Supplement to Estimation of Bounds on Potential Outcomes For Decision Making

Maggie Makar [1] Fredrik Johansson [2] John Guttag [1] David Sontag [1]

1. Additional definitions

The following definitions will be useful to prove our main statements.

Definition A1. *[Restated from Shawe-Taylor & Williamson (1999)] We say that a function class $\mathcal{F}$ is sturdy if it maps X of size n to a compact subset of $\mathbb{R}^n$ for any $n \in \mathbb{N}$.*

Definition A2. *Let (X, l_∞) be a pseudo-metric space defined with respect to the l_∞ norm, and let A be a subset of X and $\epsilon > 0$. A set $U \subseteq X$ is an ϵ-cover for A if for every $a \in A$, there exists $u \in U$ such that $||a - u||_{l_\infty} \leq \epsilon$. The ϵ-covering number of A, $\mathcal{N}(\epsilon, A, d)$ is the minimal cardinality of the ϵ-cover for A.*

Definition A3. *[Restated from (Bartlett & Shawe-Taylor, 1999)] For $\gamma \in [0, \infty]$, and $\mathcal{F} \in \mathbb{R}$, we say that a set of points $\{x_i\}_{i=1}^n$ is $\gamma-$shattered by $\mathcal{F}$ if there exists $\{s_i\}_{i=1}^n \in \mathbb{R}$ such that for all binary vectors $\{\sigma_i\}_{i=1}^n$, there is a function $f \in \mathcal{F}$ satisfying:*

$$\begin{aligned} f(x_i) &\geq s_i + \gamma \qquad && if\ \sigma_i = 1 \\ f(x_i) &\leq s_i - \gamma \qquad && otherwise \end{aligned}$$

The fat-shattering dimension can be thought of as a function from the positive reals to the set of positive integers which maps γ to the largest $\gamma-$shattered set or ∞.

We define the empirical proportion overestimated as:

Definition A4. *For $f \in \mathcal{F}$, $\gamma > 0$, a sample $z = \{x_i, y_i\}_i^n$ drawn from a fixed but unknown distribution p_t, known weights $\boldsymbol{w}$, we define the empirical risk when the distribution with respect to p:*

$$\underline{\epsilon}_f^{\boldsymbol{w}}(z, \gamma) = \sum_i w(x) \mathbb{1}\{\underline{r}_f(x, y) < \gamma\}.$$

2. Proof of theorem 1

To construct the proof, we will first study the overestimation risk when there are no training set violations (Lemma A3). To extend our results to cases where there are training set violations, we rely on a technique, presented in (Shawe-Taylor & Cristianini, 2002) and used in (Schölkopf et al., 2001), which allows us to ignore small violations in the training data at the cost of a more complex function space. This function space (formally defined in definition A5) is constructed by creating an "auxiliary function" that picks specific points to have a non-zero violation. Its complexity depends on the allowable violations. By augmenting the result from lemma A3 with the auxiliary function space, we get theorem A1, a general version of theorem 1, which gives a bound on the overestimation risk for general sturdy function spaces. Finally, we give the proof for linear function spaces, which is presented in theorem 1 in the main text.

To build up to lemma A3, we restate the following two previously established results.

Lemma A1. *Due to Shawe-Taylor & Williamson (1999): Let $\mathcal{F}$ be a sturdy function class, then for each $N \in \mathbb{N}^+$ and any fixed sequence $X \in \mathcal{X}^n$ the infimum*

$$\inf\{\gamma : \mathcal{N}(\gamma, \mathcal{F}, X) < N\}$$

is attained

We assume that f_l^1, f_l^0, f_l^0 and f_u^0 belong to a sturdy function class, as defined in definition A1.

The following lemma due to Cortes et al. (2010) bounds the second moment of the weighted loss.

Lemma A2. *Due to Cortes et al. (2010). For $x \in \mathcal{X}$, a weighting function w_t on $\mathcal{X}$, a loss function ℓ, and some function $f \in \mathcal{F}$, the second moment of the importance weighted loss can be bounded as follows:*

$$\mathbb{E}_{X|T}\left[w_t^2(X)\ell_f^2(X) \mid T = t\right] \leq d_2(p||p_t)$$

We now study the overestimation error when there are no training set violations, i.e., when $D = 0$. A direct analogy can be drawn between the following lemma (lemma A3) and hard margin one-class SVMs studied in Schölkopf et al. (2001), whereas theorem 1 is analogous to the soft margin case.

Lemma A3. *Let $\mathcal{F}$ be the class of linear functions in a kernel defined feature space, $z = \{x_i, y_i\}_{i:t_i=t}$, where $x_i, y_i \sim p_t(X, Y)$, and C_t be as defined in (1). For $f_l^t \in \mathcal{F}$, and any $\gamma > 0$, let the associated $\underline{D}^{\boldsymbol{w}_t}(z, f_t^1, \gamma) = 0$. With a probability $1 - \delta$ over the draw of random samples, we* *have that:*

$$\underline{R}_{f_t^l}(\gamma) \leq \frac{4C_t(k_t + \log\frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log\frac{1}{\delta})}{n_t}}. \tag{8}$$

where, for $t \in \{0, 1\}$,

$$k_t = \left\lceil \log \mathcal{N}(\gamma, \mathcal{F}, 2n_t) \right\rceil .$$

Proof. For a given $f_l^1 \in \mathcal{F}$:

$$\begin{aligned} P\Big(\underline{R}_{f_l^1}(\gamma) - \underline{\epsilon}_{f_l^1}^{\boldsymbol{w}}(z, \gamma) > \varepsilon\Big) &= P\Big(\underline{R}_{f_l^1}(\gamma) > \varepsilon\Big) \\ &\leq 2P\Big(\underline{\epsilon}_{f_l^1}^{\boldsymbol{w}'}(z', \gamma) > \frac{\varepsilon}{2}\Big), \end{aligned}$$

where the equality follows from the fact that the empirical error on the estimation data will always be 0 by definition of γ. And the inequality follows from applying the double (ghost) sample trick. Suppose that such an f_l^1 exists. Pick a fixed k such that

$$\gamma_k = \inf\{\gamma : \mathcal{N}(\gamma, \mathcal{F}, 2n_1) \leq 2^k\} \leq \gamma\ .$$

By Lemma A1, and assumption of sturdiness, we have that this γ_k exists. Consider the γ_k-covering, U. There exists another $f_\bullet \in U$ such that the distance between f_l^1 and $f_\bullet$ is $\leq \gamma_k \leq \gamma$, meaning $f_\bullet$ satisfies:

$$P\Big(\underline{\epsilon}_{f_l^1}^{\boldsymbol{w}'}(z', \gamma) > \frac{\varepsilon}{2}\Big) = P\Big(\underline{\epsilon}_{f_\bullet}^{\boldsymbol{w}'}(z', 0) > \frac{\varepsilon}{2}\Big)$$

This limits the complexity of the function class from infinite to having a covering number $= \mathcal{C}_{\mathcal{F}}^{\gamma}$. Swapping samples between the estimation and the ghost sample, this will create a random variable $S' = \frac{1}{M}(\underline{\epsilon}_{f_\bullet}^{\boldsymbol{w}'_1}(z'_1, 0) + \ldots + \underline{\epsilon}_{f_\bullet}^{\boldsymbol{w}'_m}(z'_m, 0), + \ldots + \underline{\epsilon}_{f_\bullet}^{\boldsymbol{w}'_M}(z'_M, 0))$ for $M = 2^{n_1}$, where the subscripts of $\boldsymbol{w}'$ and z' denote the sample index. Note that $\mathbb{E}_{x\sim p_t}[S'] = \underline{R}_{f_\bullet}(0)$ and let S denote $S' - \mathbb{E}_{x\sim p_t}[S']$, with $\mathbb{E}_{x\sim p_t}[S] = 0$. Let $\sigma^2(S) = \mathbb{E}[S^2] = \mathbb{E}[(S' - \mathbb{E}_{x\sim p_t}[S'])^2]$. By Lemma A2, we have that $\sigma^2(S') \leq d_2(p||p_1) - \underline{R}_{f_\bullet}(0)^2$. By Bernstein's inequality:

$$P\Big(\underline{R}_{f_\bullet}(0) - \underline{\epsilon}_{f_\bullet}^{w'}(z', 0) > \frac{\varepsilon}{2}\Big) \leq \exp\Big(\frac{-3n_1\varepsilon^2}{24\sigma^2(S) + 4C_1\varepsilon}\Big),$$

and a union bound over the function space:

$$\begin{aligned} &P\Big(\underline{R}_{f_\bullet}(0) - \underline{\epsilon}_{f_\bullet}^{w'}(z', 0) > \frac{\varepsilon}{2}\Big) \leq \\ &\mathcal{N}(\gamma, \mathcal{F}, 2n_1) \exp\Big(\frac{-3n_1\varepsilon^2}{24\sigma^2(S) + 4C_1\varepsilon}\Big) \end{aligned}$$

Putting it all together:

$$\begin{aligned} &P\Big(\underline{R}_{f_l^1}(\gamma) - \underline{\epsilon}_{f_l^1}^{w}(z, \gamma) > \varepsilon\Big) \\ &\leq 2P\Big(\underline{R}_{f_\bullet}(0) - \underline{\epsilon}_{f_\bullet}^{w'}(z', 0) > \frac{\varepsilon}{2}\Big) \\ &\leq 2\mathcal{N}(\gamma, \mathcal{F}, 2n_1) \exp\Big(\frac{-3n_1\varepsilon^2}{24\sigma^2(S) + 4C_1\varepsilon}\Big) \end{aligned}$$

Setting $\delta(\epsilon)$ to match the upper bound, inverting w.r.t. ϵ and removing the (negative) term $\underline{R}_{f_\bullet}(0)^2$ from the right-hand side, we get that stated bound with probability $1 - \delta$. □

Next, we define the auxiliary function space, which will allow us to study non-zero training set violations.

Definition A5. *[Restated from (Schölkopf et al., 2001), definition 13] Let $L(\mathcal{X})$ be the set of real valued, non-negative functions f on $\mathcal{X}$ with support supp(f) countable, that is the functions in in $L(\mathcal{X})$ are non-zero for at moust countably many points. We define the inner product of two functions $f, g \in L(\mathcal{X})$ by:*

$$f \cdot g \sum_{x \in supp(f)} f(x)g(x).$$

The 1-norm on $L(\mathcal{X})$ is defined by $||f||_1 = \sum_{x\in supp(f)} f(x)$. Let $L^D(\mathcal{X}) := \{f \in L(\mathcal{X}) : ||f||_1 \leq D\}$. Define a transformation, or embedding of $\mathcal{X}$ into the product space $\mathcal{X} \times L(\mathcal{X})$ as follows:

$$\begin{aligned} \varpi &: \mathcal{X} \to \mathcal{X} \times L(\mathcal{X}) \\ \varpi &: x \to (x, \Delta_x), \end{aligned}$$

where

$$\Delta_x = \begin{cases} 1, & y = x, \\ 0, & otherwise \end{cases}$$

For a function $f \in \mathcal{F}$ a set of training examples z of size n, define the function $g_f \in L(\mathcal{X})$

$$g_f(\mathbf{y}) := \sum_{x,y\in z} w_1(x) \min\{0, \gamma - \underline{r}_{f_l^1}(x, y)\}\Delta_x(\mathbf{y}),$$

where $\mathbf{y} = \{y_i\}_{i=1}^n$

We can now state the risk of overestimation for general sturdy functions.

Theorem A1. *Let $\mathcal{F}$ be any sturdy function class defined over input space $\mathcal{X}$, $z = \{x_i, y_i\}_{i:t_i=t}$, where $x_i, y_i \sim p_t(X, Y)$, and C_t be as defined in (1). For $f_l^t \in \mathcal{F}$, and any $\gamma > 0$, let the associated $\underline{D}^{\boldsymbol{w}_t}(z, f_t^1, \gamma) = D > 0$. With a probability $1 - \delta$ over the draw of random samples, we have that:*

$$\underline{R}_{f_t^l}(\gamma) \leq \frac{4C_t(k_t + \log\frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log\frac{1}{\delta})}{n_t}}. \tag{9}$$

where, for $t \in \{0, 1\}$,

$$k_t = \left\lceil \log \mathcal{N}(\gamma/2, \mathcal{F}, 2n_t) + \log \mathcal{N}(\gamma/2, L^D(\mathcal{X}), 2n_t) \right\rceil .$$

Proof sketch. The proof extends lemma A3, replacing the function class $\mathcal{F}$ with the function class of the augmented space, that is $\mathcal{F} + L(\mathcal{X}) := \{f + g : f \in \mathcal{F}, g \in L(\mathcal{X})\}$. The details of the proof are identical to theorem 14 in Schölkopf et al. (2001), and are hence omitted.

The following lemma, restated from Shawe-Taylor & Cristianini (2002) gives a bound on the auxiliary function complexity for linear functions (defined in kernel spaces).

Lemma A4. *Due to Shawe-Taylor & Cristianini (2002). For $D > 0$, all $\gamma > 0$:*

$$\begin{aligned}&\log \mathcal{N}(\gamma, L^D(\mathcal{X}), n)\\ &\leq \left\lfloor \frac{D}{2\gamma} \right\rfloor \log\left(\frac{\exp(n + {}^{D}\!/_{2\gamma} - 1)}{{}^{D}\!/_{2\gamma}}\right)\end{aligned}$$

Finally, by replacing the auxiliary function term from theorem A1 (that is $\log \mathcal{N}({}^{\gamma}\!/_{2}, L^D(\mathcal{X}), 2n_t)$) with its bound for linear functions acquired from lemma A4 (that is $\log \frac{\exp(n_t + {}^{D}\!/_{\gamma} - 1)}{{}^{D}\!/_{\gamma}}$), we get the proof for theorem 1.

3. Risk of overestimation of ITE

The risk of overestimation for the ITE can be stated as a simple extension of theorem 1. We define the ITE as $\tau(x) = Y(x, 1) - Y(x, 0)$, where $Y(x, t)$ is the potential outcome under treatment $T = t$, for patient with characteristics $X = x$. We use $\tilde{\tau}_l(x)$ to denote $f_l^1(x) - f_u^0(x)$, where f_l^1, f_u^0 are some estimates of the lower bound for the outcome under treatment and the upper bound of the outcome under non-treatment respectively. In addition, we define:

$$\overline{r}_f(x, y) = f(x) - y,$$

and for $z_t = \{x_i, y_i\}_{i:t_i = t}$, define

$$\overline{D}^{\boldsymbol{w}_t}(z, f_u^t, \gamma) = \sum_{x,y \in z} w_t(x) \min\{0, \gamma - \overline{r}_{f_u^t}(x, y)\}$$

Corollary A1. *Let $\mathcal{F}$ be the class of linear functions in a kernel defined feature space, $z_t = \{x_i, y_i\}_{i:t_i = t}$, where $x_i, y_i \sim p_t(X, Y)$, and C_t be as defined in expression (1). For $f_l^1, f_u^0 \in \mathcal{F}$, and any $\gamma > 0$, let the associated $\underline{D}^{\boldsymbol{w}_1}(z_1, f_l^1, \gamma) = D_1 > 0$, and $\overline{D}^{\boldsymbol{w}_0}(z_0, f_u^0, \gamma) = D_0 > 0$ Define $\tilde{\tau}_l := f_l^1 - f_u^0$. With probability $1 - \delta$ over random samples, we have that:*

$$\begin{aligned}\underline{R}_{\hat{\tau}_l}(\gamma) \leq \sum_t &\frac{4C_t(k_t + \log\frac{1}{\delta})}{3n_t}\\ &+ \sqrt{\frac{8d_2(p||p_t)(k_t + \log\frac{1}{\delta})}{n_t}}.\end{aligned} \tag{10}$$

where, for $t \in \{0, 1\}$,

$$k_t = \left\lceil \log \mathcal{N}({}^{\gamma}\!/_{2}, \mathcal{F}, 2n_t) + \log \mathcal{N}({}^{\gamma}\!/_{2}, L^{D_t}(\mathcal{X}), 2n_t) \right\rceil.$$

Proof. Consider the event:

$$E = \{x : \tau(x) < \tilde{\tau}_l(x) - 2\gamma\}$$

where $x \sim p$. Note that event E implies that one of the following two events must hold:

$$E_1 = \{(x, y) : \underline{r}_{f_l^1}(x, y) < \gamma\}$$

for $t = 1$.

$$E_0 = \{(x, y_0) : \overline{r}_{f_u^0}(x, y) < \gamma\}$$

for $t = 0$.

Note that $p(E_1) = \underline{R}_{f_l^1}(\gamma)$. So, theorem A1 implies that

$$p(E_1) \leq \frac{4C_t(k_t + \log\frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log\frac{1}{\delta})}{n_t}}$$

for k_t as defined in theorem A1. Similarly $p(E_0) = \overline{R}(f_u^0)$, and by a similar construction can obtain the bound on $p(E_0)$. Using a union bound we have that

$$\begin{aligned}p(E) = p(E_1 \cup E_0) &= p(E_1) + p(E_0) - p(E_1 \cap E_0)\\ &\leq p(E_1) + p(E_0),\end{aligned}$$

which completes the proof. □

4. Proof of Theorem 2

To build up to the proof of theorem 2, we first seek a bound on the fat-shattering dimension of functions defined in definition 5. This bound is constructed in a similar spirit to theorem 1.6 in (Bartlett & Shawe-Taylor, 1999). Specifically, to get a bound on the fat-shattering dimension, we rely on the lemmas A5 and A6. The former shows that the sum of any shattered set is far from the remainder of that set, the latter shows that the same sums cannot be too far apart.

Lemma A5. *Let $\mathcal{F}_u, \mathcal{F}_l, A, B$ be as defined in definition 5. Let $I = \{x_i\}_{i=1}^n$, where $x_i \sim p(X, Y)$.For a fixed $\gamma > 0$, if I is $\gamma-$shattered by $\mathcal{F}_l$ then every subset $I' \in I$ satisfies:*

$$\min_{q \in \{p,2\}} \left\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \right\|_q \geq \frac{2n\gamma}{A + B}$$

Proof. If I is γ shattered by $\mathcal{F}_l$, denote the corresponding "witness" vector by $\{s_i\}_{i=1}^n$, then for all $\boldsymbol{\sigma} = \{\sigma_1 \ldots \sigma_i \ldots \sigma_n\}$ there is an f with $\|f_l\| \leq A$ such that $\sigma_i \cdot (\theta^\top x_i - s_i) \geq \gamma$ for $i = 1 \ldots n$. Suppose that:

$$\sum_{i \in I'} s_i \geq \sum_{i \in I \setminus' I} s_i \tag{11}$$

Then fix $\sigma_i = 1$ if $i \in I'$. In that case we have that

$$\langle f_l, x_i \rangle \geq s_i + \gamma \qquad \forall i \in I' \tag{12}$$
$$\langle f_l, x_i \rangle < s_i - \gamma \qquad \forall i \in I \setminus I'. \tag{13}$$

Pick $f_u \in \mathcal{F}_u$ such that $||f_u - f_l||_p = B' \leq B$, and:

$$\langle f_u - f_l, x_i \rangle \geq s_i + \gamma \qquad \forall i \in I' \tag{14}$$
$$\langle f_u - f_l, x_i \rangle < s_i - \gamma \qquad \forall i \in I \setminus I'. \tag{15}$$

Showing that such a function exists is trivial: simply take $f_u := f_l$. For that we have $||f_u - f_l|| = 0 \leq B$, which means that the function does exist in $\mathcal{F}_u$.

From expression 12, we have that:

$$\langle f_l, \sum_{i \in I'} x_i \rangle = \sum_{i \in I'} \langle f_l, x_i \rangle \geq \sum_{i \in I'} s_i + Card(I')\gamma,$$

where $Card(.)$ denotes the cardinality. Similarly for $I \setminus I'$, we have that

$$\langle f_l, \sum_{i \in I \setminus I'} x_i \rangle < \sum_{i \in I \setminus I'} s_i + Card(I \setminus I')\gamma$$

Combining the expressions for I' and $I \setminus I'$, and from expression 11:

$$\langle f_l, \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \rangle \geq n\gamma. \tag{16}$$

We now construct the same arguments for the distance. Let $f_d := f_u - f_l$. From expression 14, we have that:

$$\langle f_d, \sum_{i \in I'} x_i \rangle = \sum_{i \in I'} \langle f_d, x_i \rangle \geq \sum_{i \in I'} s_i + Card(I')\gamma,$$

and from expression 15:

$$\langle f_d, \sum_{i \in I \setminus I'} x_i \rangle < \sum_{i \in I \setminus I'} s_i + Card(I \setminus I')\gamma$$

Combining the two, and from expression 11:

$$\langle f_d, \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \rangle \geq n\gamma. \tag{17}$$

Putting expressions 16 and 17 together,

$$\langle f_l, \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \rangle \tag{18}$$
$$+ \langle f_d, \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \rangle \geq 2n\gamma. \tag{19}$$

Note that by Cauchy-Schwartz,

$$\begin{aligned}
\langle f_l, \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \rangle &\leq \|f_l\| \Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\| \\
&\leq A \Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\| \\
&\leq A \min_{q \in \{p,2\}} \Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\|_q .
\end{aligned}$$

and,

$$\begin{aligned}
\langle f_d, \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \rangle &\leq \|f_d\|_p \Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\|_p \\
&\leq B' \Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\|_p \\
&\leq B \Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\|_p \\
&\leq B \min_{q \in \{p,2\}} \Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\|_q .
\end{aligned}$$

For expression 18 to hold:

$$\begin{aligned}
&A \min_{q \in \{p,2\}} \Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\|_q \\
&+ B \min_{q \in \{p,2\}} \Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\|_q \geq 2n\gamma \\
&(A+B) \min_{q \in \{p,2\}} \Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\|_q \geq 2n\gamma \\
&\min_{q \in \{p,2\}} \Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\|_q \geq \frac{2n\gamma}{(A+B)},
\end{aligned}$$

which completes the proof.

Lemma A6. *Let $\mathcal{F}_u, \mathcal{F}_l, r$ be as defined in definition 5. Let $I = \{x_i\}_{i=1}^n$, where $x_i \sim p(X,Y)$.For a fixed $\gamma > 0$, if I is $\gamma-$shattered by $\mathcal{F}_l$ then every subset $I' \in I$ satisfies:*

$$\Big\| \sum_{i \in I'} x_i - \sum_{i \in I \setminus I'} x_i \Big\| \leq \sqrt{n} r$$

The proof is identical to Lemma 1.3 in (Bartlett & Shawe-Taylor, 1999), and is hence omitted.

Lemma A7. *Let $\mathcal{F}_u, \mathcal{F}_l, A, B, r$ be as defined in definition 5. For a fixed $\gamma > 0$, the $\gamma-$fat shattering dimension of $\mathcal{F}_l$ can be bounded as follows:*

$$fat(\gamma, \mathcal{F}_l) \leq \left(\frac{r \cdot (A+B)}{2\gamma} \right)^2$$

Combining the results from Lemmas A6 and A5, we get that:

$$\frac{2n\gamma}{A+B} \le \min_{q\in\{p,2\}} \left\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \right\|_q \le \left\| \sum_{i\in I'} x_i - \sum_{i\in I\setminus I'} x_i \right\| \le \sqrt{n} r,$$

which gives us that:

$$\sqrt{n} \le \frac{r(A+B)}{2\gamma},$$

which completes the proof.

Theorem A2. *Let $\mathcal{F}_l^t$, $\mathcal{F}_u^t$, A, B, and r be as defined in definition 5, z, and D as defined in theorem 1,and C_t be as defined in expression (1). For $f_l^t \in \mathcal{F}_l^t$, $f_u^t \in \mathcal{F}_u^t$ and any $\gamma > 0$, with a probability $1-\delta$ over the draw of random samples, we have that:*

$$\underline{R}_{f_t^l}(\gamma) \le \frac{4C_t(k_t + \log\frac{1}{\delta})}{3n_t} + \sqrt{\frac{8d_2(p||p_t)(k_t + \log\frac{1}{\delta})}{n_t}}. \tag{20}$$

where, for $t \in \{0,1\}$,

$$k_t = \left\lceil \left(\frac{2r(A+B)}{\gamma}\right)^2 \log\left(\frac{8n_t(b-a)^2}{\gamma^2}\right) \log\left(\frac{4en_t(b-a)\gamma}{r^2(A+B)^2}\right) + \frac{D}{\gamma}\log\frac{e(n_t + {}^{D}/_{\gamma} - 1)}{{}^{D}/_{\gamma}} \right\rceil .$$

Using Corollary 3.8 (Shawe-Taylor et al., 1998), we can $\log\mathcal{N}(\gamma/2, \mathcal{F}, 2n_t)$ by its fat shattering dimension. Combining the results from lemma A7 and theorem 1, we get the final result.

5. Equivalence to quantile regression

Consider the following problem

$$\begin{aligned}
\underset{f_u, f_l}{\text{minimize}} \quad & \ell_{\tilde{w}}^{(1)}(f_u(x_i), f_l(x_i)) \\
\text{subject to} \quad & \sum_{i:t_i=t} \tilde{w}_{t_i} \max[y_i - f_u(x_i), 0] \le \beta \\
& \sum_{i:t_i=t} \tilde{w}_{t_i} \max[f_l(x_i) - y_i, 0] \le \beta \\
& f_u(x_i) \ge f_l(x_i), \;\; \forall i : t_i = t
\end{aligned} \tag{21}$$

Theorem A3. *Assume that (21) is strictly convex and has a strictly feasible solution. Then, for any fixed quantile $t \in (0.5, 1)$, there is a parameter $\beta \ge 0$ such that the minimizer of (21) with weighted absolute loss and the minimizer of the weighted quantile loss, for quantiles $(t, 1-t)$ with non-crossing constraints, are equal and have false coverage rate $1-q$.*

Proof. Problem (21) with absolute loss $\ell(y, y') = |y - y'|$ can be stated as

$$\begin{aligned}
\underset{f_u, f_l}{\text{minimize}} \quad & \sum_{i:t_i=t} \tilde{w}_{t_i} |f_u(x_i) - f_l(x_i)| \\
\text{subject to} \quad & \sum_{i:t_i=t} \tilde{w}_{t_i} \max[y_i - f_u(x_i), 0] \le \beta \\
& \sum_{i:t_i=t} \tilde{w}_{t_i} \max[f_l(x_i) - y_i, 0] \le \beta \\
& f_u(x_i) \ge f_l(x_i), \;\; \forall i : t_i = t
\end{aligned}$$

Let $Q_\beta(f_u, f_l) = \tilde{w}_{t_i}|f_u(x_i) - f_l(x_i)|$ denote the objective and F the feasibility region. Introducing Lagrange multipliers for the first two constraints, we obtain the regularized objective

$$\begin{aligned}
L(f_u, f_l, \lambda_u, \lambda_l) = & \sum_{i:t_i=t} \tilde{w}_{t_i} |f_u(x_i) - f_l(x_i)| \\
& + \frac{\lambda_u}{n} \sum_{i=1}^{n} \max(y_i - f_u(x_i), 0) - \beta \\
& + \frac{\lambda_l}{n} \sum_{i=1}^{n} \max(f_l(x_i) - y_i, 0) - \beta
\end{aligned}$$

and by convexity and strict feasibility, strong duality holds through Slater's condition,

$$\min_{u,l \in F} Q_\beta(u, l) = \max_{\lambda_u, \lambda_l \ge 0} \min_{u \ge l} L(u, l, \lambda_u, \lambda_l) \; .$$

By strict convexity, for each $\beta \ge 0$, the minimizers u^*, l^* on either side are equal for the maximizers λ_u^*, λ_l^*. Now, consider the following objective, equivalent in minima to $\tilde{L}(f_u, f_l, \lambda_u, \lambda_l)$,

$$\begin{aligned}
\tilde{L}(f_u, f_l, \lambda_u, \lambda_l) := & \sum_{i:t_i=t} \tilde{w}_{t_i} |f_u(x_i) - f_l(x_i)| \\
& + \lambda_u \sum_{i:t_i=t} \tilde{w}_{t_i} \max(y_i - f_u(x_i), 0) \\
& + \lambda_l \sum_{i:t_i=t} \tilde{w}_{t_i} \max(f_l(x_i) - y_i, 0)
\end{aligned}$$

We can separate $\tilde{L}$ into terms for which $y_i \ge f_u(x_i)$ and $y_i \ge f_l(x_i)$ respectively, adding and subtracting $\sum_i y_i$

$$\begin{aligned}
& \tilde{L}(f_u, f_l, \lambda_u, \lambda_l) \\
& = (\lambda_u - 1) \sum_{y_i \ge u(x_i)} \tilde{w}_{t_i}(y_i - f_u(x_i)) - \sum_{y_i < f_u(x_i)} \tilde{w}_{t_i}(y_i - f_u(x_i)) \\
& + (1 - \lambda_l) \sum_{y_i \ge f_l(x_i)} \tilde{w}_{t_i}(y_i - f_l(x_i)) - \sum_{y_i < f_l(x_i)} \tilde{w}_{t_i}(y_i - f_l(x_i))
\end{aligned}$$

Now, let $\lambda_u = \lambda_l = 1/(1-q)$ for $q \in (0,1)$, which means

$(1-q) \geq 0$. Multiplying by $(1-q)$ leaves us with

$$
\begin{aligned}
&\tilde{L}(f_u, f_l, \lambda_u, \lambda_l) \\
&\propto \sum_{y_i \geq f_u(x_i)} q \cdot \tilde{w}_{t_i}(y_i - f_u(x_i)) + \\
&\sum_{y_i < f_u(x_i)} (q-1) \cdot \tilde{w}_{t_i}(y_i - f_u(x_i)) \\
&+ \sum_{y_i \geq f_u(x_i)} (1-q) \cdot \tilde{w}_{t_i}(y_i - f_l(x_i)) \\
&+ \sum_{y_i < f_u(x_i)} (-q) \cdot \tilde{w}_{t_i}(y_i - f_l(x_i)) \\
&\propto \sum_{i:t_i=t} \tilde{w}_{t_i} \max[q(y_i - f_u(x_i)), (q-1)(y_i - f_u(x_i)] \\
&+ \sum_{i:t_i=t} \tilde{w}_{t_i} \max[(1-q)(y_i - f_l(x_i)), (-q)(y_i - f_l(x_i)] \\
&= \sum_{i:t_i=t} \rho^{(q)}_{\tilde{w}_{t_i}}(y_i - f_u(x_i)) + \rho^{(1-q)}_{\tilde{w}_{t_i}}(y_i - f_l(x_i)) \ ,
\end{aligned}
$$

where $\rho^{(q)}_{\tilde{w}}$ is the weighted quantile loss for quantile q. Recalling that our original problem had the constraint $f_u(x_i) \geq f_l(x_i)$, we recover the non-crossing constraint. □

6. Cross-validation algorithm

Define Ω denote a set of candidate hyperparameters. Suppose we have M possible hyperparameters, cross-validating BP proceeds as follows:

Algorithm 1 BP cross-validation for M sets of hyperparameters, and required FCR = ν

```
**Input:** $\mathcal{D} = \{x_i, t_i, y_i, w_i\}$, $p$, $\nu$, $\{\Omega\}^M$
**Output:** $\Omega^*$
Split $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{validate}}$
**for** $m = 1$ **to** $M$ **do**
    Use $\mathcal{D}_{\text{train}}$ to solve problem (6) or (7)
    Estimate $\hat{\nu}^{(m)}$, and $||\widehat{\text{IW}}||_p^{(m)}$ on $\mathcal{D}_{\text{validate}}$
**end for**
Define $M' = \{m : \hat{\nu}^{(m)} \leq \nu\}$
Set $\Omega^* := \min_{m \in M'} ||\widehat{\text{IW}}||_p^{(m)}$
```

7. Experiments

7.1. Cross-validation details

For our BP method, we have 5 hyperparameters to pick. These are α, the regularization parameter, the kernel bandwidth, β_u and β_l which are the allowed violations. The last parameter, $\gamma_{BP} > 0$, as described in section 5.3. Note that the kernel bandwidth is only relevant for the experiments done on the ACIC data, but not the IST experiments since a linear kernel is used in the latter.

For the kernel regression (KR), we first split the training data into 2. On the first half, we do the typical 3-fold cross-validation to pick the model that minimizes the weighted empirical error. This allows us to pick the kernel bandwidth, and a regularization parameter the is multiplied by the L2 norm of the weights. Again, the kernel bandwidth is only relevant for the experiments done on the ACIC data, but not the IST experiments since a linear kernel is used in the latter. The intervals are then estimated in one of two ways. For KR-MI, we use the second part of the training data to estimate the residuals. We follow algorithm 2 in (Lei et al., 2018) to get the final interval estimates. For KR-γ, we use the second half of the training data to estimate the FCR, $\hat{\nu}_{\gamma_{\text{KR}}}$, with γ_{KR} defined as the "shifting" parameter, where $\tilde{f}_u^{KR}(x_i) = \tilde{\mu}_t(x_i) + \gamma_{KR}$ and $\tilde{f}_l^{KR}(x_i) = \tilde{\mu}_t(x_i) - \gamma_{KR}$, for $\tilde{\mu}_t(x_i)$ being the predicted response value. We then pick the smallest γ_{KR} that does not violated the required FCR.

For the Gaussian process (GP), we pick the kernel bandwidth, the noise level added to the diagonal of the kernel. For BART models, we use the BartMachine package in R (Kapelner & Bleich, 2016). We do 3 fold cross-validation to pick the parameter k, which controls the prior probability that $\mathbb{E}(y|x)$ is contained in the interval (ymin,ymax), based on a normal distribution. We set the number of trees to be 200, since that did not seem to affect the results. For the CMGP, we pick the lengthscale of the RBF kernels of the two response surfaces as well as the variance and correlation parameters.

7.2. Additional IST details

Figure 4 shows the histogram of the ages in the training data for the treated and the control population. Ages$>$ 70 were downsampled to introduce a confounding effect.

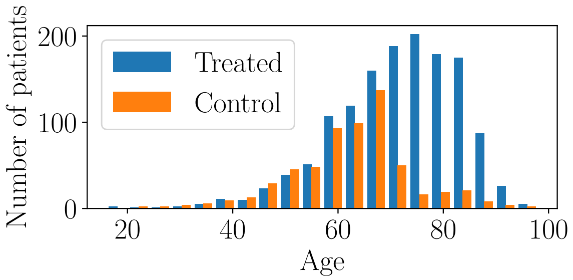


Figure 4. Distribution of data in the IST experiment

7.3. Additional IST results (heteroskedasticity)

In this section we analyze the performance of our model when the well-behavedness assumption is violated, specifically when there is heteroskedasticity. We use the IST data, and follow the same train/test splits as is done in the main paper. Here, we focus on the outcome under treatment, $Y(1)$ only. Specifically, we generate the outcome under treatment as $Y(1) = x^2 + \epsilon$, where x is the age rescaled to fall between -2, 2, and ϵ_i is drawn from a Gaussian distribution with mean 0 and standard deviation = 0.1 if $x \leq 0$, and from a Gaussian distribution with mean 0 and standard deviation = $0.1 + x$ otherwise. We set the required FCR to be ≤ 0.01. Since our main aim is to analyze how the different models perform when when heteroskedasticity occurs, we focus only on tightness of bounds as an objective.

Figure 3 shows the results from averaged over 20 simulations. It shows that of all the models that achieve the required FCR, BP-D-L2 achieves the tightest intervals. Figure 5 shows why: neither BP-D-L2 and QR (equivalent to BP-D-L1) make assumptions about well-behavedness of the residual distribution. They git adaptive intervals, which are tight when the heteroskedastic noise is low, and loose when it is high.

Table 3. IST heteroskedasticity results. Table shows results averaged over 20 simulations 5.

Model	FCR	Mean IW	Max IW
BP-D-L2	0.007 (0.5)	5.55 (0.56)	10.68 (2.35)
QR/BP-D-L1	0.006 (0.31)	6.49 (0.96)	11.63 (2.37)
KR-γ	0.065 (0.86)	3.98 (0.06)	3.98 (0.06)
KR-CI	0.007 (0.52)	6.94 (0.69)	6.94 (0.69)

7.4. ACIC results including CCI

Figure 6 is similar to figure 3 presented in the main paper but includes the performance of CCI models.

7.5. Additional ACIC results

We consider a larger sample size than that presented in the main paper. Instead of sampling $n = 200$ for training and validation of the main model, we sample $n = 1000$. In this setting, we are better able to fit the true outcomes since the larger sample size affords us the ability to fit more complex models. Figure 7 shows the results. Once again we see that our models outperform all kernel based methods. Here we see that BART-$gamma$ achieves a tighter interval width than our model for the same level of FCR violation. This highlights the strength of tree based models in that they fit highly adaptive "kernels".

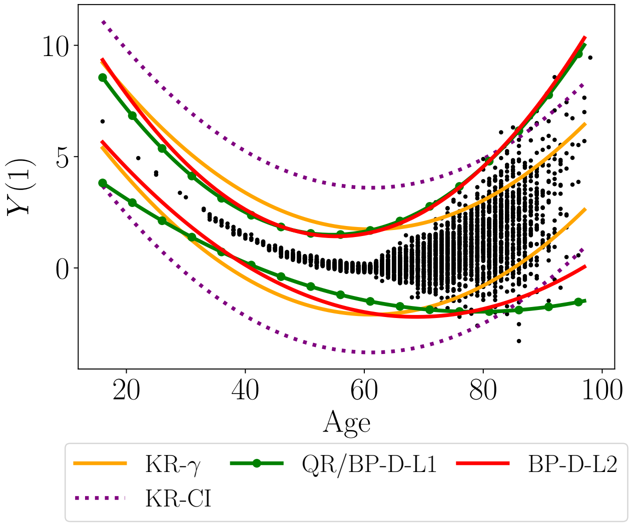


Figure 5. IST heteroskedasticity results. Plot shows results from a single simulation. Black dots show potential outcomes on the test set, lines show fitted values. The plot show that BP-D-L2 and QR (equivalent to BP-D-L1) are the only ones that are able to fit *adaptive* intervals (wider where there is high heteroskedasticity). BP-D-L2 achieves the tightest intervals on average.

References

Bartlett, P. and Shawe-Taylor, J. Generalization performance of support vector machines and other pattern classifiers. *Advances in Kernel methods—support vector learning*, pp. 43–54, 1999.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems 23*, pp. 442–450. Curran Associates, Inc., 2010.

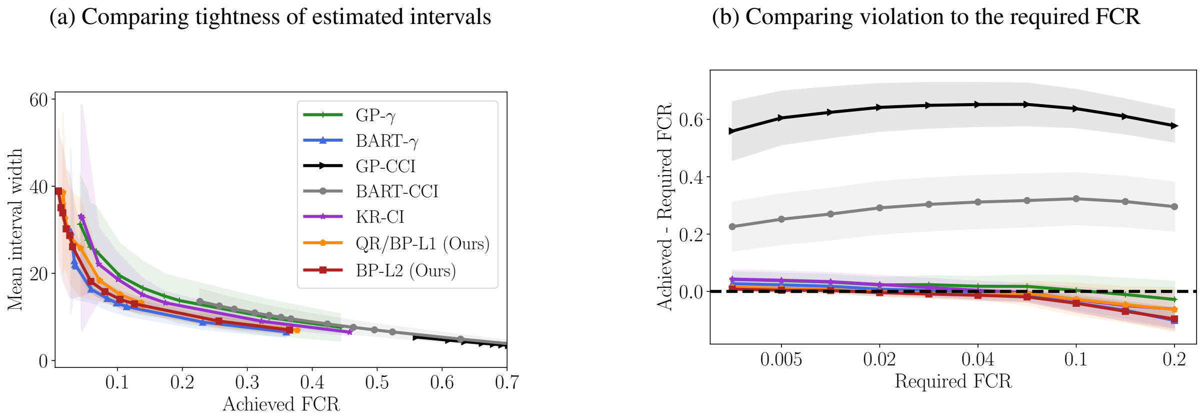
Kapelner, A. and Bleich, J. bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40, 2016. doi: 10.18637/jss.v070.i04.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7): 1443–1471, July 2001. ISSN 0899-7667.

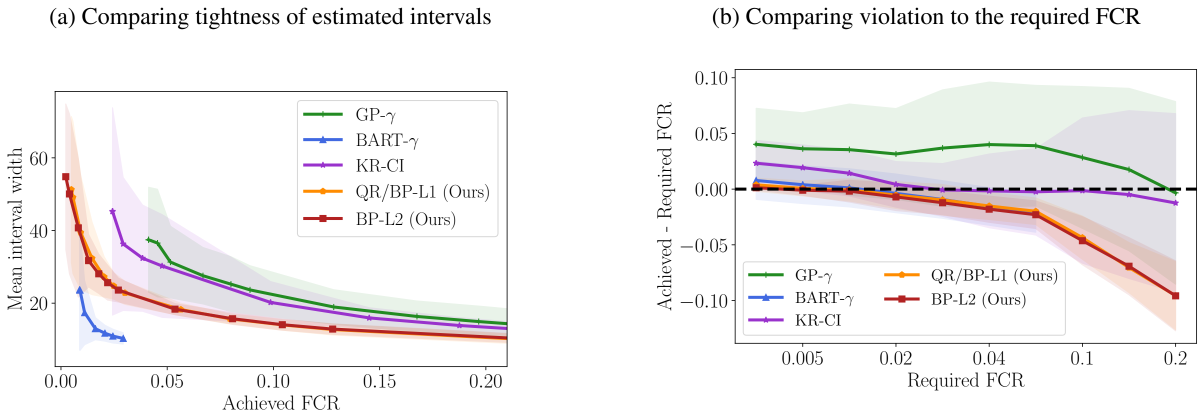
Shawe-Taylor, J. and Cristianini, N. On the generalisation of soft margin algorithms. *IEEE Transactions on Information Theory*, 48(10):2721–2735, 2002.

Shawe-Taylor, J. and Williamson, R. C. Generalization performance of classifiers in terms of observed cover-


(a) Comparing tightness of estimated intervals

(b) Comparing violation to the required FCR

Figure 6. ACIC results. Plots show results averaged over 20 simulations. Plot 6a shows the mean interval width for different values of the achieved FCR on a held-out test set. Plot 6b shares the same legend as plot 6a, and shows the violation of the required FCR (= achieved - required) at different values of required FCR. Models above the dotted black line are in violation of the required FCR. The two plots show that BP achieves a mean interval width comparable to that of BART but at a lower violation of the required FCR. BP outperforms all kernel-based methods in terms of mean interval width and violation to the required FCR. CCI methods achieve the worst violations.


(a) Comparing tightness of estimated intervals

(b) Comparing violation to the required FCR

Figure 7. ACIC results. Plots show results averaged over 20 simulations. Plot 7a shows the mean interval width for different values of the achieved FCR on a held-out test set. Plot 7b shares the same legend as plot 7a, and shows the violation of the required FCR (= achieved - required) at different values of required FCR. Models above the dotted black line are in violation of the required FCR. The two plots show that BP achieves a mean interval width comparable to that of BART but at a lower violation of the required FCR. BP outperforms all kernel-based methods in terms of mean interval width and violation to the required FCR. CCI methods achieve the worst violations.

ing numbers. In *Computational Learning Theory*, pp. 274–285, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., and Anthony, M. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998.