Electronic Companion

EC.1. K-PT MIO Formulation

In this section, we provide an adapted version of the formulation from Kallus (2017) that we use in our experiments in Section 6.4. As we will soon discuss, we adapt the original formulation for both brevity and so that it better aligns with the notation we use. Extending the notation from the problem statement, we construct a perfect binary tree and number the nodes 1 through $2^{d+1}-1$ in the order in which they appear on a breadth-first-first search. We let $R_{nm} \in \{1,-1\}$ capture the relationship of node n with its ancestor(s) $m \in \mathcal{A}(n)$, i.e., R_{nm} equals 1 iff we use m's right branch to reach n, and -1 otherwise.

We now define the decision variables used in the K-PT formulation. For every branching node $n \in \mathcal{B}$ and feature $f \in \mathcal{F}$, we let the binary variable b_{nf} indicate if feature f is selected for branching at node n. We let the binary variable χ_{in} equal 1 iff datapoint i goes left on branching node n. Further, we define membership variables $\lambda_{in} \in \{0,1\}$, which takes value 1 iff datapoint i flows to terminal node $n \in \mathcal{T}$. For $k \in \mathcal{K}$, we let $w_{nk} \in \{0,1\}$ equal 1 iff treatment k is selected at terminal node n. Also for each terminal node n, define ρ_n to be the average treatment outcome of all datapoints in that node. Let η_{in} be the product of λ_{in} and ρ_n, which captures the average outcome associated with datapoint i.

In order to linearize η, Kallus (2017) introduces big-M constraints. To this end, let $\overline{Y}_i = Y_i - \min_{j\in\mathcal{I}} Y_j$, $\overline{Y}_{max} = \max_i \overline{Y}_i$, and $M = \overline{Y}_{max}(\max_{k\in K} \sum_{i\in\mathcal{I}} \mathbb{I}[K_i = k])$. The approach now reads:

$$\text{maximize} \quad \sum_{i\in\mathcal{I}} \sum_{n\in\mathcal{T}} \eta_{in} \tag{EC.1a}$$

$$\text{subject to} \quad \lambda_{in} \le \frac{1+R_{nm}}{2} - R_{nm}\chi_{im} \qquad \forall i\in\mathcal{I}, n\in\mathcal{T}, m\in a(n) \tag{EC.1b}$$

$$\lambda_{in} \ge 1 - \sum_{\substack{m\in a(n)\\ R_{mn}=1}} \chi_{im} + \sum_{\substack{m\in a(n)\\ R_{mn}=-1}} (-1+\chi_{im}) \qquad \forall i\in\mathcal{I}, n\in\mathcal{T} \tag{EC.1c}$$

$$\sum_{f\in\mathcal{F}} b_{nf} = 1 \qquad \forall n\in\mathcal{B} \tag{EC.1d}$$

$$\chi_{in} = \sum_{\substack{f\in\mathcal{F}:\\ X_f^i=0}} b_{nf} \qquad \forall i\in\mathcal{I}, n\in\mathcal{B} \tag{EC.1e}$$

$$\eta_{in} \le \overline{Y}_{max}\lambda_{in}, \eta_{in} \le \rho_n \qquad \forall i\in\mathcal{I}, p\in\mathcal{T} \tag{EC.1f}$$

$$\eta_{in} \ge \rho_n - \overline{Y}_{max}(1-\lambda_{in}) \qquad \forall i\in\mathcal{I}, p\in\mathcal{T} \tag{EC.1g}$$

$$\sum_{i:K_i=k} (\eta_{in} - \lambda_{in}\overline{Y}_i) \le M(1-w_{nk}) \qquad \forall p\in\mathcal{T}, k\in\mathcal{K} \tag{EC.1h}$$

$$\sum_{i:K_i=k} (\eta_{in} - \lambda_{in}\overline{Y}_i) \ge M(w_{nk}-1) \qquad \forall p\in\mathcal{T}, k\in\mathcal{K} \tag{EC.1i}$$

$$\sum_{k \in K} w_{nk} = 1 \qquad \forall n \in \mathcal{T} \qquad \text{(EC.1j)}$$

$$w_{nk} \in \{0,1\} \qquad \forall n \in \mathcal{T}, k \in \mathcal{K} \qquad \text{(EC.1k)}$$

$$b_{nf} \in \{0,1\} \qquad \forall n \in \mathcal{B}, f \in \mathcal{F} \qquad \text{(EC.1l)}$$

$$\lambda_{in} \in \{0,1\} \qquad \forall i \in \mathcal{I}, n \in \mathcal{T} \qquad \text{(EC.1m)}$$

$$\chi_{in} \in \{0,1\} \qquad \forall n \in \mathcal{B}, i \in \mathcal{I} \qquad \text{(EC.1n)}$$

$$\eta_{in} \in \mathbb{R}_+ \qquad \forall n \in \mathcal{T}, i \in \mathcal{I} \qquad \text{(EC.1o)}$$

$$\rho_n \in \mathbb{R}_+ \qquad \forall n \in \mathcal{T}. \qquad \text{(EC.1p)}$$

The objective function (EC.1a) sums the associated averages over all datapoints and terminal nodes. Constraints (EC.1b) and (EC.1c) mode the flow of datapoints from the root down to the terminal nodes. Constraints (EC.1d) state that, at each branching node, the tree must branch on exactly one feature, whereas (EC.1e) defines χ using the associated branching decisions. Constraints (EC.1f) and (EC.1g) both use big-M constraints to linearize η. Constraints (EC.1h) and (EC.1i) ensure that the predicted outcome for a given treatment at each leaf of the tree is indeed the average observed outcome. Lastly, constraint (EC.1j) ensures that exactly one treatment is assigned at every terminal node.

Note that we have here converted the formulation from Kallus (2017) to a maximization problem to account for the different representation of outcome Y (higher values preferred). Further, we make the following changes from the formulation presented in Kallus (2017): *(i)* constraint (EC.1c) is slightly different, since the original constraint does not result in a correct behavior; *(ii)* for simplicity, we removed constraints inspired from Yıldız and Vielma (2013) from the formulation and replaced it with constraint (EC.1d), which serves the same function; *(iii)* we removed constraint $\sum_{i:K_i=k} \lambda_{in} \geq N_{\min}$ from the formulation to allow for an equivalent comparison with our proposed methods, and in tuning the parameter for our experiments, we did not find statistical improvements.

EC.2. B-PT MIO Formulation

Bertsimas et al. (2019) briefly discuss a coordinate descent algorithm to learn an optimal decision tree with the loss function form

$$\min_T \text{error}(T, D) + \alpha \cdot \text{complexity}(T),$$

where T is the tree being optimized, D is the training data, the "error" function measures how well the tree fits training data, and the second term is a penalization term that controls the trade-off

between the quality of the tree and the tree's complexity/size. The algorithm iterates through different branching decisions and treatment assignment options until no possible improcements are found (i.e., the tree is a local minimum). This process is repeated for different randomly generated starting trees, and "the lowest objective function is selected as the final solution". The documentation for this method is limited and there is no existing codebase allowing us to replicate this algorithm. Hence, in this section, we describe the MIO implementation that we use in our experiments for B-PT, which is adapted from Kallus (2017) to optimize the objective function proposed by Bertsimas et al. (2019).

Building from Kallus (2017), Bertsimas et al. (2019) added a regularization term that penalizes trees whose leaves have high variance. Let β_{nk} be the average outcome over all datapoints that were assigned treatment k in terminal node n. Let g_i be the *empirical* average of the outcomes at the terminal node where datapoint i lands over all datapoints that were assigned the same treatment as i (i.e. $g_i = \beta_{nK_i}$). Finally, let θ control the regularization strength. The formulation becomes:

$$\text{maximize } \theta \sum_{i\in\mathcal{I}}\sum_{n\in T}\eta_{in} - (1-\theta)\sum_{i\in\mathcal{I}}(Y_i - g_i)^2 \tag{EC.2a}$$

$$\text{subject to } \lambda_{in} \le \frac{1+R_{nm}}{2} - R_{nm}\chi_{im} \qquad \forall i\in\mathcal{I}, n\in\mathcal{T}, m\in a(n) \tag{EC.2b}$$

$$\lambda_{in} \ge 1 - \sum_{\substack{m\in a(n)\\ R_{mn}=1}}\chi_{im} + \sum_{\substack{m\in a(n)\\ R_{mn}=-1}}(-1+\chi_{im}) \qquad \forall i\in\mathcal{I}, n\in\mathcal{T} \tag{EC.2c}$$

$$\sum_{f\in\mathcal{F}} b_{nf} = 1 \qquad \forall n\in\mathcal{B} \tag{EC.2d}$$

$$\chi_{in} = \sum_{\substack{f\in\mathcal{F}:\\ X^i_f=0}} b_{nf} \qquad \forall i\in\mathcal{I}, n\in\mathcal{B} \tag{EC.2e}$$

$$\eta_{in} \le \overline{Y}_{max}\lambda_{in}, \eta_{in} \le \rho_n \qquad \forall i\in\mathcal{I}, p\in\mathcal{T} \tag{EC.2f}$$

$$\eta_{in} \ge \rho_n - \overline{Y}_{max}(1-\lambda_{in}) \qquad \forall i\in\mathcal{I}, p\in\mathcal{T} \tag{EC.2g}$$

$$\sum_{i:K_i=k}(\eta_{in} - \lambda_{in}\overline{Y}_i) \le M(1-w_{nk}) \qquad \forall p\in T, k\in\mathcal{K} \tag{EC.2h}$$

$$\sum_{i:K_i=k}(\eta_{in} - \lambda_{in}\overline{Y}_i) \ge M(w_{nk}-1) \qquad \forall p\in T, k\in\mathcal{K} \tag{EC.2i}$$

$$\sum_{k\in\mathcal{K}} w_{nk} = 1 \qquad \forall n\in\mathcal{T} \tag{EC.2j}$$

$$g_i - \beta_{nK_i} \le M(1-\lambda_{in}) \qquad \forall i\in\mathcal{I}, n\in\mathcal{T} \tag{EC.2k}$$

$$g_i - \beta_{nK_i} \ge M(\lambda_{in}-1) \qquad \forall i\in\mathcal{I}, n\in\mathcal{T} \tag{EC.2l}$$

$$w_{nk}\in\{0,1\} \qquad \forall n\in\mathcal{T}, k\in\mathcal{K} \tag{EC.2m}$$

$$b_{nf}\in\{0,1\} \qquad \forall n\in\mathcal{B}, f\in\mathcal{F} \tag{EC.2n}$$

$$\lambda_{in}\in\{0,1\} \qquad \forall i\in\mathcal{I}, n\in\mathcal{T} \tag{EC.2o}$$

$$\chi_{in} \in \{0,1\} \qquad \forall n \in \mathcal{B}, i \in \mathcal{I} \qquad \text{(EC.2p)}$$

$$\eta_{in} \in \mathbb{R}_+ \qquad \forall n \in \mathcal{T}, i \in \mathcal{I} \qquad \text{(EC.2q)}$$

$$\rho_n \in \mathbb{R}_+ \qquad \forall n \in \mathcal{T} \qquad \text{(EC.2r)}$$

$$g_i \in \mathbb{R}_+ \qquad \forall n \in \mathcal{T}. \qquad \text{(EC.2s)}$$

The objective function (EC.2a) now penalizes high variance. All constraints remain the same with the exception of (EC.2k) and (EC.2l), which force g_i to be β_{nK_i} when datapoint i lands in terminal node n. There are no additional constraints that define β_{nk} because the optimal solution for $\min_{\beta_{nk}} \sum_{i=1}^{m}(a_i - \beta_{nk})^2$ is $\beta^*_{nk} = \sum_{i=1}^{m} \frac{a_i}{m}$, where $\{a_1, \ldots, a_m\}$ are datapoints in terminal node n that were assigned treatment k in the data.

EC.3. Analysis of K-PT and B-PT

We present simple examples where K-PT and B-PT fail to predict the optimal prescriptive tree.

Example EC.1. Consider the following data distribution. Let the covariate vector $X = (X^1, X^2) \in \{0,1\}^2$, where X^1 and X^2 are independent and Bernoulli distributed with parameter 0.5. We interpret X^1 to denote the severity of a patient's condition, where $X^1 = 0$ (resp. $X^1 = 1$) indicates a healthy (resp. sick) patient. X^2 is a noise variable, unrelated to both the outcomes and the treatment assignments. There are two possible treatments, where $K = 1$ (resp. $K = 0$) indicates that the patient was treated (resp. not treated). Treatments are assigned according to a conditionally randomized experiment, with $\mathbb{P}(K = 0|X^1 = 0) = 0.9$ and $\mathbb{P}(K = 0|X^1 = 1) = 0.1$, i.e., 90% of healthy patients do not receive the treatment, and 90% of sick patients do. Thus, the assumptions of conditional exchangeability and positivity hold in this setting, see Section 1.2.1 for definitions and pointers to the relevant literature. The above numbers imply that $\mathbb{P}(K = 0) = \mathbb{P}(K = 1) = 0.5$ in the population. Table EC.1 shows the expected values of the potential outcomes in dependence of the covariate values. For simplicity, we assume that: *1)* the potential outcomes in Table EC.1 have zero variance, meaning there is no uncertainty in the outcomes conditional on X; and *2)* we have sufficient historical data at our disposal so that we can work with the true values in (8) rather than with their empirical estimates provided in the definition of $Q_{\mathcal{I}}^{\text{K-PT}}(\pi)$.

X^1	$\mathbb{E}(Y(0))$	$\mathbb{E}(Y(1))$
0	1	0.8
1	0	0.2

Table EC.1 Companion table to Example EC.1. Expectation of potential outcomes in dependence of the covariate values. A value $X^1 = 0$ (resp. $X^1 = 1$) of the covariate indicates a healthy (resp. sick) patient. Larger values of the outcomes are preferred.

From Table EC.1, it can be seen that the (unique) optimal policy treats sick patients only. In particular, this optimal policy can be modeled by a prescriptive tree of depth one (i.e., a tree with a single branching node and two leafs) that branches on X^1, does not treat any of the healthy patients (left leaf) and treats all sick patients (right leaf). The expected outcome under this policy is given by

$$\mathbb{E}(Y(0)|X^1=0)\mathbb{P}(X^1=0)+\mathbb{E}(Y(1)|X^1=1)\mathbb{P}(X^1=1)=1\times\frac{1}{2}+0.2\times\frac{1}{2}=0.6.$$

One may similarly calculate the expected outcomes under a policy that branches on the noise feature X^2. In this case, no matter which treatment is assigned at each leaf, the resulting policies would have a lower expected outcome of 0.5. These candidate policies and the corresponding expected outcomes under each policy are illustrated in the top row of Figure EC.1.

The optimal policies described above can be obtained by an omniscient decision-maker who knows the true counterfactuals $\mathbb{E}(Y(k))$. In practice, however, decision-makers can only obtain $\mathbb{E}(Y|K=k)$ from the observational data. Naturally, incorrect estimates could potentially lead to poor decisions. In particular, we now argue that K-PT indeed produces suboptimal trees due to estimation bias. Consider the tree of depth one that maximizes the objective function of K-PT given by (8). Suppose that we branch on covariate X^j at the sole branching node of the tree, $j\in\{1,2\}$. Then, at the leaf where $X^j=x^j\in\{0,1\}$, K-PT estimates the counterfactual outcomes $\mathbb{E}(Y(k)|X^j=x^j)$ as $\mathbb{E}(Y|K=k,\ X^j=x^j)$. However, these estimates can be incorrect. For example,

$$\mathbb{E}(Y(0)|X^2=0)=\sum_{x^1\in\{0,1\}}\mathbb{E}(Y(0)|X^1=x^1,X^2=0)\mathbb{P}(X^1=x^1|X^2=0)=1\times\frac{1}{2}+0\times\frac{1}{2}=0.5,$$

whereas

$$\begin{aligned}\mathbb{E}(Y|K=0,X^2=0)&=\sum_{x^1\in\{0,1\}}\mathbb{E}(Y|K=0,X^1=x^1,X^2=0)\mathbb{P}(X^1=x^1|K=0,X^2=0)\\&=\sum_{x^1\in\{0,1\}}\mathbb{E}(Y|K=0,X^1=x^1)\frac{\mathbb{P}(K=0|X^1=x^1)\mathbb{P}(X^1=x^1)}{\mathbb{P}(K=0)}=0.9.\end{aligned}$$

The difference between these two quantities is intuitive. Individuals that were not treated ($K=0$) in the data are more likely to be healthy (with a large associated outcome). The K-PT estimate $\mathbb{E}(Y|K=0,X^2=0)$ thus *overestimates* $\mathbb{E}(Y(0)|X^2=0)=0.5$. Similarly, since the vast majority of individuals that were treated ($K=1$) in the data are sick (with low outcomes), the K-PT estimate $\mathbb{E}(Y|K=1,X^2=0)=0.26$ is an *underestimator* for $\mathbb{E}(Y(1)|X^2=0)=0.5$. Repeating this calculation for all combinations of K and X^2 gives us $\mathbb{E}(Y|K=0,X^2=0)=\mathbb{E}(Y|K=0,X^2=1)=\mathbb{E}(Y|K=0)$ and similarly for $\mathbb{E}(Y|K=1,X^2=0)=\mathbb{E}(Y|K=1,X^2=1)=\mathbb{E}(Y|K=1)$. Hence, when splitting on X^2, the K-PT estimator will choose to not treat anyone ($K=0$) because the estimator's expected values are $0.9>0.26$, see the middle right subfigure of Figure EC.1.

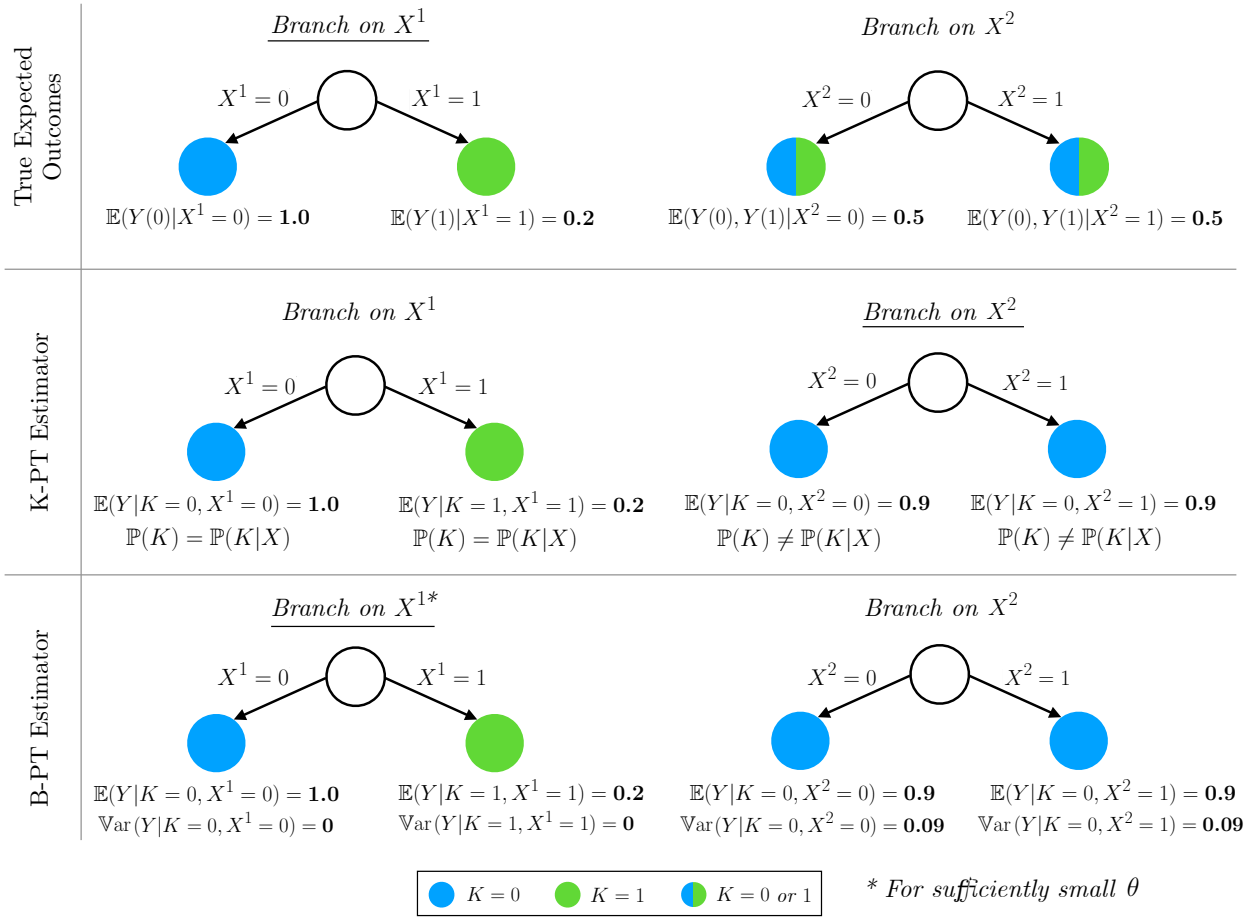

Figure EC.1 **Companion figure to Example EC.1. Each row shows the estimates of performance of different prescriptive trees (branch on X^1–left– or X^2–right–) under different estimators. In each case, only the best treatment assignments at each leaf (as measured by the corresponding estimator) are shown. In this example, the true optimal policy (top left) is to branch on X^1 and treat only those who are sick. Indeed, the expected outcome of the best tree branching on X^1 (top left) is $1.0 \cdot 0.5 + 0.2 \cdot 0.5 = 0.6$, which is larger than the expected outcome of the best X^2-branching tree (top right), which is $0.5 < 0.6$. However, the method from Kallus (2017) (middle), which is discussed in Section 5.1, will prefer to branch on X^2 (noise covariate) and not treat any patient (middle right). Indeed, it estimates the outcomes of the best X^1- (resp. X^2-) branching tree as 0.6 (resp. 0.9). We note that Assumption 4 is not satisfied at any of the leaves of the tree deemed as best by this estimator. In this case, the method of Bertsimas et al. (2019), see Section 5.2, is able to identify the truly optimal tree (bottom left); it correctly branches on X^1 provided θ is sufficiently small. Note that this method penalizes the variances corresponding to all treatments, but for simplicity the figure only depicts variances corresponding to the chosen treatment.**

We can conduct a similar analysis for the tree that branches on X^1. In this case, K-PT estimates $\mathbb{E}[Y|K = 0, X^1 = 0] = 1$ precisely as stated in Table EC.1, since the historical policy assigns treatments solely based on feature X^1. Repeating the decisions for all combinations of K and X^1 gives us a tree where K-PT will choose to treat $(K = 1)$ for those with $X^1 = 1$ and not treat otherwise, see the middle left subfigure of Figure EC.1. When choosing the optimal branching decision of the

two options, K-PT will **incorrectly** choose to branch on X^2 because it estimates a higher expected outcome ($(0.9+0.9)/2=0.9$) than when branching on X^1 ($(1.0+0.2)/2=0.6$).

Note that in a depth 1 tree that only branches on X^1, Assumption 4 is satisfied and the estimated outcome of 0.6 provided by K-PT is correct. However, K-PT prefers the alternative tree which branches on the noise variable and does not satisfy Assumption 4, in this case resulting in an overly optimistic estimation of the outcomes. We conclude that if not *all feasible trees* in the problem satisfy Assumption 4, the method from Kallus (2017) may fail.

B-PT suffers from the same problem. Figure EC.1 (bottom) shows the estimated outcomes and variances associated with each candidate tree. Since the tree that branches on the relevant feature X^1 has smaller variance in this case, it will be chosen by B-PT provided that θ is sufficiently small. ■

EXAMPLE EC.2. We now consider a variant of Example EC.1 where we add to all potential outcomes the quantity $(1-X^2)$, see Table EC.2. As before, $\mathbb{P}(K=0|X^1=0)=0.9$ and $\mathbb{P}(K=0|X^1=1)=0.1$, i.e., 90% of healthy patients do not receive the treatment, and 90% of sick patients do, also implying that $\mathbb{P}(K=0)=\mathbb{P}(K=1)=0.5$. Since X^2 now affects the values of the outcomes, we have $\mathbb{E}(Y|K=0,X^2=1)\neq\mathbb{E}(Y|K=0,X^2=0)$ and similarly for $K=1$. However, this change does *not* affect the value $\mathbb{E}(Y(1)-Y(0)|X)$ and is thus *irrelevant* in deciding which treatment is preferable. Therefore, under suitable identifiability conditions, there exists an optimal prescriptive tree of depth one, which still splits on X^1. Figure EC.2 (top) shows the outcomes, for a depth one tree, under the two possible branching decisions if the counterfactuals are known – the best tree indeed branches on the only predictive feature X^1. However, when using estimates $\mathbb{E}[Y|K=k,X=x]$ for the counterfactuals (Figure EC.2, bottom), the tree that branches on the irrelevant feature X^2 has both a higher estimated outcome and a smaller variance. Therefore, B-PT will always choose to branch on X^2 and not treat anyone in the population, no matter the choice of θ. In this case B-PT has even more incentive to select a policy based on the irrelevant feature and both K-PT and B-PT will select a suboptimal tree. ■

X^1	X^2	$\mathbb{E}(Y(0))$	$\mathbb{E}(Y(1))$
0	0	2	1.8
1	0	1	1.2
0	1	1	0.8
1	1	0	0.2

Table EC.2 Companion table to Example EC.2. Expectation of the potential outcomes in dependence of the covariate values. Similar to Example EC.1 (Continued), we let the potential outcomes in Table EC.1 have zero variance, implying that potential outcomes conditioned on the covariate values are deterministic ~~perfectly known~~.

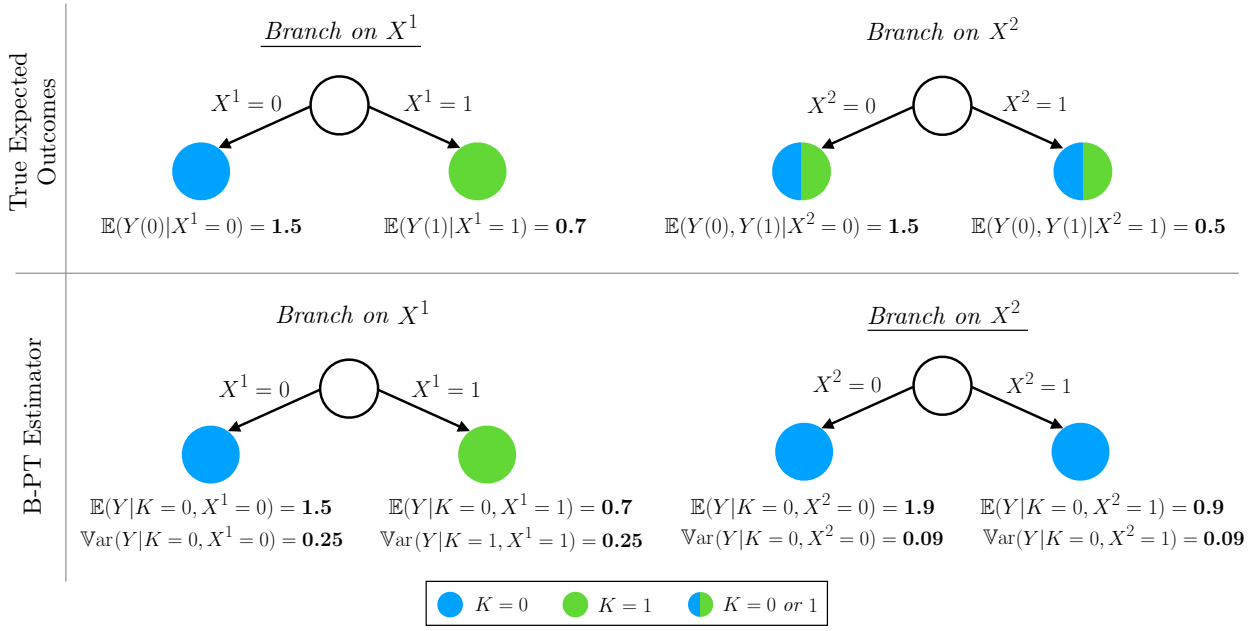

Figure EC.2 Companion figure to Example EC.2. The layout and interpretation of the figure exactly parallel that in Figure EC.1. The optimal tree (top left) has the objective $0.5(1.5) + 0.5(0.7) = 1.1$, which is higher than the X^2-branching tree (top right), $0.5(1.5) + 0.5(0.5) = 1.0 < 1.1$. However, B-PT will incorrectly branch on X^2 no matter the regularization strength θ, since the X^1-branching tree (bottom left) has a lower estimated value and also higher variance.

EC.4. Warfarin Dosage

We present an equation from Consortium (2009) that we use in our experiments in Section 6.4 to determine the optimal warfarin dose for a patient. In particular, it allows us to generate patient counterfactuals and evaluate the learned policies for our experiments. Note that VKORC1 and CYP2C9 denote genotypes. If W is the optimal weekly warfarin dosage, then the equation becomes:

$$\begin{aligned}\sqrt{W} = {} & 5.6044 - 0.2614 \times \text{Age in decades} + 0.0087 \times \text{Height in cm} + 0.0128 \times \text{Weight in kg} \\ & - 0.8677 \times \text{VKORC1 A/G} - 1.6974 \times \text{VKORC1 A/A} - 0.4854 \times \text{VKORC1 genotype unknown} \\ & - 0.5211 \times \text{CYP2C9*1/*2} - 0.9357 \times \text{CYP2C9*1/*3} - 1.0616 \times \text{CYP2C9*2/*2} \\ & - 1.9206 \times \text{CYP2C9*2/*3} - 2.3312 \times \text{CYP2C9*3/*3} - 0.2188 \times \text{CYP2C9 genotype unknown} \\ & - 0.1092 \times \text{Asian race} - 0.2760 \times \text{Black or African American} - 0.1032 \times \text{Missing or Mixed race} \\ & + 1.1816 \times \text{Enzyme inducer status} - 0.5503 \times \text{Amiodarone status}\end{aligned} \tag{EC.3}$$

EC.5. Experiment Results (Raw)

In this section, we provide tables that contain the raw results for the experiments summarized in Section 6.4.

Table EC.3 Companion table to the experiments on synthetic data in Section 6.4.1. The first and second number correspond to the average and standard deviation of the optimality gap, solving time, out-of-sample (OOS) regret, and OOS probability of correct treatment assignment (OOSP) across 5 random samples.

Depth	Method	Model	Gap	Solving Time (s)	OOS Regret	OOSP (%)
1	IPW	DT	0.00 ± 0.00	1.34 ± 0.22	123.18 ± 95.63	66.25 ± 15.84
1	IPW	Log	0.00 ± 0.00	1.31 ± 0.31	138.12 ± 114.65	63.89 ± 18.00
1	DM	LR	0.00 ± 0.00	0.76 ± 0.40	64.22 ± 38.06	75.28 ± 9.87
1	DM	Lasso	0.00 ± 0.00	0.57 ± 0.09	200.68 ± 72.26	53.44 ± 11.00
1	DR	DT, LR	0.00 ± 0.00	0.86 ± 0.46	65.76 ± 40.68	75.01 ± 10.13
1	DR	DT, Lasso	0.00 ± 0.00	2.08 ± 0.82	77.04 ± 70.23	73.70 ± 12.83
1	DR	Log, LR	0.00 ± 0.00	0.95 ± 0.56	78.78 ± 55.88	72.98 ± 11.56
1	DR	Log, Lasso	0.00 ± 0.00	1.79 ± 1.01	111.47 ± 68.44	67.16 ± 11.88
1	K-PT	-	0.00 ± 0.00	1.97 ± 0.76	161.83 ± 103.62	60.10 ± 17.46
1	B-PT	-	0.00 ± 0.00	6.28 ± 1.36	179.04 ± 106.95	57.62 ± 17.63
1	PT	DT, LR	0.00 ± 0.00	0.01 ± 0.00	62.01 ± 37.00	75.25 ± 10.29
-	CF	-	0.00 ± 0.00	0.75 ± 0.04	76.03 ± 57.20	73.13 ± 12.05
-	CT	-	0.00 ± 0.00	0.29 ± 0.02	92.20 ± 74.63	71.53 ± 13.64
-	R&C	LR	0.00 ± 0.00	0.23 ± 0.00	51.81 ± 32.87	77.11 ± 10.09

Table EC.4 Companion table to the experiments on synthetic data in Section 6.4.2. The first and second number correspond to the average and standard deviation of the optimality gap, solving time, out-of-sample (OOS) regret, and OOS probability of correct treatment assignment (OOSP) across 25 random samples.

Depth	Method	Model	Gap	Solving Time (s)	OOS Regret	OOSP (%)
2	IPW	DT	0.00 ± 0.00	255.45 ± 65.13	317.39 ± 82.90	77.10 ± 5.98
2	DM	RF/Log	0.00 ± 0.00	192.03 ± 64.13	288.79 ± 88.46	79.16 ± 6.38
2	DR	DT, RF/Log	0.00 ± 0.00	284.58 ± 83.83	279.32 ± 87.00	79.85 ± 6.28
2	K-PT	-	0.00 ± 0.03	5435.72 ± 2792.71	522.12 ± 264.43	62.33 ± 19.08
2	B-PT	-	52.30 ± 43.78	14333.53 ± 674.39	601.39 ± 239.26	56.61 ± 17.26
2	PT	DT, RF/Log	0.00 ± 0.00	2.35 ± 0.03	257.73 ± 76.39	81.40 ± 5.51
3	IPW	DT	13.56 ± 4.15	14405.89 ± 0.50	246.35 ± 48.07	82.23 ± 3.47
3	DM	RF/Log	2.10 ± 1.05	14406.33 ± 0.55	252.92 ± 69.22	81.75 ± 4.99
3	DR	DT, RF/Log	8.08 ± 2.38	14406.65 ± 0.70	243.00 ± 78.18	82.47 ± 5.64
4	IPW	DT	9.51 ± 2.91	14412.07 ± 1.17	221.00 ± 39.47	84.05 ± 2.85
4	DM	RF/Log	1.48 ± 0.73	14412.46 ± 1.05	243.23 ± 72.31	82.45 ± 5.22
4	DR	DT, RF/Log	6.26 ± 1.90	14413.10 ± 1.57	210.95 ± 48.01	84.78 ± 3.46
-	CF	-	0.00 ± 0.00	3.67 ± 0.19	304.47 ± 94.22	78.03 ± 6.80
-	CF (untuned)	-	0.00 ± 0.00	3.75 ± 0.33	542.52 ± 334.07	60.86 ± 24.10
-	CT	-	0.00 ± 0.00	2.31 ± 0.32	459.35 ± 223.37	66.86 ± 16.12
-	R&C	Best	0.00 ± 0.00	0.95 ± 0.19	225.89 ± 163.98	83.70 ± 11.83
-	R&C	Log	0.00 ± 0.00	0.12 ± 0.01	261.47 ± 198.86	81.14 ± 14.35
-	R&C	RF	0.00 ± 0.00	0.83 ± 0.03	253.59 ± 125.24	81.70 ± 9.04

Table EC.5 Mean and standard deviation statistics over 5 random samples in the synthetic experiments, with a constraint that treatment $K=1$ is available to at most C% in the population.

Depth	Method	Model	Budget Range (C)	Gap	Solving Time (s)	OOS Regret	OOSP (%)
1	DR	LR, DT	0.05-0.09	0.00 ± 0.00	0.16 ± 0.03	158.32 ± 49.39	58.35 ± 9.03
1	DR	LR, DT	0.10-0.14	0.00 ± 0.00	0.22 ± 0.13	132.80 ± 43.61	61.54 ± 9.04
1	DR	LR, DT	0.15-0.19	0.00 ± 0.00	0.23 ± 0.10	113.91 ± 44.06	65.44 ± 9.48
1	DR	LR, DT	0.20-0.24	0.00 ± 0.00	0.20 ± 0.11	85.58 ± 32.88	70.01 ± 8.02
1	DR	LR, DT	0.25-0.29	0.00 ± 0.00	0.23 ± 0.13	77.34 ± 32.68	72.25 ± 8.12
1	DR	LR, DT	0.30-0.34	0.00 ± 0.00	0.23 ± 0.13	69.97 ± 34.71	73.71 ± 8.56
1	DR	LR, DT	0.35-0.40	0.00 ± 0.00	0.23 ± 0.14	66.32 ± 36.18	75.04 ± 8.86
1	DR	LR, Log	0.05-0.09	0.00 ± 0.00	0.19 ± 0.11	163.39 ± 46.47	57.67 ± 8.51
1	DR	LR, Log	0.10-0.14	0.00 ± 0.00	0.22 ± 0.15	139.03 ± 41.97	60.64 ± 8.44
1	DR	LR, Log	0.15-0.19	0.00 ± 0.00	0.23 ± 0.13	119.83 ± 45.13	64.53 ± 9.18
1	DR	LR, Log	0.20-0.24	0.00 ± 0.00	0.22 ± 0.16	93.20 ± 40.82	68.85 ± 8.22
1	DR	LR, Log	0.25-0.29	0.00 ± 0.00	0.24 ± 0.16	85.11 ± 42.53	70.99 ± 8.61
1	DR	LR, Log	0.30-0.34	0.00 ± 0.00	0.24 ± 0.16	77.97 ± 44.93	72.40 ± 9.07
1	DR	LR, Log	0.35-0.40	0.00 ± 0.00	0.26 ± 0.18	74.07 ± 46.99	73.80 ± 9.65
1	DR	Lasso, DT	0.05-0.09	0.00 ± 0.00	0.29 ± 0.13	172.73 ± 59.95	56.89 ± 10.02
1	DR	Lasso, DT	0.10-0.14	0.00 ± 0.00	0.33 ± 0.17	150.23 ± 63.71	60.15 ± 10.97
1	DR	Lasso, DT	0.15-0.19	0.00 ± 0.00	0.37 ± 0.19	129.98 ± 70.03	63.92 ± 12.14
1	DR	Lasso, DT	0.20-0.24	0.00 ± 0.00	0.46 ± 0.24	110.86 ± 71.28	67.15 ± 12.24
1	DR	Lasso, DT	0.25-0.29	0.00 ± 0.00	0.47 ± 0.27	98.02 ± 64.04	69.78 ± 11.28
1	DR	Lasso, DT	0.30-0.34	0.00 ± 0.00	0.46 ± 0.30	84.12 ± 60.11	72.13 ± 10.78
1	DR	Lasso, DT	0.35-0.40	0.00 ± 0.00	0.41 ± 0.23	79.02 ± 64.46	73.62 ± 11.83
1	DR	Lasso, Log	0.05-0.09	0.00 ± 0.00	0.27 ± 0.14	180.83 ± 65.74	55.80 ± 10.47
1	DR	Lasso, Log	0.10-0.14	0.00 ± 0.00	0.29 ± 0.16	164.75 ± 74.07	57.94 ± 11.64
1	DR	Lasso, Log	0.15-0.19	0.00 ± 0.00	0.34 ± 0.18	148.86 ± 75.23	60.86 ± 12.30
1	DR	Lasso, Log	0.20-0.24	0.00 ± 0.00	0.34 ± 0.20	130.36 ± 73.55	63.94 ± 12.10
1	DR	Lasso, Log	0.25-0.29	0.00 ± 0.00	0.40 ± 0.25	124.98 ± 73.15	65.19 ± 12.03
1	DR	Lasso, Log	0.30-0.34	0.00 ± 0.00	0.36 ± 0.23	117.06 ± 76.89	66.58 ± 12.74
1	DR	Lasso, Log	0.35-0.40	0.00 ± 0.00	0.36 ± 0.22	114.78 ± 76.44	67.13 ± 12.96

Table EC.6 Mean and standard deviation statistics over 25 random samples in the warfarin experiments, with a constraint that the difference between estimated outcomes between White and non-White patients is at most δ.

Depth	Method	Model	Fairness (δ)	Gap	Solving Time (s)	Est. Outcome Disparity	Actual Disparity	OOS Regret	OOSP (%)
2	DM	RF/Log	0.01	0.00 ± 0.00	316.31 ± 112.04	-0.01 ± 0.01	0.01 ± 0.10	326.02 ± 106.02	0.76 ± 0.08
2	DM	RF/Log	0.02	0.00 ± 0.00	284.67 ± 125.32	-0.02 ± 0.01	-0.00 ± 0.10	321.20 ± 107.86	0.77 ± 0.08
2	DM	RF/Log	0.03	0.00 ± 0.00	290.46 ± 130.17	-0.02 ± 0.02	-0.01 ± 0.10	312.77 ± 105.23	0.77 ± 0.08
2	DM	RF/Log	0.04	0.00 ± 0.00	280.67 ± 116.49	-0.03 ± 0.02	-0.02 ± 0.09	303.54 ± 99.35	0.78 ± 0.07
2	DM	RF/Log	0.05	0.00 ± 0.00	271.45 ± 123.14	-0.03 ± 0.03	-0.03 ± 0.09	297.82 ± 96.61	0.79 ± 0.07
2	DM	RF/Log	0.06	0.00 ± 0.00	268.73 ± 151.18	-0.03 ± 0.03	-0.03 ± 0.09	294.58 ± 93.16	0.79 ± 0.07
2	DM	RF/Log	0.07	0.00 ± 0.00	241.31 ± 111.07	-0.04 ± 0.04	-0.04 ± 0.09	290.84 ± 92.17	0.79 ± 0.07
2	DM	RF/Log	0.08	0.00 ± 0.00	234.19 ± 105.39	-0.04 ± 0.04	-0.04 ± 0.09	290.78 ± 92.97	0.79 ± 0.07

EC.6. Policytree Overfitting

In Figure EC.3, we illustrate that Policytree (PT) – despite optimizing the same objective as our doubly robust method (DR) – leads to worse out-of-sample performance because it learns from continuous features (as opposed to our method, which takes in discrete features). While PT's performance on the training set is more often than not better than DR, it does not generalize well on the testing set; in contrast, our method maintains similar performance between the training and testing sets.

EC.7. Scale of Proposed MIO

Table EC.7 shows the growth of the number of constraints, continuous variables, and binary variables of formulations (5), (6), and (7) without the additional constraints described in Section 4. This

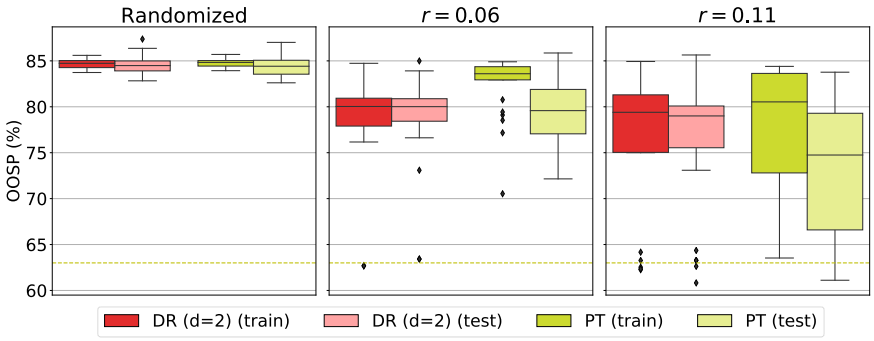

Figure EC.3 A comparison of the doubly robust method's performance to the formulation proposed by Zhou et al. (2023) (PT), on the training and testing sets. Both methods are over trees of depth $d = 2$.

growth is a function of the characteristics of the observational data (number of treatment options, size of covariates, etc.) and tree depth. It also assumes that we relax the integrality requirement on z and w, see Observation 1.

Table EC.7 Summary of growth of the number of constraints and decision variables in formulations (5), (6), and (7).

Variant	# Constraints	# Continuous vars.	# Binary vars.
Integer w and z	$\mathcal{O}(2^d\|\mathcal{K}\|\|\mathcal{X}^{\mathcal{I}}\|)$	N/A	$\mathcal{O}(2^{d+1}\|\mathcal{K}\|\|\mathcal{X}^{I}\|) + \mathcal{O}(2^d\|\mathcal{F}\|\max_f \Theta(f))$
Relaxing w and z	$\mathcal{O}(2^d\|\mathcal{K}\|\|\mathcal{X}^{\mathcal{I}}\|)$	$\mathcal{O}(2^{d+1}\|\mathcal{K}\|\|\mathcal{X}^{\mathcal{I}}\|)$	$\mathcal{O}(2^d\|\mathcal{F}\|\max_f \Theta(f))$

EC.8. Proofs

Proof of Proposition 1. We introduce additional notation to help formalize our claim. For any tree based policy $\pi \in \Pi_d$, define

$$Q^{\mathrm{IPW}}_{\mathcal{N}}(\pi) := \mathbb{E}\left[\frac{\mathbb{I}(K = \pi(X))Y}{\hat{\mu}^{\mathcal{N}}(K,X)}\right].$$

To simplify the analysis, without loss of generality, we assume that the potential outcomes have been normalized, ensuring that $|Y(k)| \le 1$ holds for all $k \in \mathcal{K}$. We note that if all policies $\pi \in \Pi_d$ are optimal, the statements in the proposition follow immediately, and thus henceforth focus on the case where $\Pi_d \backslash \mathcal{S}^\star$ is not empty.

We begin by showing that $\hat{v}^{\mathrm{IPW}}_{\mathcal{N},\mathcal{I}} \to v^\star$ w.p.1. Fix $\epsilon > 0$ and a policy $\pi \in \Pi_d$. Since μ and $\hat{\mu}^{\mathcal{N}}$ are bounded away from 0, $\exists m_0 > 0$ and $N_0 \in \mathbb{Z}_+$ such that

$$\mu(K,X) > m_0 \text{ and } \hat{\mu}^{\mathcal{N}}(K,X) > m_0 \quad \forall K \in \mathcal{K},\ X \in \mathcal{X},\ \text{and } N \ge N_0. \tag{EC.4}$$

Define the function $g : (m_0, \infty) \to \mathbb{R}$ through $g(x) := \frac{1}{x}$. Since g is uniformly continuous on the interval (m_0, ∞), equation (EC.4) implies that there exists $\delta_1 > 0$ such that

$$\left|\hat{\mu}^{\mathcal{N}}(K,X) - \mu(K,X)\right| < \delta_1 \quad \Longrightarrow \quad \left|\frac{1}{\hat{\mu}^{\mathcal{N}}(K,X)} - \frac{1}{\mu(K,X)}\right| < \frac{\epsilon}{2} \quad \forall K \in \mathcal{K},\ X \in \mathcal{X}. \tag{EC.5}$$

Moreover, since $\hat{\mu}^{\mathcal{N}}$ converges almost surely to μ and the fact that the support set of μ is finite, it follows that there exists $N^\pi \geq N_0$ such that

$$\max_{X,K} \left|\hat{\mu}^{\mathcal{N}}(K,X) - \mu(K,X)\right| < \delta_1 \quad \text{w.p.1} \quad \forall N \geq N^\pi. \tag{EC.6}$$

Therefore, it follows from equations (EC.5) and (EC.6) that

$$\max_{X,K} \left|\frac{1}{\hat{\mu}^{\mathcal{N}}(K,X)} - \frac{1}{\mu(K,X)}\right| < \frac{\epsilon}{2} \quad \text{w.p.1} \quad \forall N \geq N^\pi.$$

We then have

$$\begin{aligned}\sup_{X,K,Y} \left|\frac{\mathbb{I}(K=\pi(X))Y}{\hat{\mu}^{\mathcal{N}}(K,X)} - \frac{\mathbb{I}(K=\pi(X))Y}{\mu(K,X)}\right| &\leq \max_{X,K} \left|\frac{\mathbb{I}(K=\pi(X))}{\hat{\mu}^{\mathcal{N}}(K,X)} - \frac{\mathbb{I}(K=\pi(X))}{\mu(K,X)}\right| \\ &< \frac{\epsilon}{2} \quad \text{w.p.1} \quad \forall N \geq N^\pi,\end{aligned}$$

where the supremum and maximum above are taken over $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, $K \in \mathcal{K}$. The first inequality above follows since $|Y| \leq 1$ and implies that

$$\begin{aligned}&\mathbb{E}\left(\frac{\mathbb{I}(K=\pi(X))Y}{\mu(K,X)} - \frac{\mathbb{I}(K=\pi(X))Y}{\hat{\mu}^{\mathcal{N}}(K,X)}\right) \\ &= \sum_{x,k,y}\left(\frac{\mathbb{I}(k=\pi(x))y}{\mu(k,x)} - \frac{\mathbb{I}(k=\pi(x))y}{\hat{\mu}^{\mathcal{N}}(k,x)}\right)\mathbb{P}(X=x,K=k,Y=y) \\ &< \frac{\epsilon}{2} \quad \text{w.p.1} \quad \forall N \geq N^\pi.\end{aligned} \tag{EC.7}$$

Note that in the equation above, for the sake of clarity and ease of notation, we employ summation instead of integration. By definition of Q^{IPW} and $Q^{\text{IPW}}_{\mathcal{N}}$ and from our conditional exchangeability assumption, see Hernán and Robins (2019), it then follows that

$$\left|Q^{\text{IPW}}(\pi) - Q^{\text{IPW}}_{\mathcal{N}}(\pi)\right| = \left|Q(\pi) - Q^{\text{IPW}}_{\mathcal{N}}(\pi)\right| < \frac{\epsilon}{2} \quad \text{w.p.1} \quad \forall N \geq N^\pi. \tag{EC.8}$$

At the same time, by the strong law of large numbers, there exists I^π such that

$$\left|Q^{\text{IPW}}_{\mathcal{N},\mathcal{I}}(\pi) - Q^{\text{IPW}}_{\mathcal{N}}(\pi)\right| < \frac{\epsilon}{2} \quad \text{w.p.1} \quad \forall I \geq I^\pi. \tag{EC.9}$$

Then, equations (EC.8) and (EC.9) imply that

$$\left|Q^{\text{IPW}}_{\mathcal{N},\mathcal{I}}(\pi) - Q(\pi)\right| \leq \epsilon \quad \text{w.p.1} \quad \forall N \geq N^\pi \text{ and } I \geq I^\pi. \tag{EC.10}$$

Since the set Π_d of decision trees of depth at most d is finite, and the union of a finite number of sets each of measure zero also has measure zero, the pointwise convergence given in equation (EC.10) also results in uniform convergence, meaning that, there exist $N_1 := \max_{\pi\in\Pi_d} N^\pi$ and $I_0 := \max_{\pi\in\Pi_d} I^\pi$ such that

$$\max_{\pi\in\Pi_d}\left|Q^{\text{IPW}}_{\mathcal{N},\mathcal{I}}(\pi) - Q(\pi)\right| \leq \epsilon \quad \text{w.p.1} \quad \forall I \geq I_0 \text{ and } N \geq N_1.$$

Since $|\hat{v}^{\mathrm{IPW}}_{\mathcal{N},\mathcal{I}} - v^\star| \leq \epsilon$ and the choice of ϵ was arbitrary, then w.p.1 $\hat{v}^{\mathrm{IPW}}_{\mathcal{N},\mathcal{I}} \to v^\star$ for all $I \geq I_0$ and $N \geq N_1$. This completes the first part of the proof.

We now show that $\hat{\mathcal{S}}^{\mathrm{IPW}}_{\mathcal{N},\mathcal{I}} \subseteq \mathcal{S}^\star$ w.p.1 for I and N large enough. Consider the difference in objective value between the best and second best policies, given by

$$\rho := \max_{\pi \in \Pi_d \setminus \mathcal{S}^\star} Q(\pi) - v^\star.$$

Since for any $\pi \in \Pi_d \setminus \mathcal{S}^\star$ it holds that $Q(\pi) < v^\star$ and since the set Π_d is finite, it follows that $\rho < 0$. Let I and N be large enough such that $\epsilon_{\mathcal{N},\mathcal{I}} < -\frac{\rho}{2}$. Then, w.p.1, $\hat{v}^{\mathrm{IPW}}_{\mathcal{N},\mathcal{I}} > v^\star + \frac{\rho}{2}$, and for any $\pi \in \Pi_d \setminus \mathcal{S}^\star$ it holds that $Q^{\mathrm{IPW}}_{\mathcal{N},\mathcal{I}}(\pi) < v^\star + \frac{\rho}{2}$. It follows that if $\pi \in \Pi_d \setminus \mathcal{S}^\star$, then $Q^{\mathrm{IPW}}_{\mathcal{N},\mathcal{I}} < \hat{v}^{\mathrm{IPW}}_{\mathcal{N},\mathcal{I}}$ and hence π does not belong to set $\hat{\mathcal{S}}^{\mathrm{IPW}}_{\mathcal{N},\mathcal{I}}$. As a result, for large enough I and N, w.p.1 the inclusion $\hat{\mathcal{S}}^{\mathrm{IPW}}_{\mathcal{N},\mathcal{I}} \subseteq \mathcal{S}^\star$ holds, which completes the proof. □

We make use of the following auxiliary lemma in the proof of Proposition 2.

LEMMA EC.1 (**Bartle (2001)**). *Suppose $f_n : D \to \mathbb{R}$ and $g_n : D \to \mathbb{R}$ are sequences of functions which converge uniformly to $f, g : D \to \mathbb{R}$, respectively. Then, if f and g are bounded, i.e., $\exists B > 0$ such that $|f(x)| < B$ and $|g(x)| < B$, $f_n g_n$ converges uniformly to fg.*

Proof of Proposition 2. We introduce additional notation to help formalize our claim. For any tree based policy $\pi \in \Pi_d$, define

$$Q^{\mathrm{DR}}_{\mathcal{N}}(\pi) := \mathbb{E}\left[\hat{\nu}^{\mathcal{N}}_{\pi(X)}(X) + (Y - \hat{\nu}^{\mathcal{N}}_{\pi(X)}(X))\frac{\mathbb{I}(K = \pi(X))}{\hat{\mu}^{\mathcal{N}}(K,X)}\right].$$

Since $\hat{\mu}(K,X)$ and $\hat{\nu}_K(X)$ are bounded, there exists $B > 0$ such that $|\hat{\mu}(K,X)| < B$ and $|\hat{\nu}_K(X)| < B \quad \forall X \in \mathcal{X}$ and $K \in \mathcal{K}$. Fix $\epsilon > 0$ and a policy $\pi \in \Pi_d$. From the definition of $\hat{\mu}$, uniform continuity of the inverse function (bounded away from 0), and the assumption that $\mathcal{X}$ is finite and $\mathcal{Y}$ is bounded, $\exists N^\pi \in \mathbb{Z}_+$ such that

$$\begin{aligned}\sup_{X,K,Y}\left|\frac{\mathbb{I}(K=\pi(X))Y}{\hat{\mu}^{\mathcal{N}}(X,K)} - \frac{\mathbb{I}(K=\pi(X))Y}{\hat{\mu}(X,K)}\right| &< \max_{X,K}\left|\frac{\mathbb{I}(K=\pi(X))}{\hat{\mu}^{\mathcal{N}}(X,K)} - \frac{\mathbb{I}(K=\pi(X))}{\hat{\mu}(K,X)}\right| \\ &< \frac{\epsilon}{4(B+1)} \quad \forall N \geq N^\pi.\end{aligned} \tag{EC.11}$$

Also, by definition of $\hat{\nu}$, $\exists N_0 \in \mathbb{Z}_+$ such that

$$\max_{X,K}\left|\hat{\nu}^{\mathcal{N}}_K(X) - \hat{\nu}_K(X)\right| < \frac{\epsilon}{4(B+1)} \quad \forall N \geq N_0. \tag{EC.12}$$

Thus, using (EC.11), (EC.12), and the assumption that $\hat{\nu}$ and $\hat{\mu}(K,X)$ are bounded, by Lemma EC.1, $\exists N^\pi_1 \geq N^\pi, N_0$ such that

$$\max_{X,K}\left|\frac{\mathbb{I}(K=\pi(X))\hat{\nu}^{\mathcal{N}}_K(X)}{\hat{\mu}^{\mathcal{N}}(K,X)} - \frac{\mathbb{I}(K=\pi(X))\hat{\nu}_K(X)}{\hat{\mu}(K,X)}\right| < \frac{2\epsilon B}{4(B+1)} \quad \forall N \geq N^\pi_1. \tag{EC.13}$$

Moreover, from (EC.11), (EC.12), and (EC.13), we can conclude that for $N \geq N_1^\pi$,

$$\sup_{X,K,Y} \left| \left\{ \hat{\nu}^{\mathcal{N}}_{\pi(X)}(X) + (Y - \hat{\nu}^{\mathcal{N}}_{\pi(X)}(X)) \frac{\mathbb{I}(K=\pi(X))}{\hat{\mu}^{\mathcal{N}}(K,X)} \right\} - \left\{ \hat{\nu}_{\pi(X)}(X) + (Y - \hat{\nu}_{\pi(X)}(X)) \frac{\mathbb{I}(K=\pi(X))}{\hat{\mu}(K,X)} \right\} \right| < \frac{\epsilon}{2}.$$

Similar to the proof of Proposition 1, for a given policy $\pi \in \Pi_d$, for $N \geq N_1^\pi$,

$$\left| Q^{\mathrm{DR}}(\pi) - Q^{\mathrm{DR}}_{\mathcal{N}}(\pi) \right| < \frac{\epsilon}{2}.$$

Finally, we know that if w.p.1, either $\hat{\mu}(k,x) = \mu(k,x)$ or $\hat{\nu}_k(x) = \nu_k(x)$, for all $x \in \mathcal{X}$ and $k \in \mathcal{K}$, then $Q(\pi) = Q^{\mathrm{DR}}(\pi)$, see for example Lunceford and Davidian (2004). Since the assumption in the previous clause is satisfied w.p.1 (as a result of the almost sure convergence of either $\hat{\mu}^{\mathcal{N}}$ or $\hat{\nu}^{\mathcal{N}}_K$), we conclude that,

$$\left| Q(\pi) - Q^{\mathrm{DR}}_{\mathcal{N}}(\pi) \right| < \frac{\epsilon}{2} \quad \text{w.p.1} \quad \forall N \geq N_1^\pi.$$

The rest of the proof is omitted as it can be derived by following the same logic as in the proof of Proposition 1. □

Proof of Proposition 3. We omit the proof due to its similarity to the proof of Proposition 2. □