Robust Importance Weighting for Covariate Shift

Henry Lam  
Columbia University  
khl2114@columbia.edu

Fengpei Li  
Columbia University  
Email fl2412@columbia.edu

Siddharth Prusty  
Columbia University  
siddharth.prusty@columbia.edu

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 and the weights are poorly estimated. 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 can be shown to either strictly outperform or match the best-known existing rates for both KMM and NR, and thus is a robust combination of both estimators. The experiments show the 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 [Heckman, 1979, Zadrozny, 2004] Sugiyama et al., 2007, Quionero-Candela et al., 2009, Tzeng et al., 2017, Jiang and Zhai, 2007, Borgwardt et al., 2006]; patterns for engineering controls fluctuate due to the non-stationarity of environments Sugiyama and 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 and Yang, 2009 or domain adaptation Jiang and 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 and Doretto, 2010, Pardoe and Stone, 2010, Schölkopf et al., 2002, Quionero-Candela et al., 2009, Sugiyama and 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 and Szepesvári, 2012, under a more realistic assumption from learning theory Cucker and 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 \( O(n_{te}^{-\frac{\theta}{\theta+1}} + n_{te}^{-\frac{\theta}{\theta+2}}) \) from the parametric rate \( O(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}}) \). 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 and Lam, 2012, partially due to an 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., 2008b 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, the residuals have smaller magnitudes than the original loss values, and the resulting reweighted estimator is thus less sensitive to the variances of weights. Then, we cancel the bias incurred by the use of residuals by a judicious compensation through the estimated regression function evaluated on the test set.

Our method shares similarities with the Doubly Robust (DR) estimator in causal inference problems Kennedy et al., 2017. However, different from DR, we do not require semi-parametric estimates of the baseline prediction (corresponding to our regression function \( g \)) and conditional probability (corresponding to our importance weight) to both converge at rates \( O(n^\alpha) \) for \( \alpha > 1/4 \). In particular, we specialize our method by using a nonparametric regression (NR) function constructed from regularized least square in RKHS Cucker and Zhou, 2007; Smale and Zhou, 2007; Sun and Wu, 2009, also known as the Tikhonov regularized learning algorithm Evgeniou et al., 2000. We show that our new estimator achieves the rate \( O(n_{tr}^{-\frac{\theta}{\theta+1}} + n_{te}^{-\frac{\theta}{\theta+2}}) \), which is superior to the best-known rate of KMM in Yu and 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 and Szepesvári, 2012 which constructs an 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 and Szechtman, 2002 and a bias reduction perspective on NR.

Another noticeable 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

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

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

Assumption 2. The Radon-Nikodym derivative \( \beta(x) \triangleq \frac{dP_{te}}{dP_{tr}}(x) \) exists and is bounded by \( B < \infty \).
Assumption 3. The covariate space $X$ is compact and the label space $Y \subseteq [0,1]$. Furthermore, there exists a kernel $K(\cdot,\cdot) : X \times X \rightarrow \mathbb{R}$ which induces an RKHS $H$ and a canonical feature map $\Phi(\cdot) : X \rightarrow H$ such that $K(x,x') = \langle \Phi(x), \Phi(x') \rangle_H$ and $\|\Phi(x)\|_H \leq R$ for some $0 < R < \infty$.

Assumption 1 is the covariate shift assumption which states the conditional distribution $P(dy|x)$ remains invariant while the marginal $P_t(x)$ and $P_e(x)$ differ. Assumptions 2 and 3 are common for establishing theoretical results. Specifically, Assumption 2 can be satisfied by restricting the support of $P_e$ and $P_t$ on a compact set, although $B$ could be potentially large.

2.1 Preliminaries and Existing Approaches

Given $n_t$ labelled training data $\{(x_{tr}^1, y_{tr}^1)\}_{j=1}^{n_t}$ and $n_e$ unlabelled test data $\{(x_{te}^1)\}_{i=1}^{n_e}$ (i.e., $\{y_{te}^1\}_{i=1}^{n_e}$ are unavailable), the goal is to estimate $\nu = \mathbb{E}[Y_{te}]$. The KMM estimator [Huang et al., 2007; Gretton et al., 2009]

$$V_{KMM} = \frac{1}{n_t} \sum_{j=1}^{n_t} \hat{\beta}(x_{tr}^j) y_{tr}^j,$$

where $\hat{\beta}(x_{tr}^j)$ are solutions of a QP that attempts to match the means of training and test sets in the feature space using weights $\hat{\beta}$:

$$\min_{\hat{\beta}} \left\{ \frac{1}{n_t} \sum_{j=1}^{n_t} \hat{\beta}(x_{tr}^j) - \frac{1}{n_e} \sum_{i=1}^{n_e} \hat{\Phi}(x_{te}^i) \right\}^2_H$$

s.t. $0 \leq \hat{\beta}_j \leq B, \forall 1 \leq j \leq n_t$.

Notice we write $\hat{\beta}_j$ as $\hat{\beta}(x_{tr}^j)$ in $V_{KMM}$ informally to highlight $\hat{\beta}_j$ as estimates of $\beta(x_{tr}^j)$. The fact that (1) is a QP can be verified by the kernel trick, as in [Gretton et al., 2009]. Indeed, define matrix $K_{ij} = K(x_{tr}^i, x_{tr}^j)$ and $\kappa_j = \sum_{i=1}^{n_t} K(x_{tr}^i, x_{tr}^j)$, optimization (1) is equivalent to

$$\min_{\hat{\beta}} \frac{1}{n_t} \hat{\beta}^T K \hat{\beta} - \frac{2}{n_t} \kappa^T \hat{\beta},$$

s.t. $0 \leq \hat{\beta}_j \leq B, \forall 1 \leq j \leq n_t$.

In practice, a constraint $\frac{1}{n_t} \sum_{j=1}^{n_t} \hat{\beta}_j - 1 \leq \epsilon$ for a tolerance $\epsilon > 0$ is included to regularize the $\hat{\beta}$ towards the RND. As in [Yu and Szepesvári, 2012], we omit them to simplify analysis. On the other hand, the NR estimator

$$V_{NR} = \frac{1}{n_e} \sum_{i=1}^{n_e} \hat{g}(x_{te}^i),$$

is based on $\hat{g}(\cdot)$, some estimate of the regression function $g(x) \triangleq \mathbb{E}[Y|x]$. Notice the conditional expectation is taken regardless of $x \sim P_t$ or $P_e$. Here, we consider a $\hat{g}(\cdot)$ that is estimated nonparametrically by regularized least square in RKHS:

$$\hat{g}_{\gamma, data}(\cdot) = \arg\min_{f \in H} \left\{ \frac{1}{m} \sum_{j=1}^{m} (f(x_{tr}^j) - y_{tr}^j)^2 + \gamma \|f\|^2_H \right\},$$

where $\gamma$ is a regularization term to be chosen and the subscript $data$ represents $\{(x_{tr}^m, y_{tr}^m)\}_{j=1}^{m}$. Using the representation theorem [Schölkopf et al., 2001], optimization problem (3) can be solved in closed form with $\hat{g}_{\gamma, data}(x) = \sum_{j=1}^{m} \alpha^r e^j K(x_{tr}^j, x)$ where

$$\alpha^r e^j = (K + \gamma I)^{-1} y_{te}^j,$$

and $y_{te}^j = [y_{te}^j, \ldots, y_{te}^j]$.

2.2 Motivation

Depending on properties of $g(\cdot)$, Yu and Szepesvári, 2012 proves different rates of KMM. The most notable case is when $g \notin H$ but rather $g(\cdot) \in$ Range($T_K^{\frac{1}{2}}$), where $T_K$ is the integral operator $< T_K f(x'), k(x, x) dx > \in L^2_{\nu}$. In this case, Yu and Szepesvári, 2012 characterize $g$ with the approximation error

$$A_2(g, F) \triangleq \inf_{\|f\|_H \leq F} \|g - f\|_{L_{\nu}^2} \leq C F^{-\frac{3}{4}},$$

and the rates of KMM drops to sub-canoncial $|V_{KMM} - \nu| = O(n_e \frac{\epsilon}{n_t} + n_e \frac{n_t}{n_e} \frac{\epsilon}{n_t})$, as opposed $O(n_e \frac{\epsilon}{n_t} + n_e \frac{\epsilon}{n_t})$ when $g \in H$. As shown in Lemma 4 in the Appendix and Theorem 4.1 of [Cucker and Zhou, 2007], (5) is almost equivalent to $g(\cdot) \in$ Range($T_K^{\frac{1}{2}}$): $g(\cdot) \in$ Range($T_K^{\frac{1}{2}}$) implies (5) while (5) leads to $g(\cdot) \in$ Range($T_K^{\frac{1}{2}}$) for any $\epsilon > 0$. We adopt the characterization $g(\cdot) \in$ Range($T_K^{\frac{1}{2}}$) as our analysis is based on related learning theory estimates. In particular, our proofs rely on these estimates and are different from [Yu and Szepesvári, 2012]. For example, in [3], $\gamma$ is used as a free parameter for controlling $\|f\|_H$, whereas [Yu and Szepesvári, 2012] uses the parameter $F$ in [3]. 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 for our analysis.

Correspondingly, the convergence rate for $V_{KMM}$ when $g(\cdot) \in$ Range($T_K^{\frac{1}{2}}$) is also shown in [Yu and Szepesvári, 2012].
Szepesvári, 2012 as $|V_{NR} - \nu| = O(n_{te}^{-2} + n_{te}^{-\frac{3g}{2}})$, with $\hat{g}$ taken as $\hat{g}_r$ data in $\Theta$ 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 $\hat{\beta}$). 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) criterion $MSE(V) = Var(V) + (Bias(V))^2$ and assume an additive model $Y = y + \mathcal{E}$ 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 $V_{KMM} \triangleq \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \beta(x_i^{tr}y_i^{tr})$ and $\beta(\cdot)$ the true RND. Since

$$E[\beta(X^{tr})Y^{tr}] = E_{x \sim P_x}[\beta(x)g(x)] = E_{x \sim P_x}[g(x)] = \nu,$$

$V_{KMM}$ 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 $t^* = \frac{Cov(V, W)}{Var(W)}$ and use $V - t^*W$ to obtain

$$\min \text{Var}(V-tW) = (1-\text{corr}^2(V, W))\text{Var}(V) \leq \text{Var}(V),$$

without altering the mean of $V$. Thus we can use

$$W = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(x_j^{tr})(\hat{g}(x_j^{tr})) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(x_i^{te})$$

with $t^* = \frac{Cov(V_{KMM}, W)}{Var(W)}$. To calculate $t^*$, suppose $X^{te}$ and $X^{tr}$ are independent, then we have

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

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, $t^* \approx 1$ which gives our estimator

$$V_R = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(x_j^{tr})(y_j^{tr} - \hat{g}(x_j^{tr})) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(x_i^{te}).$$

**Bias Reduction for NR:** Consider the NR estimator $V_{NR} \triangleq \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(x_i^{te})$. Assuming again the common case where $n_{te} \gg n_{tr}$, we have

$$\text{Var}(V_{NR}) = \frac{1}{n_{te}} \text{Var}(\hat{g}(X^{te})) \approx 0,$$

and the main source of MSE is bias $E_{x \sim P_x}[\hat{g}(x) - \hat{g}(x)]$. If we add $W = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(x_j^{tr})(y_j^{tr} - \hat{g}(x_j^{tr}))$ to $V_{NR}$, we eliminate the bias which gives the same estimator

$$V_R = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(x_j^{tr})(y_j^{tr} - \hat{g}(x_j^{tr})) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(x_i^{te}).$$

### 3 Robust Estimator

We construct a new estimator $V_R(\rho)$ that can be shown to perform robustly against both KMM and NR estimators discussed above. In our construction, we split the training set with a proportion $\rho \in [0,1]$, i.e., divide $\{X^{tr}, Y^{tr}\}_{data} \triangleq \{(x_j^{tr}, y_j^{tr})\}_{j=1}^{n_{tr}}$ into

$$\{X_{KMM}^{tr}, Y_{KMM}^{tr}\}_{data} \triangleq \{(x_j^{tr}, y_j^{tr})\}_{j=1}^{n_{tr}},$$

and

$$\{X_{NR}^{tr}, Y_{NR}^{tr}\}_{data} \triangleq \{(x_j^{tr}, y_j^{tr})\}_{j=n_{tr}+1}^{n_{tr}+n_{te}},$$

where $\{X_{KMM}^{tr}, X_{te}^{tr}\}_{data} \triangleq \{(x_j^{tr})_{j=1}^{n_{tr}}, (x_i^{te})_{i=1}^{n_{te}}\}$ is used to solve for the weight $\beta$ in (1) and $\{X_{NR}^{tr}, Y_{NR}^{tr}\}_{data}$ is used to train an NR function $\hat{g}(\cdot) = \hat{g}_{r, data}(\cdot)$ for some $\gamma$ as in (3). Finally, we define our estimator $V_R(\rho)$ as

$$V_R(\rho) = \frac{1}{|\rho_{tr}|} \sum_{j=1}^{\lceil\rho_{tr}\rceil} \beta(x_j^{tr})(y_j^{tr} - \hat{g}(x_j^{tr})) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(x_i^{te}). \quad (6)$$
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 expression \( 4 \), \( \hat{g}_{\gamma, \text{data}} \) is connected to the learning theory estimates [Cucker and Zhou, 2007]. Thus, for establishing a theoretical bound, we focus on \( \hat{g} = \hat{g}_{\gamma, \text{data}} \) 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 \( V_{KMM} \) either surpasses or achieves the better rate between \( V_{KMM} \) and \( V_{NR} \). In all theorems that follow, the big-\( O \) 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 4, 5 if we assume \( g \in \text{Range}(T_K^{\frac{\theta}{n_{tr}}}) \), the convergence rate of \( \nu \) satisfies

\[
|V_R(\rho) - \nu| = O(n_{tr}^{-\frac{\theta}{n_{tr}}} + n_{te}^{-\frac{\theta}{n_{te}}}),
\]

when \( \hat{g} \) is taken to be \( \hat{g}_{\gamma, \text{data}} \) in (6) with \( \gamma = \frac{d + 2}{d + 4} \) and \( n \triangleq \min(n_{tr}, n_{te}) \).

**Corollary 1.** Under the same setting of Theorem 6 if we choose \( \gamma = n_{tr}^{-1} \), we have

\[
|V_R(\rho) - \nu| = O(n_{tr}^{-\frac{\theta}{n_{tr}}} + n_{te}^{-\frac{\theta}{n_{te}}})
\]

and if we choose \( \gamma = n_{te}^{-1} \),

\[
|V_R(\rho) - \nu| = O(n_{tr}^{-\frac{\theta}{n_{tr}}} + n_{te}^{-\frac{1}{2}}).
\]

We remark several implications. First, although not achieving canonical, \( 8 \) is an improvement over the best-known \( O(n_{tr}^{-\frac{\theta}{n_{tr}}} + n_{te}^{-\frac{\theta}{n_{te}}}) \) rate of \( V_{KMM} \) when \( g \in \text{Range}(T_K^{\frac{\theta}{n_{tr}}}) \), 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 \( \text{Range}(T_K^{\frac{\theta}{n_{tr}}}) \) get smoother and \( \text{Range}(T_K^{\frac{\theta_2}{n_{tr}}}) \subseteq \text{Range}(T_K^{\frac{\theta_1}{n_{tr}}}) \) for \( 0 < \theta_1 < \theta_2 \), with the limiting case that \( \theta \rightarrow \infty, \theta \rightarrow 1/2 \) and \( \text{Range}(T_K^{\frac{1}{2}}) \subseteq \mathcal{H} \) (i.e., \( g \in \mathcal{H} \)) for universal kernels by Mercer’s theorem.

Second, as in Theorem 4 of [Yu and Szepesvári, 2012], the optimal tuning of \( \gamma \) that leads to \( 6 \) depends on the unknown parameter \( \theta \), which may not be adaptive in practice. However, if one simply choose \( \gamma = n_{te}^{-1} \), \( V_{KMM} \) still achieves a rate no worse than \( V_{KMM} \) as depicted in \( 8 \).

Third, also in Theorem 4 of [Yu and Szepesvári, 2012], the rate of \( V_{NR} \) is \( O(n_{tr}^{-\frac{\theta}{n_{tr}}} + n_{te}^{-\frac{\theta}{n_{te}}}) \) when \( g \in \text{Range}(T_K^{\frac{\theta}{n_{tr}}}) \), which is better on \( n_{tr} \) than \( n_{te} \). 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 \( \lim_{n \rightarrow \infty} n_{tr}^{-\frac{\theta}{n_{te}}}/n_{tr} \rightarrow 0 \). However, if so, \( V_R \) can still achieve \( O(n_{tr}^{-\frac{\theta}{n_{tr}}} + n_{te}^{-\frac{1}{2}}) \) rate in \( 7 \) 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 \( \theta \rightarrow \infty \), our estimator \( V_R \) recovers the canonical rate \( n_{tr}^{-\frac{1}{2}} \) as opposed to \( n_{te}^{-\frac{1}{2}} \) in \( V_{NR} \).

Thus, in summary, when \( g \in \text{Range}(T_K^{\frac{\theta}{n_{tr}}}) \), 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 and Szepesvári, 2012]: one is \( g \in \mathcal{H} \) and the other is \( A_{\infty}(g,F) \triangleq \inf_{\|f\|_{\mathcal{H}} \leq \|g - f\|} \leq C(\log F)^{-\alpha} \) 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 and Szepesvári, 2012] under these assumptions).

**Proposition 1.** Under Assumptions 4, 5 if \( g \in \mathcal{H} \), the convergence rate of \( V_R(\rho) \) satisfies \( |V_R(\rho) - \nu| = O(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}}) \), when \( \hat{g} \) is taken to be \( \hat{g}_{\gamma, \text{data}} \) for \( \gamma > 0 \) in \( 7 \).

**Proposition 2.** Under Assumptions 4, 5 if \( A_{\infty}(g,F) \triangleq \inf_{\|f\|_{\mathcal{H}} \leq \|g - f\|} \leq C(\log F)^{-\alpha} \) for some \( C, s > 0 \), the convergence rate of \( V_R(\rho) \) satisfies \( |V_R(\rho) - \nu| = O(\log \frac{n_{tr}n_{te}}{n_{tr} + n_{te}})^{-\alpha} \), when \( \hat{g} \) is taken to be \( \hat{g}_{\gamma, \text{data}} \) for \( \gamma > 0 \) in \( 7 \).

4 Empirical Risk Minimization

The robust estimator can handle empirical risk minimization (ERM). Given loss function \( l(x, y; \theta) :
$X \times \mathbb{R} \rightarrow \mathbb{R}$ given $\theta$ in $D$, we optimize over
\[
\min_{\theta \in D} \mathbb{E}[l'(X^{te}, Y^{te}; \theta)] = \min_{\theta \in D} \mathbb{E}_{x \sim P_{te}}[l(x; \theta)],
\]
where $l(x; \theta) \triangleq \mathbb{E}_{y|x}[l'(x, Y; \theta)]$ to find
\[
\theta^* \triangleq \arg\min_{\theta \in D} \mathbb{E}_{x \sim P_{te}}[l(X^{te}; \theta)].
\]
In practice, usually a regularization term $\Omega[\theta]$ on $\theta$ is added. For example, the KMM in [Huang et al., 2007] considers
\[
\min_{\theta \in D} \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(x_j^{tr})l'(x_j^{tr}, y_j^{tr}; \theta) + \lambda \Omega[\theta]. \tag{10}
\]
We can carry out a similar modification for $V_R$:
\[
\min_{\theta \in D} \frac{1}{n_{tot}} \sum_{j=1}^{n_{tot}} \beta(x_j^{te})l'(x_j^{te}, y_j^{te}; \theta) - \hat{l}(x_j^{te}; \theta) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{l}(x_i^{te}; \theta) + \lambda \Omega[\theta], \tag{12}
\]
with $\hat{\beta}$ based on $\{X_j^{tr}, X_j^{te}\}$ and $\hat{l}(x; \theta)$ being an estimate of $l(x; \theta)$ based on $\{X_j^{tr}, X_j^{te}\}$. For later reference, we note that a similar modification can also be used on $V_{NR}$:
\[
\min_{\theta \in D} \frac{1}{n_{tot}} \sum_{j=1}^{n_{tot}} \beta(x_j^{te})l'(x_j^{te}, y_j^{te}; \theta) + \lambda \Omega[\theta]. \tag{12}
\]
We discuss two classical learning problems by [11].

**Penalized Least Square Regression:** Consider a regression problem with $l'(x, y; \theta) = (y - \langle \theta, \Phi(x) \rangle)^2$, $\Omega[\theta] = ||\theta||_2^2$, and $y \in [0, 1]$. We have
\[
l(x; \theta) = E[Y^2|x] - 2g(x)\langle \theta, \Phi(x) \rangle + \langle \theta, \Phi(x) \rangle^2,
\]
and a candidate for $\hat{l}(x, \theta)$ is to substitute $g$ with $\hat{g}_{\gamma, data}$. Then, (11) becomes
\[
\min_{\theta \in D} \sum_{j=1}^{n_{tot}} \frac{2\beta(x_j^{te})}{n_{tot}}(y_j^{te} - \hat{g}(x_j^{te}))\langle \theta, \Phi(x_j^{te}) \rangle + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \langle \hat{g}(x_i^{te}) - \langle \theta, \Phi(x_i^{te}) \rangle \rangle^2 + \lambda ||\theta||_2^2,
\]
by adding and removing the components not involving $\theta$. Furthermore, it simplifies to the QP:
\[
\min_{\alpha \in \mathbb{R}^{|\mu^{tr}|} + \mathbb{R}^{|\mu^{te}| + n_{te}}} \begin{array}{c}
-2w_1^T K_{tot} \alpha
+ \alpha^T K_{tot} \alpha
+ \frac{(w_2 - K_{tot} \alpha)^T W_3 (w_2 - K_{tot} \alpha)}{n_{te}}
\end{array}
\]
\[
+ \frac{\mathbb{E}[l'(X^{te}, Y^{te}; \theta)]}{n_{te}}
\]
\[
+ \frac{\mathbb{E}_{x \sim P_{te}}[l(X^{te}; \theta)]}{n_{te}}.
\]
by the representation theorem [Schölkopf et al., 2001]. Here $(K_{tot})_{ij} = K(x_i^{tot}, x_j^{tot})$ and $W_3 = \text{diag}(w_3)$ where $x_i^{tot} = x_i^{tr}, (w_3)_i = \beta(x_i^{tr}) (y_i^{tr} - \hat{g}(x_i^{tr}))$, $(w_2)_i = 0, (w_3)_i = 0$ for $1 \leq i \leq |\mu^{tr}|$ and $x_i^{tot} = x_i^{te} - |\mu^{tr}|$, $(w_3)_i = 0, (w_2)_i = \hat{g}(x_i^{te} - |\mu^{tr}|)$, $(w_3)_i = 1$ for $|\mu^{tr}| + 1 \leq i \leq |\mu^{tr}| + n_{te}$.

**Penalized Logistic Regression:** Consider a binary classification problem with $y \in \{0, 1\}$, $\Omega[\theta] = ||\theta||_2^2$ and $-l'(x, y; \theta) = y \log(\frac{1}{1+\exp(\langle \theta, \Phi(x) \rangle \mathbb{I})}) + (1 - y) \log(\frac{1}{1+\exp(\langle \theta, \Phi(x) \rangle \mathbb{I})})$. Thus, we have
\[
-l(x; \theta) = -g(x)\langle \theta, \Phi(x) \rangle \mathbb{I} + \log(\frac{1}{1+\exp(\langle \theta, \Phi(x) \rangle \mathbb{I})}),
\]
and we can again substitute $g$ with $\hat{g}_{\gamma, data}$. Then, (11) becomes
\[
\min_{\theta \in D} \sum_{j=1}^{n_{tot}} \frac{\beta(x_j^{te})}{n_{tot}}(y_j^{te} - \hat{g}(x_j^{te}))\langle \theta, \Phi(x_j^{te}) \rangle \mathbb{I} + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \langle \hat{g}(x_i^{te}) - \langle \theta, \Phi(x_i^{te}) \rangle \mathbb{I} \rangle^2 + \lambda ||\theta||_2^2
\]
\[
+ \log(\frac{1}{1+\exp(\langle \theta, \Phi(x_i^{te}) \rangle \mathbb{I})}) + \log(\frac{1}{1+\exp(\langle \theta, \Phi(x_j^{te}) \rangle \mathbb{I})}),
\]
which again simplifies to, by [Schölkopf et al., 2001],
\[
\min_{\alpha \in \mathbb{R}^{|\mu^{tr}|} + \mathbb{R}^{|\mu^{te}| + n_{te}}} \begin{array}{c}
\frac{w_1^T K_{tot} \alpha}{n_{tot}}
- \frac{w_2^T K_{tot} \alpha}{n_{tot}}
+ \lambda \alpha^T K_{tot} \alpha
\end{array}
\]
\[
+ \sum_{j=1}^{n_{tot}} \log(\frac{\exp(\langle K_{tot} \alpha \rangle \mu^{tr} + j)}{1+\exp(\langle K_{tot} \alpha \rangle \mu^{tr} + j)})
\]
\[
+ \sum_{i=1}^{n_{te}} \log(\frac{\exp(\langle K_{tot} \alpha \rangle \mu^{te} + i)}{1+\exp(\langle K_{tot} \alpha \rangle \mu^{te} + i)})
\]
\[
+ \frac{\mathbb{E}[l'(X^{te}, Y^{te}; \theta)]}{n_{te}}
\]
\[
+ \frac{\mathbb{E}_{x \sim P_{te}}[l(X^{te}; \theta)]}{n_{te}}.
\]
Both (13) and (14) can be optimized efficiently by standard solvers. Notably, derived from (11), an optimal solution is in the form $\hat{\theta} = \sum_{i=1}^{n_{tot}} \hat{\alpha}_i K(x_i^{tot}, x)$ 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 $\hat{\theta} = \sum_{i=1}^{n_{tot}} \hat{\alpha}_i K(x_i^{tr}, x)$ where
\[
\hat{\alpha} = (K + n_{te} \lambda \text{diag}(\hat{\beta})^{-1})^{-1} y^{tr}
\]
(we use $\hat{\alpha} = (\text{diag}(\hat{\beta}) K + n_{te} \lambda \mathbb{I})^{-1} \text{diag}(\hat{\beta}) y^{tr}$ in experiments to avoid invertibility issues caused by the sparsity of $\hat{\beta}$), 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 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 \( \langle \Phi(x), \theta \rangle + f(x; \theta) \) with \( ||\theta||_H \leq C \) and \( l'(x; y; \theta) \in \mathcal{H} \) as \( \langle Y(x, y), \Lambda \rangle + f(x; \theta) \) with \( ||\Lambda||_H \leq C \). 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

\[
V_R(\theta) \triangleq \frac{1}{|\rho_{\text{tr}}|} \left[ \sum_{j=1}^{\rho_{\text{tr}}} \hat{\beta}(x_j^\text{tr}) (l'(x_j^\text{tr}, y_j^\text{tr}; \theta) - \hat{l}(x_j^\text{tr}; \theta)) \right] + \frac{1}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} \hat{l}(x_i^\text{te}; \theta)
\]

and \( \hat{\theta}_R \triangleq \arg\min_{\theta \in \mathcal{D}} V_R(\theta) \) satisfies

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

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 \( P_{\text{tr}} \sim \mathcal{N}(0, 0.5^2) \), \( P_{\text{te}} \sim \mathcal{N}(0, 0.3^2) \) are Gaussian distributions. The labels are generated according to \( y = -x + x^3 \) and observed with 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 our robust estimator, respectively. On the population level, the best linear fit is \( y = -0.73x \) (i.e. \( \arg\min_{\alpha_0, \beta_0} \mathbb{E}_{x \sim P_{\text{tr}}}(y - (\alpha_0 x + \beta_0))^2 \) is \( \alpha_0 = -0.73, \beta_0 = 0 \)). For simplicity, we set the intercept \( \beta_0 = 0 \) as known and compare the fitted slopes for different estimators. We use a degree-3 polynomial kernel and set \( \gamma \) in \( \hat{\gamma}_{\text{data}} \) to the default value \( n_{\text{tr}}^{-1} \). The tolerance \( \epsilon \) for \( \beta \) 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. The blue circle is the training data and the orange cross is the test data. For three different approaches, as well as an additional density-ratio-based method in [Shimodaira, 2000], the fitted slope over 20 trials are summarized in Figure 1(b). The average value is plotted in Figure 1(a) with black (KMM), green (robust) and yellow (OLS) respectively. As we see, the robust estimator outperforms the two other methods, achieving higher accuracy than KMM and unweighted OLS and recovering the slope closest to the best one in the vast majority of trials.

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 sampling bias operates jointly across multiple features. In particular, after randomly splitting the training and test sets based on different proportions, the training set is further subsampled with probability of selecting \( x_i \).
in the training set proportional to \(\exp(-\sigma_1 \|x_i - \bar{x}\|)\) for some \(\sigma_1 > 0\) and the training sample mean \(\bar{x}\). Since this is a binary classification problem and we are interested in comparing different approaches, we experiment with both the penalized least square regression and the penalized logistic regression for training sets of several sizes, i.e., the proportions of the training data are 0.3, 0.5, and 0.7 respectively, with respect to the total data. We used a Gaussian kernel \(\exp(-\sigma_2 \|x_i - x_j\|)\) for some \(\sigma_2 > 0\). The tolerance \(\epsilon\) for \(\beta\) 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 and follows KMM closely in the exceptional case, confirming our finding on its improvement over the traditional methods.

### 5.3 Simulated Dataset for Estimation

To test the performance of robust estimator on an estimation problem, we simulate data from two ten-dimensional Gaussian distributions with different, randomly generated means and covariance matrices as training and test sets. The target value is \(\nu = \mathbb{E}_{x \sim P_{\nu}}[g(x)]\) for an artificially constructed regression function \(g(x) = \sin(c_1 \|x\|^2) + (1 + \exp(c_2 \|x\|)\)^{-1}\) with random \(c_1, c_2\) and labels are observed with Gaussian noise. The Gaussian kernel \(\exp(-\sigma \|x_i - x_j\|)\) for \(\sigma > 0\) and a tolerance \(\epsilon\) for \(\hat{\beta}\) are set with exactly the same parameters as in [Gretton et al., 2009] with \(\sigma = \sqrt{5}, B = 1000\) and \(\epsilon = \frac{\sqrt{\log(n_{tr})}}{\sqrt{n_{tr}}}\). We also experiment with a different \(\gamma\) by substituting \(\hat{g}_{r_{data}}\) for a naive linear OLS fit with a lasso regularization term \(\lambda > 0\). 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 form of \(\hat{g}_{r_{data}}\) is replaced with a naive OLS fit, suggesting the robust estimator still works well under other forms of control variate functions. Moreover, we see that the robust estimator exhibits satisfactory performance even when the usual assumption \(n_{tr} < n_{te}\) is violated.

### Table 1: Average MSE for Estimation

<table>
<thead>
<tr>
<th>Hyperparameters ((\lambda, n_{tr}, n_{te}))</th>
<th>(V_N)</th>
<th>(V_{KMM})</th>
<th>(V_R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.1, 50, 500)</td>
<td>0.9970</td>
<td>0.9489</td>
<td>0.9134</td>
</tr>
<tr>
<td>(0.1, 500, 50)</td>
<td>0.9964</td>
<td>0.9294</td>
<td>0.9340</td>
</tr>
<tr>
<td>(10, 50, 500)</td>
<td>0.9962</td>
<td>0.9493</td>
<td>0.9467</td>
</tr>
<tr>
<td>(10, 500, 50)</td>
<td>0.9964</td>
<td>0.9294</td>
<td>0.9288</td>
</tr>
</tbody>
</table>

### 6 Conclusion

Motivated from variance and bias reduction, we introduced a new robust estimator for covariate shift problems which leads to improved accuracy over both KMM and NR in different 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 significant computational complexity increase. From an analytical standpoint, when the regression function lies in range spaces outside of RKHS, a promising progress is made to improve upon the well-known rate gap of KMM towards the parametric. For future work, note the canonical rate is still not achieved and it remains unclear the suitable tools for further improvement, if possible at all. Moreover, outside the KMM context with the regularized empirical regression function in RKHS, establishing the eligibility and effectiveness of other reweighting method coupled with different regression functions from learning schemes requires rigorous analysis.
Acknowledgements

We gratefully acknowledge support from the National Science Foundation under grants IIS-1849280 and CMMI-1653339/1834710.

References


