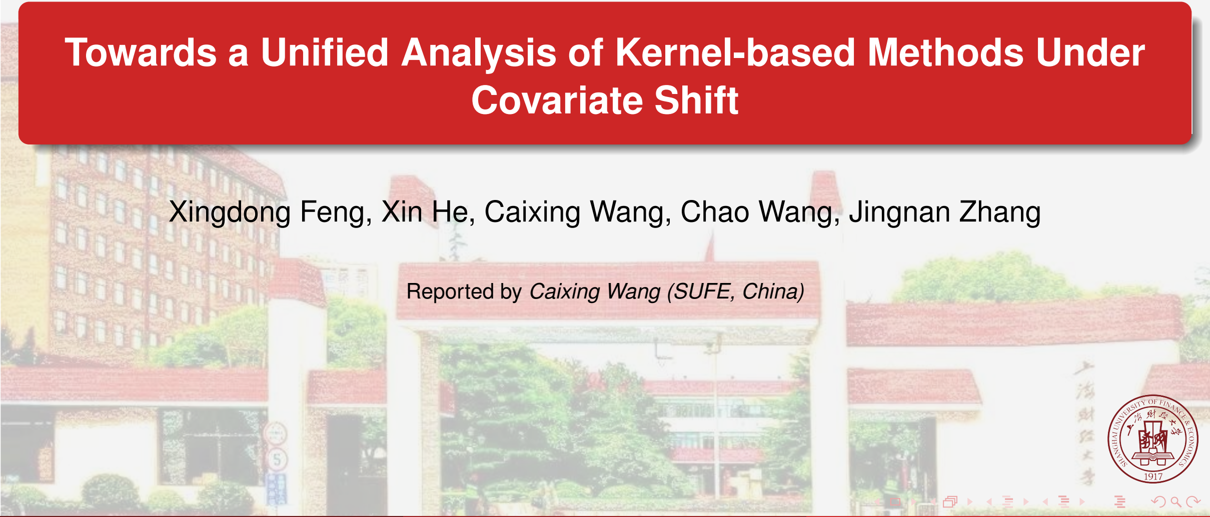
Towards a Unified Analysis of Kernel-based Methods Under Covariate Shift

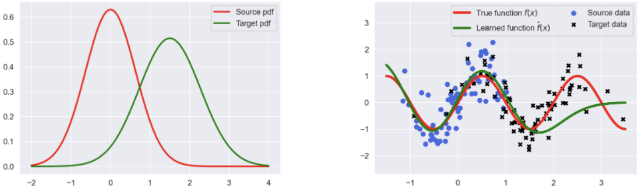


Xingdong Feng, Xin He, Caixing Wang, Chao Wang, Jingnan Zhang

Reported by *Caixing Wang (SUFE, China)*

Background

Covariate shift is a phenomenon that commonly occurs in machine learning, where the distribution of input features (covariates) changes between the source (or training) and target (or test) data, while the conditional distribution of output values given covariates remains unchanged.



(a)

(b)

Figure 1: (a) The probability density functions of normal distributions with $\mu_1 = 0, \sigma^2 = 0.4$ that the source data is driven from and $\mu_1 = 1.5, \sigma^2 = 0.6$ that the target data is driven from, respectively; (b) the learned function trained by using the source data and the true mean regression function. Note that the considered example serves as an illustration that satisfies case (ii) in Section 2.3.

Motivation

- The prediction performance can be largely degraded since the predictive function has not been trained on data that accurately represents the target environment.
- Compared to the well-studied supervised learning without such a distribution mismatch, there still exists some gap in both theoretically and numerically understanding the influence of the covariate shift under various kernel-based learning problems.

Our contribution

We propose a unified analysis of the kernel-based methods under covariate shift, which provides an insightful understanding of the influences of covariate shift on the kernel-based methods both theoretically and numerically.

- Theoretically, we show that the unweighted estimator achieves the optimal learning rates under the uniformly bounded case. Yet, the unweighted estimator is sub-optimal under the bounded second moment case. Then, we construct a weighted estimator by using an appropriate truncated ratio, which again attains a sharp convergence rate.
- Numerous experiments on synthetic data and multi-source real data with various loss functions confirm our theoretical findings.

Method

- **Classical kernel-based nonparametric estimation**

$$\widehat{f} := \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} L\big(y_i^S, f(\mathbf{x}_i^S)\big) + \lambda \|f\|_K^2$$

- **Importance ratio weighted (IRW) kernel-based nonparametric estimation**

$$\widetilde{f}^{\phi} := \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i^S) L\big(y_i^S, f(\mathbf{x}_i^S)\big) + \lambda \|f\|_K^2$$

where $\phi(\mathbf{x}) = \rho_{\mathbf{x}}^T(\mathbf{x})/\rho_{\mathbf{x}}^S(\mathbf{x})$ is the importance ratio measuring the discrepancy between distributions

Method

- **Two types of importance ratio cases**
    1. $\phi(\mathbf{x})$ is α-uniformly bounded that is $\sup_{\mathbf{x}\in\mathcal{X}}\phi(\mathbf{x})\leq\alpha$, for some positive constant α;
    2. $\phi(\mathbf{x})$'s second moment is bounded that is $E_{\mathbf{x}\sim S}[\phi^2(\mathbf{x})]\leq\beta^2$, for some constant $\beta^2\geq 1$.

Theory

Theorem 1 (Convergence rate of $\widehat{f}$ for case 1)

Under Assumptions 1-2, if the importance ratio is α-uniformly bounded, let $\lambda > c_0\delta_n^2/4$ with δ_n being the smallest positive solution to $C\sqrt{\log n}R(\sqrt{\alpha}\delta) \leq c_0\delta^2/2$, then for some constant $c_1 > 0$, with probability at least $1 - n^{-c_1}$, we have

$$\|\widehat{f} - f^*\|_T^2 \leq \alpha\left(\delta_n^2 + 2c_0^{-1}\lambda\right). \tag{1}$$

Furthermore, based on (1), we have

$$\mathcal{E}_T^L(\widehat{f}) - \mathcal{E}_T^L(f^*) \leq c_L\alpha\left(\delta_n^2 + 2c_0^{-1}\lambda\right)^{1/2}. \tag{2}$$

Theory

Theorem 2 (Convergence rate of $\widehat{f}$ for case 2)

Under Assumptions 1-2, if the importance ratio satisfies that $E_{\mathbf{x}\sim S}[\phi^2(\mathbf{x})] \leq \beta^2$, let $\lambda > c_0\delta_n^2/4$ with δ_n being the smallest positive solution to $C\sqrt{\log n}R((c_0^{-1}c_L\sqrt{\beta^2}\delta)^{1/2}) \leq c_0\delta^2/2$, then for some constant $c_2 > 0$, with probability at least $1 - n^{-c_2}$, we have

$$\|\widehat{f} - f^*\|_T^2 \leq c_0^{-1}c_L\sqrt{\beta^2}\left(\delta_n^2 + 2c_0^{-1}\lambda\right)^{1/2}. \qquad (3)$$

Furthermore, based on (2), we have

$$\mathcal{E}_T^L(\widehat{f}) - \mathcal{E}_T^L(f^*) \leq c_L\sqrt{\beta^2}\left(\delta_n^2 + 2c_0^{-1}\lambda\right)^{1/2}. \qquad (4)$$

Theory

Theorem 3 (Convergence rate of $\widehat{f}^{\phi}$ for case 2)

Under Assumptions 1-2, if the importance ratio satisfies that $E_{\mathbf{x}\sim S}[\phi^2(\mathbf{x})] \leq \beta^2$, let $\lambda > c_0\delta_n^2/4$ with δ_n being the smallest positive solution to $C\sqrt{\beta^2}\log nR(\delta) \leq c_0\delta^2/2$, and set the truncation level $\gamma_n = \sqrt{n\beta^2}$, then for some constant $c_3 > 0$, with probability at least $1 - n^{-c_3}$, we have

$$\|\widehat{f}^{\phi} - f^*\|_T^2 \leq \delta_n^2 + 2c_0^{-1}\lambda. \tag{5}$$

Furthermore, based on (5), we have

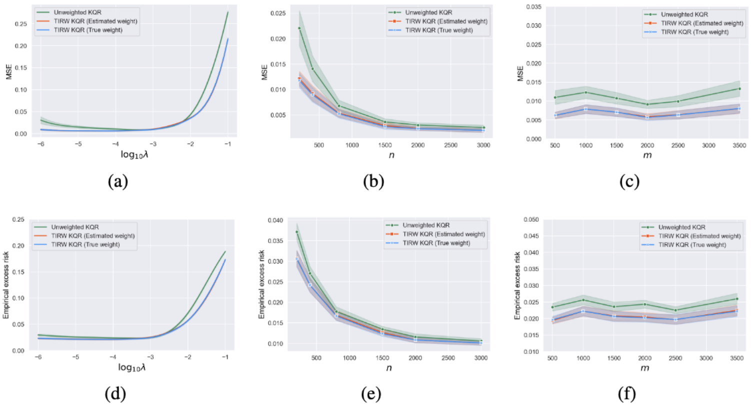
$$\mathcal{E}_T^L(\widehat{f}^{\phi}) - \mathcal{E}_T^L(f^*) \leq \frac{1}{2}c_0\delta_n^2 + 2\lambda. \tag{6}$$

Examples

Table 1: Established convergence rates under different cases

Kernel class	Uniformly bounded case		Moment bounded case	
	Unweighted estimator	TIRW estimator	Unweighted estimator	TIRW estimator
Finite rank D	$O_P(\frac{\alpha^2 D \log n}{n})$	$O_P(\frac{\alpha D \log^2 n}{n})$	$O_P((\frac{\beta^4 D \log n}{n})^{1/3})$	$O_P(\frac{\beta^2 D \log^2 n}{n})$
Polynomial decay	$O_P((\frac{\alpha^2 \log n}{n})^{\frac{2r}{2r+1}})$	$O_P((\frac{\alpha \log^2 n}{n})^{\frac{2r}{2r+1}})$	$O_P((\frac{\beta^4 \log n}{n})^{\frac{2r}{6r+1}})$	$O_P((\frac{\beta^2 \log^2 n}{n})^{\frac{2r}{2r+1}})$
Exponential decay	$O_P(\frac{\alpha^2 \log^2 n}{n})$	$O_P(\frac{\alpha \log^3 n}{n})$	$O_P((\frac{\beta^4 \log^2 n}{n})^{1/3})$	$O_P(\frac{\beta^2 \log^3 n}{n})$

Simulation



(a) (b) (c)

(d) (e) (f)

Figure 2: Averaged MSE and empirical excess risk for unweighted KQR, TIRW KQR with true weight and estimated weight, respectively. Note that in (a) and (d), the curves are plotted with respect to $\log_{10}\lambda$ with $n = 500, m = 1000$; in (b) and (e) the curves are plotted with respect to n with fixed $m = 1000, \lambda = 10^{-4}$; in (c) and (f), the curves are plotted with respect to m with fixed $n = 500, \lambda = 10^{-4}$.

Simulation

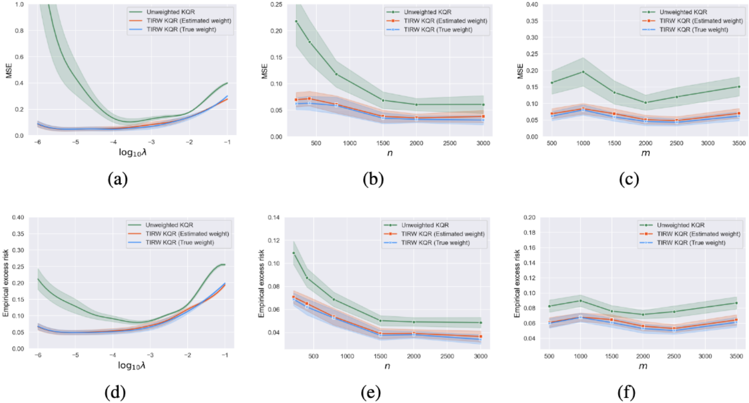


Figure 3: Averaged MSE and empirical excess risk for unweighted KQR, TIRW KQR with true weight and estimated weight, respectively. Note that in (a) and (d), the curves are plotted with respect to $\log_{10}\lambda$ with $n = 500, m = 1000$; in (b) and (e) the curves are plotted with respect to n with fixed $m = 1000, \lambda = 10^{-4}$; in (c) and (f), the curves are plotted with respect to m with fixed $n = 500, \lambda = 10^{-4}$.

Acknowledgement

Thank you all for your attention!