Doubly Robust Covariate Shift Correction

Sashank J. Reddi SJAKKAMR@CS.CMU.EDU
Machine Learning Department, School of Computer Science, Carnegie Mellon University,

Barnabás Póczos BAPOCZOS@CS.CMU.EDU
Machine Learning Department, School of Computer Science, Carnegie Mellon University,

Alex Smola ALEX@SMOLA.ORG
Machine Learning Department, School of Computer Science, Carnegie Mellon University,

Abstract

Covariate shift correction allows one to perform inference even when the distribution of the covariates on the training set does not match those on the test set. This is achieved by re-weighting observations. Such a strategy removes bias, potentially at the expense of greatly increased variance. We propose a simple strategy for removing bias while retaining small variance. It uses a biased, low variance estimate as a prior and corrects the final estimate relative to the prior. We prove that this yields an efficient estimator and demonstrate good experimental performance.

1. Introduction

Covariate shift is a common problem when dealing with real data. Quite often the experimental conditions under which a training set is generated are subtly different from the situation in which the system is deployed. For instance, in cancer diagnosis the training set may have an overabundance of diseased patients, often of a specific subtype endemic in the location where the data was gathered. Likewise, due to temporal changes in user interest the distribution of covariates in advertising systems is nonstationary. This requires increasing the weight of data related to, e.g., 'Gangnam style' when processing historic data logs.

A common approach to addressing covariate shift is to reweight data such that the reweighted distribution matches the target distribution. Suppose we observe $X := \{x_1, \ldots, x_m\}$ drawn iid from $q(x)$, typically with associated labels $Y := \{y_1, \ldots, y_m\}$ drawn from $p(y|x)$. This constitutes the 'training set'. However, we need to find a minimizer f_p^* of risk — defined in Equation 1 — with regard to $p(y|x)p(x)$, for which we only have iid draws of the covariates $X' := \{x'_1, \ldots x'_{m'}\}$. If p and q are known, importance sampling can be used for this problem:

$$\begin{aligned}\mathbf{E}_{x\sim p(x)}\mathbf{E}_{y|x}(\ell(y, f(x))) &= \int \frac{dp(x)}{dq(x)} dq(x)\mathbf{E}_{y|x}\ell(y, f(x)) \\ &= \mathbf{E}_{x\sim q(x)}\mathbf{E}_{y|x}\left[\beta(x)\ell(y, f(x))\right], \quad (1)\end{aligned}$$

where $\beta(x) := \frac{dp(x)}{dq(x)}$ and ℓ is a loss function. Correspondingly, empirical averages with respect to X and X' can be reweighted. See e.g. (Quiñonero-Candela et al., 2008; Cortes et al., 2008) and the references therein for further details. While Equation (1) allows us to correct the *bias* in the estimate, it also tends to increase the *variance* of the empirical averages considerably by weighting all observations by the Radon-Nikodym derivative β. With slight abuse of notation we denote by $\beta(X)$ the *vector* of such weights $\beta(x_1), \ldots \beta(x_m)$. As a general rule of thumb the effective sample size of such a reweighted dataset is

$$m_{\text{eff}} := \|\beta(X)\|_1^2 \,/\, \|\beta(X)\|_2^2 \,. \quad (2)$$

This quantity occurs, e.g., for a weighted average of Gaussian random variables, by deriving Chernoff bounds using the weights $\beta(X)$ (Gretton et al., 2008), or in the particle filtering context (Doucet et al., 2001).

This suggests that there may be situations where covariate shift correction may do more harm than good when applied directly: Whenever the effective sample size is tiny relative to the original problem, we might obtain an unbiased estimate, yet with such high variance that it becomes nearly useless. This situation is frequently observed in practice insofar as we encounter cases where covariate shift correction not only fails to improve generalization performance on the test set but, in fact, leads to estimates that perform worse. The problem is exacerbated by the fact that in many cases the *solutions* of the biased and the unbiased risk functionals are closer than what the distributions p and q would

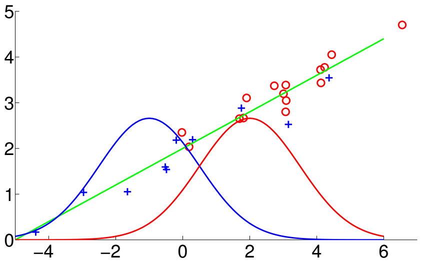

Figure 1. Example demonstrating the problem with generic covariate shift correction approaches.

suggest. Figure 1 shows what may happen. Assume that the dependence $y|x$ is linear in x, as indicated by the green line of Figure 1. In this case, inferring $y|x$ using the blue distribution q, as depicted by the blue crosses (with matching density), would lead to a perfectly accurate estimate, even if the test set is drawn according to the red distribution p. On the other hand, reweighting with $\frac{dp(x)}{dq(x)}$ would lead to a very small effective sample size since p and q are very different. While this example is obviously somewhat artificial, there exist many situations where the minimizer of the biased risk is very good. The above problem is often encountered in practice — covariate shift correction fails to improve matters due to high variance. This raises the question of how we could benefit from the low variance of the biased estimate found in q while removing bias via weighting with β.

This is precisely what doubly-robust estimators address — see, e.g., (Kang & Schafer, 2007) for an overview. They provide us with two opportunities to obtain a good estimate. In our case, these are steps 1 and 3 in the algorithm below. Whenever the unweighted estimate solves the problem, the estimate will be very good and minimizing the unbiased risk estimate will not change the final outcome by much. Conversely, whenever the unweighted estimate is useless, we still have the opportunity to amend things in the context of estimating f_p^*, due to reweighting of dataset. Briefly, our algorithm outline is the following.

Step 1: Unweighted estimate Solve the unweighted inference problem. This will give us the estimate $\hat{f}_q$. The intuition is that while $\hat{f}_q$ will not minimize the expected risk, it is often a very good proxy. Given that no reweighting was carried out, the variance for $\hat{f}_q$ is comparatively low. That is, we are using the large unweighted sample size to obtain a good starting point with high confidence.
Step 2: Covariate shift correction weights Using X and X' estimate the covariate shift correction weights. This can be done by any off-the-shelf procedure and is what typically required to obtain an unbiased risk functional (Gretton et al., 2008; Agarwal et al., 2011).

Step 3: Doubly-robust estimate: If m_{eff} is not much smaller than m, ignore step 1 and perform covariate shift corrected risk minimization using the weights from step 2. For small effective sample size use $\hat{f}_q$ as prior when performing risk minimization with respect to the unbiased risk functional. A rather strong smoother is needed since the effective sample size is much smaller and we already used the data once previously. For instance, for regression estimation this amounts to fitting the residuals of step 1. For classification this is fitting a logistic model to the tilted exponential family binomial model obtained previously.

In summary, the paper makes the following contributions. (1) We develop a simple, yet powerful, framework for doubly robust estimation in the context of covariate shift correction, which to the best of author's knowledge has not been previously explored. (2) We demonstrate the generality of the framework by providing several concrete examples. (3) We present a general theory for the framework and provide a detailed analysis in the case of kernel methods. (4) Finally, we show good experimental performance on several UCI datasets. All proofs are relegated to the appendix due to space constraints.

Related Work

There has been extensive research in covariate shift correction problem. Most of the work is directed towards estimating the weights β. Several methods have been proposed to estimate these weights by optimization and statistical techniques (Gretton et al., 2008; Agarwal et al., 2011; Tsuboi et al., 2008; Sugiyama et al., 2008). Likewise, there has been considerable work in developing doubly robust estimators for many statistical and machine learning problems, particularly in the problems involving missing data and reinforcement learning (Kang & Schafer, 2007; Dudík et al., 2011; Bang & Robins, 2005). But none of these works address the problem of our concern, namely doubly robust estimation for covariate shift correction. While few works, e.g., (Shimodaira, 2000), attempt to reduce the variance by adjusting the weights and thereby, balancing the bias-variance tradeoff, they do not tackle the problem from doubly robust estimation point of view. In fact, these methods can be used in conjunction with our approach.

2. Doubly Robust Covariate Shift Correction

2.1. Problem Formulation

We now define the problem more formally. For simplicity, we assume $m = m'$ in this paper. Our language will be that of risk minimization. For this purpose denote by $\mathcal{X}$, with $x_i \in \mathcal{X}$, the space of covariates, and by $\mathcal{Y}$, with $y_i \in \mathcal{Y}$, the space of associated labels. For any function $f : \mathcal{X} \to \mathbb{R}$, we use f_i to denote the function evaluated at point x_i. The distributions $p(x)$ and $q(x)$ are defined on $\mathcal{X}$. Moreover, $y \sim p(y|x)$. As stated in the introduction, we assume that $x_i \sim q(x)$ and $x'_i \sim p(x)$ and $y_i \sim p(y|x_i)$. Finally, we denote by $\ell : \mathcal{Y}^2 \to \mathbb{R}_0^+$ a loss function. We assume that the loss function $L-$Lipschitz and bounded above by L.[1]

Our goal is to minimize the expected risk with regard to p. Since we will only be able to measure (weighted) risks with regard to q we need to contend with the following two risk functionals: $R_p[f] := \mathbf{E}_{(x,y)\sim p}[\ell(y, f(x))]$ and $R_q[f] := \mathbf{E}_{(x,y)\sim q}[\ell(y, f(x))]$ Quite often we will deal with empirical averages, often weighted. We define

$$\widehat{R}[f|X, Y, \alpha] := \frac{1}{m}\sum_i \alpha_i \ell(y_i, f(x_i))$$

The risks for X' are defined analogously. The unweighted empirical risk is $\widehat{R}[f|X, Y] = \widehat{R}[f|X, Y, 1_m]$ where 1_m is ones vector of size m. Given a class $\mathcal{F}$ of functions $\mathcal{X} \to \mathcal{Y}$ we aim to find some f_p^* that minimizes $R_p[f]$. Unfortunately, $R_p[f]$ is not directly accessible, hence we can only approximate it via the empirical risk $\widehat{R}[f|X, Y]$, or its reweighted variant $\widehat{R}[f|X, Y, \beta]$.

Furthermore, we use a regularizer Ω to ensure that we do not overfit to the data. This regularizer plays a rather critical role in our doubly-robust approach. It quantifies the notion of 'simple' function. More specifically, we use $\Omega[f, f']$ to measure complexity of f relative to f'. By default we set $f' = 0$ with the corresponding shorthand $\Omega[f] := \Omega[f, 0]$. This views the constant null function as the simplest in the entire set. For instance, in kernel methods we have $\Omega[f, f'] := \frac{1}{2}\|f - f'\|^2$, where the norm is evaluated in a Reproducing Kernel Hilbert Space.

Finally, we introduce minimizers of expected and empirical risk, as is common in statistical learning theory (Vapnik, 1998). We use f_p^* and f_q^* to denote the minimizers of risks R_p and R_q respectively. Throughout this paper, we use the following equivalent formulations interchangeably:

$$\hat{f}_{q,\lambda} := \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R}[f|X, Y] + \lambda\Omega[f]$$

$$\hat{f}_{q,\nu} := \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R}[f|X, Y] \text{ s.t. } \Omega[f] \le \nu$$

The corresponding pair (λ, ν) and associated problem will be clear from the context. The equivalence follows from the fact that for any λ, there exists a ν such that the solution of the two problems is same. This is done merely for reasons of simplifying the theoretical analysis. This yields the following risk functionals with associated minimizers.

$$\hat{f}_{q,\lambda_q} := \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R}[f|X, Y] \text{ s.t. } \Omega[f] \le \nu_q \tag{3}$$

Here the risk functional, as defined in Equation (3) corresponds to the empirical risk minimizer when solving the inference problem with respect to the distribution $q(x)p(y|x)$. Let $\hat{\beta}$ be the estimated covariate shift weights. The next empirical risk functional is X, Y reweighted by $\hat{\beta}$ such that we obtain an unbiased estimate from p.

$$\hat{f}_{p,\lambda_p} := \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R}[f|X, Y, \hat{\beta}] \text{ s.t. } \Omega[f] \le \nu_p \tag{4}$$

Finally, the following denotes doubly robust estimator which is risk minimizer, albeit with a prior around $\hat{f}_{q,\lambda}$ rather than 0.

$$\hat{f}_{\mathrm{DR}} := \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R}[f|X, Y, \hat{\beta}] \text{ s.t. } \Omega[f, \hat{f}_{q,\lambda}] \le \nu' \tag{5}$$

Lastly, we define f^*_{q,λ_q} and f^*_{p,λ_p} to be the *constrained* minimizers of the expected risk i.e.

$$f^*_{q,\lambda_q} := \operatorname*{argmin}_{f\in\mathcal{F}} R_q[f] + \lambda_q\Omega[f] \tag{6}$$

$$f^*_{p,\lambda_p} := \operatorname*{argmin}_{f\in\mathcal{F}} R_p[f] + \lambda_p\Omega[f] \tag{7}$$

These quantities are needed since f_p^* and f_q^* might not necessarily have bounded norm in function classes that we study. For instance, in the context of kernel methods, they might be approximated in the limit by a series of kernel functions but might not be elements of the RKHS themselves.

2.2. Assumptions

It is worth mentioning the assumptions required for the application of doubly robust estimation, since they motivate our design choices.

Assumption 1 *The conditional training and test distributions are identical i.e* $p(y|x) = q(y|x)$.

This is implicit in the definition of covariate shift — if $p(y|x) \ne q(y|x)$ it would be trivial to construct counterexamples for any algorithm attempting to solve covariate shift. For instance, setting $p(y|x) = q(-y|x)$ for binary classification would lead to a maximally bad solution.

Assumption 2 *The Radon-Nikodym derivative* $\beta(x)$ *is well defined and bounded by some constant* η. *This ensures that there cannot exist sets of nonzero measures with respect to* Q *that have zero measure with respect to* P.

Again, in the absence of this assumption we could design pessimal algorithms. For instance, assume that some set S with $p(S) > 0$ has vanishing q-measure, i.e. $q(S) = 0$. In this case we could, e.g., set $y|x = 0$ for all $x \notin S$ and $y|x = C$ for $x \in S$. This would immediately imply substantial misprediction regardless of the sample size.

[1] We use the same constant L, without loss of generality.

Assumption 3 *The risk minimizer f^*_{p,λ_p} is much closer to the unweighted risk minimizer f^*_{q,λ_q} rather than the origin, i.e., $\nu_{\mathrm{DR}} = \Omega[f^*_{p,\lambda_p}, f^*_{q,\lambda_q}] \ll \Omega[f^*_{p,\lambda_p}] = \nu_p$.*

This is likely the most contentious assumption — it implies that solving the unweighted problem will be significantly beneficial for the weighted solution. An easily constructed counterexample is the mapping $y = |x|$ where p and q largely emphasize positive and negative x. However, we have never encountered such a situation in practice. It would imply that using the uncorrected estimates in the new context would be worse than random. It is the above assumption that makes our algorithm work.

2.3. Estimating Covariate Shift Weights

Before delving into a specific algorithm we need to discuss means of obtaining estimates of $\beta(X)$. A number of approaches have been proposed to obtain these estimates. We only give a brief outline of a few approaches here and refer interested readers to the appropriate references for a more thorough analysis.

Penalized Risk Minimization (PRM) The basic idea in this approach is to estimate covariate shift weights β by solving a particular regularized convex minimization problem over a function class (Nguyen et al., 2008). The rationale for the approach stems from the fact that the optima to the variational representation of KL-divergence is attained at the point $\beta(x) = \frac{p(x)}{q(x)}$ $\forall x \in \mathcal{X}$. More specifically, consider the following variational representation of KL-divergence:

$$D(p, q) = \sup_{g>0} \int \log g(x) p(x) dx - \int g(x) q(x) dx + 1.$$

This is obtained by a simple application of Legendre-Frenchel convex duality (see (Nguyen et al., 2008) for more details). More importantly for us, the supremum is attained at $g(x) = \beta(x) = p(x)/q(x)$. Let us assume that the function β belongs to RKHS $\mathcal{G}$. Since the access to distributions p and q is through their corresponding samples, we solve the following regularized empirical version of the problem:

$$\hat{\beta} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} g(x_i) - \frac{1}{m} \sum_{i=1}^{m} \log g(x'_i) + \frac{\gamma_m}{2} I^2(g)$$

where $I(g)$ is a non-negative measure of complexity for g such that $I(\beta) < \infty$. It is shown that the above estimator enjoys good statistical properties. A more detailed theoretical exposition of the estimator will follow in later sections.

Kernel Mean Matching (KMM) Another popular approach to obtain the covariate shift weights is by matching the mean embeddings in the feature space induced by a universal RKHS $\mathcal{K}$ on the domain $\mathcal{X}$ (Gretton et al., 2008). More specifically, we solve the following optimization problem

$$\min_{\hat{\beta}} \hat{L}(\hat{\beta}) := \left\| \frac{1}{m} \sum_{i=1}^{m} \hat{\beta}_i \Phi(x_i) - \frac{1}{m} \sum_{i=1}^{m} \Phi(x'_i) \right\|$$

$$\text{s.t. } 0 \leq \hat{\beta}_i \leq \eta \text{ and } \frac{1}{m} \sum_{i=1}^{m} \hat{\beta} = 1,$$

where $\Phi : \mathcal{X} \to \mathcal{K}$. Intuitively, the above procedure tries to match the mean embeddings of weighted training and test distributions. Since the RKHS is universal, matching the embeddings provides estimates for covariate shift weights β. As above, we delay the theoretical details.

It is interesting to note that while the first estimation procedure gives the function β, the KMM approach computes the function evaluated only at the training points. See e.g. (Agarwal et al., 2011) for a more detailed discussion and comparison to other approaches.

2.4. Examples

To gain a better understanding of our approach, we now present our estimators in various algorithmic settings. Let us assume, we have estimated covariate shift weights $\hat{\beta}$ via PRM, KMM or in general, any other method.

Regression The simplest setting is linear regression, possibly in a Reproducing Kernel Hilbert Space. Here the loss ℓ, the function f, and Ω are given by $f(x) = \langle w, \phi(x) \rangle$, $\ell(y, f(x)) = \frac{1}{2}(y - f(x))^2$ and $\Omega[f, f'] = \frac{1}{2} \|w - w'\|^2$, where $\phi(x)$ is a feature map. The three steps of doubly robust covariate shift correction are:

1. Solve the quadratic optimization problem below.

$$\hat{w}_{q,\lambda_q} = \arg\min_{w} \frac{1}{2} \sum_{i=1}^{m} (y_i - \langle \phi(x_i), w \rangle)^2 + \frac{\lambda_q}{2} \|w\|^2$$

2. Estimate the covariate shift correction weights $\hat{\beta}$.
3. Solve the centered weighted regression problem to obtain the doubly robust estimator $\hat{w}_{\mathrm{DR}}$.

$$\operatorname*{minimize}_{w} \frac{1}{2} \sum_{i=1}^{m} \hat{\beta}_i (y_i - \langle \phi(x_i), w \rangle)^2 + \frac{\lambda'}{2} \left\| w - \hat{w}_{q,\lambda_q} \right\|^2$$

The approach works whenever $\left\| w^*_p - w^*_q \right\| \ll \left\| w^*_p \right\|$, i.e. whenever the unbiased and the biased solutions are close compared to the overall complexity of the solutions.

SVM Classification The approach is quite analogous to the above approach, the main difference being a different loss function. This yields $f(x) = \langle w, \phi(x) \rangle$, $\ell(y, f(x)) = \max(0, 1 - yf(x))$, and $\Omega[f, f'] = \frac{1}{2} \|w - w'\|^2$. The associated algorithm is as follows:

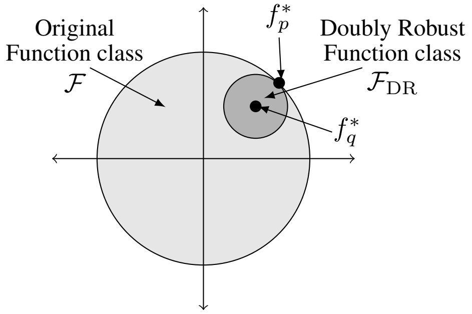

Figure 2. Pictorial representation of Doubly Robust estimation procedure. Assumption 3 implies that f_q^* is close to f_p^* than origin (as shown in the figure). While the generic covariate shift finds the weighted empirical risk minimizer over the large function class $\mathcal{F}$, doubly robust procedure optimizes over a much smaller function class $\mathcal{F}_{\mathrm{DR}}$. This leads to small variance in doubly robust procedure as as compared to generic covariate shift procedure when the effective sample size m_{eff} is small.

1. Solve a standard SVM classification problem using X, Y to obtain $\hat{w}_{q,\lambda_q}$.

$$\min_w \sum_{i=1}^m \max(0, 1 - y_i f(x_i)) + \frac{\lambda_q}{2} \|w\|^2$$

2. Estimate the covariate shift correction weights $\hat{\beta}$.
3. Solve the centered weighted SVM classification problem to obtain the doubly robust estimator $\hat{w}_{\mathrm{DR}}$.

$$\min_w \sum_{i=1}^m \hat{\beta}_i \max(0, 1 - y_i f(x_i)) + \frac{\lambda'}{2} \left\|w - \hat{w}_{q,\lambda_q}\right\|^2$$

Regression Tree The nontrivial challenge here is to define what it means to use an existing tree as a prior. We obtain the following algorithm:

1. Compute a Regression Tree $\hat{f}_{q,\lambda_q}$ using X, Y with suitable pruning strategy λ_q.
2. Estimate the covariate shift correction weights $\hat{\beta}$.
3. Compute the prediction errors $\epsilon_i := y_i - \hat{f}_{q,\lambda_q}(x_i)$. Train a second regression tree δf using $(x_i, \epsilon_i, \hat{\beta}_i)$ as covariates, labels, and sample weights. Output the corrected tree $\hat{f}_{\mathrm{DR}} := \hat{f}_{q,\lambda_q} + \delta f$.

Analogous modifications are possible for Gaussian Process estimates where we use stage 1 estimates as prior, or for neural networks. Given the generality, our analysis proceeds in two steps — we first derive a general metatheorem, followed by an application to kernel methods.

3. Theoretical Analysis

3.1. A Metatheorem

The analysis mirrors the algorithmic approach — it consists of three key components. Firstly, we need to show that $\hat{f}_{q,\lambda_q}$ is close to the expected risk minimizer f^*_{q,λ_q}. For this we can take advantage of the large sample size inherent in X, Y, thus providing us with a biased, low variance guess of a solution not too far away from f^*_{p,λ_p}. Secondly, we need to bound the uniform convergence behavior of the covariate shift corrected risk functional relative to the expected risk. Finally, we need a uniform convergence bound for a weighted risk minimization problem, such as those in (Gretton et al., 2008).

Theorem 1 *Assume that the following two conditions hold with probability at least $1 - \delta$ when $m > m_0$:*

1. **[Bound A]** *Suppose there exists a function $A(\lambda, m, \delta)$ such that for all functions $f \in \mathcal{F}$*

$$\Omega[f, \hat{f}_{q,\lambda_q}] \leq \Omega[f, f^*_{q,\lambda_q}] + A(\lambda_q, m, \delta). \quad (8)$$

That is, the solution of the unweighted empirical risk minimization problem is close *to f^*_{q,λ_q}.*

2. *Let $\widehat{T}[f]$ be an estimate of the risk $R_p[f]$, and $\hat{f}$ and $\hat{f}_{T,\lambda}$ be minimizers of $\widehat{R}[f|X, Y, \hat{\beta}]$ and $\widehat{T}[f]$ respectively subject to the constraint $\Omega[f, f_0] \leq \nu$.*

[Bound B] *Suppose we have the following relationship between these minimizers.*

$$R_p[\hat{f}] \leq R_p[\hat{f}_{T,\lambda}] + B(f_0, \hat{f}_{q,\lambda}, \hat{f}_{T,\lambda}, \lambda, \delta) \quad (9)$$

That is, the minimizer of $\widehat{T}[f]$ is close to the empirical risk minimizer.
[Bound C] *Additionally, suppose we also have*

$$\sup_{\substack{f \text{ s.t.} \\ \Omega[f,f_0] \leq \nu}} \left|\widehat{T}[f] - R_p[f]\right| \leq C(f_0, \lambda, m, \beta, \delta). \quad (10)$$

Note that the key emphasis here is on the algorithm *that was used to obtain $\hat{\beta}$ and complexity of the function class, e.g., via a Rademacher average.*

Let $\nu' = \nu_{\mathrm{DR}} + A(\lambda_q, m, \delta/3)$. The following bound holds with probability at least $1 - \delta$:

$$R_p[\hat{f}_{\mathrm{DR}}] \leq R_p[f^*_{p,\lambda_p}] + B(\hat{f}_{q,\lambda_q}, \hat{f}_{\mathrm{DR}}, \hat{f}_{T,\lambda'}, m, \delta/3) + 2C(\hat{f}_{q,\lambda_q}, \nu', \beta, \delta/3). \quad (11)$$

We can similarly prove that the bound for standard covariate shift procedure is $B(0, \hat{f}_{p,\lambda_p}, \hat{f}_{T,\lambda_p}, m, \delta/2) + 2C(0, \nu_p, \beta, \delta/2)$. The doubly robust estimator bound

(Equation 11) is better than a generic covariate shift correction bound whenever ν' is much smaller than ν_p, which is roughly Assumption 3. We will instantiate these bounds in the next section.

3.2. Theoretical Analysis for Kernel Methods

In this section, we present theoretical analysis for kernel methods. The underlying property of these algorithms, which simplifies our analysis is that of proportional quantification of the risk with change in the distribution of the training data. As mentioned earlier, the function $\Omega[f, f']$ is $\|f - f'\|^2/2$ in this case. While we only focus on kernel methods, the theoretical analysis can be generalized to a broader class along similar lines. Let $\mathbf{K}$ denote the kernel matrix corresponding to the training points X. We rely on Rademacher averages for our uniform convergence bounds (Bartlett & Mendelson, 2002). As a first step, we analyze the theoretical guarantees for the unweighted minimizer.

Unweighted Estimator

We analyze the estimated unweighted risk minimizer in this section. In particular, we derive the relationship between the risks of the true and estimated unweighted risk minimizers (see appendix for proof).

Theorem 2 *Let $\hat{f}_{q,\lambda_q}$ and f^*_{q,λ_q} be as defined in Equations* (3) *and* (6) *respectively, and $\|p - q\|_1 \le \epsilon$. Then*

$$R_p[\hat{f}_{q,\lambda_q}] \le R_p[f^*_{q,\lambda_q}] + \frac{4L\nu_q}{m}\sqrt{\mathrm{tr}(\mathbf{K})} + 6L\sqrt{\frac{\log(2/\delta)}{2m}} + 2L\epsilon$$

The above result gives a generalization bound for the estimated unweighted risk minimizer $\hat{f}_{q,\lambda_q}$. Our bounds support the fact that the estimator is asymptotically biased, an undesirable property for an estimator.

Bound A: This is achieved in two steps — firstly we need to invoke uniform convergence of the regularized risk functional on the domain of interest and secondly, we appeal to strong convexity of risk. While this assumption is used here for simplicity, we can obtain theoretical guarantees without this assumption along the lines of (Kifer et al., 2012).

Lemma 1 *Suppose $\hat{f}_{q,\lambda_q}$ and f^*_{q,λ_q} are as defined in Equations* (3) *and* (6) *respectively. Then the following holds with probability at least $1 - \delta$:*

$$\left\|\hat{f}_{q,\lambda_q} - f^*_{q,\lambda_q}\right\| \le \sqrt{\frac{4L}{\lambda_q}\left(\frac{2\nu_q}{m}\sqrt{\mathrm{tr}(\mathbf{K})} + 3\sqrt{\frac{\log(2/\delta)}{2m}}\right)}$$

We next present generalization bounds for standard covariate shift and doubly robust estimators using two approaches: (a) Penalized Risk Minimization (PRM) (b) Kernel Mean Matching (KMM) (given in the appendix).

Generalization bounds for PRM

Our main goal is to obtain a relation between the risks of $\hat{f}_{p,\lambda_p}$ (or $\hat{f}_{\mathrm{DR}}$) and f^*_{p,λ_p}. This is accomplished by first invoking uniform convergence on the true weighted risk minimizer and then showing the risk minimizer based on estimated weights is not far away from it, using distribution stability analysis. We define the following empirical risk minimization problems:

$$\hat{f}_{\hat{\beta},\lambda} := \operatorname*{argmin}_f \widehat{R}[f|X, Y, \hat{\beta}] + \lambda\Omega[f, f_0] \qquad (12)$$

$$\hat{f}_{\beta,\lambda} := \operatorname*{argmin}_f \widehat{R}[f|X, Y, \beta] + \lambda\Omega[f, f_0] \qquad (13)$$

We instantiate Theorem 1 with function $\widehat{T}[f] = R[f|X, Y, \beta]$, i.e., empirical risk with true weights β.

Bound C: The following result provides bound C of the metatheorem.

Lemma 2 *Suppose f^*_{p,λ_p} is as defined in Equation* (7). *We have the following with probability at least $1 - \delta$:*

$$\sup_{f:\|f - f_0\| \le \nu} \left|R[f|X, Y, \beta] - R_p[f^*_{p,\lambda_p}]\right| \le$$
$$\eta L\left(\frac{2\nu}{m}\sqrt{\mathrm{tr}(\mathbf{K})} + 3\sqrt{\frac{\log(2/\delta)}{2m}} + \frac{1}{m}\sqrt{\sum_{i=1}^{m} f_0^2(x_i)}\right)$$

Bound B: We now derive a relation between $\hat{f}_{\hat{\beta},\lambda}$ and $\hat{f}_{\beta,\lambda}$, i.e., Bound B of the metatheorem. The key insight here is that due to distribution stability property of kernel methods (Cortes et al., 2008), the empirical risk minimizer with shifted weights is close to the one with true weights. The following result formalizes this insight.

Lemma 3 *Suppose $\hat{f}_{\hat{\beta},\lambda}$ and $\hat{f}_{\beta,\lambda}$ are as defined in Equations* (12) *and* (13) *respectively, and $\beta \in \mathcal{G}$. Let the regularization parameter $\gamma_m = cm^{-2/(2+\tau)}$ for some $\tau > 0$ and a constant c. Then we have the following with probability at least $1 - \delta$:*

$$\left|R_p[\hat{f}_{\hat{\beta},\lambda}] - R_p[\hat{f}_{\beta,\lambda}]\right| \le \frac{2\kappa^2 L^2}{\lambda}\left(\sqrt{\eta\gamma_m} + \eta\sqrt[4]{\frac{8}{m}\log\left(\frac{2}{\delta}\right)}\right)$$

We now state the main results about generalization bounds for PRM, which follow as corollaries of Theorem 1.

Theorem 3 *Suppose $\hat{f}_{p,\lambda_p}$ and f^*_{p,λ_p} are as defined in Equations* (4) *and* (7) *respectively, and $\beta \in \mathcal{G}$. Let the regularization parameter for PRM be $\gamma_m = cm^{-2/(2+\tau)}$ for some $\tau > 0$ and a constant c. Then we have the following with probability at least $1 - \delta$.*

$$R_p[\hat{f}_{p,\lambda_p}] \le R_p[f^*_{p,\lambda_p}] + \Delta_{W,S} + \Delta_{W,R}. \qquad (14)$$

$\Delta_{W,S}$ and $\Delta_{W,R}$, representing the covariate shift and function complexity parts of the bound, are defined below.

$$\Delta_{W,S} = \frac{2\kappa^2 L^2}{\lambda}\left(\sqrt{\eta\gamma_m} + \eta\sqrt[4]{\frac{8}{m}\log\left(\frac{4}{\delta}\right)}\right)$$

$$\Delta_{W,R} = 2\eta L\left(\frac{2\nu}{m}\sqrt{\mathrm{tr}(\mathbf{K})} + 3\sqrt{\frac{1}{2m}\log\left(\frac{4}{\delta}\right)}\right)$$

Theorem 4 *Suppose $\hat{f}_{\mathrm{DR}}$ and f^*_{p,λ_p} are as defined in Equations (5) and (7) respectively, and $\beta \in \mathcal{G}$. Let the regularization parameter for PRM be $\gamma_m = cm^{-2/(2+\tau)}$ for some $\tau > 0$ and a constant c. Then we have the following with probability at least $1 - \delta$.*

$$R_p[\hat{f}_{\mathrm{DR}}] \leq R_p[f^*_{p,\lambda_p}] + \Delta_{\mathrm{DR},S} + \Delta_{\mathrm{DR},R}. \qquad (15)$$

$\Delta_{\mathrm{DR},S}$ and $\Delta_{\mathrm{DR},R}$, denoting the covariate shift and function complexity parts of the bound, are defined below.

$$\nu' = \nu_{DR} + \sqrt{\frac{4L}{\lambda_q}\left(\frac{2\nu_q\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + 3\sqrt{\frac{\log(6/\delta)}{2m}}\right)}$$

$$\Delta_{\mathrm{DR},S} = \frac{2\kappa^2 L^2}{\lambda'}\left(\sqrt{\eta\gamma_m} + \eta\sqrt[4]{\frac{8}{m}\log\left(\frac{6}{\delta}\right)}\right)$$

$$\Delta_{\mathrm{DR},R} = 2\eta L\left(\frac{2\nu'\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + 3\sqrt{\frac{\log(6/\delta)}{2m}} + \frac{\|\hat{f}_{q,\lambda_q}\|_2}{m}\right)$$

Discussion on the Generalization Bounds

In order to understand the benefit of our doubly robust estimator, we make a qualitative comparison of the various generalization bounds in this section. We only compare the bounds for PRM because the bounds for KMM (in Appendix) yield similar conclusions. From Assumption 3, we have $\nu' \ll \nu_p$ and $\lambda' > \lambda_p$, provided bound A is small. When the variance of $\hat{f}_{q,\lambda_q}$ is small, it is easy to see that $\Delta_{\mathrm{DR},R} \ll \Delta_{W,R}$ (in Equations 14 and 15). Moreover, we perform better than generic covariate shift in both the cases — good and weak estimate of weights. This is due to the fact that the bounds $\Delta_{W,S}$ and $\Delta_{\mathrm{DR},S}$ are of similar magnitude, and $\Delta_{\mathrm{DR},R} \ll \Delta_{W,R}$. Therefore, these bounds emphasize the doubly robust nature of our approach. Before ending our discussion, we need to make it explicit that our analysis only compares the upper bounds and hence, needs to be interpreted with caution. Nonetheless, our empirical evaluation, in the next section, supports our theoretical analysis and provides a compelling case to use our estimators in practice.

4. Experiments

We present our empirical results in this section. We apply doubly robust covariate shift correction to a broad range of UCI datasets and a real-world dataset to demonstrate its performance. In particular, we show that it is effective both for classification and regression settings, both for linear methods (by using a Support Vector Classifier) and non-linear approaches (by using a Regression Tree). Moreover, we investigate its efficacy using a proprietary classification problem of a large internet company.

For our experiments we compare the performance of unweighted (referred to as Unweighted), weighted (referred to as Weighted) and doubly robust (referred to as DoublyRobust) empirical estimators. That is, Unweighted ignores the problem of covariate shift correction; Weighted uses the weights computed by KLIEP (Sugiyama et al., 2008) with Gaussian kernel. For simplicity we use a reduced rank expansion with 100 basis functions in our experiments. The bandwidth of the kernel is chosen by cross-validation; Finally, DoublyRobust refers to the results obtained by our method.

Synthetic Data: This experiment is meant to provide a comparison of Weighted and DoublyRobust approaches when varying effective sample size m_{eff}. The data for this experiment is generated based on a polynomial objective (Gretton et al., 2008)

$$y = -x + x^3 + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, 0.3)\,. \qquad (16)$$

We set $p(x) = \mathcal{N}(0, 1)$ and use as biasing distribution $p(x) = \mathcal{N}(\mu, 0.3)$ where μ is adjusted such that we obtain different effective samples sizes. 300 training and test samples are drawn. We use linear regression with standard ℓ_2 penalization.

Figure 4 shows the root mean square error (RMSE) ratio of DoublyRobust to Weighted. It can be seen that DoublyRobust outperforms Weighted for lower values of m_{eff} and is marginally worse for higher values of m_{eff}. The latter is not surprising, since DoublyRobust makes use of the data thrice rather than twice.

Real Data: For a more realistic comparison we apply our method to several UCI[2] and benchmark[3] datasets. To control the amount of bias we use PCA to obtain the leading principal component. The projections onto the first principal component are then used to construct a subsampling distribution q. Let t_0 and t_1 be the minimum and the maximum of the projected values respectively. Let σ_{PC} be the standard deviation of the projected values. We then subsample using their projected values according to normal distribution $\mathcal{N}(t_0 + \alpha(t_1 - t_0), 0.5\sigma_{\mathrm{PC}})$. Varying the value of α changes the m_{eff} of the training data by shifting q relative to p. The value $\alpha \in (0, 1)$ is independently set for each

[2] http://archive.ics.uci.edu/ml/datasets.html

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

classification	Unweighted	Weighted	DoublyRobust
hill	1.00 (± 0.05)	1.03 (± 0.04)	**0.98** (± 0.03)
splice	1.00 (± 0.01)	0.98 (± 0.01)	**0.97** (± 0.01)
german	1.00 (± 0.04)	1.08 (± 0.05)	**0.96** (± 0.03)
diabetes	1.00 (± 0.05)	0.89 (± 0.02)	**0.85** (± 0.01)
ionosphere	1.00 (± 0.04)	0.82 (± 0.01)	**0.79** (± 0.01)
cod-rna	1.00 (± 0.05)	1.05 (± 0.03)	**0.94** (± 0.02)
ijcnn	1.00 (± 0.03)	0.99 (± 0.02)	**0.96** (± 0.02)
breast-cancer	1.00 (± 0.03)	**0.96** (± 0.02)	0.97 (± 0.02)
fourclass	**1.00** (± 0.03)	1.04 (± 0.02)	1.03 (± 0.02)
australian	1.00 (± 0.04)	1.02 (± 0.03)	**0.97** (± 0.03)
sonar	1.00 (± 0.05)	0.98 (± 0.04)	**0.97** (± 0.04)
spambase	1.00 (± 0.05)	0.99 (± 0.03)	**0.98** (± 0.03)

regression	Unweighted	Weighted	DoublyRobust
abalone	1.00 (± 0.01)	0.97 (± 0.03)	**0.95** (± 0.01)
mg	1.00 (± 0.04)	1.04 (± 0.03)	**0.97** (± 0.03)
enuite	1.00 (± 0.04)	0.95 (± 0.03)	**0.93** (± 0.02)
space	1.00 (± 0.05)	0.98 (± 0.04)	**0.94** (± 0.03)
mpg	1.00 (± 0.03)	**0.93** (± 0.02)	0.94 (± 0.03)
bodyfat	1.00 (± 0.04)	**0.96** (± 0.03)	0.97 (± 0.03)
cadata	**1.00** (± 0.03)	1.11 (± 0.04)	1.03 (± 0.04)
housing	1.00 (± 0.02)	0.99 (± 0.04)	**0.97** (± 0.03)

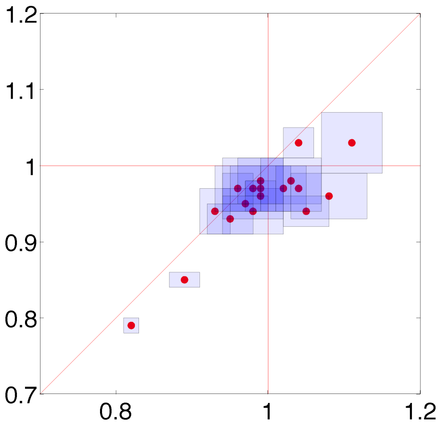

Figure 4. Relative performance of SVM classifiers and regression trees on UCI datasets. We normalize the unweighted performance to 1 and report relative variance. DoublyRobust consistently outperforms the other estimators. Error bars are obtained using 30 trials for each experiment. The graph on the RHS summarizes these results. We combine both regression and classification results since their behavior is entirely analogous. Boxes indicate the extent of uncertainty, with a red solid dot in the middle. The horizontal / vertical lines at 1 indicate whenever covariate shift performs better / worse than its unweighted counterpart. The straight diagonal line indicates whenever DoublyRobust outperforms Weighted. As can be seen, our method is much less susceptible to an increase in variance.

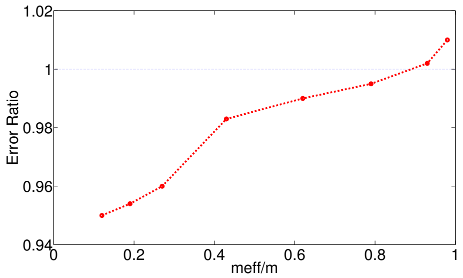

Figure 3. Comparison of Weighted and DoublyRobust using a synthetic dataset. We plot the error ratio as a function of the effective sample size. Note that a ratio of 1 implies that there is no covariate shift. As can be seen, our method improves the most whenever the increase in variance is the highest. This is consistent with the fact that it acts as variance reducer.

dataset in such a way that the effective sample size m_{eff} is less than 1/3 of the training data.

For classification, we use support vector machines with a linear kernel. As mentioned earlier, $\Omega[f, f'] = \frac{1}{2}\left\|w - w'\right\|^2$, i.e., the correction is additive in feature space. The regularization parameters are chosen separately for each empirical estimator by cross validation. We report the classification error $\Pr\left\{yf(x) < 0\right\}$. We normalize the errors with the Unweighted error.

For regression we apply regression trees to several UCI datasets. We report the square error loss for these experiments. As explained earlier, we first train a regression tree on the unweighted dataset and then build a differential regression tree on the residual with restricted tree depth in order to train the doubly robust regression tree.

The results are reported in Figure 4. We report the average RMSE error and the standard deviation over 30 trials for each experiment. The errors in both the above cases are normalized by the error of Unweighted. In both the tasks, it can be clearly seen that DoublyRobust outperforms both Unweighted and Weighted on most of the datasets. Note that neither Unweighted nor Weighted are significantly better than each other. On the other hand, our approach consistently outperforms both. This is in line with our intuition that the unweighted solution is an excellent variance reducer. Overall, we conclude that the proposed method is promising for covariate shift correction problem.

Application: We also tested the proposed method on a proprietary dataset using logistic regression. The sample size was $m = 20,555$ with $d = 30$ dimensions. Due to extreme covariate shift the effective sample size was only $m_{\text{eff}} = 28$. When normalizing the performance for the uncorrected estimate to 1, we obtained a relative error of 1.192 for covariate shift corrected risk minimization. A heuristic approach of hand tuning an upper-bound the covariate shift correction weights improved matters somewhat to a relative error of 0.986. Doubly robust covariate shift achieved a relative error of 0.975.

5. Conclusion

In this paper we proposed an intuitive and easy-to-use strategy for improving covariate shift correction. It addresses a key issue that plagues many covariate shift correction algorithms, namely that the variance increases considerably whenever samples are reweighted. It achieves this goal by using the unweighted solution as a variance-reducing proxy for the correct weighted solution. This is a rather general strategy and has been used with great success, e.g. as control variate, in the context of reinforcement learning (Sutton & Barto, 1998).

Our approach is particularly simple insofar as it requires essentially no additional code to use — all that is required in practice is to allow for reweighting and offset-correction in a linear model, a decision tree, or any other estimator that might be at hand. Of particular importance is the fact that we found our approach never to be harmful, something that cannot be said in general for covariate shift correction.

References

Agarwal, D., Li, L., and Smola, A.J. Linear-time estimators for propensity scores. *Artificial Intelligence and Statistics AISTATS*, 15:93–100, 2011. JMLR proceedings track.

Bang, H and Robins, J M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–973, 2005.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh., A. Sample selection bias correction theory. In *Proceedings of the International Conference on Algorithmic Learning Theory*, volume 5254 of *Lecture Notes in Computer Science*, pp. 38–53. Springer, 2008.

Doucet, Arnaud, de Freitas, Nando, and Gordon, Neil. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

Dudík, Miroslav, Langford, John, and Li, Lihong. Doubly robust policy evaluation and learning. In *Proceedings of the International Conference on Machine Learning*, pp. 1097–1104, 2011.

Gretton, A., Smola, A. J., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Dataset shift in machine learning. In Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. (eds.), *Covariate Shift and Local Learning by Distribution Matching*, pp. 131–160, Cambridge, MA, 2008. MIT Press.

Kang, J.D.Y. and Schafer, J.L. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Kifer, Daniel, Smith, Adam D., and Thakurta, Abhradeep. Private convex optimization for empirical risk minimization with applications to high-dimensional regression. In *COLT*, pp. 25.1–25.40, 2012.

Nguyen, X.L., Wainwright, M., and Jordan, M. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems 20*, pp. 1089–1096. MIT Press, Cambridge, MA, 2008.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. (eds.). *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2008.

Shimodaira, H. Improving predictive inference under convariance shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90, 2000.

Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems 20*, pp. 1433–1440, Cambridge, MA, 2008.

Sutton, R.S. and Barto, A.G. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., and Sugiyama, M. Direct density ratio estimation for large-scale covariate shift adaptation. In *SDM*, pp. 443–454. SIAM, 2008.

Vapnik, V. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

Yu, Y. and Szepesvári, C. Analysis of kernel mean matching under covariate shift. In *ICML*, 2012.

Appendix

Throughout this paper, we use on Rademacher averages for uniform convergence results. We define the empirical Rademacher average of a function class $\mathcal{F}$ as:

$$\mathrm{Rad}_m(\mathcal{F}) = \mathbf{E}\left[\sup_{f\in\mathcal{F}} \frac{2}{m}\left|\sum_{i=1}^{m}\sigma_i f(x_i)\right|\right]$$

where $\sigma_1, \cdots, \sigma_m$ are independent uniform $\{\pm 1\}$. For the case of kernel methods, the Rademacher average of function class $\mathcal{F} = \{f : \|f\| \leq \nu\}$ is:

$$\mathrm{Rad}_m(\mathcal{F}) = \frac{2\nu}{m}\sqrt{\mathrm{tr}(\mathbf{K})}.$$

where $\mathbf{K}$ is the kernel matrix corresponding to the training points X.

Following is an useful property of Rademacher average when the function class is translated by a function (refer (Bartlett & Mendelson, 2002) for details).

Lemma 4 *For any function class $\mathcal{F}$, we have*

$$Rad_m(\mathcal{F}+h) \leq Rad_m(F) + 2\sqrt{\frac{\sum_{i=1}^{m} h^2(x_i)}{m^2}}$$

for any function h.

A. Metatheorem

Proof of Theorem 1

Proof Consider the following conditions: (1) It is possible to obtain a good estimator of $f^*_{q,\lambda}$ from X, Y and that the covariate shift correction algorithm works i.e. Equation 8 holds (2) The minimizer of $\hat{T}[f]$ is close to the estimator $\hat{f}_{\mathrm{DR}}$ i.e Equation 9 is satisfied (3) Finally, Equation refeq:bound-c holds for a suitably small ν_0. Splitting probabilities of $\delta/3$ over violations of these conditions and then using union bound over these conditions, we get that with at least probability $1-\delta$, all the three conditions hold. More specifically, we invoke Equation (10) with $f_0 = \hat{f}_{q,\lambda}$ and $\nu' = \nu_{\mathrm{DR}} + A(\lambda, m, \delta/3)$. From Equation (8) it follows that under Assumption 3 with probability at least $1-\delta/3$ the risk minimizer with regard to p satisfies $\Omega[f, \hat{f}_{q,\lambda_q}] \leq \nu'$. Figure 5 shows a pictorial representation of the proof. Therefore, we have,

$$\begin{aligned}
R_p[\hat{f}_{T,\lambda'}] &\leq \widehat{T}[\hat{f}_{T,\lambda'}] + C(\hat{f}_{q,\lambda}, \lambda', m, \beta, \delta/3)\\
&\leq \widehat{T}[f^*_{p,\lambda_p}] + C(\hat{f}_{q,\lambda}, \lambda', m, \beta, \delta/3)\\
&\leq R_p[f^*_{p,\lambda_p}] + 2C(\hat{f}_{q,\lambda}, \lambda', m, \beta, \delta/3)
\end{aligned}$$

The first and third steps follow from Equation (10). The second step follows from the fact that f^*_{p,λ_p} satisfies $\Omega[f^*_{p,\lambda_p}, \hat{f}_{q,\lambda_q}] \leq \nu'$. Finally using Equation (9), we get the required result. ■

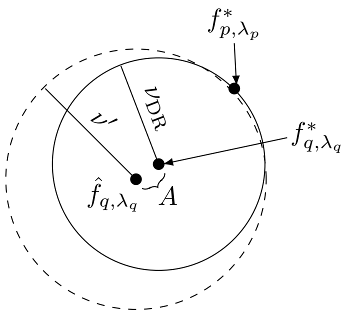

Figure 5. Pictorial representation of metatheorem and intuition behind doubly robust estimation. The estimate $\hat{f}_{q,\lambda_q}$ is close to f^*_{q,λ_q} with high probability and hence, close to the risk minimizer we are concerned about, namely f^*_{p,λ_p}

.

B. Unweighted Estimator

The following result provides a bound on the difference of risk with respect to the distributions p and q.

Lemma 5 *Suppose we have $\|p-q\|_1 \leq \epsilon$ and $\|E_{Y|X}[\ell(Y, f(X))]\|_\infty \leq L$ then $|R_p[f] - R_q[f]| \leq L\epsilon$.*

Proof The proof directly follows from the definition, as shown below.

$$\begin{aligned}
|R_p[f] - R_q[f]| &= \left|\int_{\mathcal{X}\times\mathcal{Y}} \ell(y, f(x))(p(x) - q(x))dydx\right|\\
&= \left|\int_{\mathcal{X}} \mathbb{E}_{Y|X}[\ell(y, f(x))](p(x) - q(x))dx\right|\\
&\leq L\|p-q\|_1 = L\epsilon.
\end{aligned}$$

The first step follows from the definition of risk. The second step follows from the on conditional expectation. ■

It should be noted that the above bound is tight. For example, suppose the distributions p and q have disjoint supports and $\mathbb{E}_{Y|X}[\ell(Y, f(X))]$ is the constant function L then $|R_p[f] - R_q[f]| = L\epsilon$.

Proof of Theorem 2

Proof From the standard Rademacher analysis of kernel methods (e.g., (Bartlett & Mendelson, 2002)) on function class $\mathcal{F} = \{\|f\| \leq \nu_q\}$, we have the following:

$$|R[f|X, Y] - R_q[f]| \leq \frac{2L\nu_q}{m}\sqrt{\mathrm{tr}(\mathbf{K})} + 3L\sqrt{\frac{\log(2/\delta)}{2m}}. \quad (17)$$

for all $f \in \mathcal{F}$. Therefore,

$$\begin{aligned} & R_q[\hat{f}_{q,\lambda_q}] \\ & \le R_q[\hat{f}_{q,\lambda_q}|X,Y] + \frac{2L\nu_q}{m}\sqrt{\operatorname{tr}(\mathbf{K})} + 3L\sqrt{\frac{\log(2/\delta)}{2m}} \\ & \le R_q[f^*_{q,\lambda_q}] + \frac{4L\nu_q}{m}\sqrt{\operatorname{tr}(\mathbf{K})} + 6L\sqrt{\frac{\log(2/\delta)}{2m}}. \end{aligned}$$

The second inequality follows from the fact that $R_q[\hat{f}_{q,\lambda_q}|X,Y] \le R_q[f^*_{q,\lambda_q}]$ and Equation (17). Now using the relation between the R_p and R_q in Lemma 5, we have the result. ■

Proof of Lemma 1

Proof From Theorem 2, we know that the risks of $\hat{f}_{q,\lambda_q}$ and f^*_{q,λ_q} are close. Now by using the relationship between minimizers and the objective value of two strongly convex functions, from Lemma 6, we have the required result. ■

The following result relates between the minimizers of two strongly convex function with the difference in their respective objective function values.

Lemma 6 *Assume that we have two strongly convex auxiliary functions u, v with $\|u - v\|_\infty \le \epsilon$ and with modulus of strong convexity λ. Then their minima x_u, x_v satisfy $\|x_u - x_v\| \le \sqrt{2\epsilon/\lambda}$.*

Proof By the assumption of proximity and by strong convexity we have

$$\begin{aligned} u(x_u) + \epsilon &\ge v(x_u) \\ &\ge v(x_v) + \langle x_u - x_v, \partial_x v(x_v)\rangle + \frac{\lambda}{2}\left\|x_u - x_v\right\|^2 \\ &\ge v(x_v) + \frac{\lambda}{2}\left\|x_u - x_v\right\|^2 \end{aligned}$$

The same also holds with u and v interchanged. Summing both inequalities and subtracting identical terms yields that $2\epsilon \ge \lambda \left\|x_u - x_v\right\|^2$. Rearrangement of terms proves the claim. ■

C. Generalization bounds for PRM

Proof of Lemma 2

Proof Using standard Rademacher average bound for kernel methods from (Bartlett & Mendelson, 2002), we have the following:

$$\mathrm{Rad}_m(\{f : \|f\| \le \nu\}) = \frac{2\nu}{m}\sqrt{\operatorname{tr}(\mathbf{K})}$$

Combining the above result with Lemma 4 gives

$$\mathrm{Rad}_m(f : \|f - f_0\| \le \nu) = \frac{2\nu}{m}\sqrt{\operatorname{tr}(\mathbf{K})} + \frac{1}{m}\|f_0\|_2.$$

Finally, using the standard Rademacher bounds, again from (Bartlett & Mendelson, 2002), we get the required result. ■

Lemma 7 *Suppose $\hat{f}_{\hat{\beta},\lambda}$ and $\hat{f}_{\beta,\lambda}$ are as defined in Equation (12) and Equation (13) respectively. We have the following relationship between these risk minimizers:*

$$\|\hat{f}_{\hat{\beta},\lambda} - \hat{f}_{\beta,\lambda}\| \le \frac{2\kappa L}{\lambda m}\|\hat{\beta} - \beta\|_1$$

where $\|\phi(x)\| \le \kappa \; \forall x \in \mathcal{X}$.

Proof We define the following function:

$$\begin{aligned} S[f] = \left\langle \partial_f R[\hat{f}_{\beta,\lambda}|X,Y,\beta] - \partial_f R[\hat{f}_{\hat{\beta},\lambda}|X,Y,\hat{\beta}], f - \hat{f}_{\hat{\beta},\lambda} \right\rangle \\ + \frac{\lambda}{2}\|f - \hat{f}_{\hat{\beta},\lambda}\|^2. \end{aligned}$$

We note that by construction, $S[\hat{f}_{\hat{\beta},\lambda}] = 0$. Also, note that $\partial_f S[\hat{f}_{\beta,\lambda}] = 0$ by definition of $\hat{f}_{\hat{\beta},\lambda}$ and $\hat{f}_{\beta,\lambda}$. Using the above and second order condition, it is easy to see that $\hat{f}_{\beta,\lambda}$ is the minima of $S[f]$. By above facts, we get

$$\begin{aligned} \left\langle \partial_f R[\hat{f}_{\beta,\lambda}|X,Y,\beta] - \partial_f R[\hat{f}_{\hat{\beta},\lambda}|X,Y,\hat{\beta}], \hat{f}_{\beta,\lambda} - \hat{f}_{\hat{\beta},\lambda} \right\rangle \\ + \frac{\lambda}{2}\|f - \hat{f}_{\hat{\beta},\lambda}\|^2 \le 0. \end{aligned}$$

Now by adding and subtracting $\langle \partial_f R[\hat{f}_{\hat{\beta},\lambda}|X,Y,\beta], \hat{f}_{\beta,\lambda} - \hat{f}_{\hat{\beta},\lambda}\rangle$ we obtain the following:

$$\begin{aligned} & \frac{\lambda}{2}\|\hat{f}_{\beta,\lambda} - \hat{f}_{\hat{\beta},\lambda}\|^2 \\ & + \left\langle \partial_f R[\hat{f}_{\beta,\lambda}|X,Y,\beta] - \partial_f R[\hat{f}_{\hat{\beta},\lambda}|X,Y,\beta], \hat{f}_{\beta,\lambda} - \hat{f}_{\hat{\beta},\lambda} \right\rangle \\ & + \left\langle \partial_f R[\hat{f}_{\hat{\beta},\lambda}|X,Y,\beta] - \partial_f R[\hat{f}_{\beta,\lambda}|X,Y,\hat{\beta}], \hat{f}_{\beta,\lambda} - \hat{f}_{\hat{\beta},\lambda} \right\rangle \le 0. \end{aligned}$$

Since ℓ is convex, the second term is non-negative. This leads to the fact that following expression is negative:

$$\begin{aligned} \left\langle \partial_f R[\hat{f}_{\beta,\lambda}|X,Y,\beta] - \partial_f R[\hat{f}_{\hat{\beta},\lambda}|X,Y,\hat{\beta}], \hat{f}_{\beta,\lambda} - \hat{f}_{\hat{\beta},\lambda} \right\rangle \\ + \frac{\lambda}{2}\|\hat{f}_{\beta,\lambda} - \hat{f}_{\hat{\beta},\lambda}\|^2. \end{aligned} \tag{18}$$

We observe that

$$\begin{aligned} & \left\langle \partial_f R[\hat{f}_{\hat{\beta},\lambda}|X,Y,\beta] - \partial_f R[\hat{f}_{\beta,\lambda}|X,Y,\hat{\beta}], \hat{f}_{\beta,\lambda} - \hat{f}_{\hat{\beta},\lambda} \right\rangle \\ & \le \|\partial_f R[\hat{f}_{\beta,\lambda}|X,Y,\beta] - \partial_f R[\hat{f}_{\hat{\beta},\lambda}|X,Y,\hat{\beta}]\|\|\hat{f}_{\beta,\lambda} - \hat{f}_{\hat{\beta},\lambda}\| \\ & \le \frac{\kappa L}{m}\|\hat{\beta} - \beta\|_1\|\hat{f}_{\beta,\lambda} - \hat{f}_{\hat{\beta},\lambda}\| \end{aligned}$$

The first step follows from Cauchy-Schwartz inequality. The second step follows from Lipschitz continuity of ℓ and definition of κ. Summing the above inequality with Equation 18, we get

$$\|\hat{f}_{\beta,\lambda} - \hat{f}_{\hat{\beta},\lambda}\| \leq \frac{2\kappa L}{\lambda m}\|\hat{\beta} - \beta\|_1$$

■

We next prove the following result, which is useful in proving generalization bounds for PRM. We define (generalized) hellinger distance between two functions non-negative functions f and g under distribution q and its empirical version as:

$$h_q(f,g)^2 = \int (\sqrt{f(x)} - \sqrt{g(x)})^2 q(x)dx$$

$$\hat{h}_q(f,g)^2 = \frac{1}{m}\sum_i \left(\sqrt{f_i} - \sqrt{g_i}\right)^2.$$

Lemma 8 *Suppose $\hat{\beta}$ is the estimator for β obtained using PRM. Let $\gamma_m = cm^{-2/2+\tau}$ for some $\tau > 0$ and a constant c. We have the following:*

$$\frac{1}{m}\left\|\hat{\beta} - \beta\right\|_1 \leq \sqrt{\eta\gamma_m} + \eta\sqrt[4]{\frac{8}{m}\log\left(\frac{2}{\delta}\right)}$$

with probability at least $1 - \delta$.

Proof The following holds:

$$\left[\frac{1}{m}\left\|\hat{\beta} - \beta\right\|_1\right]^2$$

$$= \frac{1}{m^2}\left[\sum_i \left(\left|\sqrt{\hat{\beta}_i} - \sqrt{\beta_i}\right|\right)\left(\sqrt{\hat{\beta}_i} + \sqrt{\beta_i}\right)\right]^2$$

$$\leq \frac{1}{m^2}\left[\sum_i \left(\sqrt{\hat{\beta}_i} - \sqrt{\beta_i}\right)^2\right]\left[\sum_i \left(\sqrt{\hat{\beta}_i} + \sqrt{\beta_i}\right)^2\right]$$

$$\leq \frac{1}{m}\hat{h}_q(\hat{\beta},\beta)^2 4m\eta = 4\eta\hat{h}_q^2(\hat{\beta},\beta).$$

The first and second step follow from Cauchy-Schwartz inequality and the fact that RND is bounded above by η respectively. Now using Hoeffding bound, we have the following relation between $h_q(\hat{\beta},\beta)$ and its empirical estimate $\hat{h}_q(\hat{\beta},\beta)$:

$$\hat{h}_q(\hat{\beta},\beta) \leq h_q(\hat{\beta},\beta) + \sqrt[4]{\frac{\eta^2}{2m}\log\left(\frac{2}{\delta}\right)}$$

with probability $1 - \delta$. Combining the above two facts and the bound on hellinger distance $h_q(\hat{\beta},\beta)$ in Theorem 2 of (Nguyen et al., 2008) gives us the required result. ■

Proof of Lemma 3

Proof We have the following:

$$\left|R_p[\hat{f}_{\hat{\beta},\lambda}] - R_p[\hat{f}_{\beta,\lambda}]\right|$$

$$= \left|\mathbf{E}_{X,Y}\left[\ell\left(Y, \hat{f}_{\hat{\beta},\lambda}(X)\right) - \ell\left(Y, \hat{f}_{\beta,\lambda}(X)\right)\right]\right|$$

$$\leq \kappa L\|\hat{f}_{\hat{\beta},\lambda} - \hat{f}_{\beta,\lambda}\| \leq \frac{2\kappa^2 L^2}{\lambda m}\|\hat{\beta} - \beta\|_1$$

$$\leq \frac{2\kappa^2 L^2}{\lambda}\left(\sqrt{\eta\gamma_m} + \eta\sqrt[4]{\frac{8}{m}\log\left(\frac{2}{\delta}\right)}\right)$$

The first inequality follows from L-Lipschitz nature of the loss function ℓ and Cauchy-Schwartz inequality. The second and third inequalities follow from Lemma 7 and Lemma 8 respectively. ■

Proof of Theorem 3

Proof The result follows on similar lines as Theorem 1 by using $f_0 = 0$. Here, since we do not require bound A, we split probability $\delta/2$ over bounds B and C. The bounds B and C are obtained from Lemma 3 and Lemma 2 respectively. ■

Proof of Theorem 4

Proof The result follows from Theorem 1 by instantiating bounds A, B and C using Lemma 1, Lemma 3 and Lemma 2 respectively. ■

D. Covariate Shift Bounds using KMM

In this section, as promised earlier, we provide theoretical details of KMM appoach. We use two different instantiations — using distribution stability and mean embedding properties — of function $\widehat{T}[f]$ to derive generalization bounds for KMM.

Bounds using Distribution Stability

We first use distribution stability to derive the generalization bounds for KMM. Similar to the PRM case, we use $\widehat{T}[f] = \widehat{R}[f|X,Y,\beta]$ here.

Bound C: Bound C is exactly same as in PRM i.e. Lemma 2.

Bound B: For Bound B, we can borrow bounds from (Cortes et al., 2008). In case of generic covariate shift, it amounts to straightforward application of the following result. Doubly robust case is less trivial. We need to prove that the stability bounds hold even in the case of regularization with a prior. This amounts to essentially identical result as above. In particular, we need to argue that Theorem 1 of (Cortes et al., 2008) holds for a given prior. The proof essentially follows the argument with minor modifications in the regularization. Hence, we do not provide it here.

Lemma 9 *Suppose $\hat{f}_{\hat{\beta},\lambda}$ and $\hat{f}_{\beta,\lambda}$ are as defined in Equation (12) and Equation (13) respectively. Let $sigma(\mathbf{K})$ denote the condition number of matrix $\mathbf{K}$. Let the regularization parameter $\gamma_m = m^{-2/(2+\tau)}$ for some $\tau > 0$, then we have the following with probability at least $1-\delta$:*

$$\begin{aligned}&\left|R_p[\hat{f}_{\hat{\beta},\lambda}] - R_p[\hat{f}_{\beta,\lambda}]\right| \\ &\leq \frac{L^2\kappa^{3/2}\sigma^{1/2}(\mathbf{K})}{\lambda}\sqrt{\frac{\eta^2+1}{m}}\left(1+\sqrt{2\log\left(\frac{2}{\delta}\right)}\right)\end{aligned}$$

We now state the main results about generalization bounds for KMM, which follow as corollary of Theorem 1.

Theorem 5 *Suppose $\hat{f}_{p,\lambda_p}$ and f^*_{p,λ_p} are as defined in Equation (4) and Equation (7) respectively. Then we have the following with probability at least $1-\delta$:*

$$R_p[\hat{f}_{p,\lambda_p}] \leq R_p[f^*_{p,\lambda_p}] + \Delta_{W,S} + \Delta_{W,R} \tag{19}$$

where $\Delta_{W,S}$ and $\Delta_{W,R}$, representing the covariate shift and function complexity parts of the bound, are defined as below.

$$\Delta_{W,S} = \frac{L^2\kappa^{3/2}\sigma^{1/2}(\mathbf{K})}{\lambda_p}\sqrt{\frac{\eta^2+1}{m}}\left(1+\sqrt{2\log\left(\frac{4}{\delta}\right)}\right)$$

$$\Delta_{W,R} = \frac{4\eta L\nu_p\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + 6\eta L\sqrt{\frac{\log(4/\delta)}{2m}}$$

Theorem 6 *Suppose $\hat{f}_{\mathrm{DR}}$ and f^*_{p,λ_p} are as defined in Equation (5) and Equation (7) respectively. Then we have the following with probability at least $1-\delta$:*

$$R_p[\hat{f}_{\mathrm{DR}}] \leq R_p[f^*_{p,\lambda_p}] + \Delta_{\mathrm{DR},S} + \Delta_{\mathrm{DR},R} \tag{20}$$

where $\Delta_{\mathrm{DR},S}$ and $\Delta_{\mathrm{DR},R}$, representing the covariate shift and function complexity parts of the bound, are defined as below.

$$\nu' = \nu_{DR} + \sqrt{\frac{4L}{\lambda_q}\left(\frac{2\nu_q\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + 3\sqrt{\frac{\log(6/\delta)}{2m}}\right)}$$

$$\Delta_{\mathrm{DR},S} = \frac{L^2\kappa^{3/2}\sigma^{1/2}(\mathbf{K})}{\lambda'}\sqrt{\frac{\eta^2+1}{m}}\left(1+\sqrt{2\log\left(\frac{6}{\delta}\right)}\right)$$

$$\Delta_{\mathrm{DR},R} = \frac{4\eta L\nu'\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + 6\eta L\sqrt{\frac{\log(6/\delta)}{2m}} + \frac{\eta L}{m}\|\hat{f}_{q,\lambda_q}\|_2$$

Bounds using Mean Embedding

Consider the following empirical risk minimization problems:

$$\hat{f} := \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R}[f|X,Y] \text{ s.t. } f \in \mathcal{F} \tag{21}$$

Recall, our goal here is to obtain a bound on $|R_p[\hat{f}] - R_p[f^*_p]|$. Here, we use a different approach to obtain the required generalization bound along the lines of (Gretton et al., 2008). In particular, here we instantiate metatheorem with $\widehat{T}[f] = \widehat{R}[f|X,Y,\beta]$ i.e. empirical risk with shift weights. This specific instantiation leads to a trivial bound B i.e. $B=0$.

Bound C: We now focus our attention towards obtaining the bound C in this case. First, similar to (Gretton et al., 2008), we require assumptions on the loss function.

Assumption 4 *We assume that $l(x,f) = \mathbf{E}_{y|x}[\ell(y,f(x))]$ belongs to a RKHS on $\mathcal{X}$ such that $\|\ell(x,f)\| \leq \rho$ when $f \in \mathcal{F}$. In addition, we also assume $l(x,y,f) = \ell(y,f(x))$ is an element of an RKHS with feature map Θ such that $\|\Theta\| \leq \kappa$ and $\|l(x,y,f)\| \leq \rho$ when $f \in \mathcal{F}$.* [4]

Under these assumptions, we can directly appeal to Corollary 1.9 and Lemma 1.5 to obtain the main result of this section.

Theorem 7 *Suppose Assumptions 4 are satisfied for function class $\mathcal{F} = \|f - f_0\| \leq \nu$. Then, we have the following with probability at least $1-\delta$:*

$$\sup_{f:\|f-f_0\|\leq\nu} |R_p[f] - R_p[f^*_{p,\lambda_p}]| \leq \Delta_K + \Delta_R$$

where $m_{\text{eff}} = \|\hat{\beta}\|_1^2/\|\hat{\beta}\|_2^2$ is the effective sample size, and Δ_K and Δ_R, defined below, represent the covariate shift correction and function complexity portion of the bound.

$$\begin{aligned}\Delta_K &= \frac{\rho\kappa(2+\sqrt{2\log(2/\delta)})}{\sqrt{m_{\text{eff}}}} \\ &\quad + \frac{2\rho\kappa\sqrt{(\eta^2+1)}(1+\sqrt{2\log(2/\delta)}}{\sqrt{m}}\end{aligned}$$

$$\Delta_R = \frac{2L\nu\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + 3L\sqrt{\frac{\log(2/\delta)}{2m}} + \frac{\sqrt{\sum_{i=1}^m f_0^2(x_i)}}{m}$$

Proof Using Corollary 1.9 and Lemma 1.5 of (Gretton et al., 2008), we have the following:

$$\begin{aligned}\left|\widehat{R}[f|X,Y,\hat{\beta}] - \widehat{R}[f|X']\right| &\leq \frac{\rho\kappa(2+\sqrt{2\log(2/\delta)})}{\sqrt{m_{\text{eff}}}} \\ &\quad + 2\rho\kappa(1+\sqrt{2\log 2/\delta})\sqrt{(\eta^2+1)/m}\end{aligned}$$

[4] We use the same constants κ and ρ in both the cases, without loss of generality.

for all $f \in \mathcal{F}$. Now using standard uniform convergence bounds we have:

$$\left|\widehat{R}[f|X'] - R_p[f]\right| \leq \frac{2L\nu\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + 3L\sqrt{\frac{\log(2/\delta)}{2m}}$$

Summing both the equations above and by using triangle inequality, we get the required result. ■

We can now use the above bounds to derive generalization bounds for weighted and doubly robust estimators.

Theorem 8 *Suppose $\hat{f}_{p,\lambda_p}$ and f^*_{p,λ_p} are as defined in Equation (4) and Equation (7) respectively. Also, suppose the function class $\mathcal{F} = \{f : \|f\| \leq \nu_p\}$ satisfies Assumption 4 with parameter ρ_p. Let $m_{\mathrm{eff}} = \|\hat{\beta}\|_1^2/\|\hat{\beta}\|_2^2$ be the effective sample size. Then we have the following with probability at least $1-\delta$:*

$$R_p[\hat{f}_{p,\lambda_p}] \leq R_p[f^*_{p,\lambda_p}] + \Gamma_{W,S} + \Gamma_{W,R}$$

where $\Gamma_{W,S}$ and $\Gamma_{W,R}$, representing the covariate shift and function complexity parts of the bound, are defined as below.

$$\Gamma_{W,S} = \frac{2\rho_p\kappa(2+\sqrt{2\log(2/\delta)})}{\sqrt{m_{\mathrm{eff}}}} + \frac{4\rho_p\kappa\sqrt{(\eta^2+1)}(1+\sqrt{2\log(2/\delta)}}{\sqrt{m}}$$

$$\Gamma_{W,R} = \frac{4L\nu_p\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + 6L\sqrt{\frac{\log(2/\delta)}{2m}}$$

Proof The result follows is similar to Theorem 1 by using $f_0 = 0$. In this case, Bound B is trivial. Bound C is obtained from Lemma 3 and Lemma 2 respectively. ■

Theorem 9 *Suppose $\hat{f}_{\mathrm{DR}}$ and f^*_{p,λ_p} are as defined in Equation (5) and Equation (7) respectively. Also, suppose the function class $\mathcal{F} = \{f : \|f - \hat{f}_{q,\lambda_q}\| \leq \nu'\}$ satisfies Assumption 4 with parameter ρ', where ν' is as defined below.. Let $m_{\mathrm{eff}} = \|\hat{\beta}\|_1^2/\|\hat{\beta}\|_2^2$ be the effective sample size. Then we have the following with probability at least $1-\delta$:*

$$R_p[\hat{f}_{\mathrm{DR}}] \leq R_p[f^*_{p,\lambda_p}] + \Gamma_{\mathrm{DR},S} + \Gamma_{\mathrm{DR},R}$$

where $\Gamma_{W,S}$ and $\Gamma_{W,R}$, representing the covariate shift and function complexity parts of the bound, are defined as below.

$$\nu' = \nu_{DR} + \sqrt{\frac{4L}{\lambda_q}\left(\frac{2\nu_q\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + 3\sqrt{\frac{\log(6/\delta)}{2m}}\right)}$$

$$\Gamma_{\mathrm{DR},S} = \frac{2\rho'\kappa(2+\sqrt{2\log(6/\delta)})}{\sqrt{m_{\mathrm{eff}}}} + \frac{4\rho'\kappa\sqrt{(\eta^2+1)}(1+\sqrt{2\log(6/\delta)}}{\sqrt{m}}$$

$$\Gamma_{\mathrm{DR},R} = \frac{4L\nu'\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + 6L\sqrt{\frac{\log(6/\delta)}{2m}} + \frac{L}{m}\|\hat{f}_{q,\lambda_p}\|_2$$

Proof The result follows from Theorem 1 by instantiating bounds A, B and C using Lemma 1, trivial bound 0 and Lemma 2 respectively. ■

It is easy to see that generalization bound for doubly robust estimator is much better than the generic bound when $\rho' \ll \rho_p$, which is roughly the rationale behind the Assumption 3. Note the dependence of bounds on the effective sample size. It is also worthwhile to mention that while we derive generalization bounds under aformentioned assumptions, it is not hard to extend it to a more general setting by employing ideas from the recent work of (Yu & Szepesvári, 2012).