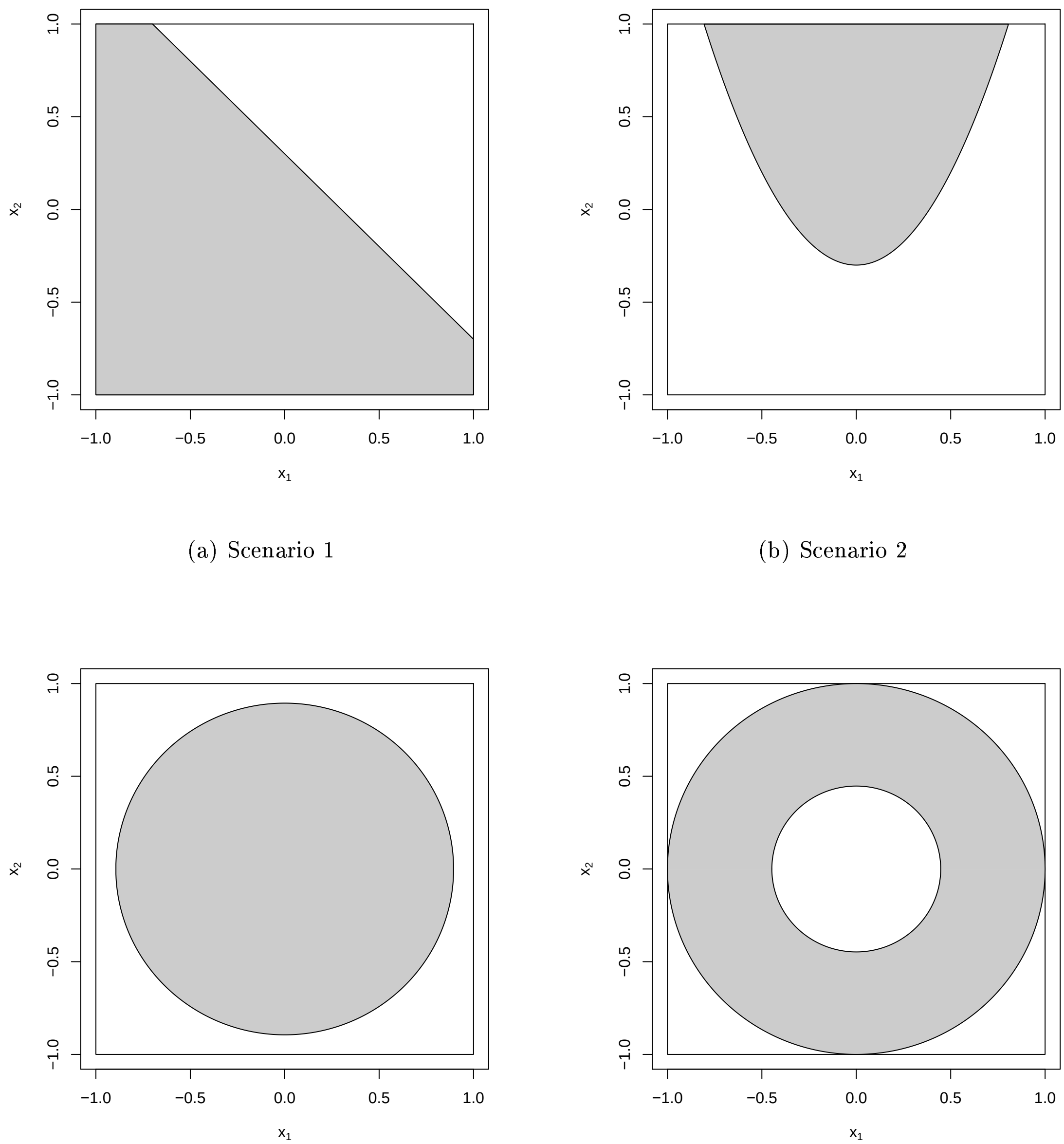
Supplement to "Residual Weighted Learning for Estimating Individualized Treatment Rules"

Xin Zhou, Nicole Mayer-Hamblett, Umer Khan and Michael R. Kosorok

June 22, 2015

*Xin Zhou is a PhD student, Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599 (email: xinzhou@live.unc.edu). Nicole Mayer-Hamblett is Associate Professor, Department of Pediatrics, and Adjunct Associate Professor, Department of Biostatistics, University of Washington (email: nicole.hamblett@seattlechildrens.org). Umer Khan is a statistician, Cystic Fibrosis Foundation Therapeutics Development Network, Seattle Childrens Hospital (email: umer.khan@seattlechildrens.org). Michael R. Kosorok is W. R. Kenan, Jr. Distiguished Professor and Chair, Department of Biostatistics, and Professor, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (email: kosorok@unc.edu).

1 Additional simulation result



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Figure 1: True optimal ITRs in simulation studies. For subjects in the shade area, the best treatment is 1; for subjects in the white area, the best treatment is -1.

Table 1: Mean (std) of treatment matching factors evaluated on the training data for 5 simulation scenarios with 5 covariates.

	$n = 100$	$n = 400$
	Scenario 0	
ℓ_1-PLS	1.00 (0.04)	1.00 (0.01)
OWL-Linear	1.03 (0.07)	1.00 (0.01)
OWL-Gaussian	1.14 (0.24)	1.02 (0.06)
RWL-Linear	1.00 (0.04)	1.00 (0.01)
RWL-Gaussian	1.00 (0.06)	1.00 (0.03)
	Scenario 1	
ℓ_1-PLS	0.99 (0.10)	0.99 (0.05)
OWL-Linear	1.10 (0.08)	1.04 (0.04)
OWL-Gaussian	1.15 (0.13)	1.05 (0.05)
RWL-Linear	0.99 (0.09)	0.99 (0.04)
RWL-Gaussian	0.99 (0.08)	0.99 (0.04)
	Scenario 2	
ℓ_1-PLS	1.00 (0.09)	1.00 (0.04)
OWL-Linear	1.06 (0.09)	1.03 (0.04)
OWL-Gaussian	1.18 (0.20)	1.10 (0.07)
RWL-Linear	1.00 (0.08)	1.00 (0.04)
RWL-Gaussian	1.01 (0.07)	1.00 (0.04)
	Scenario 3	
ℓ_1-PLS	1.00 (0.05)	1.00 (0.01)
OWL-Linear	1.02 (0.06)	1.00 (0.02)
OWL-Gaussian	1.32 (0.17)	1.14 (0.05)
RWL-Linear	1.01 (0.06)	1.01 (0.04)
RWL-Gaussian	1.04 (0.06)	1.02 (0.04)
	Scenario 4	
ℓ_1-PLS	1.00 (0.08)	1.00 (0.03)
OWL-Linear	1.07 (0.10)	1.02 (0.04)
OWL-Gaussian	1.37 (0.34)	1.27 (0.24)
RWL-Linear	1.00 (0.06)	1.00 (0.03)
RWL-Gaussian	1.00 (0.08)	1.02 (0.04)

2 Proofs

Proof of Lemma 2.1

Proof. Note that,

$$\mathbb{E}\left(\frac{\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)}{\pi(A,\boldsymbol{X})}\Big|\boldsymbol{X}\right) = \mathbb{E}\left(\mathbb{I}\big(d(\boldsymbol{X}) \neq 1\big)|\boldsymbol{X}, A=1\right) + \mathbb{E}\left(\mathbb{I}\big(d(\boldsymbol{X}) \neq -1\big)|\boldsymbol{X}, A=-1\right) = 1. \quad (1)$$

The desired result follows easily. □

Proof of Theorem 2.2

Proof. For any measurable function g,

$$\begin{aligned} Var\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\right) &= Var\left(\frac{R-\tilde{g}(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\right) + Var\left(\frac{\tilde{g}(\boldsymbol{X})-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\right) \\ &\quad +2Cov\left(\frac{R-\tilde{g}(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big), \frac{\tilde{g}(\boldsymbol{X})-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\right). \end{aligned}$$

It suffices to show that the covariance term is zero. Applying (1), we have

$$\mathbb{E}\left(\frac{R-\tilde{g}(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\Big|\boldsymbol{X}\right) = \mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\Big|\boldsymbol{X}\right) - \tilde{g}(\boldsymbol{X})\mathbb{E}\left(\frac{\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)}{\pi(A,\boldsymbol{X})}\Big|\boldsymbol{X}\right) = 0.$$

Note that

$$\tilde{g}(\boldsymbol{X}) = \mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\Big|\boldsymbol{X}\right) = \mathbb{E}(R|\boldsymbol{X}, A=1)\mathbb{I}\big(d(\boldsymbol{X}) \neq 1\big) + \mathbb{E}(R|\boldsymbol{X}, A=-1)\mathbb{I}\big(d(\boldsymbol{X}) \neq -1\big). \quad (2)$$

Thus we have,

$$\begin{aligned} \mathbb{E}\left(\frac{R-\tilde{g}(\boldsymbol{X})}{\pi(A,\boldsymbol{X})^2}\mathbb{I}\big(A \neq d(\boldsymbol{X})\big)\Big|\boldsymbol{X}\right) &= \mathbb{E}\left(R-\tilde{g}(\boldsymbol{X})|\boldsymbol{X}, A=1\right)\mathbb{I}\big(d(\boldsymbol{X}) \neq 1\big)/\pi(1,\boldsymbol{X}) \\ &\quad +\mathbb{E}\left(R-\tilde{g}(\boldsymbol{X})|\boldsymbol{X}, A=-1\right)\mathbb{I}\big(d(\boldsymbol{X}) \neq -1\big)/\pi(-1,\boldsymbol{X}) = 0. \end{aligned}$$

The desired result follows easily. □

Proof of Theorem 3.1

Proof. Given $\boldsymbol{X}=\boldsymbol{x}$, for any measurable function f, similar reasoning to that used in the proof of Lemma 2.1 yields,

$$\mathbb{E}\left(\frac{T\big(Af(\boldsymbol{X})\big)}{\pi(A,\boldsymbol{X})}\Big|\boldsymbol{X}=\boldsymbol{x}\right) = 2.$$

Then the conditional T-risk is

$$
\begin{aligned}
& \mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Af(\boldsymbol{X}))\big|\boldsymbol{X}=\boldsymbol{x}\right) \\
= \; & \mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)T(f(\boldsymbol{x}))+\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)T(-f(\boldsymbol{x}))-2g(\boldsymbol{x}) \\
= \; & \left(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)\right)T(f(\boldsymbol{x})) \\
& +2\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)-2g(\boldsymbol{x}).
\end{aligned} \tag{3}
$$

If $\mathbb{E}\left[R|\boldsymbol{X}=\boldsymbol{x},A=1\right]-\mathbb{E}\left[R|\boldsymbol{X}=\boldsymbol{x},A=-1\right]>0$, any function $f(\boldsymbol{x})\geq 1$ minimizes the conditional T-risk; similarly, if $\mathbb{E}\left[R|\boldsymbol{X}=\boldsymbol{x},A=1\right]-\mathbb{E}\left[R|\boldsymbol{X}=\boldsymbol{x},A=-1\right]<0$, any function $f(\boldsymbol{x})\leq -1$ minimizes the conditional T-risk. For either case, $sign(f^*_{T,g})=d^*$.

For the second part, by applying (3),

$$
\begin{aligned}
& \mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Ad^*(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right)-\mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Af^*_{T,g}(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right) \\
= \; & \left(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)\right)\big(T(d^*(\boldsymbol{x}))-T(f^*_{T,g}(\boldsymbol{x}))\big)=0.
\end{aligned}
$$

The desired result follows by taking expectations on both sides. □

Proof of Theorem 3.2

Proof. Given $\boldsymbol{X}=\boldsymbol{x}$. By applying (3), for any measurable function f, we have

$$
\begin{aligned}
& \mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Af(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right)-\mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Af^*_{T,g}(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right) \\
= \; & \left(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)\right)\big(T(f(\boldsymbol{x}))-T(f^*_{T,g}(\boldsymbol{x}))\big).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& \mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}(A\neq sign(f(\boldsymbol{X})))|\boldsymbol{X}=\boldsymbol{x}\right)-\mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}(A\neq d^*(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right) \\
= \; & \left(\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)-\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)\right)\left(\mathbb{I}(sign(f(\boldsymbol{x}))\neq 1)-\mathbb{I}(d^*(\boldsymbol{x})\neq 1)\right).
\end{aligned}
$$

From the proof of Theorem 3.1, when $\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)>\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)$, $f^*_{T,g}(\boldsymbol{x})\geq 1$ and $d^*(\boldsymbol{x})=1$, so $T(f^*_{T,g}(\boldsymbol{x}))=0$ and $\mathbb{I}(d^*(\boldsymbol{x})\neq 1)=0$; when $\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)<\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)$, $f^*_{T,g}(\boldsymbol{x})\leq -1$ and $d^*(\boldsymbol{x})=-1$, so $T(f^*_{T,g}(\boldsymbol{x}))=2$ and $\mathbb{I}(d^*(\boldsymbol{x})\neq 1)=1$. Note that, for any measurable function f, $1\geq T(f(\boldsymbol{x}))-\mathbb{I}(sign(f(\boldsymbol{x}))\neq 1)\geq 0$. Thus it is easy to check that when $\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=1)>\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x},A=-1)$,

$$
T(f(\boldsymbol{x}))-T(f^*_{T,g}(\boldsymbol{x}))\geq\mathbb{I}(sign(f(\boldsymbol{x}))\neq 1)-\mathbb{I}(d^*(\boldsymbol{x})\neq 1),
$$

and when $\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=1) < \mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=-1)$,

$$T(f(\boldsymbol{x})) - T(f^*_{T,g}(\boldsymbol{x})) \le \mathbb{I}(sign(f(\boldsymbol{x})) \neq 1) - \mathbb{I}(d^*(\boldsymbol{x}) \neq 1).$$

So, for either case, we have

$$\begin{aligned} & \mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}(A \neq sign(f(\boldsymbol{X})))|\boldsymbol{X}=\boldsymbol{x}\right) - \mathbb{E}\left(\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}(A \neq d^*(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right) \\ \le \; & \mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Af(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right) - \mathbb{E}\left(\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}T(Af^*_{T,g}(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}\right). \end{aligned}$$

The desired result follows by taking expectations on both sides. □

Proof of Theorem 3.3

Proof. Let $L(h,b) = (R-g(\boldsymbol{X}))T(A(h(\boldsymbol{X})+b))/\pi(A,\boldsymbol{X})$. For simplicity, we denote f_{D_n,λ_n}, h_{D_n,λ_n} and b_{D_n,λ_n} by f_n, h_n and b_n, respectively. By the definition of h_{D_n,λ_n} and b_{D_n,λ_n}, we have, for any $h \in \mathcal{H}_K$ and $b \in \mathbb{R}$,

$$\mathbb{P}_n(L(h_n,b_n)) \le \mathbb{P}_n(L(h_n,b_n)) + \frac{\lambda_n}{2}||h_n||_K^2 \le \mathbb{P}_n(L(h,b)) + \frac{\lambda_n}{2}||h||_K^2,$$

where $\mathbb{P}_n$ denotes the empirical measure of the observed data. Then, $\limsup_n \mathbb{P}_n(L(h_n,b_n)) \le \mathbb{P}(L(h,b)) = \mathcal{R}_{T,g}(h+b)$ with probability 1. This implies $\limsup_n \mathbb{P}_n(L(h_n,b_n)) \le \inf_{h\in\mathcal{H}_K, b\in\mathbb{R}} \mathcal{R}_{T,g}(h+b) \le \mathbb{P}(L(h_n,b_n))$ with probability 1. It suffices to show $\mathbb{P}_n(L(h_n,b_n)) - \mathbb{P}(L(h_n,b_n)) \to 0$ in probability.

We first obtain a bound for $||h_n||_K$. Since $\mathbb{P}_n(L(h_n,b_n)) + \lambda_n||h_n||_K^2/2 \le \mathbb{P}_n(L(h,b)) + \lambda_n||h||_K^2/2$, for any $h \in \mathcal{H}_K$ and $b \in \mathbb{R}$, we can choose $h=0$ and $b=0$ to obtain, $\mathbb{P}_n(L(h_n,b_n)) + \lambda_n||h_n||_K^2/2 \le \mathbb{P}_n((R-g(\boldsymbol{X}))/\pi(A,\boldsymbol{X}))$. Note that $0 \le T(u) \le 2$. We thus have,

$$\lambda_n||h_n||_K^2 \le 2\mathbb{P}_n(|R-g(\boldsymbol{X})|/\pi(A,\boldsymbol{X})) \le 2M_0.$$

Let $M_1 = \sqrt{2M_0}$. Then the $\mathcal{H}_K$ norm of $\sqrt{\lambda_n}h_n$ is bounded by M_1.

Next we obtain a bound for b_n. We claim that there is a global solution (h_n, b_n) such that $h_n(\boldsymbol{x}_i)+b_n \in [-1,1]$ for some i. Suppose there is a global solution (h'_n, b'_n) such that $|h'_n(\boldsymbol{x}_i)+b'_n| > 1$ for all i. Let $\delta = |h'_n(\boldsymbol{x}_{i_0}) + b'_n| = \min_{1\le i\le n}|h'_n(\boldsymbol{x}_i)+b'_n| > 1$. Then let $h_n = h'_n$ and $b_n = b'_n - (\delta-1)sign(h'_n(\boldsymbol{x}_{i_0}) + b'_n)$. It is easy to check that $h_n(\boldsymbol{x}_{i_0}) + b_n = 1$ if $h'_n(\boldsymbol{x}_{i_0}) + b'_n > 1$, and $h_n(\boldsymbol{x}_{i_0})+b_n = -1$ if $h'_n(\boldsymbol{x}_{i_0})+b'_n < -1$; furthermore when $i \neq i_0$, $h_n(\boldsymbol{x}_i)+b_n \ge 1$ if $h'_n(\boldsymbol{x}_i)+b'_n > 1$, and $h_n(\boldsymbol{x}_i) + b_n \le -1$ if $h'_n(\boldsymbol{x}_i) + b'_n < -1$. So $T(h_n(\boldsymbol{x}_i) + b_n) = T(h'_n(\boldsymbol{x}_i) + b'_n)$ for all i. Hence

(h_n, b_n) is a global solution and satisfies our claim. Now if a solution (h_n, b_n) satisfies our claim, we then have,

$$|b_n| \leq 1 + |h_n(\boldsymbol{x}_{i_0})| \leq 1 + ||h_n||_\infty.$$

Note that $||h||_\infty \leq C_K ||h||_K$. We have,

$$|\sqrt{\lambda_n} b_n| \leq \sqrt{\lambda_n} + C_K \sqrt{\lambda_n} ||h_n||_K.$$

Since $\lambda_n \to 0$, and C_K and $\sqrt{\lambda_n}||h_n||_K$ are both bounded, we have $|\sqrt{\lambda_n} b_n|$ is bounded too. Let the bound be M_2, *i.e.* $|\sqrt{\lambda_n} b_n| \leq M_2$.

Note that the class $\{\sqrt{\lambda_n} h : ||\sqrt{\lambda_n} h||_K \leq M_1\}$ is a Donsker class. So $\{\sqrt{\lambda_n}(h+b) : ||\sqrt{\lambda_n} h||_K \leq M_1, |\sqrt{\lambda_n} b| \leq M_2\}$ is also P-Donsker. Consider the function

$$T_\lambda(u) = \begin{cases} 2\sqrt{\lambda} & \text{if} \quad u < -\sqrt{\lambda}, \\ 2\sqrt{\lambda} - \frac{1}{\sqrt{\lambda}}(\sqrt{\lambda} + u)^2 & \text{if} \quad -\sqrt{\lambda} \leq u < 0, \\ \frac{1}{\sqrt{\lambda}}(\sqrt{\lambda} - u)^2 & \text{if} \quad 0 \leq u < \sqrt{\lambda}, \\ 0 & \text{if} \quad u \geq \sqrt{\lambda}. \end{cases}$$

We have $T_\lambda(\sqrt{\lambda} u) = \sqrt{\lambda} T(u)$. Since $T_\lambda(u)$ is a Lipschitz continuous function with Lipschitz constant equal to 2, and $\frac{R-g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}$ is bounded, the class $\{\sqrt{\lambda_n} L(h, b) : ||\sqrt{\lambda_n} h||_K \leq M_1, |\sqrt{\lambda_n} b| \leq M_2\}$ is also P-Donsker. Therefore,

$$\sqrt{n\lambda_n}(\mathbb{P}_n - \mathbb{P}) L(h_n, b_n) = O_p(1).$$

Consequently, from $n\lambda_n \to \infty$, $\mathbb{P}_n(L(h_n, b_n)) - \mathbb{P}(L(h_n, b_n)) \to 0$ in probability. $\square$

Proof of Lemma 3.4

Proof. Fix any $0 < \epsilon < 1$. $d^*(\boldsymbol{x}) = sign(\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) - \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1))$ is measurable. Since μ is regular, using Lusin's theorem in measure theory, we know that $d^*(\boldsymbol{x})$ can be approximated by a continuous function $f'(\boldsymbol{x}) \in C(\mathcal{X})$ such that $\mu(f'(\boldsymbol{x}) \neq d^*(\boldsymbol{x})) \leq \frac{\epsilon}{4M}$. Thus

$$\begin{aligned} & \mathbb{E}\left(\frac{R - g(\boldsymbol{X})}{\pi(A, \boldsymbol{X})} T(Af'(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right) - \mathbb{E}\left(\frac{R - g(\boldsymbol{X})}{\pi(A, \boldsymbol{X})} T(Ad^*(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right) \\ = \; & (\mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) - \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1)) \left(T(f'(\boldsymbol{x})) - T(d^*(\boldsymbol{x}))\right). \end{aligned}$$

Then,

$$
\begin{aligned}
&\mathcal{R}_{T,g}(f') - \mathcal{R}^*_{T,g} = |\mathcal{R}_{T,g}(f') - \mathcal{R}_{T,g}(d^*)| \\
= \;& \Big| \int (\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=1) - \mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=-1)) \big(T(f'(\boldsymbol{x})) - T(d^*(\boldsymbol{x}))\big)\mu(d\boldsymbol{x}) \Big| \\
\leq \;& \int \Big| (\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=1) - \mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=-1)) \Big| \Big| T(f'(\boldsymbol{x})) - T(d^*(\boldsymbol{x})) \Big| \mathbb{I}(f'(\boldsymbol{x}) \neq d^*(\boldsymbol{x}))\mu(d\boldsymbol{x}).
\end{aligned}
$$

Since $|R| \leq M$ and $0 \leq T(u) \leq 2$,

$$
\mathcal{R}_{T,g}(f') - \mathcal{R}^*_{T,g} < \epsilon
$$

Since K is universal, there exist a function $f'' \in \mathcal{H}_K$ such that $||f'' - f'||_\infty < \frac{\epsilon}{4M}$. Note that $T(\cdot)$ is Lipschitz continuous with Lipschitz constant 2. Similarly,

$$
\begin{aligned}
&|\mathcal{R}_{T,g}(f'') - \mathcal{R}_{T,g}(f')| \\
= \;& \Big| \int (\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=1) - \mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=-1)) \big(T(f''(\boldsymbol{x})) - T(f'(\boldsymbol{x}))\big)\mu(d\boldsymbol{x}) \Big| \\
\leq \;& 2\int \Big| (\mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=1) - \mathbb{E}(R|\boldsymbol{X}=\boldsymbol{x}, A=-1)) \Big| \Big| f'(\boldsymbol{x}) - f'(\boldsymbol{x}) \Big| \mu(d\boldsymbol{x}) < \epsilon.
\end{aligned}
$$

By combining the two inequalities, we have

$$
\mathcal{R}_{T,g}(f'') - \mathcal{R}^*_{T,g} < 2\epsilon.
$$

Noting that $f'' \in \mathcal{H}_K$ and letting $\epsilon \to 0$, we obtain the desired result. □

Proof of Theorem 3.6

Proof. The proof follows the idea in Devroye et al. (1996, Theorem 7.2). Since the proof is very similar, we only provide a sketch to save space.

Let $b = 0.b_1b_2b_3\cdots$ be a real number on $[0,1]$ with the given binary expansion, and let B be a random variable uniformly distributed on $[0,1]$ with expansion $B = 0.B_1B_2B_3\cdots$. Let us restrict ourselves to a random variable $\boldsymbol{X}$ with the support $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots\}$ where $\boldsymbol{x}_i \in \mathcal{X}$. For simplicity, we recode the support of $\boldsymbol{X}$ as $\{1, 2, \cdots\}$. Let

$$
P(\boldsymbol{X} = i) = p_i, \quad i \geq 1, \tag{4}
$$

where $p_1 \geq p_2 \geq \cdots > 0$, and $\sum_{i=n+1}^{\infty} p_i \geq \max(8c_n, 32np_{n+1})$ for every n. Such p_i's exist by Devroye et al. (1996, Lemma 7.1). Let $A \in \{1, -1\}$ be a binomial variable with $\pi(A, \boldsymbol{X}) = 0.5$. For a given b, set $R = AM$ if $b_{\boldsymbol{X}} = 1$, and $R = -AM$ if $b_{\boldsymbol{X}} = 0$. Then the Bayes rule is $d^*(\boldsymbol{X}) = (2 * b_{\boldsymbol{X}} - 1)$. Thus each $b \in [0,1]$ describes a different distribution of $(\boldsymbol{X}, A, R)$. Introduce the shortened notation $D_n = \{(\boldsymbol{X}_1, A_1, R_1), \cdots, (\boldsymbol{X}_n, A_n, R_n)\}$. Let d_n be a rule generated by data D_n. Define $d_n^i = d_n(i)$ for $i = 1, \cdots, n$. Let $\Delta\mathcal{R}_n(b)$ be the excess risk of the rule d_n for the distribution parametrized by b, and $\Delta\mathcal{R}_n(B)$ be the excess risk of the rule d_n for the random distribution.

$$
\begin{aligned}
\Delta\mathcal{R}_n(B) &= \mathbb{E}\left[\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}(A \neq d_n(\boldsymbol{X}))\Big|B\right] - \mathbb{E}\left[\frac{R}{\pi(A,\boldsymbol{X})}\mathbb{I}(A \neq d^*(\boldsymbol{X}))\Big|B\right] \\
&= \mathbb{E}\Big[\big(\mathbb{E}(R|B,\boldsymbol{X},A=1) - \mathbb{E}(R|B,\boldsymbol{X},A=-1)\big)\big(\mathbb{I}(d_n(\boldsymbol{X}) \neq 1) - \mathbb{I}(d^*(\boldsymbol{X}) \neq 1)\big)|B\Big] \\
&= 2M\mathbb{E}\big(\mathbb{I}(d_n(\boldsymbol{X}) \neq d^*(\boldsymbol{X}))\big|B\big) \\
&= 2M\mathbb{E}\big(\mathbb{I}(d_n(\boldsymbol{X}) \neq 2B_{\boldsymbol{X}} - 1)\big).
\end{aligned}
$$

Let $L_n(B) = \mathbb{E}\big(\mathbb{I}(d_n(\boldsymbol{X}) \neq 2B_{\boldsymbol{X}} - 1)\big)$. Then we have,

$$L_n(B) = \sum_{i=1}^{\infty} p_i \mathbb{I}(d_n^i \neq 2B_i - 1).$$

Following the same arguments used in Devroye et al. (1996, Theorem 7.2), we have

$$P(L_n(B) < 2c_n|D_n) \leq P(\sum_{i=n+1}^{\infty} p_i B_i < 2c_n) \leq e^{-2n}.$$

Hence we have

$$
\begin{aligned}
\sup_b \inf_n \mathbb{E}\left(\frac{L_n(b)}{2c_n}\right) &\geq \mathbb{E}\left(\mathbb{E}\left(\inf_n\left(\frac{L_n(b)}{2c_n}\Big|\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots\right)\right)\right) \\
&\geq \mathbb{E}\left(1 - \sum_{i=1}^{\infty}\mathbb{E}(P(L_n(B) < 2c_n|D_n)|\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots)\right) \\
&\geq 1 - \sum_{i=1}^{\infty} e^{-2n} = \frac{e^2-2}{e^2-1} > \frac{1}{2}.
\end{aligned}
$$

Here we are omitting many steps. Refer to Devroye et al. (1996, Theorem 7.2) for details. The conclusion is that there exists a b for which $\Delta\mathcal{R}_n(b) \geq 2Mc_n$, $n = 1, 2, \cdots$. □

Proof of Theorem 3.7

Proof. Define the random variable $S = \frac{R - g(\boldsymbol{X})}{\pi(A,\boldsymbol{X})}$. We consider a probability measure on the triplet $(\boldsymbol{X}, A, S)$ instead of on $(\boldsymbol{X}, A, R)$. Let $D_n = \{\boldsymbol{X}_i, A_i, S_i\}_{i=1}^n$ be independent random variables with the same distribution as $(\boldsymbol{X}, A, S)$. Let $\mathbb{P}_n$ be the empirical measure on D_n. For simplicity, we denote f_{D_n,λ_n}, h_{D_n,λ_n} and b_{D_n,λ_n} by f_n, h_n and b_n, respectively. Let $(\tilde{h}_{\lambda_n}, \tilde{b}_{\lambda_n})$ be a solution of the following optimization problem:

$$\min_{h \in \mathcal{H}_K, b \in \mathbb{R}} \frac{\lambda_n}{2}||h||_K^2 + \mathcal{R}_{T,g}(h+b).$$

Let $L(h,b) = ST(A(h(\boldsymbol{X})+b))$. Then

$$
\begin{aligned}
& \mathcal{R}_{T,g}(f_n) - \mathcal{R}_{T,g}(f^*_{T,g}) \\
\leq\ & (\mathbb{P}-\mathbb{P}_n)L(h_n,b_n) + (\frac{\lambda_n}{2}||h_n||^2 + \mathbb{P}_n L(h_n,b_n)) - (\frac{\lambda_n}{2}||\tilde{h}_{\lambda_n}||^2 + \mathbb{P}_n L(\tilde{h}_{\lambda_n},\tilde{b}_{\lambda_n})) \\
& +(\mathbb{P}_n-\mathbb{P})L(\tilde{h}_{\lambda_n},\tilde{b}_{\lambda_n}) + \mathcal{A}(\lambda_n) \\
\leq\ & (\mathbb{P}-\mathbb{P}_n)L(h_n,b_n) + (\mathbb{P}_n-\mathbb{P})L(\tilde{h}_{\lambda_n},\tilde{b}_{\lambda_n})) + \mathcal{A}(\lambda_n).
\end{aligned}
$$

We first estimate the second term by the Hoeffding inequality (Steinwart and Christmann 2008, Theorem 6.10). Since $|L(h,b)| \leq 2M_0$, we thus have, with probability at least $1-\delta/2$,

$$
(\mathbb{P}_n-\mathbb{P})L(\tilde{h}_{\lambda_n},\tilde{b}_{\lambda_n}) \leq M_0\sqrt{\frac{2\log\frac{2}{\delta}}{n}}. \tag{5}
$$

By the arguments used in the proof of Theorem 3.3, we have $||h_n||_K \leq \sqrt{\frac{2M_0}{\lambda_n}}$ and $|b_n| \leq 1 + C_K\sqrt{\frac{2M_0}{\lambda_n}}$. Then let $\mathcal{F} = \{(h,b) \in \mathcal{H}_K \times \mathbb{R} : ||h||_K \leq \sqrt{\frac{2M_0}{\lambda_n}}, |b| \leq 1 + C_K\sqrt{\frac{2M_0}{\lambda_n}}\}$. Let $\tilde{L}(h,b) = S\Big[T(A(h(\boldsymbol{X})+b)) - 1\Big]$. For the first term, $(\mathbb{P}-\mathbb{P}_n)L(h_n,b_n)$,

$$
\begin{aligned}
(\mathbb{P}-\mathbb{P}_n)L(h_n,b_n) &\leq \sup_{(h,b)\in\mathcal{F}} (\mathbb{P}-\mathbb{P}_n)L(h,b) \\
&= \sup_{(h,b)\in\mathcal{F}} (\mathbb{P}-\mathbb{P}_n)\tilde{L}(h,b) + (\mathbb{P}-\mathbb{P}_n)L(0,0).
\end{aligned}
$$

When an $(\boldsymbol{x}_i, a_i, s_i)$ triplet changes, the random variable $\sup_{(h,b)\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\tilde{L}(h,b)$ can change by no more than $\frac{2M_0}{n}$. McDiarmid's inequality (Bartlett and Mendelson 2002, Theorem 9) then implies that with probability at least $1-\delta/4$,

$$
\sup_{(h,b)\in\mathcal{F}} (\mathbb{P}-\mathbb{P}_n)\tilde{L}(h,b) \leq \mathbb{E}\sup_{(h,b)\in\mathcal{F}} (\mathbb{P}-\mathbb{P}_n)\tilde{L}(h,b) + M_0\sqrt{\frac{2\log(4/\delta)}{n}}.
$$

A similar argument, together with the fact that $\mathbb{E}\mathbb{P}_n L(0,0) = \mathbb{P}L(0,0)$, shows that with probability at least $1-\delta/2$,

$$
(\mathbb{P}-\mathbb{P}_n)L(h_n,b_n) \leq \mathbb{E}\sup_{(h,b)\in\mathcal{F}} (\mathbb{P}-\mathbb{P}_n)\tilde{L}(h,b) + 2M_0\sqrt{\frac{2\log(4/\delta)}{n}}.
$$

Let $D'_n = \{\boldsymbol{X}'_i, A'_i, S'_i\}_{i=1}^n$ be an independent sample with the same distribution as $(\boldsymbol{X}, A, S)$. Let $\mathbb{P}'_n$ denote the empirical measure on D'_n. Let σ be a uniform $\{\pm 1\}$-valued random variable, and

$\sigma_1, \ldots, \sigma_n$ be n independent copies of σ. Then we have

$$\begin{aligned}
\mathbb{E} \sup_{(h,b)\in\mathcal{F}} (\mathbb{P}-\mathbb{P}_n)\tilde{L}(h,b) &= \mathbb{E} \sup_{(h,b)\in\mathcal{F}} \mathbb{E}\Big(\mathbb{P}'_n\tilde{L}(h,b) - \mathbb{P}_n\tilde{L}(h,b)\Big|D_n\Big) \\
&\leq 2\mathbb{E} \sup_{(h,b)\in\mathcal{F}} \mathbb{P}_n\sigma\tilde{L}(h,b) \\
&\leq 2\mathbb{E}\mathbb{E}\Big(\sup_{(h,b)\in\mathcal{F}} |\mathbb{P}_n\sigma\tilde{L}(h,b)|\Big|S_i, A_i, i=1,\ldots,n\Big) \\
&\leq \frac{16M_0}{n}\mathbb{E} \sup_{(h,b)\in\mathcal{F}} \Big|\sum_{i=1}^n \sigma_i\big(h(\boldsymbol{X}_i)+b\big)\Big|.
\end{aligned}$$

The last inequality is due to the contraction inequality (Ledoux and Talagrand 1991, Corollary 3.17). The preceding can be further majorized by using Lemma 22 in Bartlett and Mendelson (2002),

$$\begin{aligned}
\mathbb{E} \sup_{(h,b)\in\mathcal{F}} (\mathbb{P}-\mathbb{P}_n)\tilde{L}(h,b) &\leq \frac{16M_0}{n}\mathbb{E} \sup_{(h,b)\in\mathcal{F}} \Big|\sum_{i=1}^n \sigma_i h(\boldsymbol{X}_i)\Big| + \frac{16M_0}{n}(1+C_K\sqrt{\frac{2M_0}{\lambda_n}})\mathbb{E}\Big|\sum_{i=1}^n \sigma_i\Big| \\
&\leq \frac{16M_0}{\sqrt{n}}\sqrt{\frac{2M_0}{\lambda_n}}C_K + \frac{16M_0}{\sqrt{n}}(1+C_K\sqrt{\frac{2M_0}{\lambda_n}}) \\
&= \frac{16M_0}{\sqrt{n}}(1+2C_K\sqrt{\frac{2M_0}{\lambda_n}}).
\end{aligned}$$

Then we have that with probability at least $1-\delta/2$,

$$(\mathbb{P}-\mathbb{P}_n)L(h_n,b_n) \leq \frac{16M_0}{\sqrt{n}}(1+2C_K\sqrt{\frac{2M_0}{\lambda_n}}) + 2M_0\sqrt{\frac{2\log(4/\delta)}{n}}. \tag{6}$$

By the assumption, (5), and (6), we obtain that with probability at least $1-\delta$,

$$\mathcal{R}_{T,g}(f_n) - \mathcal{R}^*_{T,g} \leq M_0\sqrt{\frac{2\log(2/\delta)}{n}} + \frac{16M_0}{\sqrt{n}}(1+2C_K\sqrt{\frac{2M_0}{\lambda_n}}) + 2M_0\sqrt{\frac{2\log(4/\delta)}{n}} + c\lambda_n^{\beta}.$$

Let $\lambda_n = n^{-\frac{1}{2\beta+1}}$. By Theorem 3.2, we obtain the final result that with probability at least $1-\delta$,

$$\mathcal{R}(sign(f_n)) - \mathcal{R}^* \leq \tilde{c}\sqrt{\log(4/\delta)}n^{-\frac{\beta}{2\beta+1}}.$$

Here $\tilde{c} = M_0\left(16+3\sqrt{2}+32C_K\sqrt{2M_0}\right)+c$. This completes the proof. □

Proof of Lemma 3.9

Proof. We first introduce a lemma. It is revised from Lemma 4.1 in Steinwart and Scovel (2007) to adapt to our settings for individualized treatment rules.

Lemma 2.1. *Let $\mathcal{X}$ be the closed unit ball of the Euclidean space $\mathbb{R}^p$, and P be a distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{M}$ with regular marginal distribution on $\boldsymbol{X}$. Recall $\delta(\boldsymbol{x}) = \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = 1) - \mathbb{E}(R|\boldsymbol{X} = \boldsymbol{x}, A = -1)$ for $\boldsymbol{x} \in \mathcal{X}$. On $\acute{\mathcal{X}} := 3\mathcal{X}$ we define*

$$\acute{\delta}(\boldsymbol{x}) = \begin{cases} \delta(\boldsymbol{x}) & \text{if } ||\boldsymbol{x}|| \leq 1, \\ \delta\left(\dfrac{\boldsymbol{x}}{||\boldsymbol{x}||}\right) & \text{otherwise,} \end{cases}$$

where $||\cdot||$ is the Euclidean norm. We also write $\acute{\mathcal{X}}^+ = \{\boldsymbol{x} \in \acute{\mathcal{X}} : \acute{\delta}(\boldsymbol{x}) > 0\}$, and $\acute{\mathcal{X}}^- = \{\boldsymbol{x} \in \acute{\mathcal{X}} : \acute{\delta}(\boldsymbol{x}) < 0\}$. Finally let $B(\boldsymbol{x}, r)$ denote the open ball of radius r about $\boldsymbol{x}$ in $\mathbb{R}^p$. Then for $\boldsymbol{x} \in \mathcal{X}^+$, we have $B(\boldsymbol{x}, \tau_{\boldsymbol{x}}) \subset \acute{\mathcal{X}}^+$, and for $\boldsymbol{x} \in \mathcal{X}^-$, we have $B(\boldsymbol{x}, \tau_{\boldsymbol{x}}) \subset \acute{\mathcal{X}}^-$.

The proof is simple, and is the same as that of Lemma 4.1 in Steinwart and Scovel (2007). We omit the proof here. In the lemma, the support is enlarged to ensure that all balls of the form $B(\boldsymbol{x}, \tau_{\boldsymbol{x}})$ are contained in the enlarged support. We return to the proof of Lemma 3.9.

Let $L_2(\mathbb{R}^p)$ be the L_2-space on $\mathbb{R}^p$ with respect to Lebesque measure, and $\mathcal{H}_\sigma(\mathbb{R}^p)$ be the RKHS of the Gaussian RBF kernel K_σ. The linear operator $V_\sigma : L_2(\mathbb{R}^p) \to \mathcal{H}_\sigma(\mathbb{R}^p)$ defined by

$$V_\sigma \ell(\boldsymbol{x}) = \frac{(2\sigma)^{d/2}}{\pi^{d/4}} \int_{\mathbb{R}^p} e^{-2\sigma^2||\boldsymbol{x}-\boldsymbol{y}||^2} \ell(\boldsymbol{y}) d\boldsymbol{y}, \qquad \ell \in L_2(\mathbb{R}^p),\ \boldsymbol{x} \in \mathbb{R}^p,$$

is an isometric isomorphism (Steinwart et al. 2006). Thus we have,

$$\mathcal{A}(\lambda) \leq \inf_{\ell \in L_2(\mathbb{R}^p)} \frac{\lambda}{2} ||\ell||^2_{L_2(\mathbb{R}^p)} + \mathcal{R}_{T,g}(V_\sigma \ell) - \mathcal{R}^*_{T,g}. \tag{7}$$

With the notation of Lemma 2.1 we fix a measurable $\acute{f}_P : \acute{\mathcal{X}} \to [-1, 1]$ that satisfies $\acute{f}_P = 1$ on $\acute{\mathcal{X}}^+$, $\acute{f}_P = -1$ on $\acute{\mathcal{X}}^-$, and $\acute{f}_P = 0$ otherwise. For $\ell := (\sigma^2/\pi)^{p/4} \acute{f}_P$, we immediately obtain,

$$||\ell||_{L_2(\mathbb{R}^p)} \leq \left(\frac{81\sigma^2}{\pi}\right)^{p/4} \theta(p), \tag{8}$$

where $\theta(p)$ denotes the volume of $\mathcal{X}$. As shown in the proof of Theorem 3.2, we have

$$\mathcal{R}_{T,g}(V_\sigma \ell) - \mathcal{R}^*_{T,g} = \mathbb{E}(|\delta(\boldsymbol{x})| \cdot |T(V_\sigma \ell(\boldsymbol{x})) - T(d^*(\boldsymbol{x}))|) \leq 2\mathbb{E}(|\delta(\boldsymbol{x})| \cdot |V_\sigma \ell(\boldsymbol{x}) - d^*(\boldsymbol{x})|).$$

Following the same derivations as in the proof of Theorem 2.7 of Steinwart and Scovel (2007), we also obtain

$$|V_\sigma \ell(\boldsymbol{x}) - d^*(\boldsymbol{x})| \leq 8e^{-\sigma^2 \tau^2_{\boldsymbol{x}}/(2p)}.$$

The geometric noise assumption yields

$$\mathcal{R}_{T,g}(V_\sigma \ell) - \mathcal{R}^*_{T,g} \leq 16\mathbb{E}(|\delta(\boldsymbol{x})| e^{-\sigma^2 \tau^2_{\boldsymbol{x}}/(2p)}) \leq 16C(2p)^{qp/2} \sigma^{-qp}. \tag{9}$$

Combining (7), (8) and (9) yields

$$\mathcal{A}(\lambda) \leq \left(\frac{81\sigma^2}{\pi}\right)^{p/2} \theta^2(p)\lambda/2 + 16C(2p)^{qp/2}\sigma^{-qp}.$$

The desired result now follows by taking $\sigma = \lambda^{-\frac{1}{(q+1)p}}$. $\square$

References

Bartlett, P. L. and Mendelson, S. "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results." *Journal of Machine Learning Research*, 3:463–482 (2002).

Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer (1996).

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*. Springer (1991).

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer (2008).

Steinwart, I., Hush, D. R., and Scovel, C. "An Explicit Description of the Reproducing Kernel Hilbert Spaces of Gaussian RBF Kernels." *IEEE Transactions on Information Theory*, 52(10):4635–4643 (2006).

Steinwart, I. and Scovel, C. "Fast rates for support vector machines using Gaussian kernels." *The Annals of Statistics*, 35(2):575–607 (2007).