{\Sigma}_{XX}^t\sqrt{\mathbf{\Sigma}_{XX}^s}}$ and $\mathbf{\Sigma}_{XX}^s$ and $\mathbf{\Sigma}_{XX}^t$ are the covariance matrices of P_X^s and P_X^t, respectively. Recent work shows that the Bures metric is also related to the Riemannian geometry, as it can be taken as the metric on PSD manifold [3]. Though the Bures metric defines a metric on $\mathbb{S}^+(d)$, it cannot reflect discrepancy between distributions P_X^s and P_X^t.

The kernel Bures metric [37] generalizes the PSD setting in Bures metric to the infinite-dimensional RKHS $\mathcal{H}$. Let $\mathbb{S}^+(\mathcal{H}_\mathcal{X}) \subseteq \mathcal{H}_\mathcal{X} \times \mathcal{H}_\mathcal{X}$ be the set of all positive, self-adjoint, and trace-class operators on $\mathcal{H}_\mathcal{X}$ with kernel $k_\mathcal{X}$, the kernel Bures metric $d_{\mathrm{KB}}(\cdot, \cdot)$ on $\mathbb{S}^+(\mathcal{H}_\mathcal{X})$ is written as:

$$d_{\mathrm{KB}}^2(\mathbf{R}_{XX}^s, \mathbf{R}_{XX}^t) = \mathrm{tr}\left(\mathbf{R}_{XX}^s + \mathbf{R}_{XX}^t - 2\mathbf{R}_{XX}^{st}\right),$$

where $\mathbf{R}_{XX}^{st} = \sqrt{\sqrt{\mathbf{R}_{XX}^s}\mathbf{R}_{XX}^t\sqrt{\mathbf{R}_{XX}^s}}$ and $\mathbf{R}_{XX}^s, \mathbf{R}_{XX}^t$ are the covariance operators of P_X^s and P_X^t on $\mathcal{H}_\mathcal{X}$, respectively. Note the kernel Bures is exactly the transport cost in RKHS when the push-forward measures $\phi \# P_X^s$ and $\phi \# P_X^t$ are Gaussian [37]. Zhang *et al*. [37] prove that if the measurable space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ is locally compact and Hausdorff, the embedding $P_X^s \mapsto \mathbf{R}_{XX}^s, \forall P_X^s \in \Pr(\mathcal{X})$ is injective. It turns out that $d_{\mathrm{KB}}(\cdot, \cdot)$ defines a metric on $\Pr(\mathcal{X})$, which no longer holds for the Bures metric. With this property, the kernel Bures metric can be used to quantify the discrepancy between two distributions.

3.2. Conditional Kernel Bures Metric

To introduce conditional distribution to OT, we develop the kernel covariance embedding property for conditional distributions and apply it to the kernel Bures metric. The CKB metric for conditional distributions is now defined.

Definition 1 *The Conditional Kernel Bures (CKB) metric between two conditional distributions $P_{X|Y}^s, P_{X|Y}^t \in \Pr(\mathcal{X}|\mathcal{Y})$ is defined as*

$$d_{\mathrm{CKB}}^2(\mathbf{R}_{XX|Y}^s, \mathbf{R}_{XX|Y}^t) = \mathrm{tr}\left(\mathbf{R}_{XX|Y}^s + \mathbf{R}_{XX|Y}^t - 2\mathbf{R}_{XX|Y}^{st}\right), \qquad (2)$$

where $\mathbf{R}_{XX|Y}^{st} = \sqrt{\sqrt{\mathbf{R}_{XX|Y}^s}\mathbf{R}_{XX|Y}^t\sqrt{\mathbf{R}_{XX|Y}^s}}$.

Proposition 1 *CKB $d_{\mathrm{CKB}}(\cdot, \cdot)$ defines a metric on $\mathbb{S}^+(\mathcal{H}_\mathcal{X})$.*

Recall that the conditional covariance operator $\mathbf{R}_{XX|Y}$ is also positive, self-adjoint, and trace-class on $\mathcal{H}_\mathcal{X}$ [8].

Thus, we can deduce from Proposition 1 that the CKB metric is well-defined on the conditional covariance operators.

The injective property of mean embedding $\mathbb{E}_X[\phi(X)]$ [13] and covariance embedding $\mathbf{R}_{XX}$ [37] in RKHS give the theoretical insights into how two distributions are matched via the defined metrics, *e.g.*, MMD and kernel Bures metric. Similarly, we also make connection between the CKB metric and conditional distributions. Note that though the above embedding properties are well studied, they only consider connections between the operators and the marginal distributions. As the embedding property between the covariance operators and conditional distributions is unexplored, our work focuses on extending the CKB metric to a metric on conditional distributions $\Pr(\mathcal{X}|\mathcal{Y})$. For convenience, we denote the set of measures that satisfy the 3-splitting property [37] by $\Pr^s(\mathcal{X}|\mathcal{Y}=y)$ and the direct sum by $\mathcal{H}_\mathcal{X} \oplus \mathcal{H}_\mathcal{Y}$.

Theorem 1 *Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ be the locally compact and Hausdorff measurable space and k be c_0-universal kernel. Assuming that $(\phi(X), \psi(Y))$ is a Gaussian random variable in $\mathcal{H}_\mathcal{X} \oplus \mathcal{H}_\mathcal{Y}$. For any $P^s_{X|Y}, P^t_{X|Y} \in \Pr^s(\mathcal{X}|\mathcal{Y})$, we have*

$$d_{\text{CKB}}(\mathbf{R}^s_{X X|Y}, \mathbf{R}^t_{X X|Y}) = 0 \quad \Longrightarrow \quad P^s_{X|Y} = P^t_{X|Y}.$$

The above theorem shows that the CKB metric $d_{\text{CKB}}(\cdot,\cdot)$ defines a metric on $\Pr(\mathcal{X}|\mathcal{Y})$ if some conditions are satisfied. Note that the CKB metric is exactly the minimized OT cost between two conditional distributions since $\phi\#P^s_{(X,Y)}$ and $\phi\#P^t_{(X,Y)}$ are also Gaussian. Thus, it can be used to measure the discrepancy between two conditional distributions. The condition c_0-universal [31] in Theorem 1 is satisfied by many common kernels, *e.g.*, Gaussian kernel and Laplacian kernel. The assumption of Gaussian random variable can be taken as the extension of Gaussian distribution which takes values in RKHS [17]. Recall that the feature maps $\phi(\cdot)$ and $\psi(\cdot)$ are implicit, so the conditional covariance operator $\mathbf{R}_{XX|Y}$ is not formulable in practical computation of the CKB metric. To present an explicit formulation of the CKB metric, we use the kernel trick, *i.e.*, $\langle \phi(x), \phi(x') \rangle_{\mathcal{H}_\mathcal{X}} = k_\mathcal{X}(x, x')$, to avoid the explicit nonlinear maps in the next section.

3.3. Empirical Estimation of the Conditional Kernel Bures Metric

Let $\mathcal{D}^s = \{(\mathbf{x}^s_i, \mathbf{y}^s_i)\}_{i=1}^n$ and $\mathcal{D}^t = \{(\mathbf{x}^t_j, \mathbf{y}^t_j)\}_{j=1}^m$ be two sets of samples, which are assumed to be drawn i.i.d. from P^s_{XY} and P^t_{XY}, respectively. Note that $x^{s/t}_i \in \mathbb{R}^d, y^{s/t}_i \in \mathbb{R}^c$, and we map the data $x^{s/t}_i$ (resp. $y^{s/t}_i$) to the RKHS $\mathcal{H}_\mathcal{X}$ (resp. $\mathcal{H}_\mathcal{Y}$) with the implicit feature map ϕ (resp. ψ). Let $\mathbf{K}^{s/t}_{XX}$, $\mathbf{K}^{s/t}_{YY}$ and $\mathbf{K}^{ts}_{XX}$ be the explicit kernel matrices computed as $(\mathbf{K}^{s/t}_{XX})_{ij} = k_\mathcal{X}(\mathbf{x}^{s/t}_i, \mathbf{x}^{s/t}_j)$, $(\mathbf{K}^{s/t}_{YY})_{ij} = k_\mathcal{Y}(\mathbf{y}^{s/t}_i, \mathbf{y}^{s/t}_j)$ and $(\mathbf{K}^{ts}_{XX})_{ij} = k_\mathcal{X}(\mathbf{x}^t_i, \mathbf{x}^s_j)$, respectively. Denote the feature map matrices by $\mathbf{\Phi}_{s/t}$ and $\mathbf{\Psi}_{s/t}$. Their cross-covariance matrices can be written as $\hat{\mathbf{R}}^s_{XY} = \frac{1}{n}\mathbf{\Phi}_s\mathbf{H}_n\mathbf{\Psi}^T_s$, $\hat{\mathbf{R}}^t_{XY} = \frac{1}{m}\mathbf{\Phi}_t\mathbf{H}_m\mathbf{\Psi}^T_t$, where $\mathbf{H}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is the $n \times n$ centering matrix, and $\mathbf{1}_n$ is n-dimensional vector with all elements equal to 1. As the covariance matrix $\hat{\mathbf{R}}^{s/t}_{YY}$ is always rank-deficient under the finite sample case, we regularize it as

$$\hat{\mathbf{R}}_{XX|Y} = \hat{\mathbf{R}}_{XX} - \hat{\mathbf{R}}_{XY}\left(\hat{\mathbf{R}}_{YY} + \varepsilon\mathbf{I}\right)^{-1}\hat{\mathbf{R}}_{YX}, \quad (3)$$

where $\varepsilon > 0$ is the regularization parameter. Denote the matrices

$$\mathbf{B}_s \triangleq \mathbf{I}_n - \frac{1}{n\varepsilon}\left[\mathbf{G}^s_Y - \mathbf{G}^s_Y\left(\mathbf{G}^s_Y + \varepsilon n\mathbf{I}_n\right)^{-1}\mathbf{G}^s_Y\right],$$
$$\mathbf{B}_t \triangleq \mathbf{I}_m - \frac{1}{m\varepsilon}\left[\mathbf{G}^t_Y - \mathbf{G}^t_Y\left(\mathbf{G}^t_Y + \varepsilon m\mathbf{I}_m\right)^{-1}\mathbf{G}^t_Y\right],$$

where

$$\mathbf{G}^s_{X/Y} = \mathbf{H}_n\mathbf{K}^s_{XX/YY}\mathbf{H}_n, \quad \mathbf{G}^t_{X/Y} = \mathbf{H}_m\mathbf{K}^t_{XX/YY}\mathbf{H}_m$$

are the centralized kernel matrices. With the decomposition $\mathbf{B}_{s/t} = \mathbf{C}_{s/t}\mathbf{C}^T_{s/t}$, the conditional covariance operator $\hat{\mathbf{R}}^s_{XX|Y}$ can be reformulated as ($\hat{\mathbf{R}}^t_{XX|Y}$ is the same)

$$\hat{\mathbf{R}}^s_{XX|Y} = \frac{1}{n}\mathbf{\Phi}_s\mathbf{H}_n\mathbf{C}_s\left(\mathbf{\Phi}_s\mathbf{H}_n\mathbf{C}_s\right)^T. \quad (4)$$

Proposition 2 *If $k_\mathcal{Y}$ is positive definite kernel, then $\mathbf{B}_s$ and $\mathbf{B}_t$ are positive definite for any $\varepsilon > 0$. Especially, we have*

$$\mathbf{B}_s = \varepsilon n\left(\mathbf{G}^s_Y + \varepsilon n\mathbf{I}_n\right)^{-1}, \quad \mathbf{B}_t = \varepsilon m\left(\mathbf{G}^t_Y + \varepsilon m\mathbf{I}_m\right)^{-1}.$$

Remark 1 *Proposition 2 shows that $\mathbf{B}_{s/t}$ is positive definite with a positive definite kernel $k_\mathcal{Y}$ (*e.g.*, Gaussian kernel and Laplacian kernel), so the decomposition $\mathbf{B}_{s/t} = \mathbf{C}_{s/t}\mathbf{C}^T_{s/t}$ always exists. But, such a decomposition is not unique,* e.g.*, Cholesky factorization and eigendecomposition. Here we compute $\mathbf{C}_s$ based on the Eigenvalue Decomposition (EVD) of $\mathbf{B}_s$ as ($\mathbf{C}_t$ is the same)*

$$\mathbf{B}_s = \mathbf{U}_s\mathbf{D}_s\mathbf{U}^T_s = \mathbf{U}_s\sqrt{\mathbf{D}_s}\left(\mathbf{U}_s\sqrt{\mathbf{D}_s}\right)^T = \mathbf{C}_s\mathbf{C}^T_s,$$

where $\mathbf{U}_s$ and $\mathbf{D}_s$ are the eigenvector and eigenvalue matrices of $\mathbf{B}_s$, respectively.

The reformulation Eq. (4) affords an explicit insight into the conditional covariance operator. As $\mathbf{B}_s$ is computed from the gram matrix $\mathbf{G}^s_Y$, $\mathbf{C}_s$ is highly related to the conditional variable Y. Compared with the covariance operator on RKHS $\hat{\mathbf{R}}^s_{XX} = \mathbf{\Phi}_s\mathbf{H}_n\mathbf{\Phi}^T_s/n$, the feature map $\mathbf{\Phi}_s$ in conditional covariance operator $\hat{\mathbf{R}}^s_{XX|Y}$ is transformed by the modified centering matrix $\mathbf{H}_n\mathbf{C}_s$ which contains the conditional information. Based on the above reformulation,

the following theorem provides the explicit computation of the CKB metric. Note that the reformulation Eq. (4) is included in the proof of Theorem 2, and all proofs of theorems, propositions are provided in the supplementary material.

Theorem 2 *The empirical estimation of the CKB metric is computed as*

$$\begin{aligned}&\hat{d}^2_{\text{CKB}}(\hat{\mathbf{R}}^s_{XX|Y}, \hat{\mathbf{R}}^t_{XX|Y})\\ =&\varepsilon \text{tr}\left[\mathbf{G}^s_X \left(\varepsilon n \mathbf{I}_n + \mathbf{G}^s_Y\right)^{-1}\right] + \varepsilon \text{tr}\left[\mathbf{G}^t_X \left(\varepsilon m \mathbf{I}_m + \mathbf{G}^t_Y\right)^{-1}\right]\\ &- \frac{2}{\sqrt{nm}} \left\| \left(\mathbf{H}_m \mathbf{C}_t\right)^T \mathbf{K}^{ts}_{XX} \left(\mathbf{H}_n \mathbf{C}_s\right) \right\|_*, \end{aligned} \tag{5}$$

where $\|\cdot\|_*$ is the nuclear norm.

Remark 2 *The computational complexity of the CKB metric consists of three terms shown in Eq.* (5)*. As for the first term, the cost of the kernel matrices and matrix inverse are about $\mathcal{O}((c+d+n)n^2)$. Similarly, the cost of the second term is about $\mathcal{O}((c+d+m)m^2)$. As for the third term, the cost of kernel matrix, EVD and nuclear norm is about $\mathcal{O}(nmd + n^3 + m^3 + \min(mn^2, m^2 n))$. Thus, the computational complexity of the CKB metric is about $\mathcal{O}(\max(c,d,m,n)(n^2+m^2+mn))$, where d and c are the feature dimension and number of classes, respectively.*

3.4. Convergence Analysis

In this section, we focus on the convergence of the empirical estimation of the CKB metric. This convergence theorem is based on the properties of trace-class operator on the Hilbert space and the asymptotic theory of the conditional covariance operator established by Fukumizu *et al*. [8]. Let $\hat{\mathbf{R}}^{(n)}_{XX|Y}$ be the conditional covariance operator drawn i.i.d. from distribution P_{XY} with sample size n which is computed as Eq. (3), Proposition 7 in [8] shows that the estimator $\hat{\mathbf{R}}^{(n)}_{XX|Y}$ converges to $\mathbf{R}_{XX|Y}$ in probability. Moreover, it shows that the sequence $|\text{tr}(\hat{\mathbf{R}}^{(n)}_{XX|Y}) - \text{tr}(\mathbf{R}_{XX|Y})|$ is bounded in probability at rate $\frac{1}{\varepsilon_n \sqrt{n}}$.

With the consistency of conditional covariance operator, we now establish the asymptotic theory for the CKB metric. Assuming that the conditional covariance operators are specified by the source and target domains, we define $n' = \min\{n, m\}$ and the squared CKB metric as $\hat{D}^{(n')}_{\text{CKB}} = \hat{d}^2_{\text{CKB}}(\mathbf{R}^{s^{(n)}}_{XX|Y}, \mathbf{R}^{t^{(m)}}_{XX|Y})$ and $D_{\text{CKB}} = d^2_{\text{CKB}}(\mathbf{R}^s_{XX|Y}, \mathbf{R}^t_{XX|Y})$. The convergence of $\hat{D}^{(n')}_{\text{CKB}}$ is dominated by the convergence of three terms in Eq. (2). Specifically, the convergence of first two terms are concluded from the consistency of conditional covariance operator, and the third term can be deduced to the convergence in trace-norm on the Hilbert space. We present the convergence theorem of the CKB metric as follows.

Theorem 3 *Let the regularization parameter ε in Eq.* (3) *be a series related to n',* i.e.*, $\varepsilon_{n'}$. Assuming $\varepsilon_{n'}$ satisfies that $\varepsilon_{n'} \to 0$ and $\varepsilon_{n'}\sqrt{n'} \to \infty$ ($n' \to \infty$), then we have*

$$|\hat{D}^{(n')}_{\text{CKB}} - D_{\text{CKB}}| \to 0 \quad (n' \to \infty)$$

in probability with rate $(\frac{1}{\varepsilon'_n \sqrt{n'}})^{\frac{1}{2}}$.

Theorem 3 shows that the empirical estimation error of the CKB metric converges to 0 as $n \to \infty$ in probability. Specifically, the estimation error $|\hat{D}^{(n')}_{\text{CKB}} - D_{\text{CKB}}|$ is bounded in probability at rate $(\frac{1}{\varepsilon'_n \sqrt{n'}})^{\frac{1}{2}}$. Compared to the rate $\frac{1}{\varepsilon_n \sqrt{n}}$ of the conditional covariance operator, the square root rate $(\frac{1}{\varepsilon'_n \sqrt{n'}})^{\frac{1}{2}}$ of the CKB metric comes from the convergence rate of the cross term, *i.e*., $\mathbf{R}^{st}_{XX|Y}$.

4. Unsupervised Domain Adaptation

In this section, we tackle the UDA problem by describing the domains as conditional distributions and minimizing the conditional distribution discrepancy under the CKB metric.

4.1. Conditional Distribution Matching Network

For UDA problems, $\mathcal{D}^s = \{(\mathbf{x}^s_i, \mathbf{y}^s_i)\}^n_{i=1}$ is taken as the source domain and $\mathcal{D}^t = \{\mathbf{x}^t_j\}^m_{j=1}$ the unlabeled target domain, where $\mathbf{x}^{s/t}_i$ represent the observations and $\mathbf{y}^s_i \in \mathbb{R}^K$ the one-hot labels with K classes. The primary task is to generalize the classifier $C : \mathbf{x} \mapsto \mathbf{y}$ trained on both $\mathcal{D}^s$ and $\mathcal{D}^t$ to predict the $\mathbf{y}^t_i$. Previous UDA methods assume that the target distribution is shifted from the source distribution (*i.e*., $P^s_X \neq P^t_X$) and generalize C by minimizing the distribution discrepancy. This assumption only considers the feature distribution, but ignores the discriminant information from the labels. Here we consider the shift of conditional distribution $P_{X|Y}$, which will help the adaptation model to incorporate discriminant information. To learn a conditional distribution matching model, we first design a feature extractor F based on Deep Neural Networks (DNNs), which aims to align the conditional distributions of the domains, *i.e*., $P^s_{X|Y}$ and $P^t_{X|Y}$. Then the classifier $C : F(\mathbf{x}) \mapsto \mathbf{y}$ will be trained on the aligned features. Denote the extracted features by $\mathbf{Z}^{s/t} = [F(\mathbf{x}^{s/t}_1), \ldots, F(\mathbf{x}^{s/t}_{n/m})]$ and the soft predictions by $\hat{\mathbf{Y}}^{s/t} = [C(\mathbf{z}^{s/t}_1), \ldots, C(\mathbf{z}^{s/t}_{n/m})]$, where $\sum^K_{i=1} \hat{y}^{s/t}_{ij} = 1$. The detailed network architecture is provided in the supplementary material.

The flowchart of the proposed method is shown in Figure 2. It aligns the source and target domains to the conditional invariant space by minimizing the CKB metric between the extracted features, *i.e*., $\mathbf{Z}^s$ and $\mathbf{Z}^t$. Based on the conditional invariant features, a discriminative classifier is learned by applying the entropy-based criterion to both domains.

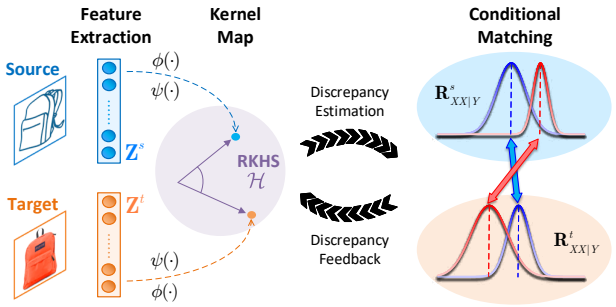


Figure 2. Flowchart of the conditional matching model. The features are mapped into the RKHS, and the conditional distributions of the domains are represented by their conditional covariance operators in RKHS. Then the conditional distribution discrepancy is estimated based on the CKB metric, and the adaptation model is optimized according to the discrepancy feedback.

A well-aligned feature space is more preferable for training classifier. Meanwhile, a more accurate classifier leads to a more precise estimation of the CKB metric and fewer misaligned sample pairs. Therefore, the two processes can benefit from each other and enhance the transferability and discriminability of the model alternatively.

In general, the proposed network is trained based on three loss terms. First, the cross-entropy function is applied to the labeled source data, which builds a basic network for classification. The cross-entropy loss $\mathcal{L}_{\text{CE}}$ is written as

$$\mathcal{L}_{\text{CE}} = \sum_{i=1}^{K} \sum_{j=1}^{n} -y^s_{ij} \log \hat{y}^s_{ij}.$$

Then the entropy $\mathcal{L}_{\text{Ent}}$ is applied to the target prediction:

$$\mathcal{L}_{\text{Ent}} = \sum_{i=1}^{K} \sum_{j=1}^{m} -\hat{y}^t_{ij} \log \hat{y}^t_{ij}.$$

This term has been proved to be effective in the semi-supervised learning and unsupervised learning [12]. For UDA, it preserves the intrinsic structure of the target domain and reduces the uncertainty of the target prediction.

To match the conditional distributions of two domains, the CKB metric is applied to the deep features learned by the nonlinear mapping F. Thus, the kernel matrices $\mathbf{K}^{s/t}_{XX}, \mathbf{K}^{ts}_{XX}$ and feature maps $\mathbf{\Phi}_{s/t}$ are computed from the deep features $\mathbf{Z}^{s/t}$ hereinafter, *i.e.*, $k(\mathbf{z}_i, \mathbf{z}_j)$ and $\phi(\mathbf{z}_i)$. In terms of the conditional variable Y, the kernel matrix $\mathbf{K}^s_{YY}$ and feature map $\mathbf{\Psi}_s$ are computed from the source labels $\mathbf{y}^s_i$. As the ground-truth labels $\mathbf{y}^t_i$ of the target samples are unknown, we use the pseudo labels $\hat{\mathbf{y}}^t_i$ to approximate them and compute the feature map as $\hat{\mathbf{\Psi}}_t$. The CKB loss is computed according to Eq. (5) as

$$\mathcal{L}_{\text{CKB}} = \hat{d}^2_{\text{CKB}}(\hat{\mathbf{R}}^s_{XX|Y}, \hat{\mathbf{R}}^t_{XX|Y}).$$

Let λ_1 and λ_2 be the trade-off parameters, the objective function of the conditional alignment model is written as

$$\min_{F,C} \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{Ent}} + \lambda_2 \mathcal{L}_{\text{CKB}}. \tag{6}$$

According to Theorem 1, the domain conditional distributions are aligned (*i.e.*, $P^s_{X|Y} = P^t_{X|Y}$) when $\mathcal{L}_{\text{CKB}} = 0$. Further, if the marginal distributions P^s_Y and P^t_Y are also aligned, then the domain joint distribution matching is also achieved as $P_{XY} = P_{X|Y} P_Y$. Since the target distribution P^t_Y is unknown, we can apply the marginal matching constraint to the label distribution estimated from the classifier's predictions. Specifically, the marginal discrepancy can be approximated by the MMD between $\mathbf{\Psi}_s$ and $\tilde{\mathbf{\Psi}}_t$, *i.e.*, $\mathcal{L}_{\text{MMD}} = \|\mathbf{\Psi}_s \mathbf{1}_n/n - \tilde{\mathbf{\Psi}}_t \mathbf{1}_m/m\|^2_{\mathcal{H}_\mathcal{Y}}$, where $\tilde{\mathbf{\Psi}}_t$ is computed from the soft predictions $\tilde{\mathbf{y}}^t_i$. Finally, the joint distribution alignment loss is the sum of $\mathcal{L}_{\text{MMD}}$ and $\mathcal{L}_{\text{CKB}}$, and the objective function is written as

$$\min_{F,C} \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{Ent}} + \lambda_2 (\mathcal{L}_{\text{CKB}} + \mathcal{L}_{\text{MMD}}). \tag{7}$$

In summary, $\mathcal{L}_{\text{MMD}}$ and $\mathcal{L}_{\text{CKB}}$ aim to integrate the samples from different domains by mitigating the conditional or joint distribution discrepancies, and the first two terms enhance the model's discriminability by using the label and prediction information from both domains.

4.2. Implementation Details

We train the proposed model with back-propagation in the mini-batch manner. As $\mathcal{L}_{\text{CKB}}$ refers to the inverse of the kernel matrices $\mathbf{G}^s_Y$ and $\mathbf{G}^t_Y$, we treat $\hat{\mathbf{Y}}^t$ in $\mathcal{L}_{\text{CKB}}$ as constant to make the optimization stable. Thus, $\mathbf{G}^s_Y$ and $\mathbf{G}^t_Y$ are independent of the network parameter and there are no gradients refer to them. The regularization parameter ϵ of inverse in Eq. (5) is set as 10^{-2} empirically. In terms of the kernel function, Gaussian kernel $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\|\mathbf{x} - \mathbf{x}'\|^2_2/\sigma^2\right)$ is adopted, and the parameter σ^2 is set as the mean of the all square Euclidean distances $\|\mathbf{x} - \mathbf{x}'\|^2_2$ that refer to the corresponding kernel matrix. The kernel parameters σ are adaptively updated for each minibatch. Thanks to the smoothness of the Gaussian kernel, the gradients of the network parameters always exist. The proposed methods in Eq. (6) and Eq. (7) are respectively abbreviated as **CKB** and **CKB+MMD** hereinafter.

5. Experiment

The proposed methods are evaluated and compared with the SOTA methods on four UDA datasets.

ImageCLEF-DA [5] consists of 3 domains with 12 common classes, *i.e.*, *Caltech* (**C**), *ImageNet* (**I**), *Pascal* (**P**), where each domain include 600 images.

Office-Home [34] contains 15500 images from 4 domains with 65 classes, *i.e.*, *Art* (**Ar**), *Clipart* (**Cl**), *Product* (**Pr**) and *Real-World* (**Rw**).

Office10 [10] consists of 4 domains with 10 classes, *i.e.*, *Amazon* (**A**), *Caltech* (**C**), *DSLR* (**D**) and *Webcam* (**W**).

Digits Recognition Follow the protocol in [15], we conduct the adaptation task between the handwritten digit datasets MNIST (**M**) and USPS (**U**).

Table 1. Accuracies (%) on Office-Home (ResNet-50), Image-CLEF-DA (ResNet-50) and Office10 (AlexNet).

Office-Home	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Mean
Source [14]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [21]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [9]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
KGOT [37]	36.2	59.4	65.0	48.6	56.5	60.2	52.1	37.8	67.1	59.0	41.9	72.0	54.7
CDAN+E [22]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
ETD [20]	51.3	71.9	**85.7**	57.6	69.2	**73.7**	57.8	51.2	**79.3**	70.2	57.5	82.1	67.3
DMP [24]	52.3	73.0	77.3	64.3	72.0	71.8	63.6	52.7	78.5	72.0	57.7	81.6	68.1
CKB	**54.7**	**74.4**	77.1	63.7	**72.2**	71.8	64.1	51.7	78.4	**73.1**	58.0	82.4	68.5
CKB+MMD	54.2	74.1	77.5	**64.6**	**72.2**	71.0	**64.5**	**53.4**	78.7	72.6	**58.4**	**82.8**	**68.7**

Image-CLEF-DA	I→P	P→I	I→C	C→I	C→P	P→C	Mean
Source [14]	74.8 ± 0.3	83.9 ± 0.1	91.5 ± 0.3	78.0 ± 0.2	65.5 ± 0.3	91.2 ± 0.3	80.7
DAN [21]	74.5 ± 0.4	82.2 ± 0.2	92.8 ± 0.2	86.3 ± 0.4	69.2 ± 0.4	89.8 ± 0.4	82.5
DANN [9]	75.0 ± 0.3	86.0 ± 0.3	96.2 ± 0.4	87.0 ± 0.5	74.3 ± 0.5	91.5 ± 0.6	85.0
KGOT [37]	76.3	83.3	93.5	87.5	74.8	89.0	84.1
CDAN+E [22]	77.7 ± 0.3	90.7 ± 0.2	97.7 ± 0.3	91.3 ± 0.3	74.2 ± 0.2	94.3 ± 0.3	87.7
ETD [20]	**81.0**	91.7	**97.9**	93.3	79.5	95.0	89.7
DMP [24]	80.7 ± 0.1	92.5 ± 0.1	97.2 ± 0.1	90.5 ± 0.1	77.7 ± 0.2	96.2 ± 0.2	89.1
CKB	80.7 ± 0.1	**93.7** ± 0.1	97.0 ± 0.1	**93.5** ± 0.2	79.2 ± 0.1	**97.0** ± 0.1	**90.2**
CKB+MMD	80.7 ± 0.2	92.2 ± 0.1	96.5 ± 0.1	92.2 ± 0.2	**79.9** ± 0.2	96.7 ± 0.1	89.7

Office10	A→C	A→D	A→W	C→A	C→D	C→W	D→A	D→C	D→W	W→A	W→C	W→D	Mean
Source [18]	82.7	85.4	78.3	91.5	88.5	83.1	80.6	74.6	99.0	77.0	69.6	100.0	84.2
GFK [10]	78.1	84.7	76.3	89.1	88.5	80.3	89.0	78.4	99.3	83.9	76.2	100.0	85.3
CORAL [32]	85.3	80.8	76.3	91.1	86.6	81.1	88.7	80.4	99.3	82.1	78.7	100.0	85.9
OT-IT [7]	83.3	84.1	77.3	88.7	90.5	88.5	83.3	84.0	98.3	88.9	79.1	99.4	87.1
KGOT [37]	85.7	86.6	82.4	91.4	92.4	87.1	91.8	**85.6**	99.3	89.7	85.0	100.0	89.7
DMP [24]	86.6	90.4	**91.3**	92.8	93.0	88.5	91.4	85.3	97.7	91.9	85.6	100.0	91.2
CKB	87.0	**93.6**	90.2	**93.4**	**93.6**	90.8	**92.7**	83.5	**100.0**	92.4	84.3	**100.0**	91.8
CKB+MMD	**87.5**	93.0	89.8	93.3	91.7	**92.9**	92.3	83.4	99.7	**92.8**	**85.8**	**100.0**	**91.9**

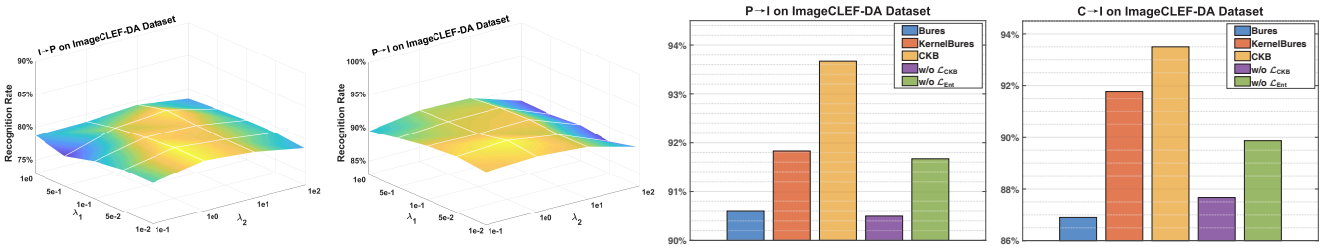
5.1. Results

Comparison. Several state-of-the-art UDA approaches are used to compare with the proposed methods, and the results are shown in Table 1-2. From the results on Office-Home in Table 1, we observe that the CKB+MMD method outperforms the compared methods in average accuracy, and the relaxed variant CKB also achieves the accuracy of 68.5%. The experiment results on ImageCLEF-DA are shown in the middle of Table 1. The CKB method improves the mean accuracy to 90.2% by further considering the discrepancy between the conditional distributions. The results show that the higher the accuracy of target predictions, the more effective the CKB alignment, *e.g.*, tasks **P** → **I** and **P** → **C**. Table 1 shows the results on Office10 dataset. OT-IT and KGOT methods achieve the accuracy of 87.1% and 89.7%, which show the superiority of the OT theory in distribution matching. CKB+MMD method achieves Top-1 accuracy in most tasks and improves the mean accuracy to 91.8%. Table 2 shows the results on digits recognition tasks. The proposed models surpass the advanced OT-based method ETD and achieves the highest accuracy in all tasks.

Table 2. Accuracies (%) on Digits (LeNet).

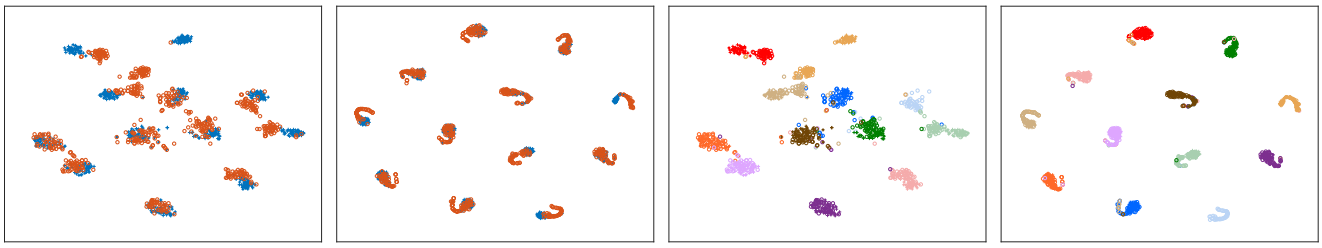
Method	M→U	U→M
Source [15]	82.2 ± 0.8	69.6 ± 3.8
DANN [9]	95.7 ± 0.1	90.0 ± 0.2
CyCADA [15]	95.6 ± 0.4	96.5 ± 0.2
DeepJDOT [4]	95.7	96.4
ETD [20]	96.4 ± 0.3	96.3 ± 0.1
CKB	96.3 ± 0.1	**96.6** ± 0.4
CKB+MMD	**96.6** ± 0.1	96.3 ± 0.1

Hyper-parameter. We investigate the selection of hyper-parameters λ_1 and λ_2 on ImageCLEF-DA dataset. The optimal λ_1 and λ_2 are respectively searched from [1e-2,5e-2,1e-1,5e-1,1e0] and [1e-1,1e0,1e1,1e2]. Figure 3 (a)-(b) show the results of grid search, we observe that the model is stable for different hyper-parameter values and (λ_1, λ_2)=(5e-1,1e0) is optimal among all settings.



(a) Hyper-parameter (b) Hyper-parameter (c) Ablation (d) Ablation

Figure 3. (a)-(b): Grid search for hyper-parameters λ_1 and λ_2. (c)-(d): Ablation analysis.



(a) Before Adaptation (b) After Adaptation (c) Before Adaptation (d) After Adaptation

Figure 4. Feature visualization of the source-only and CKB models via t-SNE [25] on Image-CLEF **C** → **I** task. '+': source domain, '∘': target domain. (a)-(b): Features colored by domains. (c)-(d): Features colored by classes.

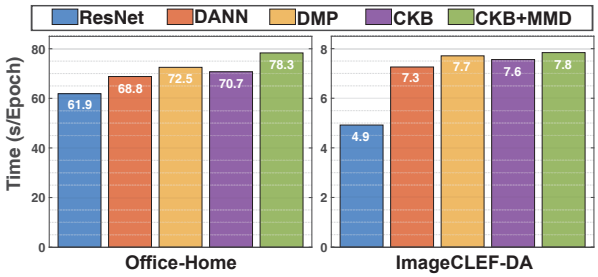


Figure 5. Time comparison.

Ablation. We compare the CKB metric with the Bures and Kernel Bures metrics [37], and evaluate the effectiveness of the loss terms in Eq. (6) on ImageCLEF-DA dataset. The model without CKB alignment loss and target entropy loss are abbreviated as w/o $\mathcal{L}_{\text{CKB}}$ and w/o $\mathcal{L}_{\text{Ent}}$, respectively. The results in Figure 3 (c)-(d) show that the CKB metric is superior to the Bures and Kernel Bures metric, which proves that the conditional operators help the model to obtain the discriminant information from the labels and predictions.

Visualization. To evaluate the aligned features quantitatively, we use t-SNE [25] to visualize the features of the source-only model (before adaptation) and the CKB model (after adaptation) on Image-CLEF **C** → **I** task. From Figure 4 (a), we observe that the conditional distribution is still shifted in the source-only model. In Figure 4 (b), all clusters are well-aligned by the CKB method. Figure 4 (c)-(d) show the features colored by classes, we observe that the CKB model achieves the inter-class separability and intra-class compactness on the target domain.

Time Comparison. We conduct the time comparison experiments on Office-Home and Image-CLEF-DA datasets. The results in Figure 5 suggest that CKB model is faster than CKB+MMD and DMP, which demonstrates that the conditional discrepancy metric is more efficient than the structure learning model DMP. As the proposed models are trained in mini-batch manner, the time complexity of the CKB metric is only about $\mathcal{O}(db_s^2)$, where b_s is the batch size. Thus the CKB metric does not introduce much complexity compared to the DNNs. Results show that CKB model only takes 10s longer than ResNet while improving the accuracy significantly by 22% on Office-Home dataset.

6. Conclusion

In this paper, we consider the conditional distribution shift problem in classification. Theoretically, we extend OT in RKHS by introducing the conditional variable, and prove that the proposed CKB metric defines a metric on the conditional distributions. An empirical estimation is derived to provide an explicit computation of the CKB metric, and its asymptotic theory is established for the consistency. By applying the CKB metric to DNNs, we propose a conditional distribution matching network which alleviates the shift of conditional distributions and preserves the intrinsic structures of both domains simultaneously. Extensive experimental results show the superiority of the proposed models in UDA problems.

Acknowledgement

This work is supported in part by the National Natural Science Foundation of China under Grants 61976229, 61906046, 11631015, and 12026601.

References

[1] Charles R Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, pages 137–144, 2007.

[3] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

[4] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *ECCV*, pages 447–463, 2018.

[5] Barbara Caputo, Henning Müller, Jesus Martinez-Gomez, Mauricio Villegas, Burak Acar, Novi Patricia, Neda Marvasti, Suzan Üsküdarlı, Roberto Paredes, Miguel Cazorla, et al. Imageclef 2014: Overview and analysis of the results. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 192–211, 2014.

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

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

[8] Kenji Fukumizu, Francis R Bach, Michael I Jordan, et al. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

[9] 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.

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

[11] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *ICML*, pages 2839–2848, 2016.

[12] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, pages 529–536, 2005.

[13] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13(3):723–773, 2012.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, pages 1989–1998, 2018.

[16] Guoliang Kang, Lu Jiang, Yunchao Wei, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for single-and multi-source domain adaptation. *IEEE TPAMI*, 2020.

[17] Ilja Klebanov, Ingmar Schuster, and TJ Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[19] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE TPAMI*, 2020.

[20] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Pengfei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *CVPR*, 2020.

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

[22] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, pages 1640–1650, 2018.

[23] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217, 2017.

[24] You-Wei Luo, Chuan-Xian Ren, Dai Dao-Qing, and Hong Yan. Unsupervised domain adaptation via discriminative manifold propagation. *IEEE TPAMI*, 2020.

[25] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11):2579–2605, 2008.

[26] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE TNN*, 22(2):199–210, 2010.

[27] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2009.

[28] Chuan-Xian Ren, Xiao-Lin Xu, and Hong Yan. Generalized conditional domain adaptation: A causal perspective with low-rank translators. *IEEE TCYB*, 50(2):821–834, 2018.

[29] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[30] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *ICML*, pages 961–968, 2009.

[31] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *JMLR*, 12(7), 2011.

[32] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016.

[33] Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *AAAI*, volume 34, pages 5940–5947, 2020.

[34] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017.

[35] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *ICML*, page 114, 2004.

[36] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *ICML*, pages 819–827, 2013.

[37] Zhen Zhang, Mianzhi Wang, and Arye Nehorai. Optimal transport in reproducing kernel hilbert spaces: Theory and applications. *IEEE TPAMI*, 2019.

[38] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *ICML*, pages 7523–7532, 2019.

[39] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. Deep subdomain adaptation network for image classification. *IEEE TNNLS*, 2020.