# Robustified Importance Sampling for Covariate Shift

**Anonymous authors**
Paper under double-blind review

## Abstract

In many learning problems, the training and testing data follow different distributions and a particularly common situation is the *covariate shift*. To correct for sampling biases, most approaches, including the popular kernel mean matching (KMM), focus on estimating the importance weights between the two distributions. Reweighting-based methods, however, are exposed to high variance when the distributional discrepancy is large. On the other hand, the alternate approach of using nonparametric regression (NR) incurs high bias when the training size is limited. In this paper, we propose and analyze a new estimator that systematically integrates the residuals of NR with KMM reweighting, based on a control-variate perspective. The proposed estimator is shown to either outperform or match the best-known existing rates for both KMM and NR, and thus is a robust combination of both estimators. The experiments shows our estimator works well in practice.

## 1 Introduction

Traditional machine learning implicitly assumes training and test data are drawn from the same distribution. However, mismatches between training and test distributions occur frequently in reality. For example, in clinical trials the patients used for prognostic factor identification may not come from the target population due to sample selection bias (Huang et al. (2007); Gretton et al. (2009)); incoming signals used for natural language and image processing, bioinformatics or econometric analyses change in distribution over time and seasonality (Sugiyama et al. (2007); Jiang & Zhai (2007); Quionero-Candela et al. (2009); Tzeng et al. (2017); Borgwardt et al. (2006); Heckman (1979); Zadrozny (2004)); patterns for engineering controls fluctuate due to the non-stationarity of environments (Sugiyama & Kawanabe (2012); Hachiya et al. (2008)).

Many such problems are investigated under the *covariate shift* assumption (Shimodaira (2000)). Namely, in a supervised learning setting with covariate $X$ and label $Y$, the marginal distribution of $X$ in the training set $P_{tr}(x)$, shifts away from the marginal distribution of the test set $P_{te}(x)$, while the conditional distribution $P(y|x)$ remains invariant in both sets. Because test labels are either too costly to obtain or unobserved, it could be uneconomical or impossible to build predictive models only on the test set. In this case, one is obliged to utilize the invariance of conditional probability to adapt or transfer knowledge from the training set, termed as transfer learning (Pan & Yang (2009)) or domain adaptation (Jiang & Zhai (2007); Blitzer et al. (2006)). Intuitively, to correct for covariate shift (i.e., cancel the bias from the training set), one can reweight the training data by assigning more weights to observations where the test data locate more often. Indeed, the key to many approaches addressing covariate shift is the estimation of importance sampling weights, or the Radon-Nikodym derivative (RND) of $dP_{te}/dP_{tr}$ between $P_{te}$ and $P_{tr}$ (Sugiyama et al. (2008a); Bickel et al. (2007); Kanamori et al. (2012); Cortes et al. (2008); Yao & Doretto (2010); Pardoe & Stone (2010); Schölkopf et al. (2002); Quionero-Candela et al. (2009); Sugiyama & Kawanabe (2012)). Among them is the popular kernel mean matching (KMM) (Huang et al. (2007); Quionero-Candela et al. (2009)), which estimates the importance weights by matching means in a reproducing kernel Hilbert space (RKHS) and can be implemented efficiently by quadratic programming (QP).

Despite the demonstrated efficiency in many covariate shift problems (Sugiyama et al. (2008a); Quionero-Candela et al. (2009); Gretton et al. (2009)), KMM can suffer from high variance, due to several reasons. The first one regards the RKHS assumption. As pointed out in Yu & Szepesvári (2012), under a more realistic assumption from learning theory (Cucker & Zhou (2007)), when the true regression function does not lie in the RKHS but a general range space indexed by a smoothness parameter $\theta > 0$, KMM degrades to sub-canonical rate $\mathcal{O}(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{\theta}{2\theta+4}})$ from the parametric rate $\mathcal{O}(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}})$. Second, if the discrepancy between the training and testing distributions is large (e.g., test samples concentrate on regions where few training samples are located), the RND becomes unstable and leads to high resulting variance (Blanchet & Lam (2012)), partially due to a induced sparsity as most weights shrink towards zero while the non-zero ones surge to huge values. This is an intrinsic challenge for reweighting methods that occurs even if the RND is known in closed-form. One way to bypass it is to identify model misspecification (Wen et al. (2014)), but as mentioned in Sugiyama et al. (2008b), the cross-validation for model selection needed in many related methods often requires the importance weights to cancel biases and the necessity for reweighting remains.

In this paper we propose a method to reduce the variance of KMM in covariate shift problems. Our method relies on an estimated regression function and the application of the importance weighting on the *residuals* of the regression. Intuitively, these residuals have smaller magnitudes than the original loss values, and the resulting reweighted estimator thus becomes less sensitive to the variances of weights. Then, we cancel the bias incurred by the use of residuals by judiciously compensation through the estimated regression function evaluated on the test set.

We specialize our method by using a nonparametric regression (NR) function constructed from regularized least square in RKHS (Cucker & Zhou (2007); Smale & Zhou (2007); Sun & Wu (2009)), also known as the Tikhonov regularized learning algorithm (Evgeniou et al. (2000)). We show that our new estimator achieves the rate $\mathcal{O}(n_{tr}^{-\frac{\theta}{2\theta+2}} + n_{te}^{-\frac{\theta}{2\theta+2}})$, which is superior to the best-known rate of KMM in Yu & Szepesvári (2012), with the same computational complexity of KMM. Although the gap to the parametric rate is yet to be closed, the new estimator certainly seems to be a step towards the right direction. To put into perspective, we also compare with an alternate approach in Yu & Szepesvári (2012) which constructs a NR function using the training set and then predicts by evaluating on the test set. Such an approach leads to a better dependence on the test size but worse dependence on the training size than KMM. Our estimator, which can be viewed as an ensemble of KMM and NR, achieves a convergence rate that is either superior or matches both of these methods, thus in a sense robust against both estimators. In fact, we show our estimator can be motivated both from a variance reduction perspective on KMM using control variates (Nelson (1990); Glynn & Szechtman (2002)) and a bias reduction perspective on NR.

Another noticable feature of the new estimator relates to data aggregation in empirical risk minimization (ERM). Specifically, when KMM is applied in learning algorithms or ERMs, the resulting optimal solution is typically a finite-dimensional span of the training data mapped into feature space (Schölkopf et al. (2001)). The optimal solution of our estimator, on the other hand, depends on both the training and testing data, thus highlighting a different and more efficient information leveraging that utilizes both data sets simultaneously.

The paper is organized as follows. Section 2 reviews the background on KMM and NR that motivates our estimator. Section 3 presents the details of our estimator and studies its convergence property. Section 4 generalizes our method to ERM. Section 5 demonstrates experimental results.

## 2 Background and Motivation

### 2.1 Assumptions

Denote $P_{tr}$ to be the probability measure for training variables $X^{tr}$ and $P_{te}$ for test variables $X^{te}$.

**Assumption 1.** $P_{tr}(dy|\boldsymbol{x}) = P_{te}(dy|\boldsymbol{x})$.

**Assumption 2.** *The Radon-Nikodym derivative* $\beta(\boldsymbol{x}) \triangleq \frac{dP_{te}}{dP_{tr}}(\boldsymbol{x})$ *exists and is bounded by* $B < \infty$.

**Assumption 3.** *The covariate space* $\mathcal{X}$ *is compact and the label space* $\mathcal{Y} \subseteq [0,1]$. *Furthermore, there exists a kernel* $K(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *which induces a RKHS* $\mathcal{H}$ *and a canonical feature map* $\Phi(\cdot) : \mathcal{X} \to \mathcal{H}$ *such that* $K(\boldsymbol{x},\boldsymbol{x}') = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle_{\mathcal{H}}$ *and* $\|\Phi(\boldsymbol{x})\|_{\mathcal{H}} \leq R$ *for some* $0 < R < \infty$.

In particular, Assumption 1 is the covariate shift assumption which states the conditional distributions $P(dy|\boldsymbol{x})$ remains invariant while the marginal $P_{tr}(\boldsymbol{x})$ and $P_{te}(\boldsymbol{x})$ shift. Assumptions 2 and

3 are common for establishing theoretical results. Specifically, Assumption 2 can be satisfied by restricting the support of $P_{te}$ and $P_{tr}$ on a compact set, although $B$ could be potentially large.

### 2.2 Problem Setup and Existing Approaches

Given $n_{tr}$ labelled training data $\{(\boldsymbol{x}_j^{tr}, \boldsymbol{y}_j^{tr})\}_{j=1}^{n_{tr}}$ and $n_{te}$ unlabelled test data $\{\boldsymbol{x}_i^{te}\}_{i=1}^{n_{te}}$ (i.e., $\{y_i^{te}\}_{i=1}^{n_{te}}$ are unavailable), the goal is to estimate $\nu = \mathbb{E}[Y^{te}]$. The KMM estimator (Huang et al. (2007); Gretton et al. (2009)) is $V_{KMM} = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \hat{\beta}(\boldsymbol{x}_j^{tr}) y_j^{tr}$, where $\hat{\beta}(\boldsymbol{x}_j^{tr})$ are solutions of a QP that attempts to match the means of training and test sets in the feature space using weights $\hat{\boldsymbol{\beta}}$:

$$\min_{\hat{\boldsymbol{\beta}}} \quad \left\{ \hat{L}(\hat{\boldsymbol{\beta}}) \triangleq \| \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \hat{\beta}_j \Phi(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \Phi(\boldsymbol{x}_i^{te}) \|_{\mathcal{H}}^2 \right\} \quad \text{s.t. } 0 \le \hat{\beta}_j \le B, \forall 1 \le j \le n_{tr}. \tag{1}$$

Notice we write $\hat{\beta}_j$ as $\hat{\beta}(\boldsymbol{x}_j^{tr})$ in $V_{KMM}$ informally to highlight $\hat{\beta}_j$ as estimates of $\beta(\boldsymbol{x}_j^{tr})$. The fact that (1) is a QP can be verified by the kernel trick, as in Gretton et al. (2009). Define matrix $K_{ij} = K(\boldsymbol{x}_i^{tr}, \boldsymbol{x}_j^{tr})$ and $\kappa_j \triangleq \frac{n_{tr}}{n_{te}} \sum_{i=1}^{n_{te}} K(\boldsymbol{x}_j^{tr}, \boldsymbol{x}_i^{te})$, optimization (1) is equivalent to

$$\min_{\hat{\boldsymbol{\beta}}} \quad \frac{1}{n_{tr}^2} \hat{\boldsymbol{\beta}}^T \boldsymbol{K} \hat{\boldsymbol{\beta}} - \frac{2}{n_{tr}^2} \boldsymbol{\kappa}^T \hat{\boldsymbol{\beta}} \quad \text{s.t. } 0 \le \hat{\beta}_j \le B, \forall 1 \le j \le n_{tr}. \tag{2}$$

In practice, a constraint $\left| \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \hat{\beta}_j - 1 \right| \le \epsilon$ for a tolerance $\epsilon > 0$ is included to regularize the $\hat{\boldsymbol{\beta}}$ towards the RND. As in Yu & Szepesvári (2012), we omit them to simplify analysis. On the other hand, the NR estimator $V_{NR} = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te})$ is based on $\hat{g}(\cdot)$, some estimate of the regression function $g(\boldsymbol{x}) \triangleq \mathbb{E}[Y|\boldsymbol{x}]$. Notice the conditional expectation is taken regardless of $\boldsymbol{x} \sim P_{tr}$ or $P_{te}$. Here, we consider a $\hat{g}(\cdot)$ that is estimated nonparametrically by regularized least square in RKHS:

$$\hat{g}_{\gamma,data}(\cdot) = \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{j=1}^{m} (f(\boldsymbol{x}_j^{tr}) - y_j^{tr})^2 + \gamma \|f\|_{\mathcal{H}}^2 \right\}, \tag{3}$$

where $\gamma$ is a regularization term to be chosen and the subscript $data$ represents $\{(\boldsymbol{x}_j^{tr}, y_j^{tr})\}_{j=1}^m$. Using the kernel trick and the representation theorem (Schölkopf et al. (2001)), optimization problem (3) can be solved in closed form with $\hat{g}_{\gamma,data}(\boldsymbol{x}) = \sum_{j=1}^m \alpha_j^{reg} K(\boldsymbol{x}_j^{tr}, \boldsymbol{x})$ where

$$\boldsymbol{\alpha}^{reg} = (\boldsymbol{K} + \gamma \boldsymbol{I})^{-1} \boldsymbol{y}^{tr} \quad \text{and} \quad \boldsymbol{y}^{tr} = [y_1^{tr}, ..., y_m^{tr}]. \tag{4}$$

### 2.3 Motivation

Depending on properties of $g(\cdot)$, Yu & Szepesvári (2012) proves different rates of KMM. The most notable case is when $g \notin \mathcal{H}$ but rather $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, where $\mathcal{T}_K$ is the integral operator $(\mathcal{T}_K f)(x') = \int_{\mathcal{X}} K(x', x) f(x) P_{tr}(dx)$ on $\mathscr{L}_{P_{tr}}^2$. In this case, Yu & Szepesvári (2012) characterize $g$ with the approximation error

$$\mathcal{A}_2(g, F) \triangleq \inf_{\|f\|_{\mathcal{H}} \le F} \|g - f\|_{\mathscr{L}_{P_{tr}}^2} \le C F^{-\frac{\theta}{2}}, \tag{5}$$

and the rates of KMM drops to sub-canonical $|V_{KMM} - \nu| = \mathcal{O}(n_{tr}^{-\frac{\theta}{(2\theta+4)}} + n_{te}^{-\frac{\theta}{(2\theta+4)}})$, as opposed to $\mathcal{O}(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}})$ when $g \in \mathcal{H}$. As shown in Lemma 4 and Theorem 4.1 of Cucker & Zhou (2007)), (5) is almost equivalent to $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$: $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ implies (5) while (5) leads to $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4} - \epsilon})$ for any $\epsilon > 0$. We will adopt the characterization $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ as our analysis is based on related learning theory estimates.

Correspondingly, the convergence rate for $V_{NR}$ when $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ is also shown in Yu & Szepesvári (2012) as $|V_{NR} - \nu| = \mathcal{O}(n_{te}^{-\frac{1}{2}} + n_{tr}^{-\frac{3\theta}{12\theta+16}})$, with $\hat{g}$ taken as $\hat{g}_{\gamma,data}$ in (3) and $\gamma$ chosen optimally. The rate of $V_{KMM}$ is usually better than $V_{NR}$ due to labelling cost (i.e. $n_{tr} < n_{te}$). However, in practice the performance of $V_{KMM}$ is not always better than $V_{NR}$. This could be partially explained by the hidden dependence of $V_{KMM}$ on potentially large $B$, but more importantly, without variance reduction, KMM is subject to the negative effects of unstable importance sampling weights (i.e. the $\hat{\boldsymbol{\beta}}$). On the other hand, the training of $\hat{g}$ requires labels hence can only be done on training set. Consequently, without reweighting, when estimating the test quantity $\nu$, the rate of $V_{NR}$ suffers from the bias.

This motivates the search for a robust estimator which does not require prior knowledge on the performance of $V_{KMM}$ or $V_{NR}$ and can, through a combination, reach or even surpass the best performance among both. For simplicity, we use the mean squared error (MSE) criteria MSE$(V) =$ Var$(V)+($Bias$(V))^2$ and assume an additive model $Y = g(X)+\mathcal{E}$ where $\mathcal{E} \sim \mathcal{N}(0, \sigma^2)$ is independent with $X$ and other errors. Under this framework, we motivate a remedy from two perspectives:

Variance reduction for KMM: Consider an idealized KMM with $V_{KMM} \triangleq \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}_j^{tr}) y_j^{tr}$ with $\beta(\cdot)$ being the true RND. Since $\mathbb{E}[\beta(X^{tr})Y^{tr}] = \mathbb{E}_{\boldsymbol{x}\sim P_{tr}}(\beta(\boldsymbol{x})g(\boldsymbol{x})) = \mathbb{E}_{\boldsymbol{x}\sim P_{te}}[g(\boldsymbol{x})] = \nu$, $V_{KMM}$ is unbiased and the only source of MSE becomes the variance. It then follows from standard control variates that, given an estimator $V$ and a zero-mean random variable $W$, we can set $t^\star = \frac{\text{Cov}(V,W)}{\text{Var}(W)}$ and use $V - t^\star W$ to obtain $\min_t \text{Var}(V - tW) = (1-\text{corr}^2(V,W))\text{Var}(V) \leq \text{Var}(V)$ without altering the mean of $V$. Thus we can use $W = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}_j^{tr})(\hat{g}(\boldsymbol{x}_j^{tr})) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te})$ with $t^\star = \frac{\text{Cov}(V_{KMM},W)}{\text{Var}(W)}$. To calculate $t^\star$, suppose $X^{te}$ and $X^{tr}$ are independent, then

$$\begin{aligned}\text{Cov}(V_{KMM}, W) =& \frac{1}{n_{tr}} \text{Cov}(\beta(X^{tr})Y^{tr}, \beta(X^{tr})\hat{g}(X^{tr})) \\ =& \frac{1}{n_{tr}} \text{Cov}(\beta(X^{tr})g(X^{tr}), \beta(X^{tr})\hat{g}(X^{tr})) \approx \frac{1}{n_{tr}} \text{Var}(\beta(X^{tr})\hat{g}(X^{tr})),\end{aligned}$$

if $\hat{g}$ is close enough to $g$. On the other hand, in the usual case where $n_{te} \gg n_{tr}$,

$$\text{Var}(W) = \frac{1}{n_{tr}} \text{Var}(\beta(X^{tr})\hat{g}(X^{tr})) + \frac{1}{n_{te}} \text{Var}(\hat{g}(X^{te})) \approx \frac{1}{n_{tr}} \text{Var}(\beta(X^{tr})\hat{g}(X^{tr})).$$

Thus, $t^\star \approx 1$ which gives our estimator $V_R = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}_j^{tr})(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr})) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te})$.

Bias reduction for NR: Consider the NR estimator $V_{NR} \triangleq \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te})$. Assuming again the common case where $n_{te} \gg n_{tr}$, we have Var$(V_{NR}) = \frac{1}{n_{te}}$Var$(\hat{g}(X^{te})) \approx 0$, and thus the main source of MSE is the bias $\mathbb{E}_{\boldsymbol{x}\sim P_{te}}[g(\boldsymbol{x}) - \hat{g}(\boldsymbol{x})]$. If we add $W = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}_j^{tr})(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr}))$ to $V_{NR}$, we eliminate the bias which gives $V_R = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}_j^{tr})(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr})) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te})$.

## 3 Robust Estimator

We construct a new estimator $V_R(\rho)$ that can be demonstrated to perform robustly against both the KMM and NR estimators discussed above. In our construction, we split the training set with a proportion $0 \leq \rho \leq 1$, i.e., divide $\{\boldsymbol{X}^{tr}, \boldsymbol{Y}^{tr}\}_{data} \triangleq \{(\boldsymbol{x}_j^{tr}, y_j^{tr})\}_{j=1}^{n_{tr}}$ into $\{\boldsymbol{X}_{KMM}^{tr}, \boldsymbol{Y}_{KMM}^{tr}\}_{data} \triangleq \{(\boldsymbol{x}_j^{tr}, y_j^{tr})\}_{j=1}^{\lfloor \rho n_{tr} \rfloor}$ and $\{\boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}\}_{data} \triangleq \{(\boldsymbol{x}_j^{tr}, y_j^{tr})\}_{j=\lfloor \rho n_{tr} \rfloor+1}^{n_{tr}}$. Then we use $\{\boldsymbol{X}_{KMM}^{tr}, \boldsymbol{X}^{te}\}_{data} \triangleq \{\{\boldsymbol{x}_j^{tr}\}_{j=1}^{\lfloor \rho n_{tr} \rfloor}, \{\boldsymbol{x}_i^{te}\}_{i=1}^{n_{te}}\}$ to solve for the weight $\hat{\boldsymbol{\beta}}$ in (1) and use $\{\boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}\}_{data}$ to train an NR function $\hat{g}(\cdot) = \hat{g}_{\gamma,data}(\cdot)$ for some $\gamma$ as in (3). Finally, we define our estimator $V_R(\rho)$ as

$$V_R(\rho) \triangleq \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr})(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr})) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te}). \tag{6}$$

First, we remark the parameter $\rho$ controlling the splitting of data serves mainly for theoretical considerations. In practice, the data can be used for both purposes simultaneously. Second, as mentioned, many $\hat{g}$ other than (3) could be considered for control variate. However, aside from the availability of closed-form expressions (4), $\hat{g}_{\gamma,data}$ is connected to the learning theory estimates Cucker & Zhou (2007). Thus, for establishing a theoretical bound, we focus on $\hat{g} = \hat{g}_{\gamma,data}$ for now.

Our main result is the convergence analysis with respect to $n_{tr}$ and $n_{te}$ which rigorously justified the previous intuition. In particular, we show that $V_R$ either surpasses or achieves the better rate between $V_{KMM}$ and $V_{NR}$. In all theorems that follow, the Big-$\mathcal{O}$ notations can be interpreted either as $1-\delta$ high probability bound or a bound on expectation. The proofs are left in the Appendix.

**Theorem 1.** *Under Assumptions 1-3, if $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, the convergence rate of $V_R(\rho)$ satisfies*

$$|V_R(\rho) - \nu| = \mathcal{O}(n_{tr}^{-\frac{\theta}{2\theta+2}} + n_{te}^{-\frac{\theta}{2\theta+2}}), \tag{7}$$

*when $\hat{g}$ is taken to be $\hat{g}_{\gamma,data}$ in (6) with $\gamma = n^{-\frac{\theta+2}{\theta+1}}$ and $n \triangleq \min(n_{tr}, n_{te})$.*

**Corollary 1.** *Under the same setting of theorem 1, if we choose $\gamma = n^{-1}$, we have*

$$|V_R(\rho) - \nu| = \mathcal{O}(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{\theta}{2\theta+4}}) \tag{8}$$

*and if we choose $\gamma = n_{tr}^{-1}$,*

$$|V_R(\rho) - \nu| = \mathcal{O}(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{1}{2}}). \tag{9}$$

We remark several implications. First, although not achieving canonical, (7) is an improvement over the best-known $\mathcal{O}(n_{tr}^{-\frac{\theta}{(2\theta+4)}} + n_{te}^{-\frac{\theta}{(2\theta+4)}})$ rate of $V_{KMM}$ when $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, especially for small $\theta$, suggesting that $V_R$ is more suitable than $V_{KMM}$ when $g$ is irregular. Indeed, $\theta$ is a smoothness parameter that measures the regularity of $g$. When $\theta$ increases, functions in $Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ get smoother and $Range(\mathcal{T}_K^{\frac{\theta_2}{2\theta_2+4}}) \subseteq Range(\mathcal{T}_K^{\frac{\theta_1}{2\theta_1+4}})$ for $0 < \theta_1 < \theta_2$, with the limiting case that $\theta \to \infty$, $\frac{\theta}{2\theta+4} \to 1/2$ and $Range(\mathcal{T}_K^{\frac{1}{2}}) \subseteq \mathcal{H}$ (i.e. $g \in \mathcal{H}$) for universal kernels by Mercer's theorem.

Second, as in Theorem 4 of Yu & Szepesvári (2012), the optimal tuning of $\gamma$ that leads to (7) depends on the unknown parameter $\theta$, which may not be adaptive in practice. However, if one simply choose $\gamma = n^{-1}$, $V_R$ still achieves a rate no worse than $V_{KMM}$ as depicted in (8).

Third, also in Theorem 4 of Yu & Szepesvári (2012), the rate of $V_{NR}$ is $\mathcal{O}(n_{te}^{-\frac{1}{2}} + n_{tr}^{-\frac{3\theta}{12\theta+16}})$ when $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, which is better on $n_{te}$ but not $n_{tr}$. Since usually $n_{tr} < n_{te}$, the rate of $V_{KMM}$ generally excels. Indeed, in this case the rate of $V_{NR}$ beats $V_{KMM}$ only if $\lim_{n\to\infty} n_{te}^{\frac{6\theta+8}{3\theta+6}}/n_{tr} \to 0$. However, if so, $V_R$ can still achieve $\mathcal{O}(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{1}{2}})$ rate in (9) which is better than $V_{NR}$, by simply taking $\gamma = n_{tr}^{-1}$, i.e., regularizing the training process more when the test set is small. Moreover, as $\theta \to \infty$, our estimator $V_R$ recovers the canonical rate $n_{tr}^{-\frac{1}{2}}$ as opposed to $n_{tr}^{-\frac{1}{4}}$ in $V_{NR}$.

Thus, in summary, when $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, our estimator $V_R$ outperforms both $V_{KMM}$ and $V_{NR}$ across the relative sizes of $n_{tr}$ and $n_{te}$. The outperformance over $V_{KMM}$ is strict when $\gamma$ is chosen dependent on $\theta$, and the performance is matched when $\gamma$ is chosen robustly without knowledge of $\theta$.

For completeness, we consider two other characterizations of $g$ discussed in Yu & Szepesvári (2012): one is $g \in \mathcal{H}$ and the other is $\mathcal{A}_\infty(g, F) \triangleq \inf_{\|f\|_{\mathcal{H}} \le F} \|g - f\| \le C(\log F)^{-s}$ for some $C, s > 0$ (e.g., $g \in H^s(\mathcal{X})$ with $K(\cdot,\cdot)$ being the Gaussian kernel, where $H^s$ is the Sobolev space with integer $s$). The two assumptions are, in a sense, more extreme (being optimistic or pessimistic). The next two results show that the rates of $V_R$ in these situations match the existing ones for $V_{KMM}$ (the rates for $V_{NR}$ are not discussed in Yu & Szepesvári (2012) under these assumptions).

**Proposition 1.** *Under Assumptions 1-3, if $g \in \mathcal{H}$, the convergence rate of $V_R(\rho)$ satisfies $|V_R(\rho) - \nu| = \mathcal{O}(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}})$, when $\hat{g}$ is taken to be $\hat{g}_{\gamma,data}$ for $\gamma > 0$ in (6).*

**Proposition 2.** *Under Assumptions 1-3, if $\mathcal{A}_\infty(g, F) \triangleq \inf_{\|f\|_{\mathcal{H}} \le F} \|g - f\| \le C(\log F)^{-s}$ for some $C, s > 0$, the convergence rate of $V_R(\rho)$ satisfies $|V_R(\rho) - \nu| = \mathcal{O}\left(\log^{-s} \frac{n_{tr} n_{te}}{n_{tr} + n_{te}}\right)$, when $\hat{g}$ is taken to be $\hat{g}_{\gamma,data}$ for $\gamma > 0$ in (6).*

## 4 Empirical Risk Minimization

The robust estimator can handle empirical risk minimization (ERM). Given loss function $l'(x, y; \theta) : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ given $\theta$ in $\mathcal{D}$, we optimize over $\min_{\theta \in \mathcal{D}} \mathbb{E}[l'(X^{te}, Y^{te}; \theta)] = \min_{\theta \in \mathcal{D}} \mathbb{E}_{\boldsymbol{x} \sim P_{te}}[l(\boldsymbol{x}; \theta)]$

where $l(\boldsymbol{x};\theta) \triangleq \mathbb{E}_{Y|\boldsymbol{x}}[l'(\boldsymbol{x},Y;\theta)]$ to find $\theta^\star \triangleq \operatorname{argmin}_{\theta\in\mathcal{D}} \mathbb{E}_{\boldsymbol{x}\sim P_{te}}[l(X^{te};\theta)]$. In practice, usually a regularization term $\Omega[\theta]$ on $\theta$ is added. For example, the KMM in Huang et al. (2007) considers

$$\min_{\theta\in\mathcal{D}} \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \hat{\beta}(\boldsymbol{x}_j^{tr}) l'(\boldsymbol{x}_j^{tr}, y_j^{tr};\theta) + \lambda\Omega[\theta]. \tag{10}$$

We can carry out a similar modification to utilize $V_R$ as

$$\min_{\theta\in\mathcal{D}} \frac{1}{\lfloor\rho n_{tr}\rfloor} \sum_{j=1}^{\lfloor\rho n_{tr}\rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr})(l'(\boldsymbol{x}_j^{tr}, y_j^{tr};\theta) - \hat{l}(\boldsymbol{x}_j^{tr};\theta)) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{l}(\boldsymbol{x}_i^{te};\theta) + \lambda\Omega[\theta], \tag{11}$$

with $\hat{\boldsymbol{\beta}}$ based on $\{\boldsymbol{X}_{KMM}^{tr}, \boldsymbol{X}^{te}\}$ and $\hat{l}(x;\theta)$ being an estimate of $l(x;\theta)$ based on $\{\boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}\}$. For later reference, we note that a similar modification can also be used to utilize $V_{NR}$:

$$\min_{\theta\in\mathcal{D}} \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{l}(\boldsymbol{x}_i^{te};\theta) + \lambda\Omega[\theta]. \tag{12}$$

Below we discuss two classical learning problems using (11).

**Penalized Least Square Regression:** Consider a regression problem with $l'(\boldsymbol{x},y;\theta) = (y - \langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}})^2$, $\Omega[\theta] = \|\theta\|_{\mathcal{H}}^2$ and $y \in [0,1]$. We have $l(\boldsymbol{x};\theta) = \mathbb{E}[Y^2|\boldsymbol{x}] - 2g(\boldsymbol{x})\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}} + \langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}}^2$, and a candidate for $\hat{l}(\boldsymbol{x},\theta)$ is to substitute $g$ with $\hat{g}_{\gamma,data}$. Then, (11) becomes

$$\min_{\theta\in\mathcal{D}} \sum_{j=1}^{\lfloor\rho n_{tr}\rfloor} -\frac{2\beta(\boldsymbol{x}_j^{tr})}{\lfloor\rho n_{tr}\rfloor}(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr}))\langle\theta,\Phi(\boldsymbol{x}_j^{tr})\rangle_{\mathcal{H}} + \frac{1}{n_{te}}\sum_{i=1}^{n_{te}}(\hat{g}(\boldsymbol{x}_i^{te}) - \langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}})^2 + \lambda\|\theta\|_{\mathcal{H}}^2,$$

by adding and removing the components not involving $\theta$. Furthermore, it simplifies to the QP:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^{\lfloor\rho n_{tr}\rfloor+n_{te}}} \frac{-2\boldsymbol{w}_1^T\boldsymbol{K}_{tot}\boldsymbol{\alpha}}{\lfloor\rho n_{tr}\rfloor} + \frac{(\boldsymbol{w}_2 - \boldsymbol{K}_{tot}\boldsymbol{\alpha})^T\boldsymbol{W}_3(\boldsymbol{w}_2 - \boldsymbol{K}_{tot}\boldsymbol{\alpha})}{n_{te}} + \lambda\boldsymbol{\alpha}^T\boldsymbol{K}_{tot}\boldsymbol{\alpha}, \tag{13}$$

by the representation theorem Schölkopf et al. (2001). Here $(\boldsymbol{K}_{tot})_{ij} = K(\boldsymbol{x}_i^{tot}, \boldsymbol{x}_j^{tot})$ and $\boldsymbol{W}_3 = \operatorname{diag}(\boldsymbol{w}_3)$ where $\boldsymbol{x}_i^{tot} = \boldsymbol{x}_i^{tr}$, $(w_1)_i = \beta(\boldsymbol{x}_i^{tr})(y_i^{tr} - \hat{g}(\boldsymbol{x}_i^{tr}))$, $(w_2)_i = 0$, $(w_3)_i = 0$ for $1 \leq i \leq \lfloor\rho n_{tr}\rfloor$ and $\boldsymbol{x}_i^{tot} = \boldsymbol{x}_{i-\lfloor\rho n_{tr}\rfloor}^{te}$, $(w_1)_i = 0$, $(w_2)_i = \hat{g}(\boldsymbol{x}_{i-\lfloor\rho n_{tr}\rfloor}^{te})$, $(w_3)_i = 1$ for $\lfloor\rho n_{tr}\rfloor + 1 \leq i \leq \lfloor\rho n_{tr}\rfloor + n_{te}$. Notice (13) has a closed-form solution $\hat{\boldsymbol{\alpha}} = (\boldsymbol{W}_3\boldsymbol{K}_{tot} + \lambda n_{te}\boldsymbol{I})^{-1}(\frac{n_{te}}{\lfloor\rho n_{tr}\rfloor}\boldsymbol{w}_1 + \boldsymbol{w}_2)$.

**Penalized Logistic Regression:** Consider a binary classification problem with $y \in \{0,1\}$, $\Omega[\theta] = \|\theta\|_{\mathcal{H}}^2$ and $-l'(\boldsymbol{x},y;\theta) = y\log(\frac{1}{1+\exp\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}}}) + (1-y)\log(\frac{\exp\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}}}{1+\exp\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}}})$. Thus, $-l(\boldsymbol{x};\theta) = -g(\boldsymbol{x})\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}} + \log(\frac{\exp\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}}}{1+\exp\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}}})$ and we substitute $g$ with $\hat{g}_{\gamma,data}$. Then, (11) becomes

$$\begin{aligned}\min_{\theta\in\mathcal{D}} &\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor} \frac{\beta(\boldsymbol{x}_j^{tr})}{\lfloor\rho n_{tr}\rfloor}(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr}))\langle\theta,\Phi(\boldsymbol{x}_j^{tr})\rangle_{\mathcal{H}} \\ &+ \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} -\hat{g}(\boldsymbol{x}_i^{te})\langle\theta,\Phi(\boldsymbol{x}_i^{te})\rangle_{\mathcal{H}} + \log(\frac{\exp\langle\theta,\Phi(\boldsymbol{x}_i^{te})\rangle_{\mathcal{H}}}{1+\exp\langle\theta,\Phi(\boldsymbol{x}_i^{te})\rangle_{\mathcal{H}}}) + \lambda\|\theta\|_{\mathcal{H}}^2.\end{aligned}$$

which again simplifies to, by Schölkopf et al. (2001), the convex program:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^{\lfloor\rho n_{tr}\rfloor+n_{te}}} \frac{\boldsymbol{w}_1^T\boldsymbol{K}_{tot}\boldsymbol{\alpha}}{\lfloor\rho n_{tr}\rfloor} - \frac{\boldsymbol{w}_2^T\boldsymbol{K}_{tot}\boldsymbol{\alpha}}{n_{te}} + \frac{\sum_{i=1}^{n_{te}}\log(\frac{\exp(\boldsymbol{K}_{tot}\boldsymbol{\alpha})_{\lfloor\rho n_{tr}\rfloor+i}}{1+\exp(\boldsymbol{K}_{tot}\boldsymbol{\alpha})_{\lfloor\rho n_{tr}\rfloor+i}})}{n_{te}} + \lambda\boldsymbol{\alpha}^T\boldsymbol{K}_{tot}\boldsymbol{\alpha}. \tag{14}$$

Both (11) and (14) can be optimized efficiently by standard solvers. Notably, (11) gives a solution in the form $\hat{\theta} = \sum_{i=1}\hat{\alpha}_i K(\boldsymbol{x}_i^{tot}, \boldsymbol{x})$ which spans on both training and test data. In contrast, the solution of (10) or (12) only spans on one of them. For example, as shown in Huang et al. (2007), the penalized least square solution for (10) is $\hat{\theta} = \sum_{i=1}\hat{\alpha}_i K(\boldsymbol{x}_i^{tr}, \boldsymbol{x})$ where $\hat{\boldsymbol{\alpha}} = (\boldsymbol{K} + n_{te}\lambda\operatorname{diag}(\hat{\boldsymbol{\beta}})^{-1})^{-1}\boldsymbol{y}^{tr}$ ( we use $\hat{\boldsymbol{\alpha}} = (\operatorname{diag}(\hat{\boldsymbol{\beta}})\boldsymbol{K} + n_{te}\lambda\boldsymbol{I})^{-1}\operatorname{diag}(\hat{\boldsymbol{\beta}})\boldsymbol{y}^{tr}$ in experiments to avoid invertibility issues caused by the sparsity of $\hat{\boldsymbol{\beta}}$), so only the training data are in the span of the feature space that constitutes $\hat{\theta}$. The aggregation of both sets suggests a more effective/robust utilization of data . We conclude with a theorem on ERM similar to Corollary 8.9 in Gretton et al. (2009), which guarantees the convergence of the solution of (11) in a simple setting.

**Theorem 2.** *Assume $l(x;\theta)$ and $\hat{l}(x;\theta) \in \mathcal{H}$ can be expressed as $\langle \Phi(x), \theta \rangle_{\mathcal{H}} + f(x;\theta)$ with $||\theta||_{\mathcal{H}} \leq C$ and $l'(x,y;\theta) \in \mathcal{H}$ as $\langle \Upsilon(x,y), \Lambda \rangle_{\mathcal{H}} + f(x;\theta)$ with $||\Lambda||_{\mathcal{H}} \leq C$. Denote this class of loss functions $\mathcal{G}$ and further assume $l(x;\theta)$ are continuous, bounded by $D$ and L-Lipschitz on $\theta$ uniformly over $x$ for $(\theta, x)$ in a compact set $\mathcal{D} \times \mathcal{X}$. Then, the ERM with $\hat{\theta}_R \triangleq \operatorname{argmin}_{\theta \in \mathcal{D}} V_R(\theta)$ and $V_R(\theta) \triangleq \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr})(l'(\boldsymbol{x}_j^{tr}, y_j^{tr}; \theta) - \hat{l}(\boldsymbol{x}_j^{tr};\theta)) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{l}(\boldsymbol{x}_i^{te};\theta)$ satisfies*

$$\mathbb{E}[l'(X_{te}, Y_{te}; \hat{\theta}_R)] \leq \mathbb{E}[l'(X_{te}, Y_{te}; \theta^\star)] + \mathcal{O}(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}}).$$

# 5 EXPERIMENTS

## 5.1 TOY DATASET REGRESSION

We first present a toy example to provide comparison with KMM. The data is generated as the polynomial regression example in Shimodaira (2000); Huang et al. (2007), where $P_{tr} \sim \mathcal{N}(0.5, 0.5^2)$, $P_{te} \sim \mathcal{N}(0, 0.3^2)$ are Gaussian distributions. The labels are generated according to $y = -x + x^3$ and observed in Gaussian noise $\mathcal{N}(0, 0.3^2)$. We sample 500 points in both training and test data and fit a linear model using ordinary least square (OLS), KMM and the robust estimator, respectively. On the population level, the best linear fit is $y = -0.73x$. For simplicity, we the intercept to 0 and compare the fitted slopes for different approaches. We used a degree-3 polynomial kernel and the $\gamma$ in $\hat{g}_{\gamma,data}$ is set to the default value $n_{tr}^{-1}$. The tolerance $\epsilon$ for $\hat{\boldsymbol{\beta}}$ is set similarly as in Huang et al. (2007) with a slight tuning to avoid an overly-sparse solution. The slope is fitted without regularization. In Figure 1[a], the red curve is the true polynomial regression function and the purple line is the best linear fit. As we see, the robust estimator outperforms the two other methods, recovering the green line closest to the best one. The performance over 20 trials are summarized in Figure 1[b].

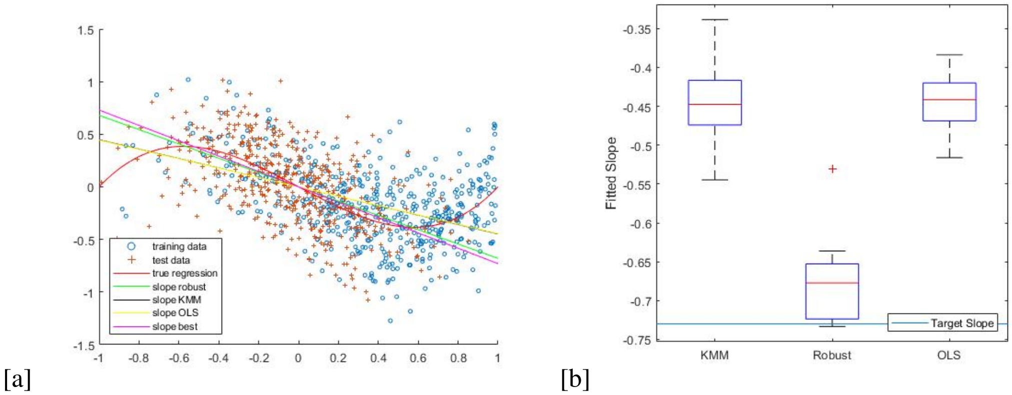

Figure 1: [a] Linear fit with OLS,KMM and Robust estimator; [b] Slope estimation performance

## 5.2 REAL WORLD DATASET FOR ERM

Next, we test our approach in ERM on a real world dataset, the breast cancer dataset from the UCI Archive. We consider the second biased sampling scheme in Huang et al. (2007) where the probability of selecting $\boldsymbol{x}_i$ into the training set depends jointly on multiple features and is proportional to $\exp(-\sigma_1 \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|)$ for some $\sigma > 0$ and the sample mean $\bar{\boldsymbol{x}}$. Since this is a binary classification problem, we can experiment with both the penalized least square regression and the penalized logistic regression for different sizes of training sets. We used a Gaussian kernel $\exp(-\sigma_2 \|\boldsymbol{x}_i - \boldsymbol{x}_j\|)$. The tolerance $\epsilon$ for $\hat{\boldsymbol{\beta}}$ is set exactly as in Huang et al. (2007). For both experiments, we choose parameters $\gamma = n_{tr}^{-1}$ as default, $\lambda = 5$ by cross-validation and $\sigma_1 = -1/100$, $\sigma_2 = \sqrt{0.5}$. Finally, we used the fitted parameters (i.e. optimal solution $\hat{\theta}$ in ERM) to predict the labels on the test set and compare with the hidden real ones. The summary of test error comparison is shown in Figure 2 where we use the term *unweighted* to denote the case for (12), *KMM* for (10) and *Robust* for (11).

The robust estimator gives the lowest test error in 5 cases out of 6, confirming our finding on its improvement over the traditional methods.

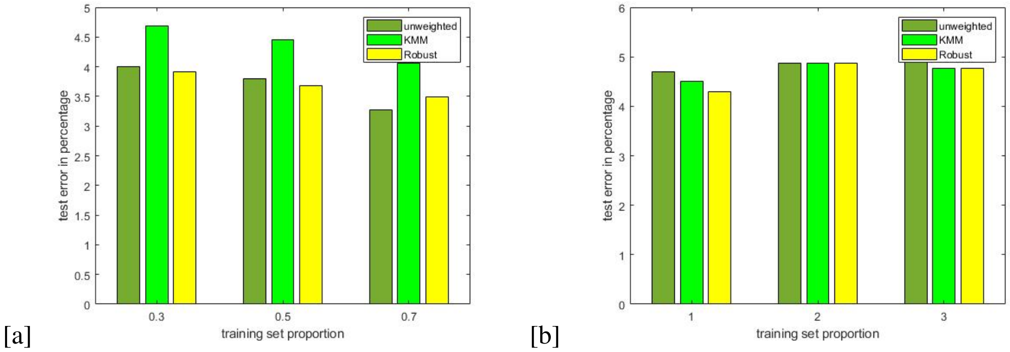

Figure 2: Classification performance for penalized [a] least square regression; [b] logistic regression

### 5.3 Simulated Dataset for Estimation

On an estimation problem, we simulate data from two ten dimensional Gaussian distribution with different, randomly generated mean and covariance matrix as training and test sets. The target value is $\nu = \mathbb{E}_{\boldsymbol{x}\sim P_{te}}[g(\boldsymbol{x})]$ for an artificial $g(x) = \sin(c_1\|\boldsymbol{x}\|_2^2) + (1+\exp(\boldsymbol{c}_2^T\boldsymbol{x}))^{-1}$ with random $c_1, \boldsymbol{c}_2$ and labels are observed with Gaussian noise. A Gaussian kernel $\exp(-\sigma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)$ and a tolerance $\epsilon$ for $\hat{\boldsymbol{\beta}}$ are set exactly as in Gretton et al. (2009) with $\sigma = \sqrt{5}$, $B = 1000$ and $\epsilon = \frac{\sqrt{n_{tr}}-1}{\sqrt{n_{tr}}}$. We also experiment with a different $\hat{g}$ by substituting $\hat{g}_{\gamma,data}$ for a naive linear OLS fit. At each iteration, we use the sample mean from $10^6$ data points (without adding noise) as the true mean and calculate the average MSE over 100 estimations for $V_R$, $V_{KMM}$ and $V_{NR}$ respectively. As shown in Table 1, the performances of $V_R$ are again consistently on par with the best case scenarios, even when the usual assumption $n_{tr} < n_{te}$ is violated.

Table 1: Average MSE for Estimation

| Hyperparameters | $V_{NR}$ | $V_{KMM}$ | $V_R$ |
|---|---|---|---|
| $\lambda = 0.1, n_{tr} = 50, n_{te} = 500$ | 0.9970 | 0.9489 | 0.9134 |
| $\lambda = 0.1, n_{tr} = 500, n_{te} = 500$ | 1.0006 | 0.9294 | 0.9340 |
| $\lambda = 0.1, n_{tr} = 500, n_{te} = 50$ | 1.0021 | 0.9245 | 0.9242 |
| $\lambda = 10, n_{tr} = 50, n_{te} = 500$ | 0.9962 | 0.9493 | 0.9467 |
| $\lambda = 10, n_{tr} = 500, n_{te} = 500$ | 0.9964 | 0.9294 | 0.9288 |
| $\lambda = 10, n_{tr} = 500, n_{te} = 50$ | 0.9965 | 0.9245 | 0.9293 |

## 6 Conclusion

Motivated both as a variance reduction on KMM and a bias reduction on NR, we introduced a new robust estimator for tackling covariate shift problems which, through a straightforward integration of traditional methods, leads to improved accuracy over both KMM and NR in many settings. From a practical standpoint, the control variates and data aggregation enable the estimation/training process to be more stable and data-efficient at no expense of increased computational complexity. From an analytical standpoint, a promising progress is made to reduce the rate gap of KMM towards the parametric when the regression function lies in range spaces outside of RKHS. For future work, note the canonical rate is still not achieved and it remains unclear the suitable tool to improve the rate further, if possible. Moreover, besides the regularized empirical regression function in RKHS, the eligibility and effectiveness of other estimated regression functions also require rigorous analysis.

## References

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pp. 81–88. ACM, 2007.

Jose Blanchet and Henry Lam. State-dependent importance sampling for rare-event simulation: An overview and recent advances. *Surveys in Operations Research and Management Science*, 17(1): 38–59, 2012.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 120–128. Association for Computational Linguistics, 2006.

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pp. 38–53. Springer, 2008.

Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in computational mathematics*, 13(1):1, 2000.

Peter W Glynn and Roberto Szechtman. Some new perspectives on the method of control variates. In *Monte Carlo and Quasi-Monte Carlo Methods 2000*, pp. 27–49. Springer, 2002.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

Hirotaka Hachiya, Takayuki Akiyama, Masashi Sugiyama, and Jan Peters. Adaptive importance sampling with automatic model selection in value function approximation. In *AAAI*, pp. 1351–1356, 2008.

James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pp. 153–161, 1979.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pp. 601–608, 2007.

Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 264–271, 2007.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012.

Mikhail Anatolevich Lifshits. *Gaussian random functions*, volume 322. Springer Science & Business Media, 2013.

Barry L Nelson. Control variate remedies. *Operations Research*, 38(6):974–992, 1990.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

David Pardoe and Peter Stone. Boosting for regression transfer. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 863–870. Omnipress, 2010.

Iosif Pinelis et al. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pp. 416–426. Springer, 2001.

Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

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.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pp. 1433–1440, 2008a.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008b.

Hongwei Sun and Qiang Wu. A note on application of integral operator in learning theory. *Applied and Computational Harmonic Analysis*, 26(3):416–421, 2009.

Hongwei Sun and Qiang Wu. Regularized least square regression with dependent samples. *Advances in Computational Mathematics*, 32(2):175–189, 2010.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Junfeng Wen, Chun-Nam Yu, and Russell Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *ICML*, pp. 631–639, 2014.

Yi Yao and Gianfranco Doretto. Boosting for transfer learning with multiple sources. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1855–1862. IEEE, 2010.

Yao-Liang Yu and Csaba Szepesvári. Analysis of kernel mean matching under covariate shift. In *ICML*, pp. 1147–1154. Omnipress, 2012.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 114. ACM, 2004.

# 7 Appendix

We mention that our proofs rely on learning theory estimates and are different from Yu & Szepesvári (2012). For example, in (3), $\gamma$ is used as a free parameter for controlling $\|f\|_{\mathcal{H}}$, whereas Yu & Szepesvári (2012) uses the parameter $F$ in (5). Although the two approaches are equivalent from an optimization viewpoint, with $\gamma$ being the Lagrange dual variable, the former approach turns out to be more suitable to analyse $V_R$.

Throughtout the proof, $h(\cdot) \in \mathcal{H}$ is assumed to be an unspecified function in the RKHS. Also, we use $\mathbb{E}_X[\cdot]$ to denote expectation over the randomness of $X$ while fixing others and $\mathbb{E}_{|X}[\cdot]$ as the conditional expectation $\mathbb{E}[\cdot|X]$.

Moreover we remark that all results involving $\hat{g}_{\gamma,data}$ can be interpreted either as a high probability bound or a bound on expectation over $\mathbb{E}_{data}$ (i.e., if we train $\hat{g}_{\gamma,\boldsymbol{X}^{tr}_{NR},\boldsymbol{Y}^{tr}_{NR}}$ using $\boldsymbol{X}^{tr}_{NR}$, $\boldsymbol{Y}^{tr}_{NR}$, then $\mathbb{E}_{data}$ means $\mathbb{E}_{\boldsymbol{X}^{tr}_{NR},\boldsymbol{Y}^{tr}_{NR}}$ ). The same interpretation applies for the results with Big-$\mathcal{O}$ notations. Finally, constants $C_2, C_2'$, $C_3$, $C_3'$ and $C_3''$ as well as similar constants introduced later which depend on $R, g(\cdot)$ or $\delta$ (for $1-\delta$ high probability bound) will be denoted by a common $C$ during proof for the ease of presentation.

## 7.1 Preliminaries

**Lemma 1.** *Under Assumption 3, for any $f \in \mathcal{H}$, we have*

$$\|f\|_\infty = \sup_{x\in\mathcal{X}} |\langle f(\cdot), \Phi(\cdot, x)\rangle_{\mathcal{H}}| \leq R\|f\|_{\mathcal{H}}. \tag{15}$$

*and consequently $\|f\|_{\mathscr{L}^2_{P_{tr}}} \leq R\|f\|_{\mathcal{H}}$ as well.*

**Lemma 2** (Azuma-Hoeffding). *Let $X_1, ..., X_n$ be independent and identically distributed random variables with $0 \leq X \leq B$, then*

$$P(|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i - \mathbb{E}[X]| > \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{B^2}}. \tag{16}$$

**Corollary 2.** *Under the same assumption of Lemma 2, with probability at least $1-\delta$,*

$$|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i - \mathbb{E}[X]| \leq B\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}. \tag{17}$$

Moreover, an important probability $1-\delta$ bound we shall use later for $\hat{L}(\boldsymbol{\beta}_{|\boldsymbol{x}^{tr}_1,\ldots,\boldsymbol{x}^{tr}_{n_{tr}}})$ follows from Yu & Szepesvári (2012) (see also Gretton et al. (2009) and Pinelis et al. (1994)) :

$$\begin{aligned}\hat{L}(\boldsymbol{\beta}_{|\boldsymbol{x}^{tr}_1,\ldots,\boldsymbol{x}^{tr}_{n_{tr}}}) =& \left\| \frac{1}{n_{tr}}\sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}^{tr}_j)\Phi(\boldsymbol{x}^{tr}_j) - \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} \Phi(\boldsymbol{x}^{te}_i)\right\|_{\mathcal{H}} \\ \leq& \sqrt{2\log\frac{2}{\delta}} R \sqrt{\left(\frac{B^2}{n_{tr}} + \frac{1}{n_{te}}\right)}.\end{aligned} \tag{18}$$

## 7.2 Learning Theory Estimates

To adopt the assumption in Yu & Szepesvári (2012); Cucker & Zhou (2007) that the true regression function $g(\cdot) \notin \mathcal{H}$ but $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, we introduce the related results from learning theory.

First, define $\zeta \triangleq \frac{\theta}{2\theta+4}$ for some $\theta > 0$ so that $0 < \zeta < 1/2$. Given $g(\cdot) \in Range(\mathcal{T}_K^{\zeta})$ and $m$ training sample $\{(\boldsymbol{x}_j, y_j)\}_{j=1}^m$ (sampled from $P_{tr}$)), we define $g_\gamma(\cdot) \in \mathcal{H} : \mathcal{X} \to \mathbb{R}$ to be

$$g_\gamma(\cdot) = \operatorname*{argmin}_{f\in\mathcal{H}} \left\{ \|f-g\|^2_{\mathscr{L}^2_{P_{tr}}} + \gamma\|f\|^2_{\mathcal{H}} \right\} \tag{19}$$

where $\|f - g\|_{\mathscr{L}^2_{P_{tr}}} = \sqrt{\mathbb{E}_{\boldsymbol{x}\sim P_{tr}}(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2}$ denotes the $\mathscr{L}^2$ norm under $P_{tr}$. On the other hand, $\hat{g}_{\gamma,data}(\cdot) \in \mathcal{H}$ is defined in (3) as

$$\hat{g}_{\gamma,data}(\cdot) = \operatorname*{argmin}_{f\in\mathcal{H}} \left\{ \frac{1}{m}\sum_{j=1}^{m}(f(\boldsymbol{x}_j) - y_j)^2 + \gamma\|f\|_{\mathcal{H}}^2 \right\}.$$

Moreover, following the notations in section 4.5 of Cucker & Zhou (2007), given Banach space $(\mathscr{L}^2_{P_{tr}}, \|\cdot\|_{\mathscr{L}^2_{P_{tr}}})$ and the kernel-induced Hilbert subspace $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, we define a $\tilde{\mathbb{K}}$-functional: $\mathscr{L}^2_{P_{tr}} \times (0,\infty) \to \mathbb{R}$ to be

$$\tilde{\mathbb{K}}(l,\gamma) \triangleq \inf_{f\in\mathcal{H}}\{\|l - f\|_{\mathscr{L}^2_{P_{tr}}} + \gamma\|f\|_{\mathcal{H}}\}$$

for $l(\cdot) \in \mathscr{L}^2_{P_{tr}}$ and $t > 0$. Moreover, for $0 < r < 1$, the interpolation space $(\mathscr{L}^2_{P_{tr}}, \mathcal{H})_r$ consists of all the elements $l(\cdot) \in \mathscr{L}^2_{P_{tr}}$ such that

$$\|l\|_r \triangleq \sup_{\gamma>0}\frac{\tilde{\mathbb{K}}(l,\gamma)}{\gamma^r} < \infty. \tag{20}$$

**Lemma 3.** *Define* $\mathbb{K} : \mathscr{L}^2_{P_{tr}} \times (0,\infty) \to \mathbb{R}$ *to be*

$$\mathbb{K}(l,\gamma) \triangleq \inf_{f\in\mathcal{H}}\{\|l - f\|^2_{\mathscr{L}^2_{P_{tr}}} + \gamma\|f\|^2_{\mathcal{H}}\}, \tag{21}$$

*then for any* $l(\cdot) \in (\mathscr{L}^2_{P_{tr}}, \mathcal{H})_r$, *we have*

$$\sup_{\gamma>0}\frac{\mathbb{K}(l,\gamma)}{\gamma^r} \leq \left(\sup_{\gamma>0}\frac{\tilde{\mathbb{K}}(l,\sqrt{\gamma})}{(\sqrt{\gamma})^r}\right)^2 = \|l\|_r^2 < \infty. \tag{22}$$

*Proof.* It follows from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}, \quad \forall a, b \geq 0$ that

$$\sqrt{\mathbb{K}(l,\gamma)} \leq \tilde{\mathbb{K}}(l,\sqrt{\gamma}). \tag{23}$$

Thus, for any $l(\cdot) \in (\mathscr{L}^2_{P_{tr}}, \mathcal{H})_r$, we have

$$\sup_{\gamma>0}\frac{\mathbb{K}(l,\gamma)}{\gamma^r} \leq \left(\sup_{\gamma>0}\frac{\tilde{\mathbb{K}}(l,\sqrt{\gamma})}{(\sqrt{\gamma})^r}\right)^2 = \|l\|_r^2 < \infty. \tag{24}$$

□

On the other hand, assuming $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, it follows from the proof of Theorem 4.1 in Cucker & Zhou (2007) that

$$g(\cdot) \in (\mathscr{L}^2_{P_{tr}}, \mathcal{H}^+)_{\frac{\theta}{\theta+2}} \tag{25}$$

where $\mathcal{H}^+$ is a closed subspace of $\mathcal{H}$ spanned by eigenfunctions of the kernel $K$ (e.g.,$\mathcal{H}^+ = \mathcal{H}$ when $P_{tr}$ is non-degenerate, see Remark 4.18 of Cucker & Zhou (2007)). Indeed, the next lemma shows we can measure smoothness through interpolation space just as range space.

**Lemma 4.** *Assuming* $P_{tr}$ *is non-degenerate on* $\mathcal{X}$. *Then if* $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, *we have* $g \in (\mathscr{L}^2_{P_{tr}}, \mathcal{H})_{\frac{\theta}{\theta+2}}$. *On the other hand, if* $g \in (\mathscr{L}^2_{P_{tr}}, \mathcal{H})_{\frac{\theta}{\theta+2}}$, *then* $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}-\epsilon})$ *for all* $\epsilon > 0$.

*Proof.* It follows from Theorem 4.1, Corollary 4.17 and Remark 4.18 of Cucker & Zhou (2007). □

Now we are ready to adopt the standard assumptions and theoretical results from learning theory in RKHS. They can be found in Cucker & Zhou (2007); Sun & Wu (2009); Smale & Zhou (2007); Yu & Szepesvári (2012). First, given $g(\cdot) \in Range(\mathcal{T}_K^{\zeta})$ and $m$ training sample $\{(\boldsymbol{x}_j, y_j)\}_{j=1}^m$ (sampled from $P_{tr}$)), it follows from Lemma 3 of Smale & Zhou (2007) (see as well Remark 3.3 and Corollary 3.2 in Sun & Wu (2009)) that

$$\|g_\gamma - g\|_{\mathscr{L}^2_{P_{tr}}} \leq C_2\gamma^{\zeta}. \tag{26}$$

Second, it follows from Theorem 3.1 in Sun & Wu (2009) as well as Smale & Zhou (2007); Sun & Wu (2010) that

$$\|g_\gamma - \hat{g}_{\gamma,data}\|_{\mathscr{L}^2_{P_{tr}}} \leq C_2'(\gamma^{-1/2}m^{-1/2} + \gamma^{-1}m^{-3/4}), \tag{27}$$

and, by triangle inequality,

$$\|g - \hat{g}_{\gamma,data}\|_{\mathscr{L}^2_{P_{tr}}} \leq C_3(\gamma^{\zeta} + \gamma^{-1/2}m^{-1/2} + \gamma^{-1}m^{-3/4}). \tag{28}$$

Notice here that by choosing $\gamma = m^{-\frac{3}{4(1+\zeta)}}$, we recover the Corollary 3.2 of Sun & Wu (2009). Finally it follows from Theorem of Smale & Zhou (2007), we have

$$\|g_\gamma - \hat{g}_{\gamma,data}\|_{\mathcal{H}} \leq C_3'\gamma^{-1}m^{-1/2}, \tag{29}$$

with $C_3' = 6R\log\frac{2}{\delta}$. In fact, if we define $\sigma^2 \triangleq \mathbb{E}_{\boldsymbol{x}\sim P_{tr}}\mathbb{E}_{Y|\boldsymbol{x}}(g(\boldsymbol{x}) - Y)^2$, then Theorem 3 of Smale & Zhou (2007) stated that $\|g_\gamma - \hat{g}_{\gamma,data}\|_{\mathcal{H}} \leq C_3''((\sqrt{\sigma^2} + \|g_\gamma - g\|_{\mathscr{L}^2_{P_{tr}}})\gamma^{-1}m^{-1/2} + \gamma^{-1}m^{-1})$.

### 7.3 Main Proofs

*Proof of Theorem 1 and Corollary 1.* If $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ (i.e. $\zeta = \frac{\theta}{2\theta+4}$) and we set $h(\cdot) = g_\gamma(\cdot)$ and $\hat{g} = \hat{g}_{\gamma,\boldsymbol{X}^{tr}_{NR},\boldsymbol{Y}^{tr}_{NR}}$ for some $\gamma > 0$, then:

$$\begin{aligned}
&V_R(\rho) - \nu \\
=&\frac{1}{\lfloor\rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}\hat{\beta}(\boldsymbol{x}_j^{tr})(y_j^{tr} - g(\boldsymbol{x}_j^{tr})) + \frac{1}{\lfloor\rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))(g(\boldsymbol{x}_j^{tr}) - h(\boldsymbol{x}_j^{tr})) \\
&+ \frac{1}{\lfloor\rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))(h(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr})) \\
&+ \frac{1}{\lfloor\rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}\beta(\boldsymbol{x}_j^{tr})(g(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr})) + \frac{1}{n_{te}}\sum_{i=1}^{n_{te}}\hat{g}(\boldsymbol{x}_i^{te}) - \nu.
\end{aligned} \tag{30}$$

To bound terms in (30), we first use Corollary 2 to conclude that with probability at least $1-\delta$,

$$|\frac{1}{\lfloor\rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}\hat{\beta}(\boldsymbol{x}_j^{tr})(y_j^{tr} - g(\boldsymbol{x}_j^{tr}))| \leq B\sqrt{\frac{1}{\lfloor\rho n_{tr}\rfloor}\log\frac{2}{\delta}} = \mathcal{O}(n_{tr}^{-1/2}). \tag{31}$$

We hold on our discussion for the second term. For the third term, since $h, \hat{g} \in \mathcal{H}$,

$$\begin{aligned}
&\left|\frac{1}{\lfloor\rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))(h(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr}))\right| \\
=&\left|\frac{1}{\lfloor\rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))\langle h - \hat{g}, \Phi(\boldsymbol{x}_j^{tr})\rangle_{\mathcal{H}}\right| \\
=&\left|\left\langle h - \hat{g}, \frac{1}{\lfloor\rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))\Phi(\boldsymbol{x}_j^{tr})\right\rangle_{\mathcal{H}}\right| \\
\leq&\|h - \hat{g}\|_{\mathcal{H}}(\hat{L}(\hat{\boldsymbol{\beta}}) + \hat{L}(\boldsymbol{\beta}_{|\boldsymbol{x}_1^{tr},\ldots,\boldsymbol{x}_{\lfloor\rho n_{tr}\rfloor}^{tr}})) \leq 2\|h - \hat{g}\|_{\mathcal{H}}\hat{L}(\boldsymbol{\beta}_{|\boldsymbol{x}_1^{tr},\ldots,\boldsymbol{x}_{\lfloor\rho n_{tr}\rfloor}^{tr}}),
\end{aligned} \tag{32}$$

by definition of (1). Thus, when taking $h = g_\gamma$ and $\hat{g} = \hat{g}_{\gamma,\boldsymbol{X}^{tr}_{NR},\boldsymbol{Y}^{tr}_{NR}}$ for some $\gamma$, we can combine (18) and (29) to guarantee, with probability $1 - 2\delta$,

$$\begin{aligned}
&\left|\frac{1}{\lfloor\rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))(h(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr}))\right| \\
\leq&\sqrt{8\log\frac{2}{\delta}}RC(1-\rho)^{-1/2}(\gamma^{-1}n_{tr}^{-1/2}) \cdot \sqrt{\left(\frac{B^2}{n_{tr}} + \frac{1}{n_{te}}\right)} \\
=&\mathcal{O}(\gamma^{-1}n_{tr}^{-1/2}(n_{tr}^{-1} + n_{te}^{-1})^{\frac{1}{2}}).
\end{aligned} \tag{33}$$

For the last term $\tau \triangleq \frac{1}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}\beta(\boldsymbol{x}_j^{tr})(g(\boldsymbol{x}_j^{tr})-\hat{g}(\boldsymbol{x}_j^{tr}))+\frac{1}{n_{te}}\sum_{i=1}^{n_{te}}\hat{g}(\boldsymbol{x}_i^{te})-\nu$, the analysis relies the splitting of data, as we notice that,

$$
\begin{aligned}
&\mathbb{E}_{|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\left[\frac{1}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}\beta(\boldsymbol{x}_j^{tr})(g(\boldsymbol{x}_j^{tr})-\hat{g}(\boldsymbol{x}_j^{tr}))+\frac{1}{n_{te}}\sum_{i=1}^{n_{te}}\hat{g}(X_i^{te})-\nu\right]\\
=&\mathbb{E}_{\boldsymbol{x}\sim P_{tr}}[\beta(\boldsymbol{x})g(\boldsymbol{x})]-\nu-\mathbb{E}_{\boldsymbol{x}\sim P_{tr}}[\beta(\boldsymbol{x})\hat{g}(\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{x}\sim P_{te}}[\hat{g}(\boldsymbol{x})]\\
=&\mathbb{E}_{\boldsymbol{x}\sim P_{te}}[g(\boldsymbol{x})]-\nu-\mathbb{E}_{\boldsymbol{x}\sim P_{te}}[\hat{g}(\boldsymbol{x})]+\mathbb{E}_{\boldsymbol{x}\sim P_{te}}[\hat{g}(\boldsymbol{x})]\\
=&0.
\end{aligned}
\tag{34}
$$

Notice for the second line follows since $\hat{g}(\cdot)$ is determined by $\{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}\}$ and thus is independent with $\{\boldsymbol{X}_{KMM}^{tr},\boldsymbol{Y}_{KMM}^{tr}\}$ or $\{\boldsymbol{X}^{te}\}$. Thus, we have

$$
\begin{aligned}
\text{Var}(\tau)=&\text{Var}(\mathbb{E}_{|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}(\tau))+\mathbb{E}[\text{Var}_{|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}(\tau)]\\
=&\mathbb{E}[\text{Var}_{|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}(\tau)]\\
=&\frac{1}{\lfloor \rho n_{tr} \rfloor}\mathbb{E}[\text{Var}_{\boldsymbol{x}\sim P_{tr}|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}(\beta(\boldsymbol{x})(g(\boldsymbol{x})-\hat{g}(\boldsymbol{x})))]+\frac{1}{n_{te}}\mathbb{E}[\text{Var}_{\boldsymbol{x}\sim P_{te}|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}(\hat{g}(\boldsymbol{x}))]\\
\leq&\frac{B^2}{\lfloor \rho n_{tr} \rfloor}\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|g-\hat{g}\|_{\mathscr{L}_{P_{tr}}^2}^2+\frac{1}{n_{te}}\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|\hat{g}\|_{\mathscr{L}_{P_{te}}^2}^2\\
\leq&\frac{B^2}{\lfloor \rho n_{tr} \rfloor}\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|g-\hat{g}\|_{\mathscr{L}_{P_{tr}}^2}^2+\frac{B}{n_{te}}\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|\hat{g}\|_{\mathscr{L}_{P_{tr}}^2}^2,
\end{aligned}
\tag{35}
$$

and we can use the Chebyshev's inequality and Lemma 1 to conclude, with probability at least $1-\delta$,

$$
|\tau|\leq\sqrt{\frac{1}{\delta}}\sqrt{\frac{B^2}{\lfloor \rho n_{tr} \rfloor}\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|g-\hat{g}\|_{\mathscr{L}_{P_{tr}}^2}^2+\frac{BR^2}{n_{te}}},
\tag{36}
$$

which becomes, by (28), probability $1-2\delta$:

$$
\begin{aligned}
|\tau|\leq&\sqrt{\frac{1}{\delta}}\sqrt{\frac{B^2}{\lfloor \rho n_{tr} \rfloor}C(1-\rho)^{-3/4}(\gamma^{\zeta}+\gamma^{-1/2}n_{tr}^{-1/2}+\gamma^{-1}n_{tr}^{-3/4})+\frac{BR^2}{n_{te}}}\\
=&\mathcal{O}((\gamma^{\zeta}+\gamma^{-1/2}n_{tr}^{-1/2}+\gamma^{-1}n_{tr}^{-3/4})n_{tr}^{-1/2}+n_{te}^{-1/2})
\end{aligned}
\tag{37}
$$

with $\zeta=\frac{\theta}{2\theta+4}$. Now, to bound the second term $\frac{1}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr})-\beta(\boldsymbol{x}_j^{tr}))(g(\boldsymbol{x}_j^{tr})-h(\boldsymbol{x}_j^{tr}))$,

$$
\begin{aligned}
&\frac{1}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}|(\hat{\beta}(\boldsymbol{x}_j^{tr})-\beta(\boldsymbol{x}_j^{tr}))(g(\boldsymbol{x}_j^{tr})-g_{\gamma}(\boldsymbol{x}_j^{tr}))|\\
\leq&\frac{B}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}|g(\boldsymbol{x}_j^{tr})-g_{\gamma}(\boldsymbol{x}_j^{tr})|\\
\leq&\Big|\frac{B}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}|g(\boldsymbol{x}_j^{tr})-g_{\gamma}(\boldsymbol{x}_j^{tr})|-B\|g-g_{\gamma}\|_{\mathscr{L}_{P_{tr}}^1}\Big|+B\|g-g_{\gamma}\|_{\mathscr{L}_{P_{tr}}^1}\\
\leq&\sqrt{\frac{1}{\delta}}\sqrt{\frac{B^2}{\rho n_{tr}}\|g-g_{\gamma}\|_{\mathscr{L}_{P_{tr}}^2}^2}+B\|g-g_{\gamma}\|_{\mathscr{L}_{P_{tr}}^2}\\
\leq&\sqrt{\frac{1}{\delta}}BC\gamma^{\zeta}\sqrt{\frac{1}{\rho n_{tr}}}+C\gamma^{\zeta}=\mathcal{O}(\gamma^{\zeta})=\mathcal{O}(\gamma^{\frac{\theta}{2\theta+4}}).
\end{aligned}
\tag{38}
$$

where $\mathscr{L}_{P_{tr}}^1$ denotes the 1-norm $\mathbb{E}_{\boldsymbol{x}\sim P_{tr}}|g(\boldsymbol{x})-g_{\gamma}(\boldsymbol{x})|$. Notice the second to last line follows from Chebyshev inequality, Cauchy-Schwarz inequality and the last line from (26).

Thus, when taking $h = g_\gamma$ and $\hat{g} = \hat{g}_{\gamma, \boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}}$ for some $\gamma > 0$, we can combine (31),(33),(37) and (38) to have,

$$\begin{aligned} |V_R(\rho) - \nu| =& \mathcal{O}(n_{tr}^{-\frac{1}{2}}) + \mathcal{O}(\gamma^{\frac{\theta}{2\theta+4}}) + \mathcal{O}(\gamma^{-1} n_{tr}^{-1/2} (n_{tr}^{-1} + n_{te}^{-1})^{\frac{1}{2}}) \\ &+ \mathcal{O}((\gamma^{\frac{\theta}{2\theta+4}} + \gamma^{-1/2} n_{tr}^{-1/2} + \gamma^{-1} n_{tr}^{-3/4}) n_{tr}^{-1/2} + n_{te}^{-1/2}) \\ =& \mathcal{O}(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}} + \gamma^{\frac{\theta}{2\theta+4}} + \gamma^{-\frac{1}{2}} n_{tr}^{-1} + \gamma^{-\frac{1}{2}} n_{tr}^{-\frac{1}{2}} n_{te}^{-\frac{1}{2}}), \end{aligned} \tag{39}$$

after simplification. Now, if we take $\gamma = n^{-\frac{\theta+2}{\theta+1}}$ where $n \triangleq \min(n_{tr}, n_{te})$, then (39) becomes

$$\begin{aligned} |V_R(\rho) - \nu| = \mathcal{O}(n^{-\frac{1}{2}} + n^{-\frac{\theta}{2(\theta+1)}} + n^{\frac{\theta+2}{2(\theta+1)}} n^{-1}) =& \mathcal{O}(n^{-\frac{\theta}{2\theta+2}}) \\ =& \mathcal{O}(n_{tr}^{-\frac{\theta}{(2\theta+2)}} + n_{te}^{-\frac{\theta}{(2\theta+2)}}), \end{aligned} \tag{40}$$

which is the statement of the theorem. However, note that if we choose $\gamma = n^{-1}$, the rate becomes $\mathcal{O}(n_{tr}^{-\frac{\theta}{(2\theta+4)}} + n_{te}^{-\frac{\theta}{(2\theta+4)}})$. Moreover if $\lim_{n \to \infty} n_{te}^{\frac{6\theta+8}{3\theta+6}} / n_{tr} \to 0$ and we choose $\gamma = n_{tr}^{-1}$, then the rate becomes $\mathcal{O}(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{1}{2}})$. □

*Proof of Proposition 1.* Fixing $\gamma > 0$, if $g \in \mathcal{H}(i.e., g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ with $\theta \to \infty))$, then by definition of $g_\gamma$ we would have:

$$\|g_\gamma\|^2_{\mathcal{H}} \leq \frac{\|g_\gamma - g\|^2_{\mathscr{L}^2_{P_{tr}}} + \gamma \|g_\gamma\|^2_{\mathcal{H}}}{\gamma} \leq \frac{\|g - g\|^2_{\mathscr{L}^2_{P_{tr}}} + \gamma \|g\|^2_{\mathcal{H}}}{\gamma} = \|g\|^2_{\mathcal{H}}, \tag{41}$$

or equivalently $\|g_\gamma\|_{\mathcal{H}} = \mathcal{O}(1)$ since the fixed true regression function $\|g\|_{\mathcal{H}} = \mathcal{O}(1)$. Thus, a simplified analysis shows:

$$\begin{aligned} V_R(\rho) - \nu =& \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) Y_j^{tr} - \nu \\ &+ \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \hat{g}(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te}) \end{aligned} \tag{42}$$

Note that the first term on the right is nothing but the $V_{KMM}$ estimator with $100 \times \rho$ percent of the training data and we shall denote it as $V_{KMM}(\rho)$ without ambiguity. For the second term, assuming $\hat{g} = \hat{g}_{\gamma, \boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}}$, is bounded by

$$\begin{aligned} & \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \hat{g}(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te}) \\ =& \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \langle \hat{g}, \Phi(\boldsymbol{x}_j^{tr}) \rangle_{\mathcal{H}} - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \langle \hat{g}, \Phi(\boldsymbol{x}_i^{n_{te}}) \rangle_{\mathcal{H}} \\ =& \left\langle \hat{g}, \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{i=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \Phi(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \Phi(\boldsymbol{x}_i^{te}) \right\rangle_{\mathcal{H}} \leq \|\hat{g}_{\gamma, \boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}}\|_{\mathcal{H}} \hat{L}(\hat{\boldsymbol{\beta}}), \end{aligned} \tag{43}$$

Then, by (42) and (43), we have

$$\begin{aligned} |V_R(\rho) - \nu| \leq& |V_{KMM}(\rho) - \nu| + \hat{L}(\hat{\boldsymbol{\beta}})(\|g_\gamma - \hat{g}_{\gamma, \boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}}\|_{\mathcal{H}} + \|g_\gamma\|_{\mathcal{H}}) \\ =& \mathcal{O}(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}}), \end{aligned} \tag{44}$$

following (41), (29) and Theorem 1 of Yu & Szepesvári (2012). □

*Proof of Proposition 2.* If $g$ only satisfies the condition $\mathcal{A}_\infty(g,F) \triangleq \inf_{\|f\|_{\mathcal{H}}\le F} \|g-f\| \le C(\log F)^{-s}$ for some $C,s>0$, then we again follow the analysis in the proof of Proposition 1 and arrive at the decomposition in (42)

$$\begin{aligned}|V_R(\rho)-\nu| \le& |V_{KMM}(\rho)-\nu| + \hat{L}(\hat{\boldsymbol{\beta}})(\|g_\gamma - \hat{g}_{\gamma,\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|_{\mathcal{H}} + \|g_\gamma\|_{\mathcal{H}}) \\ =&\mathcal{O}(\log^{-s} \frac{n_{tr}n_{te}}{n_{tr}+n_{te}}),\end{aligned} \tag{45}$$

which is the rate of $V_{KMM}$ by Theorem 3 of Yu & Szepesvári (2012). □

*Proof.* Proof of Theorem 2

Define $\epsilon \triangleq \sup_{\theta\in\mathcal{D}} \left| V_R(\theta) - \mathbb{E}[l'(X^{te},Y^{te};\theta)]\right|$, we have

$$\mathbb{E}[l'(X_{te},Y_{te};\hat{\theta}_R)] - \epsilon \le V_R(\hat{\theta}_R) \le V_R(\theta^\star) \le \mathbb{E}[l'(X_{te},Y_{te};\theta^\star)] + \epsilon. \tag{46}$$

On the other hand, we know by triangle inequality that $\epsilon$ is bounded by

$$\sup_{\theta\in\mathcal{D}} \big| \frac{1}{\lfloor \rho n_{tr}\rfloor} \sum_{j=1}^{\lfloor \rho n_{tr}\rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) l'(\boldsymbol{x}_j^{tr}, y_j^{tr};\theta) - \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} l(\boldsymbol{x}_i^{te};\theta)\big|$$

$$+ \sup_{\theta\in\mathcal{D}} \big| \frac{1}{\lfloor \rho n_{tr}\rfloor} \sum_{j=1}^{\lfloor \rho n_{tr}\rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \hat{l}(\boldsymbol{x}_j^{tr};\theta) - \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} \hat{l}(\boldsymbol{x}_i^{te};\theta)\big| + \sup_{\theta\in\mathcal{D}} \big| \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} l(\boldsymbol{x}_i^{te};\theta) - \mathbb{E}[l(X_{te};\theta)]\big|,$$

where the first term is bounded by $\mathcal{O}(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}})$ following Corollary in Gretton et al. (2009). Moreover, the second term is also $\mathcal{O}(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}})$ as in (43) or Lemma 8.7 in Gretton et al. (2009). For the last term, due to the Lipschitz and compact assumption, it follows from Theorem 19.5 of Van der Vaart (2000) (see also Example 19.7 of Van der Vaart (2000)) that function class $\mathcal{G}$ is $P_{te}$-Donsker, which means that

$$\mathbb{G}_n(\theta) \triangleq \sqrt{n_{te}}\left(\frac{1}{n_{te}}\sum_{i=1}^{n_{te}} l(\boldsymbol{x}_i^{te};\theta) - \mathbb{E}_{\boldsymbol{x}\sim P_{te}}[l(\boldsymbol{x};\theta)]\right)$$

converges in distribution to a Gaussian Process $\mathbb{G}_\infty$ with zero mean and covariance function $\mathrm{Cov}(\mathbb{G}_\infty(\theta_1),\mathbb{G}_\infty(\theta_2)) = \mathbb{E}_{\boldsymbol{x}\sim P_{te}}(l(\boldsymbol{x};\theta_1)l(\boldsymbol{x};\theta_2)) - \mathbb{E}_{\boldsymbol{x}\sim P_{te}} l(\boldsymbol{x};\theta_1)\mathbb{E}_{\boldsymbol{x}\sim P_{te}} l(\boldsymbol{x};\theta_2)$. Notice $\mathbb{G}_\infty$ can be viewed as random function in $C(\mathcal{D})$, the space of continuous and bounded function on $\theta$. Since for any $z\in C(\mathcal{D})$, the mapping $z \to \|z\|_\infty \triangleq \sup_{\theta\in\mathcal{D}} z(\theta)$ is continuous with respect to the supremum norm, it follows from the continuous-mapping theorem that $n_{te}^{\frac{1}{2}} \sup_{\theta\in\mathcal{D}} \big|\frac{1}{n_{te}}\sum_{i=1}^{n_{te}} l(\boldsymbol{x}_i^{te};\theta) - \mathbb{E}[l(X_{te};\theta)]\big|$ converges in distribution to $\|\mathbb{G}_\infty\|_\infty$ which has finite expectations based on the assumptions on $\mathcal{G}$ (see, e.g., Section 14, Theorem 1 of Lifshits (2013)). Thus, by definition of convergence in distribution, for any $\delta>0$, we can find some constant $D'$ that

$$P[\|\mathbb{G}_n\|_\infty > D'] = P[\|\mathbb{G}_\infty\|_\infty > D'] + o(1) \le \delta + o(1), \tag{47}$$

which means, we can find some $N$ such that when $n_{te} > N$,

$$P_{te}\big(\sup_{\theta\in\mathcal{D}} \left|\frac{1}{n_{te}}\sum_{i=1}^{n_{te}} l(\boldsymbol{x}_i^{te};\theta) - \mathbb{E}[l(X_{te};\theta)]\right| > n_{te}^{-\frac{1}{2}} D'\big) = P_{te}(\|\mathbb{G}_n\|_\infty > D') \le 2\delta,$$

and consequently, with probability $1-2\delta$, we have

$$\sup_{\theta\in\mathcal{D}} \big|\frac{1}{n_{te}}\sum_{i=1}^{n_{te}} l(\boldsymbol{x}_i^{te};\theta) - \mathbb{E}[l(X_{te};\theta)]\big| \le n_{te}^{-\frac{1}{2}} D'.$$

In other words, we also have

$$\sup_{\theta\in\mathcal{D}} \big|\frac{1}{n_{te}}\sum_{i=1}^{n_{te}} l(\boldsymbol{x}_i^{te};\theta) - \mathbb{E}[l(X_{te};\theta)]\big| = \mathcal{O}(n_{te}^{-\frac{1}{2}}),$$

which concludes our proof. □