Supplementary Material for:
Recursive Partitioning for Personalization using Observational Data

Nathan Kallus [1]

Omitted Proofs

Proof of Theorem 1. By Asn. 1, we have

$$\begin{aligned}
\mathbb{E}[Y \mid X=x, T=t] &= \mathbb{E}[Y(T) \mid X=x, T=t] && \text{(definition of } Y=Y(T)\text{)}\\
&= \mathbb{E}[Y(t) \mid X=x, T=t] && \text{(conditioned on } T=t\text{)}\\
&= \mathbb{E}[Y(t) \mid X=x] && \text{(Asn. 1).}
\end{aligned}$$

Consider a realization of the data and $X=x$ where convergence occurs for all $t \in [m]$. Let

$$\epsilon(x) = \inf\{\zeta : s \in [m],\, \zeta = \left(\mathbb{E}[Y \mid X=x, T=s] - \min_{t\in[m]} \mathbb{E}[Y \mid X=x, T=t]\right) > 0\},$$

where $\inf(\varnothing) = \infty$. By assumption of convergence at this realization of the data and $X=x$, we have that eventually for all $t \in [m]$, $|\hat{\mu}_{t,n_t}(x) - \mathbb{E}[Y \mid X=x, T=t]| < \epsilon(x)/2$, at which point we must necessarily also have $\hat{\tau}_n(x) \in \arg\min_{t\in[m]} \mathbb{E}[Y \mid X=x, T=t] = \arg\min_{t\in[m]} \mathbb{E}[Y(t) \mid X=x]$. By assumption of pointwise consistency and because the intersection of finitely many a.s. events is a.s., the set of such realization of the data and $X=x$ have probability 1. □

Proof of Theorem 2. First note that, given any x with $\mathbb{P}(T=t \mid X=x) > 0$, we have

$$\mathbb{E}[Y \mid X=x, T=t] = \frac{\mathbb{E}[Y\mathbb{I}[T=t] \mid X=x]}{\mathbb{P}(T=t \mid X=x)} = \mathbb{E}\left[\frac{Y\mathbb{I}[T=t]}{\phi(t,x)} \mid X=x\right] = \mathbb{E}\left[\frac{Y\mathbb{I}[T=t]}{\phi(T,X)} \mid X=x\right] = \mathbb{E}\left[\frac{Y\mathbb{I}[T=t]}{Q} \mid X=x\right].$$

Therefore, since $\mathbb{P}(T=t \mid X) > 0$ almost surely,

$$\begin{aligned}
R(\tau) = \mathbb{E}[Y(\tau(X))] &= \mathbb{E}[\mathbb{E}[Y(\tau(X)) \mid X]] && \text{(iterated expectations)}\\
&= \mathbb{E}[\mathbb{E}[Y(\tau(X)) \mid X, T=\tau(X)]] && \text{(Asn. 1)}\\
&= \mathbb{E}[\mathbb{E}[Y \mid X, T=\tau(X)]] && \text{(definition of } Y\text{)}\\
&= \mathbb{E}[\mathbb{E}[Y\mathbb{I}[T=\tau(X)]/Q \mid X]] && \text{(above observation)}\\
&= \mathbb{E}[Y\mathbb{I}[T=\tau(X)]/Q] && \text{(iterated expectations)}.
\end{aligned}$$
□

Proof of Theorem 4. We start with 1vA. Restrict to x such that $\phi(s,x) > 0\ \forall s$ (almost everywhere). Let $\mu(t,x) = \mathbb{E}[Y(t) \mid X=x]$. Under Asn. 1,

$$\begin{aligned}
\delta^{\text{tvA}}(x) &= \mathbb{E}[Y \mid X=x, T=t] - \mathbb{E}[Y \mid X=x, T\neq t]\\
&= \mathbb{E}[Y \mid X=x, T=t] - \textstyle\sum_{s\neq t} \mathbb{E}[Y \mid X=x, T=s]\,\mathbb{P}(T=s \mid X=x, T\neq t)\\
&= \mu(t,x) - \textstyle\sum_{s\neq t} \phi(s,x)\mu(s,x) / \sum_{s\neq t} \phi(s,x).
\end{aligned}$$

[1] School of Operations Research and Information Engineering and Cornell Tech, Cornell University. Correspondence to: Nathan Kallus <kallus@cornell.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Since $\phi(s,x)>0$, it's clear that $\delta^{t\text{vA}}(x)\le\delta^{s\text{vA}}(x)\ \forall s$ if and only if $\mu(t,x)\le\mu(s,x)\ \forall s$. The rest of the proof for 1vA follows the same way as Thm. 1, showing that, under the assumption of pointwise consistent estimation, the estimation gap $\sup_{t\in[m]}\left|\hat\delta_n^{t\text{vA}}(x)-\delta^{t\text{vA}}(x)\right|$ is eventually smaller than half the decision gap, $\epsilon^{1\text{vA}}(x)=\inf\{\zeta: s\in[m],\ \zeta=\left(\delta^{s\text{vA}}(x)-\min_{t\in[m]}\delta^{t\text{vA}}(x)\right)>0\}$, a.s. and for almost everywhere x.

Next, we deal with 1v1-A. Fix x. Fix any $t_m\in\arg\max_{t\in[m]}\mu(t,x)$. Let $\delta^{t\text{vmin}}(x)=\min_{s\neq t}\delta^{t\text{v}s}(x)$. If $t,s\neq t_m$, then $\delta^{t\text{vmin}}(x)-\delta^{s\text{vmin}}(x)=\mu(t,x)-\mu(s,x)$. On the other hand, for any $t\in[m]$, we always have both $\mu(t,x)-\mu(t_m,x)\le 0$ and $\delta^{t\text{vmin}}(x)-\delta^{t_m\text{vmin}}(x)\le 0$. Therefore, we have

$$\begin{aligned} t\in\arg\min_{t\in[m]}\mu(t,x) &\iff \mu(t,x)-\mu(s,x)\le 0\ \forall s\neq t \iff \mu(t,x)-\mu(s,x)\le 0\ \forall s\neq t,t_m\\ &\iff \delta^{t\text{vmin}}(x)-\delta^{s\text{vmin}}(x)\le 0\ \forall s\neq t,t_m \iff \delta^{t\text{vmin}}(x)-\delta^{s\text{vmin}}(x)\le 0\ \forall s\neq t\\ &\iff t\in\arg\min_{t\in[m]}\delta^{t\text{vmin}}(x).\end{aligned}$$

Let $\hat\delta_n^{t\text{vmin}}(x)=\min_{s\neq t}\hat\delta_{n_t+n_s}^{t\text{v}s}(x)$ and note that $\sup_{t\in[m]}\left|\hat\delta_n^{t\text{vmin}}(x)-\delta^{t\text{vmin}}(x)\right|\le\sup_{t\in[m],s\in[m]}\left|\hat\delta_{n_t+n_s}^{t\text{v}s}(x)-\delta^{t\text{v}s}(x)\right|$, which converges to zero under pointwise consistency. The rest of the proof for 1v1-A follows as above, showing that this estimation gap is eventually smaller than half the decision gap, $\epsilon^{1\text{v}1\text{-A}}(x)=\inf\{\zeta: s\in[m],\ \zeta=\left(\delta^{s\text{vmin}}(x)-\min_{t\in[m]}\delta^{t\text{vmin}}(x)\right)>0\}$, a.s. and for almost everywhere x.

Next, we deal with 1v1-B. Fix x and a realization of the data where convergence holds for all $t\neq s$. Then, eventually $\left|\hat\delta_{n_t+n_s}^{t\text{v}s}(x)-\delta^{t\text{v}s}(x)\right|\le\left|\delta^{t\text{v}s}(x)\right|/2$ for all $t\neq s$ such that $\delta^{t\text{v}s}(x)\neq 0$. That is, eventually $\mathbb{I}\left[\hat\delta_{n_t+n_s}^{t\text{v}s}(x)<0\right]=\mathbb{I}\left[\delta^{t\text{v}s}(x)<0\right]$ for all $t\neq s$ such that $\delta^{t\text{v}s}(x)\neq 0$. Restrict to such large enough n. Let $k_t(x)=\sum_{t\neq s}\mathbb{I}\left[\delta^{t\text{v}s}(x)<0\right]$, $\hat k_t(x)=\sum_{t\neq s}\mathbb{I}\left[\hat\delta_{n_t+n_s}^{t\text{v}s}(x)<0\right]$, and $k_{\min}(x)=\left|\arg\min_{t\in[m]}\mu(t,x)\right|$. Then, $t\in\arg\min_{t\in[m]}\mu(t,x)\iff k_t(x)=m-k_{\min}(x)\iff\hat k_t(x)\ge m-k_{\min}(x)\Longleftarrow t\in\arg\max_{t\in[m]}\sum_{s\neq t}\mathbb{I}\left[\hat\delta_{n_t+n_s}^{t\text{v}s}(x)<0\right]$. □

Proof of Theorem 5. By random sampling, $(X_{i_j},T_{i_j},Y_{i_j}(1),\dots,Y_{i_j}(m))$ are distributed iid as $(X,T,Y(1),\dots,Y(m))$ is in population. For $j\in[n_{\text{test}}]$, let i_{jt} be i_j's match for treatment t, or i_j if $T_{i_j}=t$. Under exact matching, $Y_{i_{jt}}(1),\dots,Y_{i_{jt}}(m)\mid X_{j_i}$ is distributed the same as $Y_{i_j}(1),\dots,Y_{i_j}(m)\mid X_{j_i},T_{j_i}=t$. By writing $\hat Y_{i_jt}=Y_{i_{jt}}=\sum_{s=1}^m\mathbb{I}\left[t=s\right]Y_{i_{js}}(s)$, we see that

$$\begin{aligned}\mathbb{E}[\hat Y_{i_j\tau(X_{i_j})}] &= \mathbb{E}\left[\mathbb{E}\left[\textstyle\sum_{s=1}^m\mathbb{I}\left[s=\tau(X_{i_j})\right]Y_{i_{js}}(s)\mid X_{i_j}\right]\right] && \text{(iterated expectation)}\\ &= \textstyle\sum_{s=1}^m\mathbb{E}\left[\mathbb{I}\left[s=\tau(X_{i_j})\right]\mathbb{E}\left[Y_{i_{js}}(s)\mid X_{j_i}\right]\right] && \text{(linearity)}\\ &= \textstyle\sum_{s=1}^m\mathbb{E}\left[\mathbb{I}\left[s=\tau(X_i)\right]\mathbb{E}\left[Y_{i_j}(s)\mid X_i,T_i=s\right]\right] && \text{(exact matching)}\\ &= \textstyle\sum_{s=1}^m\mathbb{E}\left[\mathbb{I}\left[s=\tau(X_i)\right]\mathbb{E}\left[Y_{i_j}(s)\mid X_i\right]\right] && \text{(Asn. 1)}\\ &= \mathbb{E}\left[\mathbb{E}\left[\textstyle\sum_{s=1}^m\mathbb{I}\left[s=\tau(X_i)\right]Y_{i_j}(s)\mid X_i\right]\right] && \text{(linearity)}\\ &= \mathbb{E}\left[Y_{i_j}(\tau(X_{i_j}))\right] && \text{(iterated expectation)} \quad \square\end{aligned}$$