

## A Omitted Proofs

### A.1 Proof of Theorem 1

First we note that  $W(X, T)Y - \sum_t \pi_t(X)Y(t) = \sum_t (W(X, t)\delta_{Tt} - \pi_t(X))Y(t)$ . Then analyzing each summand separately, we can obtain:

$$\begin{aligned}
& \mathbb{E}[W(X, t)\delta_{Tt}Y(t) - \pi_t(X)Y(t)] \\
&= \mathbb{E}[\mathbb{E}[W(X, t)\delta_{Tt}Y(t) - \pi_t(X)Y(t) \mid X]] \\
&= \mathbb{E}[W(X, t)\mathbb{E}[\delta_{Tt}Y(t) \mid X] - \pi_t(X)\mathbb{E}[Y(t) \mid X]] \\
&= \mathbb{E}[W(X, t)\mathbb{E}[\mathbb{E}[\delta_{Tt}Y(t) \mid X, T] \mid X] - \pi_t(X)\mathbb{E}[\mathbb{E}[Y(t) \mid X, Z] \mid X]] \\
&= \mathbb{E}[W(X, t)\mathbb{E}[\delta_{Tt}\mathbb{E}[Y(t) \mid X, T = t] \mid X] - \pi_t(X)\mathbb{E}[\mu_t(Z) \mid X]] \\
&= \mathbb{E}[W(X, t)\mathbb{E}[\delta_{Tt} \mid X]\mathbb{E}[\mu_t(Z) \mid X, T = t] - \pi_t(X)\mathbb{E}[\mu_t(Z) \mid X]] \\
&= \mathbb{E}[W(X, t)\eta_t(X)\mathbb{E}[\mu_t(Z) \mid X, T = t] - \pi_t(X)\mathbb{E}[\mu_t(Z) \mid X]] \\
&= \mathbb{E}\left[W(X, t)\eta_t(X)\nu_t(X, t) - \pi_t(X)\sum_{t'}\eta_{t'}(X)\nu_t(X, t')\right]
\end{aligned}$$

Therefore the solution to the equation  $\mathbb{E}[W(X, t)\delta_{Tt}Y(t) - \pi_t(X)Y(t)] = 0$  is given by:

$$W(X, t)\eta_t(X)\nu_t(X, t) - \pi_t(X)\sum_{t'}\eta_{t'}(X)\nu_t(X, t') = \Omega_t(X)$$

where  $\Omega_t(X)$  is any arbitrary function of  $X$  with mean zero. Solving this for  $W$  gives

$$W(X, t) = \frac{\pi_t(X)\sum_{t'}\eta_{t'}(X)\nu_t(X, t') + \Omega_t(X)}{\eta_t(X)\nu_t(X, t)},$$

and finally replacing  $X$  with  $x$  gives the required solution.

### A.2 Proof of Theorem 2

Define

$$\tau_{SAPE}^\pi = \frac{1}{n} \sum_i \sum_t \pi_t(X_i) \mu_t(Z_i).$$

Using  $(x + y)^2 \leq 2x^2 + 2y^2$ , we have

$$\mathbb{E}[(\hat{\tau}_W^\pi - \tau^\pi)^2 \mid X_{1:n}, T_{1:n}] \leq 2\mathbb{E}[(\hat{\tau}_W^\pi - \tau_{SAPE}^\pi)^2 \mid X_{1:n}, T_{1:n}] + 2\mathbb{E}[(\tau_{SAPE}^\pi - \tau^\pi)^2 \mid X_{1:n}, T_{1:n}].$$

Noting that  $\tau^\pi = \mathbb{E}[\tau_{SAPE}^\pi]$ , Assumption 3 implies that  $\mathbb{E}[(\tau_{SAPE}^\pi - \tau^\pi)^2] = O(1/n)$ . Markov's inequality yields  $\mathbb{E}[(\tau_{SAPE}^\pi - \tau^\pi)^2 \mid X_{1:n}, T_{1:n}] = O_p(1/n)$ .

Let  $CMSE(W, \mu) = \mathbb{E}[(\hat{\tau}_W^\pi - \tau_{SAPE}^\pi)^2 \mid X_{1:n}, T_{1:n}]$ . We proceed to bound  $CMSE$ . By iterating expectations we can obtain:

$$\begin{aligned}
CMSE(W, \mu) &= \mathbb{E}[(\hat{\tau}_W^\pi - \tau_{SAPE}^\pi)^2 \mid X, T] \\
&= \mathbb{E}[\mathbb{E}[(\hat{\tau}_W^\pi - \tau_{SAPE}^\pi)^2 \mid X, T, Z \mid X, T]] \\
&= \mathbb{E}[\mathbb{E}[\hat{\tau}_W^\pi - \tau_{SAPE}^\pi \mid X, T, Z]^2 \mid X, T] + \mathbb{E}[\mathbb{V}[\hat{\tau}_W^\pi - \tau_{SAPE}^\pi \mid X, T, Z] \mid X, T] \\
&= \mathbb{E}[(\frac{1}{n} \sum_{i,t} f_{it} \mu_t(Z_i))^2 \mid X, T] + \mathbb{E}[\mathbb{V}[\frac{1}{n} \sum_i W_i Y(T_i) \mid X, T, Z] \mid X, T] \\
&\leq \mathbb{E}[(\frac{1}{n} \sum_{i,t} f_{it} \mu_t(Z_i))^2 \mid X, T] + \frac{\sigma^2}{n^2} \|W\|_2^2
\end{aligned}$$

where  $\sigma$  is the bound defined in Assumption 3.

Next observe that for any (possibly correlated) random variables  $A_1, \dots, A_n$  and numbers  $p_1, \dots, p_n \in \mathbb{R}^n$  such that  $\sum_i p_i = 1$ , we have  $\mathbb{V}[\sum_i p_i A_i] \leq \max_i \mathbb{V}[A_i]$ . Given this, we can

simplify the first term above further, as follows:

$$\begin{aligned}
\mathbb{E}[(\frac{1}{n} \sum_{i,t} f_{it} \mu_t(Z_i))^2 \mid X, T] &= (\frac{1}{n} \sum_{i,t} f_{it} \mathbb{E}[\mu_t(Z_i) \mid X_i, T_i])^2 + \mathbb{V}[\frac{1}{n} \sum_{i,t} f_{it} \mu_t(Z_i) \mid X, T] \\
&= (\frac{1}{n} \sum_{i,t} f_{it} \nu_t(X_i, T_i))^2 + \frac{1}{n^2} \sum_i \mathbb{V}[\sum_t f_{it} \mu_t(Z_i) \mid X, T] \\
&\leq (\frac{1}{n} \sum_{i,t} f_{it} \nu_t(X_i, T_i))^2 + \frac{1}{n^2} \sum_i \max_t f_{it}^2 \mathbb{V}[\mu_t(Z_i) \mid X, T] \\
&\leq (\frac{1}{n} \sum_{i,t} f_{it} \nu_t(X_i, T_i))^2 + \frac{2\sigma^2}{n^2} \sum_i \max_t W_i^2 \delta_{T_i t} + \pi_t(X_i)^2 \\
&\leq (\frac{1}{n} \sum_{i,t} f_{it} \nu_t(X_i, T_i))^2 + \frac{2\sigma^2}{n^2} \sum_i (W_i^2 + 1) \\
&= (\frac{1}{n} \sum_{i,t} f_{it} \nu_t(X_i, T_i))^2 + \frac{2\sigma^2}{n^2} \|W\|_2^2 + \frac{2\sigma^2}{n}
\end{aligned}$$

where for the second inequality we used the fact that  $(x + y)^2 \leq 2x^2 + 2y^2$ . This gives us  $CMSE(W, \mu) \leq J^*(W, \mu) + O_p(1/n)$ , and combining this with the above gives  $\mathbb{E}[(\hat{\tau}_W^\pi - \tau^\pi)^2 \mid X_{1:N}, T_{1:N}] \leq 2J^*(W, \mu) + O_p(1/n)$  as required.

### A.3 Proof of Lemma 2

First, we will use the notation  $f(z; x, t)$  for the conditional measure of  $Z$  given  $X = x$  and  $T = t$ , and observe that according to Bayes rule we have:

$$\frac{f(z; x, t'')}{f(z; x, t')} = \frac{e_{t''}(z)}{e_{t'}(z)} \frac{\eta_{t'}(x)}{\eta_{t''}(x)}$$

Define  $\mathbb{E}_{xt}$  and  $\mathbb{P}_{xt}$  as shorthand for expectation and probability given  $X = x, T = t$  respectively. Then given the above, for any  $M > 0$  we can bound

$$\begin{aligned}
\nu_t(x, t'') &= \frac{\eta_{t''}(x)}{\eta_{t'}(x)} \mathbb{E}[\mu_t(Z) \mid X = x, T = t''] \\
&= \int_{\mathcal{Z}} f(z; x, t'') \mu_t(z) dz \\
&= \frac{\eta_{t''}(x)}{\eta_{t'}(x)} \int_{\mathcal{Z}} \frac{e_{t''}(z)}{e_{t'}(z)} \frac{\eta_{t'}(x)}{\eta_{t''}(x)} f(z; x, t') \mu_t(z) dz \\
&= \frac{\eta_{t'}(x)}{\eta_{t''}(x)} \int_{\mathcal{Z}} \frac{e_{t''}(z)}{e_{t'}(z)} f(z; x, t') \mu_t(z) dz \\
&\leq \frac{\eta_{t'}(x)}{\eta_{t''}(x)} \left( M \mathbb{E}_{xt'}[\mathbb{1}\{\frac{e_{t''}(z)}{e_{t'}(z)} \leq M\} \mu_t(Z)] + \mathbb{E}_{xt'}[\mathbb{1}\{\frac{e_{t''}(z)}{e_{t'}(z)} > M\} \frac{e_{t''}(z)}{e_{t'}(z)} \mu_t(Z)] \right)
\end{aligned}$$

Now we can use the fact that  $\mu_t$  is  $b$ -bounded to bound the first term by

$$\begin{aligned}
M \mathbb{E}_{xt'}[\mathbb{1}\{\frac{e_{t''}(z)}{e_{t'}(z)} \leq M\} \mu_t(Z)] &= M \nu_t(x, t') - M \mathbb{E}_{xt'}[\mathbb{1}\{\frac{e_{t''}(z)}{e_{t'}(z)} > M\} \mu_t(Z)] \\
&\leq M \nu_t(x, t') + Mb \mathbb{P}_{xt'}[\frac{e_{t''}(z)}{e_{t'}(z)} > M],
\end{aligned}$$

and in addition applying Cauchy Schwartz we can bound the second term by

$$\begin{aligned}
\mathbb{E}_{x,t'}[\mathbb{I}\{\frac{e_{t''}(z)}{e_{t'}(z)} > M\} \frac{e_{t''}(z)}{e_{t'}(z)} \mu_t(Z)] &\leq \sqrt{\mathbb{E}_{x,t'}[\mathbb{I}\{\frac{e_{t''}(z)}{e_{t'}(z)} > M\} \mu_t(Z)^2] \mathbb{E}_{x,t'}[\left(\frac{e_{t''}(z)}{e_{t'}(z)}\right)^2]} \\
&\leq \sqrt{b^2 \mathbb{P}_{x,t'}[\frac{e_{t''}(z)}{e_{t'}(z)} > M] \mathbb{E}_{x,t'}[\left(\frac{e_{t''}(z)}{e_{t'}(z)}\right)^2]} \\
&\leq b \sqrt{\mathbb{P}_{x,t'}[\frac{1}{e_{t'}(z)} > M] \mathbb{E}_{x,t'}[\left(\frac{1}{e_{t'}(z)}\right)^2]}.
\end{aligned}$$

Now define  $g(x, t) = \mathbb{E}_{x,t}[\left(\frac{1}{e_t(z)}\right)^2]$ . By Assumption 2 we know  $g(x, t)$  is finite for every  $x$  and  $t$ . Also by Markov's inequality we know that  $\mathbb{P}_{x,t}[\frac{1}{e_t(z)} > M] \leq \frac{g(x,t)}{M^2}$ . Therefore putting all of the above together we can obtain

$$\begin{aligned}
\nu_t(x, t'') &\leq \frac{\eta_{t'}(x)}{\eta_{t''}(x)} \left( M \nu_t(x, t') + \frac{2bg(x, t')}{M} \right) \\
&\leq \frac{\eta_{t'}(x)}{\eta_{t''}(x)} \left( M |\nu_t(x, t')| + \frac{2bg(x, t')}{M} \right)
\end{aligned}$$

This inequality is valid every  $M$ , so we can pick  $M$  to make it as tight as possible. Choosing  $M = \sqrt{\frac{2bg(x, t')}{\nu_t(x, t')}}$  gives us:

$$\nu_t(x, t'') \leq \frac{\eta_{t'}(x)}{\eta_{t''}(x)} \sqrt{8bg(x, t') |\nu_t(x, t')|}$$

Finally note that, since by symmetry  $\mathbb{E}[\mu(Z) \mid X, T] = -\mathbb{E}[-\mu(Z) \mid X, T]$ , we can strengthen this inequality to the following

$$|\nu_t(x, t'')| \leq \frac{\eta_{t'}(x)}{\eta_{t''}(x)} \sqrt{8bg(x, t') |\nu_t(x, t')|},$$

and noting that  $g(x, t') = \mathbb{E}[e_t^{-2}(Z) \mid X = x, T = t']$  gives us our final result.

#### A.4 Proof of Lemma 3

First note that by Assumption 6  $\mathcal{F}$  is compact. Also  $J(W, \cdot)$  is continuous for every  $W$ , since by Assumption 7 we know that the norm on each  $\mathcal{F}_t$  dominates the norm on  $\mathcal{L}^2(\mathcal{Z})$  and this continuity result would be trivial if  $\mathcal{F}_t = \mathcal{L}^2(\mathcal{Z})$ . This means that by the Extreme Value theorem we can replace the supremum over  $\mu$  with a maximum over  $\mu$  in the quantity we are bounding. Given this, we will proceed by bounding  $\min_W \max_{\mu \in \mathcal{F}} B(W, \mu)$  using von Neumann's minimax theorem to swap the minimum and the maximum, and then use this to establish the overall bound for  $J(W, \mu)$ .

Next we can observe that  $B(W, \mu)$  is linear, and therefore both convex and concave, for each of  $W$  and  $\mu$ . Next, by Assumption 6  $\mathcal{F}$  is convex and compact, and following the same argument as above  $B(W, \cdot)$  is continuous for every  $W$ . In addition,  $B(W, \mu)$  is also clearly continuous in  $W$  for fixed  $\mu$ , and the set  $\{W : \|W\|_2 \leq M\}$  is obviously compact and convex for any constant  $M$ . Thus by von Neumann's minimax theorem we have the following for every finite  $M$ :

$$\min_{\|W\|_2 \leq M} \max_{\mu \in \mathcal{F}} B(W, \mu) = \max_{\mu \in \mathcal{F}} \min_{\|W\|_2 \leq M} B(W, \mu) \quad (8)$$

Given this, we can bound  $\min_W \max_{\mu \in \mathcal{F}} \hat{J}(W, \mu)$  as follows, which is valid for any  $M$ :

$$\begin{aligned}
\min_W \max_{\mu \in \mathcal{F}} \hat{J}(W, \mu) &\leq \min_W \max_{\mu \in \mathcal{F}} B(W, \mu)^2 + \frac{1}{n^2} W^T \Sigma W \\
&\leq \min_W \max_{\mu \in \mathcal{F}} B(W, \mu)^2 + \frac{\sigma^2}{n^2} \|W\|_2^2 \\
&\leq \min_{\|W\|_2 \leq M} \max_{\mu \in \mathcal{F}} B(W, \mu)^2 + \frac{\sigma^2}{n^2} \|W\|_2^2 \\
&\leq \left( \min_{\|W\|_2 \leq M} \max_{\mu \in \mathcal{F}} B(W, \mu) \right)^2 + \frac{\sigma^2}{n^2} M^2 \\
&= \left( \max_{\mu \in \mathcal{F}} \min_{\|W\|_2 \leq M} B(W, \mu) \right)^2 + \frac{\sigma^2}{n^2} M^2 \\
&= \max_{\mu \in \mathcal{F}} \min_{\|W\|_2 \leq M} B(W, \mu)^2 + \frac{\sigma^2}{n^2} M^2
\end{aligned}$$

In these inequalities we use the fact that  $\min_W \max_{\mu} B(W, \mu)^2 = (\min_W \max_{\mu} B(W, \mu))^2$  and  $\max_{\mu} \min_W B(W, \mu)^2 = (\max_{\mu} \min_W B(W, \mu))^2$  due to the symmetry of  $\mu$  in  $\mathcal{F}$  implied by Assumption 5.

#### A.5 Proof of Lemma 4

Let  $\prod$  denote Cartesian product. First we note that without loss of generality we can prove this lemma in the case that  $\mathcal{F} = \prod_t \mathcal{F}_t$ , since in general  $F \subseteq \prod_t \mathcal{F}_t$  so  $\sup_{\mu \in \mathcal{F}} \inf_W J(W, \mu) \leq \sup_{\mu \in \prod_t \mathcal{F}_t} \inf_W J(W, \mu)$ , and it is easy to verify that all of our assumptions would still hold on the larger set  $\prod_t \mathcal{F}_t$ .

Now define the set  $\mathcal{H}^0 = \{\mu \in \prod_t \mathcal{L}^2(\mathcal{Z}) : \mathbb{E}[\nu_T(X, T)^2] = 0\}$ . Each coordinate  $\mathcal{H}_t^0$  of  $\mathcal{H}^0$  is a subspace of  $\mathcal{L}^2(\mathcal{Z})$ , so we can also define its orthogonal complement  $\mathcal{H}_t^+$ . Also, we have separability since from Section 2.1 we know that  $\mathcal{Z} \subseteq \mathbb{R}^q$ , so any function  $f \in \mathcal{L}^2(\mathcal{Z})$  can be uniquely represented as  $f = f^0 + f^+$  where  $f^0 \in \mathcal{H}_t^0$  and  $f^+ \in \mathcal{H}_t^+$ . This means that for each  $\mu_t \in \mathcal{F}_t$ , we can similarly uniquely represent  $\mu_t = \mu_t^0 + \mu_t^+$ , and we can easily extend this to a unique representation of the vector  $\mu = \mu^0 + \mu^+$ . Now in the case that  $\mathbb{E}[\nu_T(X, T)^2] = 0$  we have  $\nu_{T_i}(X_i, T_i) = 0$  almost surely for all  $i$ , and it follows from Lemma 2 that  $\nu_{T_i}(X_i, t) = 0$  almost surely also for all  $i$  and  $t$ . Therefore any component of  $\mu$  in  $\mathcal{H}^0$  has no effect on the function  $J(W, \mu)$  which we are bounding, so without loss of generality we can restrict our attention to the following space:

$$\mathcal{F}^+ = \prod_t (\mathcal{H}_t^+ \cap \mathcal{F}_t).$$

By construction the only function in  $\mathcal{F}^+$  such that  $\mathbb{E}[\nu_T(X, T)^2] = 0$  is the zero function, which we can also ignore in our bounds below, since when  $\mu = 0$  we can easily obtain  $J(W, \mu) = 0$  by choosing  $W = 0$ . Furthermore, by Assumption 7 we know that for each  $t$  the  $\mathcal{F}_t$  norm dominates the  $\mathcal{L}^2(\mathcal{Z})$  norm, so it must be the case that each space  $\mathcal{F}_t^+$  is closed, since  $\mathcal{H}_t^+$  is a closed subspace of  $\mathcal{L}^2(\mathcal{Z})$  due to it being an orthogonal complement. Thus it follows easily from Assumption 4 that  $\mathcal{F}^+$  is closed, given that its norm is an  $\mathbb{R}^m$  norm on top of the corresponding  $\mathcal{F}_t^+$  norms and  $m$  is finite.

Now, based on Lemma 3, it is sufficient to pick weights in response to  $\mu$  that control for a single mean outcome function. Instead of actually constructing a particular set of weights, we take the approach of viewing this as a convex optimization problem. Specifically, given  $\mu$ , we calculate the minimum euclidean norm of all weights that set the bias term  $B(W, \mu)$  to zero exactly. This can be formulated as the following convex optimization program

$$\begin{aligned}
&\min_W \sum_i W_i^2 \\
&\text{s.t. } \sum_i W_i \nu_{T_i}(X_i, T_i) = \sum_{i,t} \pi_t(X_i) \nu_t(X_i, T_i).
\end{aligned}$$

Given the program only has linear constraints with equality, it satisfies Slater's condition, and therefore satisfies strong duality, which we will use to find the optimal value of this program. First we calculate the Lagrangian as

$$\mathcal{L}_n(W, \lambda) = \sum_i W_i^2 + \lambda \left( \sum_i W_i \nu_{T_i}(X_i, T_i) - \sum_{i,t} \pi_t(X_i) \nu_t(X_i, T_i) \right).$$

It can easily be verified by taking derivatives that for any  $\lambda \in \mathbb{R}$  this function is minimized by setting  $W_i = -\frac{\lambda}{2} \nu_{T_i}(X_i, T_i)$ , and plugging this value in gives the dual formulation of the program as

$$\max_{\lambda} -\frac{\lambda^2}{4} \sum_i \nu_{T_i}(X_i, T_i)^2 - \lambda \sum_{i,t} \pi_t(X_i) \nu_t(X_i, T_i),$$

which is unconstrained. Again by taking derivatives we can maximize this function, and we find the maximum value is given by

$$\lambda = -2 \frac{\sum_i \pi_t(X_i) \nu_t(X_i, T_i)}{\sum_i \nu_{T_i}(X_i, T_i)^2},$$

and finally plugging this value into the dual objective function we see that the euclidean norm of the weights  $W^*$  solving the convex program above is given by

$$\|W^*\|_2^2 = \frac{(\sum_i \pi_t(X_i) \nu_t(X_i, T_i))^2}{\sum_i \nu_{T_i}(X_i, T_i)^2}.$$

Now define  $\mathbb{E}_n$  as the mean with respect to the empirical distribution of the logged data. Then this objective value can be reformulated as

$$\|W^*\|_2^2 = \frac{n \mathbb{E}_n[\sum_t \nu_t(X, T)]^2}{\mathbb{E}_n[\nu_T(X, T)^2]}.$$

Therefore choosing  $M = \|W^*\|$ , combining this result with Lemma 3 gives us

$$\begin{aligned} \min_W \sup_{\mu \in \mathcal{F}^+} J(W, \mu) &\leq \sup_{\mu \in \mathcal{F}^+} \frac{1}{n} \left( \frac{\sigma^2 \mathbb{E}_n[\sum_t \nu_t(X, T)]^2}{\mathbb{E}_n[\nu_T(X, T)^2]} \right) \\ &= \sup_{\mu \in \mathcal{F}^+} \frac{1}{n} \left( \frac{\sigma^2 \mathbb{E}_n[\sum_t \nu_t(X, T)]^2}{\mathbb{E}[\nu_T(X, T)^2] + (\mathbb{E}_n[\nu_T(X, T)^2] - \mathbb{E}[\nu_T(X, T)^2])} \right). \end{aligned}$$

Given this we will proceed by arguing that we can bound the denominator away from zero. We can note that  $\mu$  appears in both the numerator and denominator on the same scale, so without loss of generality we can further restrict our attention to  $\mu$  with fixed norm. By Assumption 8 we know that we can rescale every  $\mu \in \mathcal{F}$  to have norm  $\gamma$  for some  $\gamma > 0$ . Given this we will restrict ourselves to the set  $\mathcal{F}_\gamma^+ = \{\mu \in \mathcal{F}^+ : \|\mu\| = \gamma\}$ . Since  $\mathcal{F}_\gamma^+$  is the intersection of two closed sets it must be closed. Furthermore by Assumption 6 it is also compact, so it satisfies the conditions for the extreme value theorem. By construction  $\mathbb{E}[\nu_T(X, T)^2] > 0$  for every  $\mu \in \mathcal{F}_\gamma^+$ , so putting the above together we have  $\inf_{\mu \in \mathcal{F}_\gamma^+} \mathbb{E}[\nu_T(X, T)^2] > 0$ . We will define this value to be  $\alpha$ .

Now the numerator in the above bound is clearly bounded above by some  $\beta > 0$  uniformly over  $\mu \in \mathcal{F}$ , since by Assumption 9 we know that every  $\mu_t \in \mathcal{F}_t$  is uniformly bounded by some global constant, and therefore all  $\nu$  terms are bounded by some constant  $b$ . Given this all that remains to be shown is that  $\sup_{\mu \in \mathcal{F}} |\mathbb{E}_n[\nu_T(X, T)^2] - \mathbb{E}[\nu_T(X, T)^2]|$  converges in probability to zero. In order to show this we will define the following terms:

$$\begin{aligned} D_n &= \mathbb{E}_n[\nu_T(X, T)^2] \\ E_n &= \sup_{\mu \in \mathcal{F}} |D_n - \mathbb{E}[D_n]| \end{aligned}$$

We need to show that  $E_n$  converges uniformly to zero. Define  $D'_n$  as an arbitrary recalculation of  $D_n$  replacing  $(X_{1:n}, T_{1:n})$  with  $(X'_{1:n}, T'_{1:n})$ , which differ from the originals at most in a single

coordinate  $i$ , and define  $E'_n = \sup_{\mu \in \mathcal{F}} |D'_n - \mathbb{E}[D'_n]|$ . Furthermore as argued above all  $\nu$  terms are bounded above by some constant  $b$ , so each  $\nu_{T_i}(X_i, T_i)^2$  is bounded by  $b^2$ . Given this we can obtain

$$\begin{aligned}
|E_n - E'_n| &= \left| \sup_{\mu \in \mathcal{F}} |D_n - \mathbb{E}[D_n]| - \sup_{\mu \in \mathcal{F}} |D'_n - \mathbb{E}[D'_n]| \right| \\
&\leq \sup_{\mu \in \mathcal{F}} |(D_n - \mathbb{E}[D_n]) - (D'_n - \mathbb{E}[D'_n])| \\
&= \sup_{\mu \in \mathcal{F}} |D_n - D'_n| \\
&= \frac{1}{n} \sup_{\mu \in \mathcal{F}} |\nu_{T_i}(X_i, T_i)^2 - \nu_{T'_i}(X'_i, T'_i)^2| \\
&\leq \frac{2b^2}{n}
\end{aligned}$$

Given this we can apply McDiarmid's inequality to obtain the following bound:

$$P(|E_n - \mathbb{E}[E_n]| \leq \epsilon) \leq 2 \exp\left(-\frac{n\epsilon^2}{2b^4}\right)$$

This implies that  $E_n - \mathbb{E}[E_n] = o_p(1)$ . Next we show that  $\mathbb{E}[E_n] = o_p(1)$  also. We do this using a symmetrization argument as follows, where  $D'_n$  is defined identically to  $D_n$  using iid shadow variables  $(X'_i, T'_i)$  in place of  $(X_i, T_i)$  for each  $i$ , and  $\epsilon_i$  are iid Rademacher random variables:

$$\begin{aligned}
\mathbb{E}[E_n] &= \mathbb{E} \left[ \sup_{\mu \in \mathcal{F}} |D_n - \mathbb{E}[D_n]| \right] \\
&= \mathbb{E} \left[ \sup_{\mu \in \mathcal{F}} \left| \frac{1}{n} \sum_i \nu_{T_i}(X_i, T_i) - \mathbb{E}[\nu_{T'_i}(X'_i, T'_i)] \right| \right] \\
&\leq 2 \mathbb{E} \left[ \sup_{\mu \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i \nu_{T_i}(X_i, T_i) \right| \right] \\
&\leq 2 \sum_t \mathbb{E} \left[ \sup_{\mu \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i \delta_{T_i t} \nu_t(X_i, T_i) \right| \right] \\
&\leq 4 \sum_t \mathbb{E} \left[ \sup_{\mu \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i \nu_t(X_i, T_i) \right| \right] \\
&\leq 4 \sum_t \mathbb{E} \left[ \sup_{\mu \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i \mu_t(Z_i) \right| \right] \\
&\leq 4 \sum_t \mathcal{R}_n(\mathcal{F}_t)
\end{aligned}$$

where in the third inequality we appeal to the Rademacher comparison lemma [28, Thm. 4.12]. Thus since from Assumption 10 we know that the Rademacher complexity of each set  $\mathcal{R}_n(\mathcal{F}_n)$  vanishes, it follows that  $\mathbb{E}[E_n] = o_p(1)$ . Putting everything from above together we get

$$\min_W \sup_{\mu \in \mathcal{F}} J(W, \mu) \leq \frac{1}{n} \frac{\beta}{\alpha + o_p(1)},$$

so we have  $\min_W \sup_{\mu \in \mathcal{F}} J(W, \mu) \leq O_p(1/n)$  as required.

## A.6 Proof of Theorem 4

First, Assumption 4 follows trivially from the definition of  $\mathcal{F}^K$ . Next, Assumption 5 and Assumption 8 follow from the fact that  $\mathcal{F}^K$  consists of all functions in  $\text{span}(\mathcal{F}_t)$  with norm at most 1, as does the fact that it is a closed space. Given that  $K$  is a Mercer kernel, balls in the corresponding RKHS have finite covering number [44], and it follows easily from this that  $\mathcal{F}^K$  has

finite covering numbers, as its covering number must be bounded above by the sum of the covering numbers of the spaces  $\mathcal{F}_t^K$ . So  $\mathcal{F}^K$  is closed and totally bounded with respect to its norm, and therefore compact, which gives us Assumption [6](#). Clearly each  $\mathcal{F}_t$  is contained in  $\mathcal{L}^2(\mathcal{Z})$  since RKHSs are square integrable, and the fact that the  $K$  norm dominates the  $\mathcal{L}^2$  follows from Mercer's Theorem, which implies that  $\|f\|_K^2 = \sum_{i=1}^{\infty} f_i^2/\sigma_i$ , where  $f_i$  is the  $i$ th eigenvalue of  $f$  for some orthonormal basis of  $\mathcal{L}^2(\mathcal{Z})$ , and  $\sigma_i \geq 0$  converges to zero. This gives us Assumption [7](#). Next, by construction each  $\mathcal{F}_t^K$  consists of all functions in the RKHS up to norm 1. Therefore assuming the kernel  $K$  is bounded, it is trivial to verify that function application must be globally bounded, since for any function  $f \in \mathcal{F}_t^K$  we have  $f(x) \leq \langle f, K_x \rangle \leq \|f\| \sqrt{K(x, x)} \leq \sqrt{K(x, x)}$ , which gives us Assumption [9](#). Finally, given this characterization of  $\mathcal{F}_t^K$  as the 1-ball of the RKHS, it has vanishing Rademacher complexity [[33](#) Thm. 2.1], so we have Assumption [10](#).

### A.7 Proof of Theorem [5](#)

First we will find a closed form expression for  $\sup_{\mu \in \mathcal{F}^K} (\frac{1}{n} \sum_{i,t} f_{it} \nu_t(X_i, T_i))^2$ , where  $\nu_t$  is defined in terms of the approximate posterior  $\hat{\varphi}$ . In this derivation we will use the shorthand  $\varphi_i$  for the conditional density of  $Z_i$  given  $X_i$  and  $T_i$  under  $\hat{\varphi}$ , and  $T_K$  for the kernel intergral operator defined according to  $T_K f = \int_{\mathcal{Z}} K(\cdot, z) f(z) dz$ . In this derivation we will make use of the fact that  $\langle f, g \rangle_{\mathcal{L}^2} = \langle f, T_K g \rangle_K$  for any square integrable  $f$  and  $g$ . Given all this we can obtain:

$$\begin{aligned}
\sup_{\mu \in \mathcal{F}^K} \left( \frac{1}{n} \sum_{i,t} f_{it} \nu_t(X_i, T_i) \right)^2 &= \sum_t \sup_{\mu_t \in \mathcal{F}_t^K} \left( \frac{1}{n} \sum_i f_{it} \langle \mu_t, \varphi_i \rangle_{\mathcal{L}^2} \right)^2 \\
&= \sum_t \sup_{\mu_t \in \mathcal{F}_t^K} \left\langle \mu_t, \frac{1}{n} \sum_i f_{it} \varphi_i \right\rangle_{\mathcal{L}^2}^2 \\
&= \sum_t \sup_{\mu_t \in \mathcal{F}_t^K} \left\langle \mu_t, T_K \frac{1}{n} \sum_i f_{it} \varphi_i \right\rangle_K^2 \\
&= \sum_t \frac{\langle T_K \frac{1}{n} \sum_i f_{it} \varphi_i, T_K \frac{1}{n} \sum_i f_{it} \varphi_i \rangle_K}{\|T_K \frac{1}{n} \sum_i f_{it} \varphi_i\|_K} \\
&= \sum_t \left\langle T_K \frac{1}{n} \sum_i f_{it} \varphi_i, T_K \frac{1}{n} \sum_i f_{it} \varphi_i \right\rangle_K \\
&= \sum_t \left\langle \frac{1}{n} \sum_i f_{it} \varphi_i, T_K \frac{1}{n} \sum_i f_{it} \varphi_i \right\rangle_{\mathcal{L}^2} \\
&= \frac{1}{n^2} \sum_{i,j,t} f_{it} f_{jt} \langle \varphi_i, T_K \varphi_j \rangle_{\mathcal{L}^2} \\
&= \frac{1}{n^2} \sum_{i,j,t} f_{it} f_{jt} \int_{\mathcal{Z}} \varphi_i(z) \left( \int_{\mathcal{Z}'} K(z, z') \varphi_j(z') dz' \right) dz \\
&= \frac{1}{n^2} \sum_{i,j,t} f_{it} f_{jt} \int_{\mathcal{Z}} \int_{\mathcal{Z}'} \varphi_i(z) \varphi_j(z') K(z, z') dz' dz \\
&= \frac{1}{n^2} \sum_{i,j,t} f_{it} f_{jt} \mathbb{E}[K(Z_i, Z'_j)],
\end{aligned}$$

where the expectation in the last term is implicitly defined in terms of the approximate posterior  $\hat{\varphi}$  and conditional on the observed data.

Given this, and recalling that  $f_{it} = W_i \delta_{T_{it}} - \pi_t(X_i)$  we can derive a closed form for our minimization objective, as follows:

$$\begin{aligned}
\sup_{\mu \in \mathcal{F}^K} J(W, \mu) &= \frac{1}{n^2} \sum_{i,j,t} f_{it} f_{jt} Q_{ij} + \frac{1}{n^2} \sum_{i,j} W_i W_j \Gamma_{ij} \\
&= \frac{1}{n^2} \sum_{i,j,t} Q_{ij} (W_i W_j \delta_{T_{it}} \delta_{T_{jt}} - 2W_j \delta_{T_{jt}} \pi_t(X_i) + \pi_t(X_i) \pi_t(X_j)) \\
&\quad + \frac{1}{n^2} \sum_{i,j} W_i W_j \Gamma_{ij} \\
&= \frac{1}{n^2} \sum_{i,j} W_i W_j (Q_{ij} \delta_{T_i T_j} + \Gamma_{ij}) - \frac{2}{n^2} \sum_j W_j (\sum_i Q_{ij} \pi_{T_j}(X_i)) \\
&\quad + \frac{1}{n^2} \sum_{i,j,t} Q_{ij} \pi_t(X_i) \pi_t(X_j)
\end{aligned}$$

Finally we can conclude by noting that this corresponds to the quadratic program formulation given in the question with  $c = \frac{1}{n^2} \sum_{i,j,t} Q_{ij} \pi_t(X_i) \pi_t(X_j)$ .

## B Additional Experimentation Details

### B.1 Experiment Scenario

All our experiments were conducted using the setup described in Section 5.1. We used the following parameter values for our data-generating distribution:

$$\begin{aligned}
\alpha &= [1.0, -2.0, -1.0, 2.0, 4.0, 0.0, -2.0, -1.0, -3.0, 1.0] \\
\alpha_0 &= 0 \\
\sigma_X &= 4.0 \\
\beta &= [0.5, -0.5] \\
\beta_0 &= 0 \\
\zeta(0) &= 1.0 \\
\zeta(0)_0 &= 0 \\
\zeta(1) &= -0.5 \\
\zeta(1)_0 &= 0 \\
\sigma_Y &= 0.01
\end{aligned}$$

In addition, the policy  $\pi$  we are evaluating takes the form as described in Section 5.1 and we used the following parameter values for this policy:

$$\begin{aligned}
\psi_0 &= [-0.1, 0.2, 0.2, -0.1, -0.1, -0.1, 0.1, 0.1, 0.1, -0.1] \\
\psi_1 &= -\psi_0
\end{aligned}$$

### B.2 Method Implementation Details

In all methods where we sampled from the posterior  $\hat{\varphi}(\cdot; x, t)$ , this sampling was done using STAN [7], solving QPs and LCQPs was done using the Python package quadprog<sup>3</sup> all stochastic gradient descent (SGD) learning was performed using the Adam [25] optimizer with a learning rate of 0.001.

**OptZ** We ran Algorithm 1 with  $B = 50$ .

**IPS** Since the propensity scores  $\eta_t(x)$  are not tractable to compute analytically, we trained a neural network  $\hat{\eta}$  to estimate this function. This was done by sampling batches of  $(Z, X)$  pairs

<sup>3</sup><https://pypi.org/project/quadprog/>



from the data model, and training the network using SGD to predict the vector of probabilities  $P_T = \beta^T Z + \beta_0$  from  $X$ , using cross-entropy loss. We used a neural network with two hidden layers of size 200 for  $\hat{\eta}$ , and trained for 2000 iterations with a batch size of 32. We found in practice this training was very stable and gave accurate results.

**DirX** For each  $t$  we trained a neural network  $\hat{\rho}_t$  to predict  $\nu_t(x, t)$  by taking the set of  $(X, T, Y)$  triplets in our training data where  $T = t$ , and training the network using SGD to predict  $Y$  from  $X$  using MSE loss. Based on pilot experiments we used a network architecture with a single hidden layer of size 100, and trained using a batch size of 128. We used 80% of our data for training, and used the remaining 20% for the purpose of early stopping. We trained for a maximum of 500 epochs, or until we made no progress on development data for 20 epochs.

**DirZ** For each  $t$  we trained a neural network  $\hat{\mu}_t$  to predict  $\mu_t$ . This was done by taking the set of  $(X, T, Y)$  triplets in our training data, and for each sampling 200  $Z$  values from the posterior using our identified model given  $X$  and  $T$ . This gives us a set of  $(Z, T, Y)$  triplets 200 times as large as our original training set. We then trained each  $\hat{\mu}_t$  network by taking the set of these triplets where  $T = t$ , and optimized the network using SGD on this data predicting  $Y$  from  $Z$ . We used the same settings for this optimization as with the *direct-naive* method, except we allowed up to 1000 epochs. Note that for both training and inference we limited ourselves to sampling 200  $Z$  values per data point due to computational limitations.

## C Additional Experiment Results

### C.1 Results with Alternative Link Functions and Constraints

We present here our additional experiment results. In these results **SimplexOptZ** refers to our method using the simplex constraints discussed in Section [4.2](#).

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.39 ± .07	.24 ± .02	.36 ± .02	.81 ± .02
500	.19 ± .02	.18 ± .02	.23 ± .02	.49 ± .02
1000	.11 ± .01	.11 ± .01	.13 ± .01	.27 ± .01
2000	.08 ± .01	.08 ± .01	.09 ± .01	.17 ± .01

Table 5: Convergence of RMSE for weighted estimator using our weights, with **step** link

n	<b>DirX:OptZ</b> <sub>0.001</sub>	<b>DirX:OptZ</b> <sub>0.2</sub>	<b>DirX:OptZ</b> <sub>1.0</sub>	<b>DirX:OptZ</b> <sub>5.0</sub>
200	.57 ± .06	.42 ± .03	.39 ± .03	.43 ± .03
500	.55 ± .02	.46 ± .02	.39 ± .02	.37 ± .02
1000	.49 ± .02	.45 ± .01	.39 ± .01	.32 ± .01
2000	.48 ± .01	.47 ± .01	.42 ± .01	.34 ± .01

Table 6: Convergence of RMSE for doubly robust estimator using our weights and **DirX**, with **step** link

n	<b>DirZ:OptZ</b> <sub>0.001</sub>	<b>DirZ:OptZ</b> <sub>0.2</sub>	<b>DirZ:OptZ</b> <sub>1.0</sub>	<b>DirZ:OptZ</b> <sub>5.0</sub>
200	.41 ± .07	.29 ± .02	.50 ± .02	1.1 ± .03
500	.20 ± .02	.21 ± .02	.31 ± .02	.70 ± .02
1000	.11 ± .01	.13 ± .01	.18 ± .01	.42 ± .01
2000	.08 ± .01	.09 ± .01	.13 ± .01	.26 ± .01

Table 7: Convergence of RMSE for doubly robust estimator using our weights and **DirZ**, with **step** link

n	<b>SimplexOptZ</b> <sub>0.001</sub>	<b>SimplexOptZ</b> <sub>0.2</sub>	<b>SimplexOptZ</b> <sub>1.0</sub>	<b>SimplexOptZ</b> <sub>5.0</sub>
200	.30 ± .02	.25 ± .02	.38 ± .02	.91 ± .02
500	.18 ± .02	.19 ± .02	.24 ± .02	.54 ± .02
1000	.12 ± .01	.11 ± .01	.13 ± .01	.29 ± .01
2000	.07 ± .01	.08 ± .01	.10 ± .01	.18 ± .01

Table 8: Convergence of RMSE for weighted estimator using our weights and constraining  $W \in n\Delta^n$ , with **step** link

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.47 ± .03	2.0 ± .03	2.1 ± .03	2.3 ± .02	2.5 ± .02	.52 ± .02	2.6 ± .02
500	.48 ± .03	2.0 ± .02	2.1 ± .02	2.3 ± .02	2.6 ± .02	.48 ± .02	2.6 ± .01
1000	.39 ± .02	2.0 ± .01	2.1 ± .01	2.3 ± .01	2.5 ± .01	.48 ± .02	2.6 ± .01
2000	.40 ± .01	2.0 ± .01	2.1 ± .01	2.3 ± .01	2.5 ± .01	.45 ± .02	2.6 ± .01

Table 9: Convergence of RMSE for benchmark methods, with **step** link

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.03 ± .39	.11 ± .21	.29 ± .21	.78 ± .18
500	.09 ± .17	.10 ± .15	.17 ± .16	.47 ± .15
1000	.02 ± .11	.05 ± .09	.08 ± .09	.25 ± .09
2000	.03 ± .07	.05 ± .06	.07 ± .07	.16 ± .07

Table 10: Convergence of bias for weighted estimator using our weights, with **step** link

n	<b>DirX:OptZ</b> <sub>0.001</sub>	<b>DirX:OptZ</b> <sub>0.2</sub>	<b>DirX:OptZ</b> <sub>1.0</sub>	<b>DirX:OptZ</b> <sub>5.0</sub>
200	.43 ± .38	.35 ± .24	.31 ± .24	.37 ± .22
500	.51 ± .19	.42 ± .18	.35 ± .18	.33 ± .17
1000	.47 ± .13	.44 ± .11	.37 ± .10	.30 ± .11
2000	.47 ± .09	.46 ± .08	.41 ± .08	.33 ± .08

Table 11: Convergence of bias for doubly robust estimator using our weights and **DirX**, with **step** link

n	<b>DirZ:OptZ</b> <sub>0.001</sub>	<b>DirZ:OptZ</b> <sub>0.2</sub>	<b>DirZ:OptZ</b> <sub>1.0</sub>	<b>DirZ:OptZ</b> <sub>5.0</sub>
200	.05 ± .40	.19 ± .22	.45 ± .22	1.1 ± .21
500	.10 ± .18	.14 ± .16	.26 ± .16	.68 ± .17
1000	.04 ± .11	.09 ± .10	.15 ± .10	.41 ± .10
2000	.03 ± .07	.06 ± .07	.10 ± .07	.25 ± .07

Table 12: Convergence of bias for doubly robust estimator using our weights and **DirZ**, with **step** link

n	<b>SimplexOptZ</b> <sub>0.001</sub>	<b>SimplexOptZ</b> <sub>0.2</sub>	<b>SimplexOptZ</b> <sub>1.0</sub>	<b>SimplexOptZ</b> <sub>5.0</sub>
200	.04 ± .30	.12 ± .21	.31 ± .21	.89 ± .20
500	.08 ± .15	.10 ± .15	.18 ± .16	.51 ± .16
1000	.01 ± .12	.06 ± .09	.09 ± .09	.27 ± .10
2000	.03 ± .07	.05 ± .06	.07 ± .07	.17 ± .07

Table 13: Convergence of bias for weighted estimator using our weights and constraining  $W \in n\Delta^n$ , with **step** link

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.40 ± .25	1.9 ± .21	2.1 ± .20	2.3 ± .19	2.5 ± .18	.49 ± .18	2.6 ± .14
500	.43 ± .21	2.0 ± .16	2.1 ± .15	2.3 ± .14	2.6 ± .13	.45 ± .16	2.6 ± .12
1000	.37 ± .12	2.0 ± .10	2.1 ± .09	2.3 ± .09	2.5 ± .08	.46 ± .15	2.6 ± .11
2000	.39 ± .10	2.0 ± .08	2.1 ± .07	2.3 ± .07	2.5 ± .07	.42 ± .17	2.6 ± .11

Table 14: Convergence of bias for benchmark methods, with **step** link

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.07 ± .01	.04 ± .00	.04 ± .00	.07 ± .00
500	.04 ± .00	.03 ± .00	.03 ± .00	.04 ± .00
1000	.02 ± .00	.02 ± .00	.02 ± .00	.02 ± .00
2000	.01 ± .00	.01 ± .00	.01 ± .00	.01 ± .00

Table 15: Convergence of RMSE for weighted estimator using our weights, with **exp** link

n	<b>DirX:OptZ</b> <sub>0.001</sub>	<b>DirX:OptZ</b> <sub>0.2</sub>	<b>DirX:OptZ</b> <sub>1.0</sub>	<b>DirX:OptZ</b> <sub>5.0</sub>
200	.13 ± .01	.10 ± .01	.12 ± .01	.11 ± .01
500	.10 ± .01	.09 ± .01	.10 ± .01	.12 ± .01
1000	.08 ± .00	.08 ± .00	.08 ± .00	.10 ± .00
2000	.07 ± .00	.07 ± .00	.08 ± .00	.09 ± .00

Table 16: Convergence of RMSE for doubly robust estimator using our weights and **DirX**, with **exp** link

n	<b>DirZ:OptZ</b> <sub>0.001</sub>	<b>DirZ:OptZ</b> <sub>0.2</sub>	<b>DirZ:OptZ</b> <sub>1.0</sub>	<b>DirZ:OptZ</b> <sub>5.0</sub>
200	.15 ± .02	.15 ± .01	.25 ± .01	.44 ± .01
500	.10 ± .01	.11 ± .01	.18 ± .01	.32 ± .01
1000	.07 ± .01	.08 ± .01	.12 ± .01	.23 ± .01
2000	.04 ± .00	.05 ± .00	.08 ± .00	.16 ± .00

Table 17: Convergence of RMSE for doubly robust estimator using our weights and **DirZ**, with **exp** link

n	<b>SimplexOptZ</b> <sub>0.001</sub>	<b>SimplexOptZ</b> <sub>0.2</sub>	<b>SimplexOptZ</b> <sub>1.0</sub>	<b>SimplexOptZ</b> <sub>5.0</sub>
200	.05 ± .00	.11 ± .01	.22 ± .01	.39 ± .02
500	.04 ± .00	.08 ± .01	.15 ± .01	.28 ± .01
1000	.02 ± .00	.05 ± .00	.09 ± .00	.19 ± .00
2000	.02 ± .00	.03 ± .00	.07 ± .00	.14 ± .00

Table 18: Convergence of RMSE for weighted estimator using our weights and constraining  $W \in n\Delta^n$ , with **exp** link

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.12 ± .01	.76 ± .02	.81 ± .02	.92 ± .02	1.0 ± .02	.10 ± .01	1.0 ± .01
500	.11 ± .00	.76 ± .01	.82 ± .01	.92 ± .01	1.0 ± .01	.10 ± .01	1.0 ± .01
1000	.10 ± .00	.74 ± .01	.79 ± .01	.90 ± .01	1.0 ± .01	.09 ± .01	1.1 ± .01
2000	.10 ± .00	.73 ± .00	.78 ± .00	.88 ± .00	.99 ± .01	.10 ± .01	1.0 ± .01

Table 19: Convergence of RMSE for benchmark methods, with **exp** link

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.01 ± .06	.00 ± .04	−0.00 ± .04	−0.05 ± .04
500	.01 ± .04	.01 ± .03	.00 ± .03	−0.03 ± .03
1000	.00 ± .02	.01 ± .02	−0.00 ± .02	−0.01 ± .02
2000	.01 ± .01	.01 ± .01	.00 ± .01	.00 ± .01

Table 20: Convergence of bias for weighted estimator using our weights, with **exp** link

n	<b>DirX:OptZ</b> <sub>0.001</sub>	<b>DirX:OptZ</b> <sub>0.2</sub>	<b>DirX:OptZ</b> <sub>1.0</sub>	<b>DirX:OptZ</b> <sub>5.0</sub>
200	.07 ± .10	.08 ± .06	.10 ± .05	.11 ± .04
500	.07 ± .07	.07 ± .05	.09 ± .05	.11 ± .04
1000	.06 ± .05	.07 ± .03	.07 ± .03	.09 ± .02
2000	.06 ± .04	.06 ± .03	.07 ± .03	.09 ± .03

Table 21: Convergence of bias for doubly robust estimator using our weights and **DirX**, with **exp** link

n	<b>DirZ:OptZ</b> <sub>0.001</sub>	<b>DirZ:OptZ</b> <sub>0.2</sub>	<b>DirZ:OptZ</b> <sub>1.0</sub>	<b>DirZ:OptZ</b> <sub>5.0</sub>
200	.05 ± .14	.12 ± .10	.23 ± .09	.43 ± .09
500	.03 ± .10	.09 ± .07	.16 ± .07	.32 ± .07
1000	.02 ± .06	.06 ± .05	.10 ± .05	.23 ± .06
2000	.01 ± .04	.03 ± .03	.07 ± .03	.16 ± .04

Table 22: Convergence of bias for doubly robust estimator using our weights and **DirZ**, with **exp** link

n	<b>SimplexOptZ</b> <sub>0.001</sub>	<b>SimplexOptZ</b> <sub>0.2</sub>	<b>SimplexOptZ</b> <sub>1.0</sub>	<b>SimplexOptZ</b> <sub>5.0</sub>
200	.02 ± .05	.10 ± .06	.20 ± .08	.38 ± .11
500	.02 ± .04	.07 ± .04	.14 ± .05	.27 ± .06
1000	.01 ± .02	.04 ± .02	.08 ± .03	.19 ± .04
2000	.01 ± .01	.03 ± .02	.06 ± .02	.14 ± .02

Table 23: Convergence of bias for weighted estimator using our weights and constraining  $W \in n\Delta^n$ , with **exp** link

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.10 ± .06	.74 ± .13	.80 ± .13	.91 ± .15	1.0 ± .16	.07 ± .07	1.0 ± .08
500	.11 ± .03	.76 ± .07	.81 ± .08	.92 ± .09	1.0 ± .10	.07 ± .08	1.0 ± .09
1000	.10 ± .03	.74 ± .05	.79 ± .05	.90 ± .06	1.0 ± .06	.06 ± .07	1.0 ± .10
2000	.10 ± .02	.73 ± .03	.78 ± .03	.88 ± .04	.99 ± .04	.07 ± .07	1.0 ± .09

Table 24: Convergence of bias for benchmark methods, with **exp** link

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.47 ± .04	.35 ± .02	.39 ± .02	.48 ± .02
500	.36 ± .05	.27 ± .02	.30 ± .02	.40 ± .02
1000	.25 ± .02	.22 ± .01	.25 ± .01	.37 ± .01
2000	.14 ± .01	.14 ± .01	.17 ± .01	.27 ± .01

Table 25: Convergence of RMSE for weighted estimator using our weights, with **cubic** link

n	<b>DirX:OptZ</b> <sub>0.001</sub>	<b>DirX:OptZ</b> <sub>0.2</sub>	<b>DirX:OptZ</b> <sub>1.0</sub>	<b>DirX:OptZ</b> <sub>5.0</sub>
200	.58 ± .07	.41 ± .03	.42 ± .02	.38 ± .02
500	.37 ± .04	.33 ± .02	.35 ± .02	.37 ± .02
1000	.31 ± .02	.31 ± .02	.33 ± .02	.39 ± .01
2000	.21 ± .02	.23 ± .02	.26 ± .01	.32 ± .01

Table 26: Convergence of RMSE for doubly robust estimator using our weights and **DirX**, with **cubic** link

n	<b>DirZ:OptZ</b> <sub>0.001</sub>	<b>DirZ:OptZ</b> <sub>0.2</sub>	<b>DirZ:OptZ</b> <sub>1.0</sub>	<b>DirZ:OptZ</b> <sub>5.0</sub>
200	.49 ± .04	.42 ± .03	.54 ± .03	.76 ± .02
500	.38 ± .05	.30 ± .02	.38 ± .02	.59 ± .02
1000	.27 ± .02	.25 ± .02	.32 ± .02	.52 ± .02
2000	.16 ± .01	.16 ± .01	.22 ± .01	.39 ± .01

Table 27: Convergence of RMSE for doubly robust estimator using our weights and **DirZ**, with **cubic** link

n	<b>SimplexOptZ</b> <sub>0.001</sub>	<b>SimplexOptZ</b> <sub>0.2</sub>	<b>SimplexOptZ</b> <sub>1.0</sub>	<b>SimplexOptZ</b> <sub>5.0</sub>
200	.45 ± .04	.41 ± .03	.52 ± .03	.70 ± .03
500	.37 ± .05	.30 ± .02	.37 ± .02	.55 ± .02
1000	.26 ± .02	.24 ± .02	.31 ± .02	.50 ± .02
2000	.14 ± .01	.15 ± .01	.20 ± .01	.35 ± .01

Table 28: Convergence of RMSE for weighted estimator using our weights and constraining  $W \in n\Delta^n$ , with **cubic** link

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.46 ± .04	1.1 ± .02	1.1 ± .03	1.3 ± .03	1.4 ± .03	.36 ± .02	1.4 ± .01
500	.38 ± .02	1.1 ± .01	1.2 ± .01	1.3 ± .01	1.4 ± .01	.34 ± .02	1.4 ± .01
1000	.39 ± .01	1.1 ± .01	1.2 ± .01	1.3 ± .01	1.4 ± .01	.35 ± .02	1.4 ± .01
2000	.35 ± .01	1.1 ± .01	1.2 ± .01	1.3 ± .01	1.4 ± .01	.39 ± .02	1.4 ± .01

Table 29: Convergence of RMSE for benchmark methods, with **cubic** link

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.01 ± .47	.16 ± .31	.29 ± .27	.45 ± .16
500	−0.01 ± .36	.12 ± .24	.22 ± .21	.37 ± .15
1000	.03 ± .25	.12 ± .18	.20 ± .15	.35 ± .12
2000	.02 ± .14	.07 ± .12	.13 ± .11	.26 ± .08

Table 30: Convergence of bias for weighted estimator using our weights, with **cubic** link

n	<b>DirX:OptZ</b> <sub>0.001</sub>	<b>DirX:OptZ</b> <sub>0.2</sub>	<b>DirX:OptZ</b> <sub>1.0</sub>	<b>DirX:OptZ</b> <sub>5.0</sub>
200	.13 ± .57	.26 ± .32	.33 ± .27	.34 ± .17
500	.12 ± .35	.22 ± .25	.28 ± .21	.34 ± .15
1000	.18 ± .26	.24 ± .19	.29 ± .15	.37 ± .11
2000	.14 ± .16	.18 ± .14	.22 ± .12	.31 ± .09

Table 31: Convergence of bias for doubly robust estimator using our weights and **DirX**, with **cubic** link

n	<b>DirZ:OptZ</b> <sub>0.001</sub>	<b>DirZ:OptZ</b> <sub>0.2</sub>	<b>DirZ:OptZ</b> <sub>1.0</sub>	<b>DirZ:OptZ</b> <sub>5.0</sub>
200	.04 ± .49	.25 ± .34	.45 ± .29	.74 ± .20
500	.03 ± .38	.17 ± .25	.32 ± .21	.57 ± .16
1000	.04 ± .26	.16 ± .19	.28 ± .16	.50 ± .13
2000	.03 ± .16	.10 ± .13	.19 ± .12	.37 ± .10

Table 32: Convergence of bias for doubly robust estimator using our weights and **DirZ**, with **cubic** link

n	<b>SimplexOptZ</b> <sub>0.001</sub>	<b>SimplexOptZ</b> <sub>0.2</sub>	<b>SimplexOptZ</b> <sub>1.0</sub>	<b>SimplexOptZ</b> <sub>5.0</sub>
200	.02 ± .45	.22 ± .35	.40 ± .34	.65 ± .28
500	.01 ± .37	.15 ± .26	.29 ± .23	.52 ± .18
1000	.04 ± .25	.15 ± .19	.27 ± .17	.47 ± .14
2000	.02 ± .14	.08 ± .13	.17 ± .11	.34 ± .09

Table 33: Convergence of bias for weighted estimator using our weights and constraining  $W \in n\Delta^n$ , with **cubic** link

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.26 ± .38	1.1 ± .20	1.1 ± .20	1.2 ± .21	1.4 ± .22	.32 ± .16	1.4 ± .12
500	.31 ± .22	1.1 ± .10	1.2 ± .10	1.3 ± .10	1.4 ± .11	.29 ± .18	1.4 ± .10
1000	.37 ± .14	1.1 ± .07	1.2 ± .07	1.3 ± .08	1.4 ± .08	.32 ± .16	1.4 ± .10
2000	.34 ± .09	1.1 ± .05	1.2 ± .06	1.3 ± .06	1.4 ± .06	.34 ± .18	1.4 ± .11

Table 34: Convergence of bias for benchmark methods, with **cubic** link

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.09 ± .01	.08 ± .01	.13 ± .01	.23 ± .01
500	.06 ± .01	.06 ± .00	.09 ± .00	.16 ± .00
1000	.04 ± .01	.04 ± .00	.06 ± .00	.12 ± .00
2000	.02 ± .00	.03 ± .00	.04 ± .00	.08 ± .00

Table 35: Convergence of RMSE for weighted estimator using our weights, with **linear** link

n	<b>DirX:OptZ</b> <sub>0.001</sub>	<b>DirX:OptZ</b> <sub>0.2</sub>	<b>DirX:OptZ</b> <sub>1.0</sub>	<b>DirX:OptZ</b> <sub>5.0</sub>
200	.15 ± .01	.13 ± .01	.13 ± .01	.13 ± .01
500	.14 ± .01	.13 ± .01	.13 ± .01	.13 ± .00
1000	.13 ± .01	.13 ± .00	.13 ± .00	.13 ± .00
2000	.12 ± .00	.12 ± .00	.12 ± .00	.12 ± .00

Table 36: Convergence of RMSE for doubly robust estimator using our weights and **DirX**, with **linear** link

n	<b>DirZ:OptZ</b> <sub>0.001</sub>	<b>DirZ:OptZ</b> <sub>0.2</sub>	<b>DirZ:OptZ</b> <sub>1.0</sub>	<b>DirZ:OptZ</b> <sub>5.0</sub>
200	.11 ± .01	.11 ± .01	.18 ± .01	.35 ± .01
500	.06 ± .01	.08 ± .01	.13 ± .01	.24 ± .01
1000	.05 ± .01	.06 ± .00	.09 ± .00	.18 ± .00
2000	.03 ± .00	.04 ± .00	.06 ± .00	.12 ± .00

Table 37: Convergence of RMSE for doubly robust estimator using our weights and **DirZ**, with **linear** link

n	<b>SimplexOptZ</b> <sub>0.001</sub>	<b>SimplexOptZ</b> <sub>0.2</sub>	<b>SimplexOptZ</b> <sub>1.0</sub>	<b>SimplexOptZ</b> <sub>5.0</sub>
200	.09 ± .01	.09 ± .01	.15 ± .01	.29 ± .01
500	.06 ± .01	.07 ± .01	.10 ± .01	.19 ± .01
1000	.04 ± .01	.04 ± .00	.07 ± .00	.14 ± .00
2000	.02 ± .00	.03 ± .00	.05 ± .00	.09 ± .00

Table 38: Convergence of RMSE for weighted estimator using our weights and constraining  $W \in n\Delta^n$ , with **linear** link

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.15 ± .01	.57 ± .01	.60 ± .01	.66 ± .01	.72 ± .01	.13 ± .00	.76 ± .00
500	.15 ± .01	.57 ± .00	.60 ± .00	.66 ± .00	.72 ± .00	.13 ± .00	.76 ± .00
1000	.14 ± .00	.57 ± .00	.60 ± .00	.66 ± .00	.72 ± .00	.13 ± .00	.76 ± .00
2000	.14 ± .00	.57 ± .00	.60 ± .00	.66 ± .00	.72 ± .00	.13 ± .00	.76 ± .00

Table 39: Convergence of RMSE for benchmark methods, with **linear** link



n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.03 ± .09	.06 ± .06	.11 ± .05	.23 ± .04
500	.02 ± .05	.04 ± .05	.08 ± .04	.15 ± .04
1000	.01 ± .04	.03 ± .03	.05 ± .03	.11 ± .03
2000	.01 ± .02	.02 ± .02	.04 ± .02	.08 ± .02

Table 40: Convergence of bias for weighted estimator using our weights, with **linear** link

n	<b>DirX:OptZ</b> <sub>0.001</sub>	<b>DirX:OptZ</b> <sub>0.2</sub>	<b>DirX:OptZ</b> <sub>1.0</sub>	<b>DirX:OptZ</b> <sub>5.0</sub>
200	.12 ± .09	.12 ± .06	.12 ± .05	.13 ± .04
500	.13 ± .06	.12 ± .05	.12 ± .04	.12 ± .04
1000	.12 ± .04	.12 ± .03	.12 ± .03	.12 ± .03
2000	.12 ± .03	.12 ± .03	.12 ± .02	.12 ± .02

Table 41: Convergence of bias for doubly robust estimator using our weights and **DirX**, with **linear** link

n	<b>DirZ:OptZ</b> <sub>0.001</sub>	<b>DirZ:OptZ</b> <sub>0.2</sub>	<b>DirZ:OptZ</b> <sub>1.0</sub>	<b>DirZ:OptZ</b> <sub>5.0</sub>
200	.04 ± .10	.09 ± .07	.17 ± .06	.34 ± .05
500	.02 ± .06	.06 ± .05	.12 ± .05	.24 ± .05
1000	.02 ± .04	.05 ± .04	.08 ± .04	.18 ± .04
2000	.01 ± .03	.03 ± .03	.06 ± .03	.12 ± .02

Table 42: Convergence of bias for doubly robust estimator using our weights and **DirZ**, with **linear** link

n	<b>SimplexOptZ</b> <sub>0.001</sub>	<b>SimplexOptZ</b> <sub>0.2</sub>	<b>SimplexOptZ</b> <sub>1.0</sub>	<b>SimplexOptZ</b> <sub>5.0</sub>
200	.02 ± .08	.07 ± .06	.14 ± .06	.29 ± .06
500	.02 ± .05	.05 ± .05	.09 ± .05	.19 ± .04
1000	.01 ± .04	.03 ± .03	.06 ± .03	.14 ± .03
2000	.01 ± .02	.02 ± .02	.04 ± .02	.09 ± .02

Table 43: Convergence of bias for weighted estimator using our weights and constraining  $W \in n\Delta^n$ , with **linear** link

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.13 ± .08	.57 ± .05	.60 ± .05	.66 ± .05	.72 ± .05	.13 ± .04	.76 ± .04
500	.14 ± .05	.57 ± .04	.60 ± .03	.66 ± .03	.72 ± .03	.12 ± .04	.76 ± .02
1000	.14 ± .04	.57 ± .02	.60 ± .02	.66 ± .02	.72 ± .02	.13 ± .03	.76 ± .03
2000	.13 ± .03	.57 ± .02	.60 ± .02	.66 ± .02	.72 ± .02	.13 ± .03	.76 ± .02

Table 44: Convergence of bias for benchmark methods, with **linear** link

n	OptZ <sub>0.001</sub>	OptZ <sub>0.2</sub>	OptZ <sub>1.0</sub>	OptZ <sub>5.0</sub>
200	.94 ± .21	.31 ± .04	.52 ± .04	1.1 ± .04
500	.32 ± .07	.21 ± .03	.33 ± .03	.76 ± .03
1000	.23 ± .04	.13 ± .02	.18 ± .01	.48 ± .01
2000	.10 ± .01	.09 ± .01	.11 ± .01	.28 ± .01

Table 45: Convergence of RMSE for weighted estimator using our weights, with **step** link and  $\dim_z = 2$

n	IPS	OptX <sub>0.001</sub>	OptX <sub>0.2</sub>	OptX <sub>1.0</sub>	OptX <sub>5.0</sub>	DirX	DirZ
200	.50 ± .05	2.2 ± .04	2.3 ± .04	2.5 ± .04	2.7 ± .03	.91 ± .07	2.8 ± .04
500	.50 ± .03	2.3 ± .03	2.4 ± .03	2.6 ± .02	2.8 ± .02	.57 ± .04	2.8 ± .03
1000	.47 ± .02	2.3 ± .01	2.4 ± .01	2.6 ± .01	2.8 ± .01	.50 ± .04	2.9 ± .02
2000	.42 ± .02	2.2 ± .01	2.3 ± .01	2.5 ± .01	2.8 ± .01	.40 ± .02	2.9 ± .01

Table 46: Convergence of RMSE for benchmark methods, with **step** link and  $\dim_z = 2$

n	OptZ <sub>0.001</sub>	OptZ <sub>0.2</sub>	OptZ <sub>1.0</sub>	OptZ <sub>5.0</sub>
200	.30 ± .89	.17 ± .25	.46 ± .23	1.1 ± .20
500	.10 ± .31	.13 ± .16	.29 ± .16	.74 ± .15
1000	.02 ± .22	.07 ± .11	.15 ± .10	.47 ± .09
2000	−0.01 ± .10	.05 ± .07	.08 ± .07	.27 ± .07

Table 47: Convergence of bias for weighted estimator using our weights, with **step** link and  $\dim_z = 2$

n	IPS	OptX <sub>0.001</sub>	OptX <sub>0.2</sub>	OptX <sub>1.0</sub>	OptX <sub>5.0</sub>	DirX	DirZ
200	.39 ± .31	2.2 ± .25	2.3 ± .24	2.5 ± .22	2.7 ± .20	.83 ± .38	2.8 ± .26
500	.46 ± .20	2.2 ± .16	2.3 ± .15	2.6 ± .14	2.8 ± .13	.53 ± .21	2.8 ± .18
1000	.45 ± .11	2.3 ± .08	2.4 ± .07	2.6 ± .07	2.8 ± .06	.45 ± .21	2.9 ± .12
2000	.40 ± .11	2.2 ± .08	2.3 ± .08	2.5 ± .07	2.7 ± .07	.38 ± .11	2.9 ± .05

Table 48: Convergence of bias for benchmark methods, with **step** link and  $\dim_z = 2$

## C.2 Results with Varying Dimensionality of Hidden Confounders

We present some additional results here experimenting with varying the dimensionality of the hidden confounder  $Z$ . We experimented with increasing the dimensionality of  $Z$  to either 2 or 5, while keeping the dimensionalities of all other components equal. We extended the data generating process components described in Appendix [B.1](#) as follows:

- The  $i'$ th row of  $\alpha$  is defined by cyclically rotating the elements of the  $\alpha$  defined in Appendix [B.1](#) by  $i - 1$  places clockwise.
- The  $i'$ th row of  $\beta$  is defined as  $[0.5 - 0.05(i - 1), 0.5 + 0.05(i - 1)]$ .
- The  $i'$ th row of  $\zeta(0)$  is defined as  $1.0 - 0.3(i - 1)$ .
- The  $i'$ th row of  $\zeta(1)$  is defined as  $-0.5 + 0.3(i - 1)$ .
- The extra rows of all other matrices that need to be extended to accommodate higher dimensional  $Z$  are simply duplicated.

We present results below for our method and benchmarks, for the **step** link function. We can observe that the same overall pattern of behavior occurs, with our methods still appearing to be consistent, though with slower convergence as the dimensionality of  $Z$  increases.

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.72 ± .07	.42 ± .05	.64 ± .04	.92 ± .02
500	.47 ± .07	.31 ± .03	.47 ± .03	.79 ± .02
1000	.35 ± .04	.22 ± .02	.36 ± .02	.64 ± .01
2000	.31 ± .05	.17 ± .01	.26 ± .01	.47 ± .01

Table 49: Convergence of RMSE for weighted estimator using our weights, with **step** link and  $\dim_z = 5$

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.40 ± .06	1.1 ± .05	1.2 ± .04	1.3 ± .04	1.4 ± .04	.44 ± .04	1.5 ± .05
500	.33 ± .03	1.1 ± .03	1.2 ± .03	1.4 ± .03	1.5 ± .03	.40 ± .04	1.5 ± .04
1000	.29 ± .02	1.1 ± .02	1.2 ± .02	1.3 ± .02	1.5 ± .02	.35 ± .03	1.6 ± .02
2000	.26 ± .02	1.1 ± .01	1.1 ± .01	1.3 ± .01	1.4 ± .01	.31 ± .02	1.5 ± .01

Table 50: Convergence of RMSE for benchmark methods, with **step** link and  $\dim_z = 5$

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.03 ± .72	.32 ± .27	.61 ± .22	.91 ± .13
500	.01 ± .47	.23 ± .21	.44 ± .16	.78 ± .11
1000	−0.01 ± .35	.18 ± .13	.34 ± .11	.64 ± .08
2000	.03 ± .31	.13 ± .11	.24 ± .09	.47 ± .07

Table 51: Convergence of bias for weighted estimator using our weights, with **step** link and  $\dim_z = 5$

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.23 ± .33	1.1 ± .25	1.1 ± .23	1.3 ± .21	1.4 ± .20	.36 ± .25	1.5 ± .29
500	.24 ± .23	1.1 ± .16	1.2 ± .16	1.4 ± .15	1.5 ± .15	.32 ± .23	1.5 ± .21
1000	.25 ± .15	1.1 ± .11	1.2 ± .11	1.3 ± .11	1.5 ± .10	.31 ± .16	1.6 ± .10
2000	.24 ± .12	1.1 ± .06	1.1 ± .06	1.3 ± .06	1.4 ± .06	.29 ± .12	1.5 ± .07

Table 52: Convergence of bias for benchmark methods, with **step** link and  $\dim_z = 5$

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.35 ± .05	.20 ± .03	.26 ± .02	.70 ± .03
500	.21 ± .02	.12 ± .01	.18 ± .01	.38 ± .01
1000	.13 ± .02	.05 ± .01	.11 ± .01	.22 ± .01
2000	.07 ± .01	.03 ± .00	.07 ± .01	.16 ± .01

Table 53: Convergence of RMSE for weighted estimator using our weights, with **step** link and  $\sigma_x = 0.1$

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.18 ± .02	.33 ± .03	.44 ± .03	.76 ± .03	1.6 ± .04	.34 ± .06	3.2 ± .05
500	.07 ± .01	.18 ± .01	.25 ± .01	.45 ± .01	1.0 ± .02	.08 ± .02	3.2 ± .04
1000	.06 ± .01	.09 ± .01	.13 ± .01	.26 ± .01	.68 ± .01	.04 ± .01	3.3 ± .02
2000	.04 ± .00	.05 ± .00	.08 ± .00	.16 ± .00	.44 ± .01	.04 ± .00	3.3 ± .01

Table 54: Convergence of RMSE for benchmark methods, with **step** link and  $\sigma_x = 0.1$

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	-0.12 ± .32	.13 ± .15	.23 ± .13	.67 ± .17
500	.11 ± .18	.11 ± .06	.17 ± .06	.37 ± .06
1000	.03 ± .13	.02 ± .04	.09 ± .05	.21 ± .05
2000	-0.03 ± .06	.01 ± .03	.06 ± .03	.15 ± .03

Table 55: Convergence of bias for weighted estimator using our weights, with **step** link and  $\sigma_x = 0.1$

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	-0.01 ± .18	.29 ± .16	.41 ± .16	.74 ± .18	1.6 ± .22	.17 ± .30	3.2 ± .29
500	.01 ± .07	.17 ± .07	.24 ± .07	.45 ± .07	1.0 ± .09	.03 ± .07	3.2 ± .22
1000	-0.02 ± .06	.08 ± .04	.12 ± .04	.26 ± .05	.67 ± .07	-0.01 ± .04	3.3 ± .10
2000	.00 ± .04	.05 ± .02	.08 ± .03	.16 ± .03	.43 ± .03	.00 ± .04	3.3 ± .08

Table 56: Convergence of bias for benchmark methods, with **step** link and  $\sigma_x = 0.1$

### C.3 Results with varying confounder strength

We present some additional results here experimenting with varying the strength of relationship between  $X$  and  $Z$ . We do this by varying the  $\sigma_X$  parameter in the data generating process described in Appendix [B.1](#). Lower values of  $\sigma_X$  mean that more of the variance of  $X$  is explained by  $Z$ , corresponding to less confounding, whereas higher values of  $\sigma_X$  indicate that less of the variance of  $X$  is explained by  $Z$ , corresponding to more confounding.

We experimented with  $\sigma_X$  values in the range  $[0.1, 1.0, 4.0, 10.0]$ , noting that  $\sigma_X = 4.0$  is what was used in prior experiments, and we again used the **step** link function in this experiment. Our results are presented below. As expected as the level of confounding increases, the performance of the benchmark methods strongly decrease. However, our methods appear to be robust to the increasing level of confounding, maintaining close to zero bias with sufficiently low levels of regularization.

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.33 ± .05	.23 ± .02	.30 ± .03	.74 ± .03
500	.13 ± .01	.12 ± .01	.16 ± .01	.37 ± .02
1000	.13 ± .02	.06 ± .01	.11 ± .01	.24 ± .01
2000	.06 ± .01	.04 ± .00	.09 ± .01	.16 ± .01

Table 57: Convergence of RMSE for weighted estimator using our weights, with **step** link and  $\sigma_x = 1.0$

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.20 ± .03	2.2 ± .04	2.4 ± .04	2.7 ± .04	3.1 ± .04	.41 ± .05	3.1 ± .05
500	.12 ± .01	2.1 ± .03	2.3 ± .03	2.6 ± .03	3.0 ± .03	.22 ± .02	3.1 ± .03
1000	.08 ± .01	2.0 ± .02	2.1 ± .02	2.5 ± .01	3.0 ± .01	.09 ± .01	3.2 ± .02
2000	.05 ± .01	1.7 ± .01	1.9 ± .01	2.3 ± .01	2.8 ± .01	.06 ± .01	3.2 ± .01

Table 58: Convergence of RMSE for benchmark methods, with **step** link and  $\sigma_x = 1.0$

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	−0.15 ± .29	.17 ± .15	.26 ± .15	.72 ± .16
500	.07 ± .11	.08 ± .09	.14 ± .08	.36 ± .09
1000	.07 ± .10	.04 ± .05	.10 ± .05	.23 ± .04
2000	.02 ± .05	.02 ± .04	.08 ± .04	.16 ± .03

Table 59: Convergence of bias for weighted estimator using our weights, with **step** link and  $\sigma_x = 1.0$

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.07 ± .18	2.2 ± .24	2.4 ± .23	2.7 ± .22	3.1 ± .21	.21 ± .35	3.1 ± .30
500	.04 ± .11	2.1 ± .18	2.3 ± .18	2.6 ± .17	3.0 ± .16	.09 ± .20	3.1 ± .19
1000	.04 ± .07	1.9 ± .09	2.1 ± .08	2.5 ± .08	3.0 ± .07	.04 ± .08	3.2 ± .10
2000	.03 ± .05	1.7 ± .08	1.9 ± .07	2.3 ± .07	2.8 ± .07	.05 ± .04	3.2 ± .06

Table 60: Convergence of bias for benchmark methods, with **step** link and  $\sigma_x = 1.0$

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.39 ± .07	.24 ± .02	.36 ± .02	.81 ± .02
500	.19 ± .02	.18 ± .02	.23 ± .02	.49 ± .02
1000	.11 ± .01	.11 ± .01	.13 ± .01	.27 ± .01
2000	.08 ± .01	.08 ± .01	.09 ± .01	.17 ± .01

Table 61: Convergence of RMSE for weighted estimator using our weights, with **step** link and  $\sigma_x = 4.0$

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.47 ± .03	2.0 ± .03	2.1 ± .03	2.3 ± .02	2.5 ± .02	.52 ± .02	2.6 ± .02
500	.48 ± .03	2.0 ± .02	2.1 ± .02	2.3 ± .02	2.6 ± .02	.48 ± .02	2.6 ± .01
1000	.39 ± .02	2.0 ± .01	2.1 ± .01	2.3 ± .01	2.5 ± .01	.48 ± .02	2.6 ± .01
2000	.40 ± .01	2.0 ± .01	2.1 ± .01	2.3 ± .01	2.5 ± .01	.45 ± .02	2.6 ± .01

Table 62: Convergence of RMSE for benchmark methods, with **step** link and  $\sigma_x = 4.0$

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.03 ± .39	.11 ± .21	.29 ± .21	.78 ± .18
500	.09 ± .17	.10 ± .15	.17 ± .16	.47 ± .15
1000	.02 ± .11	.05 ± .09	.08 ± .09	.25 ± .09
2000	.03 ± .07	.05 ± .06	.07 ± .07	.16 ± .07

Table 63: Convergence of bias for weighted estimator using our weights, with **step** link and  $\sigma_x = 4.0$

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.40 ± .25	1.9 ± .21	2.1 ± .20	2.3 ± .19	2.5 ± .18	.49 ± .18	2.6 ± .14
500	.43 ± .21	2.0 ± .16	2.1 ± .15	2.3 ± .14	2.6 ± .13	.45 ± .16	2.6 ± .12
1000	.37 ± .12	2.0 ± .10	2.1 ± .09	2.3 ± .09	2.5 ± .08	.46 ± .15	2.6 ± .11
2000	.39 ± .10	2.0 ± .08	2.1 ± .07	2.3 ± .07	2.5 ± .07	.42 ± .17	2.6 ± .11

Table 64: Convergence of bias for benchmark methods, with **step** link and  $\sigma_x = 4.0$

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.60 ± .07	.37 ± .04	.53 ± .04	.97 ± .03
500	.42 ± .05	.32 ± .04	.39 ± .03	.69 ± .03
1000	.27 ± .03	.27 ± .03	.32 ± .03	.52 ± .02
2000	.20 ± .03	.18 ± .02	.24 ± .02	.40 ± .02

Table 65: Convergence of RMSE for weighted estimator using our weights, with **step** link and  $\sigma_x = 10.0$

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.95 ± .05	1.6 ± .04	1.6 ± .04	1.7 ± .04	1.8 ± .04	1.2 ± .06	1.8 ± .04
500	.91 ± .03	1.6 ± .02	1.6 ± .02	1.7 ± .02	1.9 ± .02	.96 ± .05	1.8 ± .03
1000	.91 ± .02	1.6 ± .02	1.6 ± .01	1.7 ± .01	1.9 ± .01	.92 ± .03	1.9 ± .02
2000	.92 ± .02	1.6 ± .02	1.6 ± .01	1.7 ± .01	1.9 ± .01	.92 ± .02	1.9 ± .02

Table 66: Convergence of RMSE for benchmark methods, with **step** link and  $\sigma_x = 10.0$

n	<b>OptZ</b> <sub>0.001</sub>	<b>OptZ</b> <sub>0.2</sub>	<b>OptZ</b> <sub>1.0</sub>	<b>OptZ</b> <sub>5.0</sub>
200	.01 ± .60	.23 ± .29	.48 ± .23	.95 ± .18
500	.01 ± .42	.19 ± .26	.34 ± .20	.68 ± .14
1000	.06 ± .27	.13 ± .24	.25 ± .21	.50 ± .15
2000	.03 ± .20	.10 ± .15	.19 ± .14	.38 ± .13

Table 67: Convergence of bias for weighted estimator using our weights, with **step** link and  $\sigma_x = 10.0$

n	<b>IPS</b>	<b>OptX</b> <sub>0.001</sub>	<b>OptX</b> <sub>0.2</sub>	<b>OptX</b> <sub>1.0</sub>	<b>OptX</b> <sub>5.0</sub>	<b>DirX</b>	<b>DirZ</b>
200	.92 ± .26	1.5 ± .23	1.6 ± .22	1.7 ± .22	1.8 ± .21	1.1 ± .35	1.8 ± .24
500	.89 ± .19	1.6 ± .13	1.6 ± .13	1.7 ± .12	1.9 ± .12	.92 ± .26	1.8 ± .19
1000	.91 ± .12	1.6 ± .08	1.6 ± .08	1.7 ± .08	1.9 ± .08	.91 ± .18	1.9 ± .12
2000	.92 ± .11	1.6 ± .08	1.6 ± .08	1.7 ± .08	1.9 ± .08	.91 ± .11	1.9 ± .11

Table 68: Convergence of bias for benchmark methods, with **step** link and  $\sigma_x = 10.0$