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

7 Appendix

Throughout the proofs, $h(\cdot) \in \mathcal{H}$ is assumed to be an unspecified function in the RKHS. Also, we use $\mathbb{E}_X[\cdot]$ to denote expectation over the randomness of X while fixing others and $\mathbb{E}_{|X}[\cdot]$ as the conditional expectation $\mathbb{E}[\cdot|X]$. Moreover we remark that all results involving $\hat{g}_{\gamma,data}$ can be interpreted either as a high probability bound or a bound on expectation over $\mathbb{E}_{data}$ (i.e., if we train $\hat{g}_{\gamma,\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}$ using $\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}$, then $\mathbb{E}_{data}$ means $\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}$). The same interpretation applies for the results with Big-$\mathcal{O}$ notations. Finally, constants C_2, C_2', C_3, C_3' and C_3'' as well as similar constants introduced later which depend on $R, g(\cdot)$ or δ (for $1-\delta$ high probability bound) will sometimes be denoted by a common C during the proofs for ease of presentation.

7.1 Preliminaries

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

$$\|f\|_\infty = \sup_{x\in\mathcal{X}} |\langle f(\cdot), \Phi(\cdot,x)\rangle_{\mathcal{H}}| \leq R\|f\|_{\mathcal{H}}. \tag{1}$$

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

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

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

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

$$|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i - \mathbb{E}[X]| \leq B\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}. \tag{3}$$

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

$$\begin{aligned}\hat{L}(\boldsymbol{\beta}_{|\boldsymbol{x}_1^{tr},...,\boldsymbol{x}_{n_{tr}}^{tr}}) =& \left\|\frac{1}{n_{tr}}\sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}_j^{tr})\Phi(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}}\sum_{i=1}^{n_{te}}\Phi(\boldsymbol{x}_i^{te})\right\|_{\mathcal{H}} \\ \leq& \sqrt{2\log\frac{2}{\delta}}R\sqrt{\left(\frac{B^2}{n_{tr}} + \frac{1}{n_{te}}\right)}.\end{aligned} \tag{4}$$

7.2 Learning Theory Estimates

To adopt the more realistic assumption as in [Yu and Szepesvári, 2012, Cucker and Zhou, 2007] that the true regression function $g(\cdot) \notin \mathcal{H}$ but rather $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, we need results from learning theory.

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

$$g_\gamma(\cdot) = \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ \|f - g\|^2_{\mathscr{L}^2_{P_{tr}}} + \gamma \|f\|^2_{\mathcal{H}} \right\} \tag{5}$$

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

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

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

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

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

$$\|l\|_r \triangleq \sup_{\gamma > 0} \frac{\tilde{\mathbb{K}}(l, \gamma)}{\gamma^r} < \infty. \tag{6}$$

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

$$\mathbb{K}(l, \gamma) \triangleq \inf_{f \in \mathcal{H}} \{ \|l - f\|^2_{\mathscr{L}^2_{P_{tr}}} + \gamma \|f\|^2_{\mathcal{H}} \}. \tag{7}$$

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

$$\sup_{\gamma > 0} \frac{\mathbb{K}(l, \gamma)}{\gamma^r} \le \left(\sup_{\gamma > 0} \frac{\tilde{\mathbb{K}}(l, \sqrt{\gamma})}{(\sqrt{\gamma})^r} \right)^2 = \|l\|_r^2 < \infty. \tag{8}$$

Proof. It follows from $\sqrt{a + b} \le \sqrt{a} + \sqrt{b}, \quad \forall a, b \ge 0$ that

$$\sqrt{\mathbb{K}(l, \gamma)} \le \tilde{\mathbb{K}}(l, \sqrt{\gamma}). \tag{9}$$

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

$$\sup_{\gamma > 0} \frac{\mathbb{K}(l, \gamma)}{\gamma^r} \le \left(\sup_{\gamma > 0} \frac{\tilde{\mathbb{K}}(l, \sqrt{\gamma})}{(\sqrt{\gamma})^r} \right)^2 = \|l\|_r^2 < \infty. \tag{10}$$

□

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

$$g(\cdot) \in (\mathscr{L}^2_{P_{tr}}, \mathcal{H}^+)_{\frac{\theta}{\theta+2}} \tag{11}$$

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

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

Proof. The proof follows from Theorem 4.1, Corollary 4.17 and Remark 4.18 of [Cucker and Zhou, 2007]. □

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

$$\|g_\gamma - g\|_{\mathscr{L}^2_{P_{tr}}} \leq C_2 \gamma^{\zeta}. \tag{12}$$

Second, it follows from Theorem 3.1 in [Sun and Wu, 2009] as well as [Smale and Zhou, 2007, Sun and Wu, 2010] that

$$\|g_\gamma - \hat{g}_{\gamma,data}\|_{\mathscr{L}^2_{P_{tr}}} \leq C_2'(\gamma^{-1/2} m^{-1/2} + \gamma^{-1} m^{-3/4}), \tag{13}$$

and, by the triangle inequality,

$$\|g - \hat{g}_{\gamma,data}\|_{\mathscr{L}^2_{P_{tr}}} \leq C_3(\gamma^{\zeta} + \gamma^{-1/2} m^{-1/2} + \gamma^{-1} m^{-3/4}). \tag{14}$$

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

$$\|g_\gamma - \hat{g}_{\gamma,data}\|_{\mathcal{H}} \leq C_3' \gamma^{-1} m^{-1/2}, \tag{15}$$

with $C_3' = 6R \log \frac{2}{\delta}$. In fact, if we define $\sigma^2 \triangleq \mathbb{E}_{\boldsymbol{x} \sim P_{tr}} \mathbb{E}_{Y|\boldsymbol{x}} (g(\boldsymbol{x}) - Y)^2$, then Theorem 3 of [Smale and Zhou, 2007] stated that

$$\|g_\gamma - \hat{g}_{\gamma,data}\|_{\mathcal{H}} \leq C_3''((\sqrt{\sigma^2} + \|g_\gamma - g\|_{\mathscr{L}^2_{P_{tr}}})\gamma^{-1} m^{-1/2} + \gamma^{-1} m^{-1}). \tag{16}$$

7.3 Main Proofs

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

$$\begin{aligned}
&V_R(\rho) - \nu \\
=&\frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr})(y_j^{tr} - g(\boldsymbol{x}_j^{tr})) + \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} (\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))(g(\boldsymbol{x}_j^{tr}) - h(\boldsymbol{x}_j^{tr})) \\
&+ \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} (\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))(h(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr})) \\
&+ \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \beta(\boldsymbol{x}_j^{tr})(g(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr})) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te}) - \nu.
\end{aligned} \tag{17}$$

To bound terms in (17), we first use Corollary 2 to conclude that with probability at least $1 - \delta$,

$$|\frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr})(y_j^{tr} - g(\boldsymbol{x}_j^{tr}))| \leq B \sqrt{\frac{1}{\lfloor \rho n_{tr} \rfloor} \log \frac{2}{\delta}} = \mathcal{O}(n_{tr}^{-1/2}). \tag{18}$$

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

$$
\begin{aligned}
&\left|\frac{1}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))(h(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr}))\right| \\
=&\left|\frac{1}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))\langle h - \hat{g}, \Phi(\boldsymbol{x}_j^{tr})\rangle_{\mathcal{H}}\right| \\
=&\left|\left\langle h - \hat{g}, \frac{1}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))\Phi(\boldsymbol{x}_j^{tr})\right\rangle_{\mathcal{H}}\right| \\
\leq&\|h - \hat{g}\|_{\mathcal{H}}(\hat{L}(\hat{\boldsymbol{\beta}}) + \hat{L}(\boldsymbol{\beta}_{|\boldsymbol{x}_1^{tr},\ldots,\boldsymbol{x}_{\lfloor \rho n_{tr} \rfloor}^{tr}})) \leq 2\|h - \hat{g}\|_{\mathcal{H}}\hat{L}(\boldsymbol{\beta}_{|\boldsymbol{x}_1^{tr},\ldots,\boldsymbol{x}_{\lfloor \rho n_{tr} \rfloor}^{tr}}),
\end{aligned} \tag{19}
$$

by definition of (1). Thus, when taking $h = g_\gamma$ and $\hat{g} = \hat{g}_{\gamma,\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}$ for some γ, we can combine (4) and (15) to guarantee, with probability $1 - 2\delta$,

$$
\begin{aligned}
&\left|\frac{1}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))(h(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr}))\right| \\
\leq&\sqrt{8\log\frac{2}{\delta}}RC(1-\rho)^{-1/2}(\gamma^{-1}n_{tr}^{-1/2}) \cdot \sqrt{\left(\frac{B^2}{n_{tr}} + \frac{1}{n_{te}}\right)} \\
=&\mathcal{O}(\gamma^{-1}n_{tr}^{-1/2}(n_{tr}^{-1} + n_{te}^{-1})^{\frac{1}{2}}).
\end{aligned} \tag{20}
$$

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

$$
\begin{aligned}
&\mathbb{E}_{|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\left[\frac{1}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}\beta(\boldsymbol{x}_j^{tr})(g(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr})) + \frac{1}{n_{te}}\sum_{i=1}^{n_{te}}\hat{g}(X_i^{te}) - \nu\right] \\
=&\mathbb{E}_{\boldsymbol{x}\sim P_{tr}}[\beta(\boldsymbol{x})g(\boldsymbol{x})] - \nu - \mathbb{E}_{\boldsymbol{x}\sim P_{tr}}[\beta(\boldsymbol{x})\hat{g}(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P_{te}}[\hat{g}(\boldsymbol{x})] \\
=&\mathbb{E}_{\boldsymbol{x}\sim P_{te}}[g(\boldsymbol{x})] - \nu - \mathbb{E}_{\boldsymbol{x}\sim P_{te}}[\hat{g}(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P_{te}}[\hat{g}(\boldsymbol{x})] \\
=&0.
\end{aligned} \tag{21}
$$

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

$$
\begin{aligned}
\mathrm{Var}(\tau) =&\mathrm{Var}(\mathbb{E}_{|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}(\tau)) + \mathbb{E}[\mathrm{Var}_{|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}(\tau)] \\
=&\mathbb{E}[\mathrm{Var}_{|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}(\tau)] \\
=&\frac{1}{\lfloor \rho n_{tr} \rfloor}\mathbb{E}[\mathrm{Var}_{\boldsymbol{x}\sim P_{tr}|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}(\beta(\boldsymbol{x})(g(\boldsymbol{x}) - \hat{g}(\boldsymbol{x})))] + \frac{1}{n_{te}}\mathbb{E}[\mathrm{Var}_{\boldsymbol{x}\sim P_{te}|\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}(\hat{g}(\boldsymbol{x}))] \\
\leq&\frac{B^2}{\lfloor \rho n_{tr} \rfloor}\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|g - \hat{g}\|_{\mathscr{L}_{P_{tr}}^2}^2 + \frac{1}{n_{te}}\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|\hat{g}\|_{\mathscr{L}_{P_{te}}^2}^2 \\
\leq&\frac{B^2}{\lfloor \rho n_{tr} \rfloor}\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|g - \hat{g}\|_{\mathscr{L}_{P_{tr}}^2}^2 + \frac{B}{n_{te}}\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|\hat{g}\|_{\mathscr{L}_{P_{tr}}^2}^2,
\end{aligned} \tag{22}
$$

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

$$
|\tau| \leq \sqrt{\frac{1}{\delta}}\sqrt{\frac{B^2}{\lfloor \rho n_{tr} \rfloor}\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|g - \hat{g}\|_{\mathscr{L}_{P_{tr}}^2}^2 + \frac{BR^2}{n_{te}}}, \tag{23}
$$

which becomes, by (14), with probability $1-2\delta$,

$$
\begin{aligned}
|\tau| \leq & \sqrt{\frac{1}{\delta}}\sqrt{\frac{B^2}{\lfloor \rho n_{tr}\rfloor}C(1-\rho)^{-3/4}(\gamma^{\zeta}+\gamma^{-1/2}n_{tr}^{-1/2}+\gamma^{-1}n_{tr}^{-3/4})+\frac{BR^2}{n_{te}}} \\
=&\mathcal{O}((\gamma^{\zeta}+\gamma^{-1/2}n_{tr}^{-1/2}+\gamma^{-1}n_{tr}^{-3/4})n_{tr}^{-1/2}+n_{te}^{-1/2})
\end{aligned} \tag{24}
$$

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

$$
\begin{aligned}
&\frac{1}{\lfloor \rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor \rho n_{tr}\rfloor}|(\hat{\beta}(\boldsymbol{x}_j^{tr})-\beta(\boldsymbol{x}_j^{tr}))(g(\boldsymbol{x}_j^{tr})-g_{\gamma}(\boldsymbol{x}_j^{tr}))| \\
\leq&\frac{B}{\lfloor \rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor \rho n_{tr}\rfloor}|g(\boldsymbol{x}_j^{tr})-g_{\gamma}(\boldsymbol{x}_j^{tr})| \\
\leq&\Big|\frac{B}{\lfloor \rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor \rho n_{tr}\rfloor}|g(\boldsymbol{x}_j^{tr})-g_{\gamma}(\boldsymbol{x}_j^{tr})|-B\|g-g_{\gamma}\|_{\mathscr{L}^1_{P_{tr}}}\Big|+B\|g-g_{\gamma}\|_{\mathscr{L}^1_{P_{tr}}} \\
\leq&\sqrt{\frac{1}{\delta}}\sqrt{\frac{B^2}{\rho n_{tr}}\|g-g_{\gamma}\|^2_{\mathscr{L}^2_{P_{tr}}}}+B\|g-g_{\gamma}\|_{\mathscr{L}^2_{P_{tr}}} \\
\leq&\sqrt{\frac{1}{\delta}}BC\gamma^{\zeta}\sqrt{\frac{1}{\rho n_{tr}}}+C\gamma^{\zeta}=\mathcal{O}(\gamma^{\zeta})=\mathcal{O}(\gamma^{\frac{\theta}{2\theta+4}}).
\end{aligned} \tag{25}
$$

where $\mathscr{L}^1_{P_{tr}}$ denotes the 1-norm $\mathbb{E}_{\boldsymbol{x}\sim P_{tr}}|g(\boldsymbol{x})-g_{\gamma}(\boldsymbol{x})|$. Notice the second-to-last line follows from the Chebyshev inequality, the Cauchy-Schwarz inequality, and the last line from (12).

Thus, when taking $h=g_{\gamma}$ and $\hat{g}=\hat{g}_{\gamma,\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}$ for some $\gamma>0$, we can combine (18), (20), (24) and (25) to have

$$
\begin{aligned}
|V_R(\rho)-\nu| =&\mathcal{O}(n_{tr}^{-\frac{1}{2}})+\mathcal{O}(\gamma^{\frac{\theta}{2\theta+4}})+\mathcal{O}(\gamma^{-1}n_{tr}^{-1/2}(n_{tr}^{-1}+n_{te}^{-1})^{\frac{1}{2}}) \\
&+\mathcal{O}((\gamma^{\frac{\theta}{2\theta+4}}+\gamma^{-1/2}n_{tr}^{-1/2}+\gamma^{-1}n_{tr}^{-3/4})n_{tr}^{-1/2}+n_{te}^{-1/2}) \\
=&\mathcal{O}(n_{tr}^{-\frac{1}{2}}+n_{te}^{-\frac{1}{2}}+\gamma^{\frac{\theta}{2\theta+4}}+\gamma^{-\frac{1}{2}}n_{tr}^{-1}+\gamma^{-\frac{1}{2}}n_{tr}^{-\frac{1}{2}}n_{te}^{-\frac{1}{2}}),
\end{aligned} \tag{26}
$$

after simplification. Now, if we take $\gamma=n^{-\frac{\theta+2}{\theta+1}}$ where $n\triangleq\min(n_{tr},n_{te})$, then (26) becomes

$$
\begin{aligned}
&|V_R(\rho)-\nu| \\
=&\mathcal{O}(n^{-\frac{1}{2}}+n^{-\frac{\theta}{2(\theta+1)}}+n^{\frac{\theta+2}{2(\theta+1)}}n^{-1})=\mathcal{O}(n^{-\frac{\theta}{2\theta+2}})=\mathcal{O}(n_{tr}^{-\frac{\theta}{(2\theta+2)}}+n_{te}^{-\frac{\theta}{(2\theta+2)}}),
\end{aligned} \tag{27}
$$

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

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

$$
\|g_{\gamma}\|^2_{\mathcal{H}}\leq\frac{\|g_{\gamma}-g\|^2_{\mathscr{L}^2_{P_{tr}}}+\gamma\|g_{\gamma}\|^2_{\mathcal{H}}}{\gamma}\leq\frac{\|g-g\|^2_{\mathscr{L}^2_{P_{tr}}}+\gamma\|g\|^2_{\mathcal{H}}}{\gamma}=\|g\|^2_{\mathcal{H}}, \tag{28}
$$

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

$$
\begin{aligned}
V_R(\rho) - \nu =& \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) Y_j^{tr} - \nu \\
&+ \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \hat{g}(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te})
\end{aligned} \tag{29}
$$

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

$$
\begin{aligned}
&\frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \hat{g}(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te}) \\
=& \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \big\langle \hat{g}, \Phi(\boldsymbol{x}_j^{tr}) \big\rangle_{\mathcal{H}} - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \big\langle \hat{g}, \Phi(\boldsymbol{x}_i^{n_{te}}) \big\rangle_{\mathcal{H}} \\
=& \left\langle \hat{g}, \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{i=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \Phi(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \Phi(\boldsymbol{x}_i^{te}) \right\rangle_{\mathcal{H}} \leq \|\hat{g}_{\gamma, \boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}}\|_{\mathcal{H}} \hat{L}(\hat{\boldsymbol{\beta}}),
\end{aligned} \tag{30}
$$

Then, by (29) and (30), we have

$$
\begin{aligned}
|V_R(\rho) - \nu| \leq& |V_{KMM}(\rho) - \nu| + \hat{L}(\hat{\boldsymbol{\beta}}) (\|g_\gamma - \hat{g}_{\gamma, \boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}}\|_{\mathcal{H}} + \|g_\gamma\|_{\mathcal{H}}) \\
=& \mathcal{O}(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}}),
\end{aligned} \tag{31}
$$

following (28), (15) and Theorem 1 of [Yu and Szepesvári, 2012]. □

Proof of Proposition 2. If the function g only satisfies the condition $\mathcal{A}_\infty(g, F) \triangleq \inf_{\|f\|_{\mathcal{H}} \leq F} \|g - f\| \leq C(\log F)^{-s}$ for some $C, s > 0$, then we again follow the analysis in the proof of Proposition 1 and arrive at the decomposition in (29)

$$
\begin{aligned}
|V_R(\rho) - \nu| \leq& |V_{KMM}(\rho) - \nu| + \hat{L}(\hat{\boldsymbol{\beta}}) (\|g_\gamma - \hat{g}_{\gamma, \boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}}\|_{\mathcal{H}} + \|g_\gamma\|_{\mathcal{H}}) \\
=& \mathcal{O}(\log \frac{n_{tr} n_{te}}{n_{tr} + n_{te}})^{-s},
\end{aligned} \tag{32}
$$

which is the rate of V_{KMM} by Theorem 3 of [Yu and Szepesvári, 2012]. □

Proof of Theorem 2. Define $\epsilon \triangleq \sup_{\theta \in \mathcal{D}} \left| V_R(\theta) - \mathbb{E}[l'(X^{te}, Y^{te}; \theta)] \right|$. We have

$$
\mathbb{E}[l'(X_{te}, Y_{te}; \hat{\theta}_R)] - \epsilon \leq V_R(\hat{\theta}_R) \leq V_R(\theta^\star) \leq \mathbb{E}[l'(X_{te}, Y_{te}; \theta^\star)] + \epsilon. \tag{33}
$$

On the other hand, we know by the triangle inequality that ϵ is bounded by

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

where the first term is bounded by $\mathcal{O}(n_{tr}^{-\frac{1}{2}}+n_{te}^{-\frac{1}{2}})$ following Corollary 8.9 in [Gretton et al., 2009]. Moreover, the second term is also $\mathcal{O}(n_{tr}^{-\frac{1}{2}}+n_{te}^{-\frac{1}{2}})$ as in (30) or Lemma 8.7 in [Gretton et al., 2009]. For the last term, due to the Lipschitz and compact assumption, it follows from Theorem 19.5 of [Van der Vaart, 2000] (see also Example 19.7 of [Van der Vaart, 2000]) that function class $\mathcal{G}$ is P_{te}-Donsker, which means that

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

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

$$P(\|\mathbb{G}_n\|_\infty > D') = P(\|\mathbb{G}_\infty\|_\infty > D') + o(1) \le \delta + o(1), \tag{34}$$

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

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

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

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

In other words, we also have

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

which concludes our proof. $\square$

References

[Bickel et al., 2007] Bickel, S., Brückner, M., and Scheffer, T. (2007). Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88. ACM.

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

[Blitzer et al., 2006] Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics.

[Borgwardt et al., 2006] Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57.

[Cortes et al., 2008] Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. (2008). Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer.

[Cucker and Zhou, 2007] Cucker, F. and Zhou, D. X. (2007). *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press.

[Evgeniou et al., 2000] Evgeniou, T., Pontil, M., and Poggio, T. (2000). Regularization networks and support vector machines. *Advances in computational mathematics*, 13(1):1.

[Glynn and Szechtman, 2002] Glynn, P. W. and Szechtman, R. (2002). Some new perspectives on the method of control variates. In *Monte Carlo and Quasi-Monte Carlo Methods 2000*, pages 27–49. Springer.

[Gretton et al., 2009] Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5.

[Hachiya et al., 2008] Hachiya, H., Akiyama, T., Sugiyama, M., and Peters, J. (2008). Adaptive importance sampling with automatic model selection in value function approximation. In *AAAI*, pages 1351–1356.

[Heckman, 1979] Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.

[Huang et al., 2007] Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608.

[Jiang and Zhai, 2007] Jiang, J. and Zhai, C. (2007). Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 264–271.

[Kanamori et al., 2012] Kanamori, T., Suzuki, T., and Sugiyama, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367.

[Kennedy et al., 2017] Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245.

[Lifshits, 2013] Lifshits, M. A. (2013). *Gaussian random functions*, volume 322. Springer Science & Business Media.

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

[Pan and Yang, 2009] Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

[Pardoe and Stone, 2010] Pardoe, D. and Stone, P. (2010). Boosting for regression transfer. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 863–870. Omnipress.

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

[Quionero-Candela et al., 2009] Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press.

[Schölkopf et al., 2001] Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

[Schölkopf et al., 2002] Schölkopf, B., Smola, A. J., Bach, F., et al. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[Shimodaira, 2000] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

[Smale and Zhou, 2007] Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172.

[Sugiyama and Kawanabe, 2012] Sugiyama, M. and Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

[Sugiyama et al., 2007] Sugiyama, M., Krauledat, M., and MÃžller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005.

[Sugiyama et al., 2008a] Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., and Kawanabe, M. (2008a). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440.

[Sugiyama et al., 2008b] Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. (2008b). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746.

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

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

[Tzeng et al., 2017] Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176.

[Van der Vaart, 2000] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

[Wen et al., 2014] Wen, J., Yu, C.-N., and Greiner, R. (2014). Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *ICML*, pages 631–639.

[Yao and Doretto, 2010] Yao, Y. and Doretto, G. (2010). Boosting for transfer learning with multiple sources. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1855–1862. IEEE.

[Yu and Szepesvári, 2012] Yu, Y. L. and Szepesvári, C. (2012). Analysis of kernel mean matching under covariate shift. In *ICML*, pages 1147–1154. Omnipress.

[Zadrozny, 2004] Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114. ACM.