Optimal Transport in Reproducing Kernel Hilbert Space: Theory and Applications

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (JULY 2020)

Zhen Zhang, Mianzhi Wang, and Arye Nehorai
Reporter: Fengjiao Gong

June 9, 2022

Outline

Monge's Formulation

Find a **transport map** $T : \mathbb{R}^n \to \mathbb{R}^n$ that pushes μ to ν (denoted as $\boldsymbol{T_\# \mu = v}$) to minimize the total transport cost

$$\inf_{T_\# \mu = \nu} \int_{\mathbb{R}^n} \|\vec{x} - T(\vec{x})\|_2^2 d\mu(\vec{x}) \tag{1}$$

where $\mu, \nu \in \Pr(\mathbb{R}^n)$ are two probability measures, and $\Pr(\mathbb{R}^n)$ is the set of Borel probability measures on $\mathbb{R}^n$.

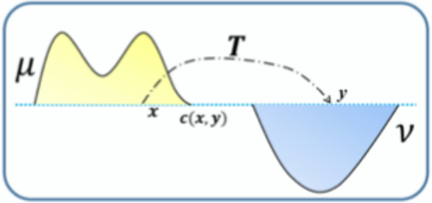


Figure 1: Illustration of the optimal transport problem.

Monge's Formulation

Find a **transport map** $T : \mathbb{R}^n \to \mathbb{R}^n$ that pushes μ to ν (denoted as $\boldsymbol{T_{\#}\mu = v}$) to minimize the total transport cost

$$\inf_{T_{\#}\mu=\nu} \int_{\mathbb{R}^n} \|\vec{x} - T(\vec{x})\|_2^2 d\mu(\vec{x}) \tag{1}$$

where

- $\Pr(\mathbb{R}^n)$ — set of Borel probability measures on $\mathbb{R}^n$
- $\mu, \nu \in \Pr(\mathbb{R}^n)$ — two probability measures

ISSUE: Existence of T cannot be guaranteed!

Kantorovich relaxation

Kantorovich's Formulation

Minimized over **all transport plans** instead of **transport map**

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\vec{x} - \vec{y}\|_2^2 d\pi(\vec{x}, \vec{y}), \tag{2}$$

- Transport plan $\pi(\vec{x}, \vec{y})$: joint probability measure describing the amount of mass transported from location $\vec{x}$ to location $\vec{y}$
- $\Pi(\mu, \nu)$: set of joint probability measures on $\mathbb{R}^n \times \mathbb{R}^n$, with marginals μ, ν
- Splitting — mass at one location can be divided and transported to multiple destinations

Kantorovich's Formulation

Minimized over **all transport plans** instead of **transport map**

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\vec{x} - \vec{y}\|_2^2 d\pi(\vec{x}, \vec{y}), \tag{2}$$

Wasserstein distance — metric on $\Pr\left(\mathbb{R}^n\right)$

$$d_{\mathrm{W}}(\mu, \nu) \triangleq \inf_{\pi \in \Pi(\mu,\upsilon)} \left[\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\vec{x} - \vec{y}\|_2^2 d\pi(\vec{x}, \vec{y})\right]^{\frac{1}{2}} \tag{3}$$

[reference] Cedric Villani, Topics in Optimal Transportation. Providence, RI, USA: American Mathematical Society, 2003, vol. 58.

OT between Gaussian Measures
Wasserstein distance

Remark 1. $(\cdot)^{\frac{1}{2}}$ denotes the principle matrix square root, i.e., for any positive semi-definite (PSD) matrix Σ, then $\mathbf{\Sigma}^{\frac{1}{2}} = (\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T)^{\frac{1}{2}} = \boldsymbol{U}\Lambda^{\frac{1}{2}}\boldsymbol{U}^T$.

[reference] D. Dowson and B. Landau, "The Fréchet distance between multivariate normal distributions," J. Multivariate Anal., vol. 12, no. 3, pp. 450–455, 1982.

OT between Gaussian Measures
Wasserstein distance

Theorem 1 . Let μ and ν be two probability measures on $\mathbb{R}^n$ with finite first and second order moments. Let $\vec{m}_\mu$ and $\vec{m}_\nu$, and $\mathbf{\Sigma}_\mu$ and $\mathbf{\Sigma}_\upsilon$ be the corresponding expectations and covariance matrices, respectively. Write

$$d_{\mathrm{GaW}}(\mu, \nu) = \left[\|\vec{m}_\mu - \vec{m}_\nu\|_2^2 + \mathrm{tr}\left(\mathbf{\Sigma}_\mu + \mathbf{\Sigma}_\upsilon - 2\mathbf{\Sigma}_{\mu\nu}\right) \right]^{\frac{1}{2}} \tag{4}$$

where $\mathbf{\Sigma}_{\mu\nu} = \left(\mathbf{\Sigma}_\mu^{\frac{1}{2}} \mathbf{\Sigma}_\nu \mathbf{\Sigma}_\mu^{\frac{1}{2}} \right)^{\frac{1}{2}}$. Then,

1) $d_{\mathrm{GaW}}(\mu, \nu) \leq d_{\mathrm{W}}(\mu, \nu)$, and

2) The equality will be valid if both μ and ν are Gaussian.

OT between Gaussian Measures
Wasserstein distance

Theorem 1 . Let μ and ν be two probability measures on $\mathbb{R}^n$ with finite first and second order moments. Let $\vec{m}_\mu$ and $\vec{m}_\nu$, and $\mathbf{\Sigma}_\mu$ and $\mathbf{\Sigma}_\upsilon$ be the corresponding expectations and covariance matrices, respectively. Write

$$d_{\mathrm{GaW}}(\mu, \nu) = \left[\|\vec{m}_\mu - \vec{m}_\nu\|_2^2 + \mathrm{tr}\left(\mathbf{\Sigma}_\mu + \mathbf{\Sigma}_\upsilon - 2\mathbf{\Sigma}_{\mu\nu}\right)\right]^{\frac{1}{2}} \tag{5}$$

- $d_{\mathrm{GaW}}(\mu, \nu)$: metric on set of all Gaussian measures — both Gaussian
- $d_{\mathrm{GaW}}(\mu, \mu)$: metric on covariance matrices — same expectations

Bures metric

OT between Gaussian Measures
Wasserstein distance

Corollary 1. Let $\mathrm{Sym}^+(n)$ be the set of all positive semi-definite matrices of size $n \times n$. Then,

$$d_{\mathrm{B}}\left(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2\right) = \left[\mathrm{tr}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\boldsymbol{\Sigma}_{12}\right)\right]^{\frac{1}{2}} \tag{6}$$

defines a metric on $\mathrm{Sym}^+(n)$.

[reference] R. Bhatia, T. Jain, and Y. Lim, "On the Bures Wasserstein distance between positive definite matrices," Expositiones Mathematicae, 2018.

OT between Gaussian Measures
Optimal transport map

Theorem 2 . Let μ and ν be two Gaussian measures on $\mathbb{R}^n$ whose covariance matrices are of full rank. Let $\vec{m}_\mu$ and $\vec{\boldsymbol{m}}_\upsilon$, and $\boldsymbol{\Sigma}_\mu$ and $\boldsymbol{\Sigma}_\upsilon$, denote the respective expectations and covariance matrices. Then the optimal transport map T_{G} between μ and ν exists, and can be written as

$$T_{\mathrm{G}}(\vec{\boldsymbol{x}}) = \boldsymbol{\Sigma}_\mu^{-\frac{1}{2}} \boldsymbol{\Sigma}_{\mu\nu} \boldsymbol{\Sigma}_\mu^{-\frac{1}{2}} \left(\vec{\boldsymbol{x}} - \vec{m}_\mu\right) + \vec{m}_\upsilon \tag{7}$$

Covariance matrices are *full-rank*.

[reference] R. Bhatia, T. Jain, and Y. Lim, "On the Bures Wasserstein distance between positive definite matrices," Expositiones Mathematicae, 2018.

OT between Gaussian Measures
Optimal transport map

Remark 2. "$\dagger$" denotes the Moore-Penrose inverse. $\mathrm{Im}(\Sigma)$ denotes the image of the linear transform Σ , i.e., $\mathrm{Im}(\boldsymbol{\Sigma}) = \{\boldsymbol{\Sigma}\vec{x}, \vec{x} \in \mathbb{R}^n\}$.

[reference]

https://en.wikipedia.org/wiki/Moore%E2%80%93Penrose_inverse

M. Gelbrich, "On a formula for the L2 Wasserstein metric between measures on Euclidean and hilbert spaces," Mathematische Nachrichten, vol. 147, no. 1, pp. 185–203, 1990.

OT between Gaussian Measures
Optimal transport map

Theorem 3. Let μ and ν be two Gaussian measures defined on $\mathbb{R}^n$. Let $\bar{\mu}$ and $\bar{\nu}$ be the corresponding centered Gaussian measures which are derived from μ and ν , respectively, by translation. Let $\boldsymbol{P}_\mu$ be the projection matrix onto $\mathrm{Im}\left(\boldsymbol{\Sigma}_\mu\right)$. Then the optimal transport map T_{G} from $\bar{\mu}$ to $P_{\mu\#\bar{\nu}}$ is linear and self-adjoint, and can be written as

$$T_{\mathrm{G}}(\overrightarrow{\boldsymbol{x}}) = \left(\boldsymbol{\Sigma}_\mu^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{\Sigma}_{\mu\nu} \left(\boldsymbol{\Sigma}_\mu^{\frac{1}{2}}\right)^{\dagger} \overrightarrow{\boldsymbol{x}} \tag{8}$$

Covariance matrix is *rank-deficient* ($\bar{\mu} \Rightarrow P_{\mu\#}\bar{\nu}$).

Optimal Transport in Reproducing Kernel Hilbert Space

1. Reproducing Kernel Hilbert Space(RKHS)
2. Kantorovich'S OT in RKHS
3. OT between Gaussian Measures in RKHS
4. Applications
   4.1 Image Classification
   4.2 Domain Adaptation
5. Experiments

Reproducing Kernel Hilbert Space(RKHS)

- **RKHS**

Function $k : \mathcal{X} \times \mathcal{X}$ is called a reproducing kernel of $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space, if k satisfies:
1) $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}$,
2) $\forall x \in \mathcal{X}, f \in \mathcal{H}, \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.
where $\mathcal{H}$ is a Hilbert space of $\mathbb{R}$-valued functions defined on nonempty set $\mathcal{X}$.

- **Feature map**

Implicit feature map $\phi : \mathcal{X} \to \mathcal{H}$

$$\phi(x) = k(\cdot, x)$$

Then

$$\langle \phi(x), \phi(y)\rangle_{\mathcal{H}} = k(x, y), \forall x, y \in \mathcal{X}$$

Kantorovich'S OT in RKHS
OT Formulation

Given $\mu, \nu \in \Pr(\mathcal{X})$, the Kantorovich optimal transport between pushforward measures $\phi_{\#}\mu$ and $\phi_{\#}\nu$ on $\mathcal{H}_{\mathcal{K}}$ is written as

$$d_{\mathrm{W}}\left(\phi_{\#}\mu, \phi_{\#}v\right) = \left[\inf_{\pi_{\mathcal{K}} \in \Pi(\phi_{\#}\mu, \phi_{\#}v)} \int_{\mathcal{H}_{\mathcal{K}} \times \mathcal{H}_{\mathcal{K}}} \|u - v\|^2_{\mathcal{H}_{\mathcal{K}}} d\pi_{\mathcal{K}}(u, v)\right]^{\frac{1}{2}} \tag{9}$$

where

- $\Pi\left(\phi_{\#}\mu, \phi_{\#}v\right)$: set of joint probability measures on $\mathcal{H}_{\mathcal{K}} \times \mathcal{H}_{\mathcal{K}}$ with marginals $\phi_{\#}\mu, \phi_{\#}\nu$
- k: positive definite kernel on $\mathcal{X} \times \mathcal{X}$
- $(\mathcal{H}_{\mathcal{K}}, \mathcal{B}_{\mathcal{H}_{\mathcal{K}}})$: reproducing kernel Hilbert space generated by k

Theorem 4. Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a Borel space, and let the reproducing kernel k be measurable. Given $\mu, \nu \in \Pr(\mathcal{X})$, we write

$$d_{\mathrm{W}}^{\mathcal{H}}(\mu, v) = \left[\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} d^2(x, y) d\pi(x, y) \right]^{\frac{1}{2}} \tag{10}$$

where $d^2(x, y) = \|\phi(x) - \phi(y)\|^2_{\mathcal{H}_{\mathcal{K}}} = k(x, x) + k(y, y) - 2k(x, y)$. Then,
1) $d_{\mathrm{W}}^{\mathcal{H}}(\mu, \nu) = d_{\mathrm{W}}\left(\phi_{\#}\mu, \phi_{\#}v\right)$, and
2) If π^* is a minimizer of (10) , then $(\phi, \phi)_{\#}\pi^*$ is a minimizer of (9), where

$$(\phi, \phi)(x, y) = (\phi(x), \phi(y))$$

and $(\phi, \phi) : \mathcal{X} \times \mathcal{X} \to \mathcal{H}_{\mathcal{K}} \times \mathcal{H}_{\mathcal{K}}$

Discrete OT

Discrete version of (10) can be written as

$$\operatorname*{Min}_{P \in U_{nm}} \operatorname{tr}\left(P^T \boldsymbol{D}\right) \tag{11}$$

where

- U_{nm} — set of $n \times m$ non-negative matrices representing probabilistic couplings with marginals $\hat{\mu}, \hat{\upsilon}$,

$$U_{nm} = \left\{ \boldsymbol{P} \in \mathbb{R}_{+}^{n \times m} \mid \boldsymbol{P}\overrightarrow{\mathbf{1}}_m = \overrightarrow{\hat{\boldsymbol{\mu}}}, \boldsymbol{P}^T\overrightarrow{\mathbf{1}}_n = \overrightarrow{\hat{\boldsymbol{\upsilon}}} \right\}$$

- $\boldsymbol{D}$ cost matrix

$$\boldsymbol{D}_{i,j} = k\left(x_i, x_i\right) + k\left(y_j, y_j\right) - 2k\left(x_i, y_j\right)$$

- Empirical histograms:

$$\hat{\mu} = \sum_{i=1}^{n} \hat{\mu}_i \delta_{x_i}, \quad \hat{\upsilon} = \sum_{j=1}^{m} \hat{\upsilon}_j \delta_{y_j}$$

OT between Gaussian Measures in RKHS

Let μ be a Borel probability measure on $\mathcal{X}$, and

- Mean: $m_\mu = E_{X\sim\mu}(\phi(X))$
- Covariance operator: $R_\mu = E_{X\sim\mu}\left((\phi(X) - m_\mu) \otimes (\phi(X) - m_\mu)\right)^2$

exist and be bounded with respect to the Hilbert norm and HilbertSchmidt norm, respectively.

Tensor product of $\mathcal{H}$:

$$(u \otimes v)w = \langle v, w\rangle_{\mathcal{H}} u, \quad \forall u, v, w \in \mathcal{H}$$

[reference] A. Gretton, O. Bousquet, A. Smola, and B. Scho€lkopf, "Measuring statistical dependence with hilbert-schmidt norms," in Proc. 16th Int. Conf. Algorithmic Learn. Theory, 2005, pp. 63–77.

OT between Gaussian Measures in RKHS

Remark 3 .
(1) The square root of an nonnegative, self-adjoint, and compact operator G is defined as $G^{\frac{1}{2}} = \sum_{i=1} \sqrt{\lambda_i(G)}\varphi_i(G) \otimes \varphi_i(G)$, where $\lambda_i(G)$ and $\varphi_i(G)$ are eigenvalues and eigenfunctions of G .
(2) The trace of an trace-class operator G on a separable Hilbert space $\mathcal{H}$ is defined as $\mathrm{tr}(G) = \sum_{i=1}^{\dim(\mathcal{H})} \langle Ge_i, e_i \rangle$, where $\{e_i\}_{i=1}^{\dim(\mathcal{H})}$ is an orthonormal system of $\mathcal{H}$.

If data distributions in RKHS (the corresponding pushforward measures) are Gaussian, the conclusions in RKHS are similar to the ones in Euclidean spaces.

OT between Gaussian Measures in RKHS
KGW distance

Proposition 1. Assume that the hypotheses in Theorem 4 hold. Let $\mu, \nu \in \Pr(\mathcal{X})$. Let m_μ and m_ν , and R_μ and R_ν , be the corresponding means and covariance operators, respectively. Write

$$d_{\mathrm{GW}}^{\mathcal{H}}(\mu, v) = \left[\|m_\mu - m_v\|_{\mathcal{H}_\mathcal{K}}^2 + \operatorname{tr}\left(R_\mu + R_v - 2R_{\mu\nu}\right)\right]^{\frac{1}{2}} \tag{12}$$

where $R_{\mu\nu} = \left(R_\mu^{\frac{1}{2}} R_\nu R_\mu^{\frac{1}{2}}\right)^{\frac{1}{2}}$. Then,

(1) $d_{\mathrm{GW}}^{\mathcal{H}}(\mu, v) \leq d_{\mathrm{W}}^{\mathcal{H}}(\mu, v)$, and

(2) The equality will be valid if both $\phi_{\#}\mu$ and $\phi_{\#}\nu$ are Gaussian.

OT between Gaussian Measures in RKHS

Corollary 2. Let $\mathrm{Sym}^+(\mathcal{H}_\mathcal{K}) \subseteq \mathcal{H}_\mathcal{K} \otimes \mathcal{H}_\mathcal{K}$ be the set of nonnegative, self-adjoint, and trace-class operators in $\mathcal{H}_\mathcal{K}$. Then

$$d_{\mathrm{B}}^{\mathcal{H}}(R_1, R_2) = \left[\mathrm{tr}\left(R_1 + R_2 - 2R_{12}\right)\right]^{\frac{1}{2}} \tag{13}$$

defines a metric on $\mathrm{Sym}^+(\mathcal{H}_\mathcal{K})$.

$d_{\mathrm{B}}^{\mathcal{H}}$ *quantifies the difference between the dispersions of data in RKHS.*
$d_{\mathrm{B}}^{\mathcal{H}}$ *quantifies the discrepancy between distributions*

OT between Gaussian Measures in RKHS

Definition 1. Let $\mu \in \Pr(\mathcal{X})$. If there exist disjoint subsets Ω_1, Ω_2, and Ω_3, satisfying $\mathcal{X} = \Omega_1 \cup \Omega_2 \cup \Omega_3$, and $\mu\left(\Omega_1\right), \mu\left(\Omega_2\right), \mu\left(\Omega_3\right) > 0$, then we say μ satisfies the 3-splitting property.

Here $\Pr^s(\mathcal{X})$ is the set of Borel measures satisfying the 3-splitting property.

OT between Gaussian Measures in RKHS

Theorem 5. Let the measurable space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be locally compact and Hausdorff. Let k be a c_0-universal reproducing kernel. Then, the embedding $\mu \to R_\mu, \forall \mu \in \mathrm{Pr}^s(\mathcal{X})$ is injective.

$d_{\mathrm{B}}^{\mathcal{H}}$ (KB) induces a metric on $\mathrm{Pr}^s(\mathcal{X})$.

[reference]
`https://en.wikipedia.org/wiki/Compact_space`
`https://en.wikipedia.org/wiki/Hausdorff_space`

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet, "Universality, characteristic kernels and RKHS embedding of measures," J. Mach. Learn. Res., vol. 12, pp. 2389–2410, Jul. 2011.

Proposition 2. Given $\mu, \nu \in \Pr(\mathcal{X})$, assume the pushforward measures $\phi_{\#}\mu$ and $\phi_{\#}\nu$ on RKHS are Gaussian. Let $\bar{\mu}_\phi$ and $\bar{\nu}_\phi$ be the respective centered measures of $\phi_{\#}\mu$ and $\phi_{\#}\nu$. Let P_μ be the projection operator on $\mathrm{Im}\,(R_\mu)$. Then the kernel Gauss-optimal transport map $T_{\mathrm{G}}^{\mathcal{H}}$ between $\bar{\mu}_\phi$ and $P_{\mu\#}\,(\bar{v}_\phi)$ is a linear and self-adjoint operator, and can be written as

$$T_{\mathrm{G}}^{\mathcal{H}}(u) = \left(R_\mu^{\frac{1}{2}}\right)^{\dagger} R_{\mu\nu} \left(R_\mu^{\frac{1}{2}}\right)^{\dagger} u, \forall u \in \mathcal{H}_{\mathcal{K}} \tag{14}$$

OT between Gaussian Measures in RKHS

Sample matrices

$$\boldsymbol{X} = [x_1, x_2, \ldots, x_n], \quad \boldsymbol{Y} = [y_1, y_2, \ldots, y_m]$$

Mapped data matrices

$$\Phi_X = [\phi(x_1), \phi(x_2), \ldots, \phi(x_n)], \quad \Phi_Y = [\phi(y_1), \phi(y_2), \ldots, \phi(y_m)]$$

Kernel matrices

$$(K_{XX})_{ij} = k(x_i, x_j), \quad (K_{XY})_{ij} = k(x_i, y_j), \quad (K_{YY})_{ij} = k(y_i, y_j)$$

Centering matrices

$$\boldsymbol{H}_n = \boldsymbol{I}_{n\times n} - \frac{1}{n}\vec{\mathbf{1}}_n \vec{1}_n^T \quad , \quad \boldsymbol{H}_m = \boldsymbol{I}_{m\times m} - \frac{1}{m}\vec{\mathbf{1}}_m \vec{\mathbf{1}}_m^T$$

OT between Gaussian Measures in RKHS

Estimated mean:

$$\hat{m}_\mu = \frac{1}{n}\Phi_X \overrightarrow{1}_n, \quad \hat{m}_v = \frac{1}{m}\Phi_Y \overrightarrow{1}_m$$

Estimated covariance operators:

$$\hat{R}_\mu = \frac{1}{n}\Phi_X H_n \Phi_X^T, \quad \hat{R}_v = \frac{1}{m}\Phi_Y H_m \Phi_Y^T$$

Remark 4. $\|\cdot\|_*$ denotes the nuclear norm, i.e., $\|A\|_* = \sum_{i=1}^r \sigma_i(A)$, where $\sigma_i(A)$ are the singular values of matrix A

KGW Distance Empirical Estimation

Proposition 3. The empirical kernel Gauss-Wasserstein distance is

$$\hat{d}_{\mathrm{GW}}^{\mathcal{H}}(\mu,\nu) = \left[\frac{1}{n}\operatorname{tr}\left(\boldsymbol{K}_{XX}\right) + \frac{1}{m}\operatorname{tr}\left(K_{YY}\right) - \frac{2}{mn}\vec{1}_n^T \boldsymbol{K}_{XY}\vec{\mathbf{1}}_m - \frac{2}{\sqrt{mn}}\left\|\boldsymbol{H}_n\boldsymbol{K}_{XY}\boldsymbol{H}_m\right\|_*\right]^{\frac{1}{2}} \tag{15}$$

The kernel Bures distance between $\hat{R}_\mu$ and $\hat{R}_\nu$ is

$$d_{\mathrm{B}}^{\mathcal{H}}\left(\hat{R}_\mu,\hat{R}_\nu\right) = \left[\frac{1}{n}\operatorname{tr}\left(\boldsymbol{K}_{XX}\boldsymbol{H}_n\right) + \frac{1}{m}\operatorname{tr}\left(\boldsymbol{K}_{YY}\boldsymbol{H}_m\right) - \frac{2}{\sqrt{mn}}\left\|\boldsymbol{H}_n\boldsymbol{K}_{XY}\boldsymbol{H}_m\right\|_*\right]^{\frac{1}{2}} \tag{16}$$

Empirical Estimation of the KGOT Map

Proposition 4. Let X and Y be data matrices sampled from μ and ν , respectively. Then the empirical projection operator on $\operatorname{Im}\left(\hat{R}_\mu\right)$ is

$$\hat{P}_\mu = \Phi_X \boldsymbol{H}_n \boldsymbol{C}_{XX}^{\dagger} \boldsymbol{H}_n \Phi_X^T \tag{17}$$

and the empirical Gauss-optimal transport map from $\bar{\mu}_\phi$ and $P_{\mu\#}\left(\bar{v}_\phi\right)$ is

$$\hat{T}_{\mathrm{G}}^{\mathcal{H}} = \sqrt{\frac{n}{m}} \Phi_X H_n C_{XX}^{\dagger} C_{XYYX}^{\frac{1}{2}} C_{XX}^{\dagger} H_n \Phi_X^T \tag{18}$$

where

$$\begin{gathered} \boldsymbol{C}_{XX} = \boldsymbol{H}_n \boldsymbol{K}_{XX} \boldsymbol{H}_n \\ \boldsymbol{C}_{XYYX} = \boldsymbol{H}_n \boldsymbol{K}_{XY} \boldsymbol{H}_m \boldsymbol{K}_{YX} \boldsymbol{H}_n \end{gathered} \tag{19}$$

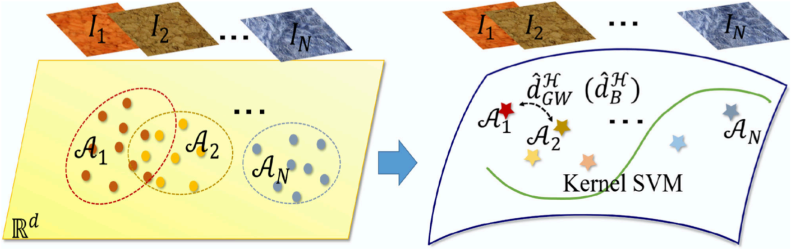
OT between Gaussian Measures in RKHS
KGOT Map

Proposition 5 .

$$\hat{T}_{\mathrm{G}}^{\mathcal{H}} \hat{R}_{\mu} \hat{T}_{\mathrm{G}}^{\mathcal{H}} = \hat{P}_{\mu} \hat{R}_{\nu} \hat{P}_{\mu} \tag{20}$$

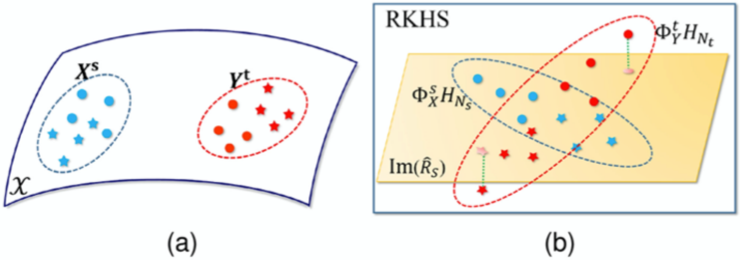
Align covariance operators in RKHS.

Application — Image Classification



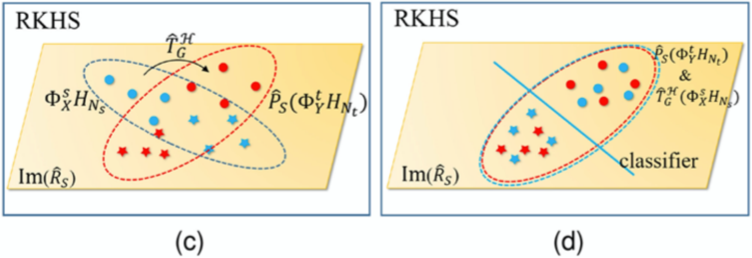
- Represent each image I_i by a collection of feature samples A_i.
- Compute the KGW (or the KB) distances between any pair of images.
- Apply kernel SVM to conduct classification.

Application – Domain Adaptation



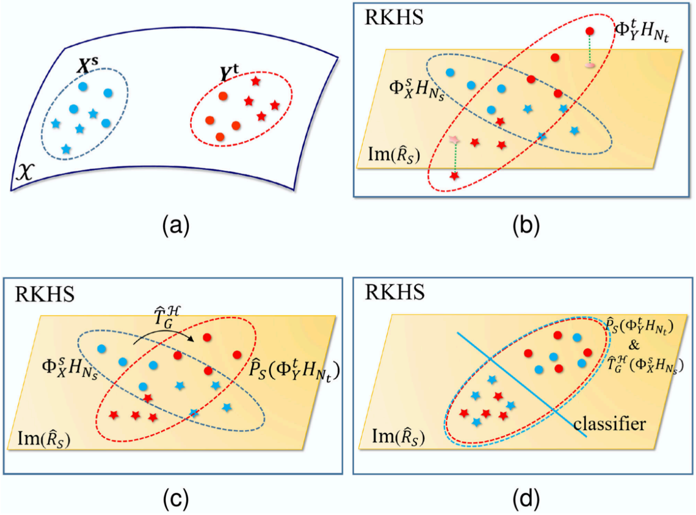
(a) The labeled dataset, $\boldsymbol{X}^s$, in the source domain and the unlabeled dataset, Y^t , in the target domain. Dots and stars represent different classes;
(b) Map X^s and Y^t to the RKHS $\mathcal{H}_{\mathcal{K}}$, and centralize the mapped data. (The centered source dataset $\Phi_X^s H_{N_s}$ lies in $\mathrm{Im}\left(\hat{R}_s\right)$);

Application – Domain Adaptation



(c) Project the target dataset $\Phi_Y^t H_{N_t}$ onto $\mathrm{Im}\left(\hat{R}_s\right)$. The projection is $\hat{P}_s\left(\Phi_Y^t H_{N_t}\right)$;
(d) Apply the KGOT map to transport the source data to the target domain. The transported data is $\hat{T}_{\mathrm{G}}^{\mathcal{H}}\left(\Phi_X^s H_{N_s}\right)$.

Domain Adaptation



Finally, train a classifier using $\hat{T}_{\mathrm{G}}^{\mathcal{H}}\left(\Phi_X^s H_{N_s}\right)$, then apply to $\hat{P}_s\left(\Phi_Y^t \boldsymbol{H}_{N_t}\right)$.

Experiments

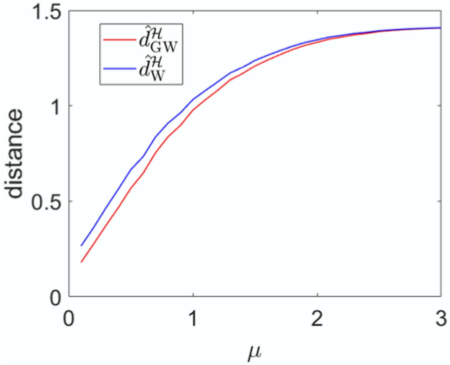
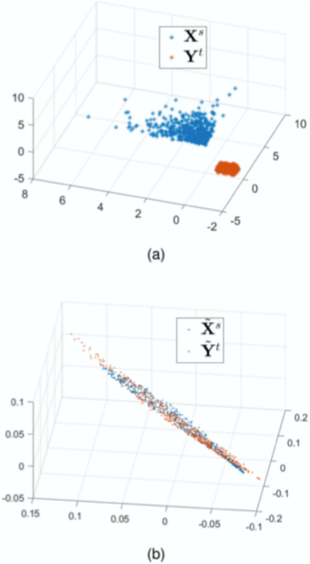


Figure 2: The estimated KGW and KW distances between Gaussian distributions $N(m\vec{1}, I)$ and $N(-m\vec{1}, I)$.

Clearly, KGW is less than KW.

Experiments



(a) Source dataset X^s , and target dataset Y^t ;
(b) Representations of datasets $\hat{T}^{\mathcal{H}}_{\mathrm{G}}\left(\Phi^s_X H_{N_s}\right)$ and $\hat{P}_s\left(\Phi^t_Y H_{N_t}\right)$ under coordinate sysmtem $(l_i)_{i=1}^3$.

Distributions of $\tilde{X^s}$ and $\tilde{Y^t}$ are close to each other.

Thanks!