Generalizing Off-Policy Learning under Sample Selection Bias (Supplementary material)

Tobias Hatt[1] **Daniel Tschernutter**[1] **Stefan Feuerriegel**[1,2]

[1]ETH Zurich, Switzerland
[2]LMU Munich, Germany

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

A MATHEMATICAL APPENDIX

A.1 PROOF OF PROPOSITION 1

We know that, with the Radon-Nikodým derivative $R = \mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{P}_{\text{Train}}$,

$$V(\pi) = \mathbb{E}\big[Y(\pi)\big] = \mathbb{E}_{\text{Train}}\big[R\,Y^{\pi}\big]. \tag{31}$$

where

$$R = \frac{\mathrm{d}\mathbb{P}(X,T,Y)}{\mathrm{d}\mathbb{P}_{\text{Train}}(X,T,Y)} = \frac{\mathrm{d}\mathbb{P}(X,T,Y)}{\mathrm{d}\mathbb{P}(X,T,Y \mid S=1)} \tag{32}$$

$$= \frac{\mathrm{d}\mathbb{P}(X,T,Y)}{\mathrm{d}\mathbb{P}(X,T,Y,S=1)}\mathbb{P}\left(S=1\right) = \frac{\mathbb{P}\left(S=1\right)}{\mathbb{P}(S=1 \mid X,T,Y)}. \tag{33}$$

□

Remark 1. With the above result for the Radon-Nikodým derivative, we can see the effect of the selection variable S: If S does not depend on X, T, and Y, then $R = 1$. Therefore, $\mathbb{P}$ would be identical to $\mathbb{P}_{\text{Train}}$ and, as a consequence, the policy value on the target population, i. e., $V_{\text{Target}}(\pi)$, would coincide with the policy value on the training data, i. e., $\mathbb{E}_{\text{Train}}\big[Y^{\pi}\big]$. If, however, S depends on X, T, and Y, then the policy value on the target population does not coincide with the policy value on the training data and, therefore, $V_{\text{Target}}(\pi) \neq \mathbb{E}_{\text{Train}}\big[Y^{\pi}\big]$.

A.2 PROOF OF THEOREM 1

Let $Z = \frac{1}{n}\sum_{i=1}^{n} R_i^*$. Then,

$$V_{\text{Target}}(\pi) \leqslant \hat{V}^*_{\text{Target}}(\pi) + \sup_{\pi\in\Pi}\left|\hat{V}^*_{\text{Target}}(\pi) - V_{\text{Target}}(\pi)\right|, \tag{34}$$

and

$$\sup_{\pi\in\Pi}\left|\hat{V}^*_{\text{Target}}(\pi) - V_{\text{Target}}(\pi)\right| \tag{35}$$

$$= \sup_{\pi\in\Pi}\left|\frac{\frac{1}{n}\sum_{i=1}^{n} R_i^*\psi_i(\pi)}{Z} - \frac{V_{\text{Target}}(\pi)}{Z} + \frac{V_{\text{Target}}(\pi)(1-Z)}{Z}\right| \tag{36}$$

$$\leqslant \frac{1}{Z}\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n} R_i^*\psi_i(\pi) - V(\pi)\right| + \sup_{\pi\in\Pi}\frac{C\left|1-Z\right|}{Z}. \tag{37}$$

We let

$$T = \sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n} R_i^*\psi_i(\pi) - V_{\text{Target}}(\pi)\right|. \tag{38}$$

Since $|Y| \leqslant C$ and, therefore, $|\mu_t(x)| \leqslant C$, $R_i^* \leqslant u$, and $1-\eta \geqslant \pi^b(x) \geqslant \eta$ (for some $\eta > 0$ due to positivity), we have that

1.) for $\psi_i^{\mathrm{DM}}(\pi)$ from (8): $T = \sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n} R_i^*(\pi(X_i)\mu_1(X_i) + (1-\pi(X_i))\mu_0(X_i)) - V(\pi)\right|$ satisfies bounded differences with $\frac{4Cu}{n}$,

2.) for $\psi_i^{\mathrm{NIPW}}(\pi)$ from (9): $T = \sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n} R_i^*(\frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^{n} W_j^{\mathrm{IPW}}}(1-2T_i)(1-T_i-\pi(X_i))Y_i) - V(\pi)\right|$, satisfies bounded differences with $\frac{4Cu}{n}\frac{1-\eta}{\eta}$,

3.) for $\psi_i^{\mathrm{DR}}(\pi)$ from (10): $T = \sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n} R_i^*(\psi_i^{\mathrm{DM}}(\pi) + W_i^{\mathrm{IPW}}(1-2T_i)(1-T_i-\pi(X_i))(Y_i-\mu_{T_i}(X_i))) - V(\pi)\right|$, satisfies bounded differences with $\frac{4Cu}{n}\frac{1+\eta}{\eta}$.

Hence, T satisfies bounded differences with $\frac{4Cu}{n}K_\psi$, where $K_\psi = 1$ for $\psi_i^{\mathrm{DM}}(\pi)$, $K_\psi = \frac{1-\eta}{\eta}$ for $\psi_i^{\mathrm{NIPW}}(\pi)$, and $K_\psi = \frac{1+\eta}{\eta}$ for $\psi_i^{\mathrm{DR}}(\pi)$.

Thus, using McDiarmid's inequality yields

$$\mathbb{P}\left(T - \mathbb{E}[T] \geqslant \epsilon\right) \leqslant \exp(-\frac{n\epsilon^2}{8C^2u^2K_\psi^2}). \tag{39}$$

Therefore, we have that

$$\mathbb{P}\left(T - \mathbb{E}[T] \leqslant \epsilon\right) \geqslant 1 - \exp(-\frac{n\epsilon^2}{8C^2u^2K_\psi^2}). \tag{40}$$

Using $p_1 = \exp(-\frac{n\epsilon^2}{8C^2u^2K_\psi^2})$ and, therefore, $\epsilon = 2CuK_\psi\sqrt{\frac{2\log(1/p_1)}{n}}$, we have that with probability at least $1 - p_1$,

$$T \leqslant \mathbb{E}[T] + 2CuK_\psi\sqrt{\frac{2\log(1/p_1)}{n}}. \tag{41}$$

Since $\mathbb{E}[R_i^*\psi_i(\pi)] = V(\pi)$, a standard symmetrization argument yields

$$\mathbb{E}[T] \leqslant \mathbb{E}\Big[\frac{1}{2^n}\Sigma_{\sigma\in\{-1,+1\}^n}\sup_{\pi\in\Pi}|\frac{1}{n}\sum_{i=1}^{n}\sigma_i R_i^*\psi_i(\pi)|\Big]. \tag{42}$$

Then, using the Rademacher comparison theorem (Thm 4.12 in Ledoux and Talagrand [2013]), this yields

$$\mathbb{E}[T] \leqslant 2CuK_\psi\mathbb{E}[\mathcal{R}_n(\Pi)], \tag{43}$$

where K_ψ is from above and depends on whether one uses $\psi_i^{\mathrm{DM}}(\pi)$, $\psi_i^{\mathrm{NIPW}}(\pi)$, or $\psi_i^{\mathrm{DR}}(\pi)$. Moreover, $\mathcal{R}_n(\Pi)$ satisfies bounded differences with constants $\frac{2}{n}$ and, hence, we can again use McDiarmid's inequality, which yields

$$\mathbb{P}\left(\mathbb{E}[\mathcal{R}_n(\Pi)] - \mathcal{R}_n(\Pi) \geqslant \epsilon\right) \leqslant \exp(\frac{-\epsilon^2 n}{2}). \tag{44}$$

Therefore, we have that

$$\mathbb{P}\left(\mathbb{E}[\mathcal{R}_n(\Pi)] - \mathcal{R}_n(\Pi) \leqslant \epsilon\right) \geqslant 1 - \exp(\frac{-\epsilon^2 n}{2}). \tag{45}$$

Using $p_2 = \exp(\frac{-\epsilon^2 n}{2})$ and, therefore, $\epsilon = \sqrt{\frac{2\log(1/p_2)}{n}}$, we have that with probability at least $1 - p_2$,

$$\mathbb{E}[\mathcal{R}_n(\Pi)] \leqslant \mathcal{R}_n(\Pi) + \sqrt{\frac{2\log(1/p_2)}{n}}. \tag{46}$$

The second term in (37) can be bounded using $0 \leqslant R_i^* \leqslant u$, $\mathbb{E}\left[R^*\right] = 1$, and Hoeffding's inequality:

$$\mathbb{P}\left(|1 - Z| \geqslant \epsilon\right) \leqslant 2\exp(-2\epsilon^2u^{-2}n). \tag{47}$$

Therefore, we have that

$$\mathbb{P}\left(|1 - Z| \leqslant \epsilon\right) \geqslant 1 - 2\exp(-2\epsilon^2u^{-2}n). \tag{48}$$

Using $p_3 = 2\exp(-2\epsilon^2u^{-2}n)$ and, therefore, $\epsilon = u\sqrt{\frac{\log(2/p_3)}{2n}}$, we have that with probability at least $1 - p_3$,

$$C|1 - Z| \leqslant Cu\sqrt{\frac{\log(2/p_3)}{2n}}. \tag{49}$$

Finally, using that $1/Z \leqslant 1/l$, we get that with probability at least $1 - p_1 - p_2 - p_3$,

$$\sup_{\pi\in\Pi}\left|\hat{V}^*_{\text{Target}}(\pi) - V_{\text{Target}}(\pi)\right| \leqslant 2C\frac{u}{l}K_\psi\mathcal{R}_n(\Pi) + 2C\frac{u}{l}K_\psi\sqrt{\frac{2\log(1/p_2)}{n}} + 2C\frac{u}{l}K_\psi\sqrt{\frac{2\log(1/p_1)}{n}} + C\frac{u}{l}\sqrt{\frac{\log(2/p_3)}{2n}}. \tag{50}$$

Let $p_1, p_2 = \delta/4$ and $p_3 = 2\delta/4$, then, using that $K_\psi \geqslant 1$, the above is bounded by $2C\frac{u}{l}K_\psi\mathcal{R}_n(\Pi) + 2C\frac{u}{l}K_\psi\sqrt{\frac{18\log(4/\delta)}{n}}$. The proof is completed by recognizing that, since the true $R^* \in \mathcal{R}$, we have that $\hat{V}^*_{\text{Target}}(\pi) \leqslant \overline{V}_{\text{Target}}(\pi)$. □

Remark 1. *We briefly explain how Theorem 1 is proven in the case in which we do not have access to the true nuisance functions (using the results from Athey and Wager [2021].*

Let $\tilde{V}^*_{Target}(\pi)$ be the estimator which uses the true nuisance functions and $\tilde{V}^*_{Target}(\pi)$ the estimator which uses estimated nuisance functions. Let $Z = \frac{1}{n}\sum_{i=1}^n R^*_i$. Then,

$$V_{Target}(\pi) \leqslant \hat{V}^*_{Target}(\pi) + \sup_{\pi\in\Pi}\left|\hat{V}^*_{Target}(\pi) - \tilde{V}^*_{Target}(\pi) + \tilde{V}^*_{Target}(\pi) - V_{Target}(\pi)\right| \tag{51}$$

$$\leqslant \hat{V}^*_{Target}(\pi) + \sup_{\pi\in\Pi}\left|\tilde{V}^*_{Target}(\pi) - \hat{V}^*_{Target}(\pi)\right| + \sup_{\pi\in\Pi}\left|\tilde{V}^*_{Target}(\pi) - V_{Target}(\pi)\right| \tag{52}$$

The term $\sup_{\pi\in\Pi}\left|\tilde{V}^*_{Target}(\pi) - V_{Target}(\pi)\right|$ can be bounded analogously to the proof above. The term $\sup_{\pi\in\Pi}\left|\tilde{V}^*_{Target}(\pi) - \hat{V}^*_{Target}(\pi)\right|$ can be bounded using Lemma 4 in Athey and Wager [2021]. The result follows the analogously to the proof above.

A.3 PROOF OF THEOREM 2

Let (i) denote the ith index of the increasing order statistics, an ordering where $\psi_{(1)}(\theta) \leqslant \ldots \psi_{(n)}(\theta)$. Hence, we address the following optimization problem

$$\max_R \frac{\sum_{i=1}^n R_{(i)}\psi_{(i)}(\theta)}{\sum_{i=1}^n R_{(i)}} \quad \text{s. t.,} \quad l \leqslant R_{(i)} \leqslant u,\, R_{(i)} \geqslant 0, \forall i = 1, \ldots, n. \tag{53}$$

We derive a closed-form solution for any of the $\psi_i(\theta)$ in (8), (9), and (10), which generalizes the solution of Kallus and Zhou [2018] to all standard policy learning methods. Since the constraint on R is linear, the above optimization problem is a linear fractional program. Hence, we can use the Charnes-Cooper transformation [Charnes and Cooper, 1962] with $\tilde{R}_{(i)} = R_{(i)}/\sum_{i=1}^n R_{(i)}$ and $t = 1/\sum_{i=1}^n R_{(i)}$, which yields

$$\max_{\tilde{R}} \sum_{i=1}^n \tilde{R}_{(i)}\psi_{(i)}(\theta) \quad \text{s.t.} \quad t\,l \leqslant \tilde{R}_{(i)} \leqslant t\,u,\, \tilde{R}_{(i)} \geqslant 0 \sum_{i=1}^n \tilde{R}_{(i)} = 1, t \geqslant 0, \forall i = 1, \ldots, n \tag{54}$$

The corresponding dual problem has the dual variables $\lambda \in \mathbb{R}$ for the normalization constraint and $w, v \in \mathbb{R}^n_+$ for the box constraints on the normalized Radon-Nikodým derivative. It is given by

$$\min_{\lambda,v,w} \lambda, \text{ s.t. } \sum_{i=1}^n v_{(i)}\,u + w_{(i)}\,l \geqslant 0,\, \lambda + v_{(i)} - w_{(i)} \geqslant \psi_{(i)}(\theta), \forall i = 1, \ldots, n\, \lambda \in \mathbb{R}, v, w \in \mathbb{R}^n_+. \tag{55}$$

At the optimal solution, only one of the primal weight bound constraints, (for nontrivial bounds $l < u$), $t\,l \leqslant R_{(i)}$ or $R_{(i)} \leqslant t\,u$ will be tight. At the optimal solution, by complementary slackness, either none or one of the nonbinding primal constraints is nonzero, i. e., either $v_{(i)}$, $w_{(i)}$, or none is nonzero. Moreover, $t = 0$ is infeasible, since $t = 0$ would imply $\tilde{R}_{(i)} = 0$ for all i, which contradicts $\sum_{i=1}^n \tilde{R}_{(i)} = 1$. Hence, $t \neq 0$. At the optimal solution, the constraint $\sum_{i=1}^n v_{(i)}\,u + w_{(i)}\,l \geqslant 0$ must be active. Otherwise, we can find a λ which is smaller than the optimal one but still feasible, and hence contradicts the optimality. As a result, at an optimal solution, we have that:

$$\sum_{i=1}^n v_{(i)}\,u + w_{(i)}\,l = 0, \tag{56}$$

$$v_{(i)} - w_{(i)} = \psi_{(i)}(\theta) - \lambda, \forall i = 1, \ldots, n. \tag{57}$$

Since $v, w \geqslant 0$, we see the following by distinction of cases. If $\psi_{(i)}(\theta) \geqslant \lambda$, then $w_{(i)} = 0$ and $v_{(i)} = \psi_{(i)}(\theta) - \lambda$. If $\psi_{(i)}(\theta) < \lambda$, then $v_{(i)} = 0$ and $w_{(i)} = \lambda - \psi_{(i)}(\theta)$.

At optimality, since (i) is the increasing order statistics, there exists some index $k \in \{1, \ldots, n\}$ such that $\psi_{(k)}(\theta) < \lambda \leqslant \psi_{(k+1)}(\theta)$. Hence, we can substitute the solution from (57) in (56) and obtain the following

$$\sum_{i=1}^k l(\lambda - \psi_{(i)}(\theta)) - \sum_{i=k+1}^n u(\psi_{(i)}(\theta) - \lambda) = 0, \tag{58}$$

and, therefore,

$$\lambda(k) = \frac{l \sum_{i=1}^{k} \psi_{(i)}(\theta) + u \sum_{i=k+1}^{n} \psi_{(i)}(\theta)}{k\, l + (n-k)\, u}. \tag{59}$$

The optimal k is given by $k^* = \inf\{k : \lambda(k) \leqslant \psi_{(k+1)}(\theta)\}$, which can be seen by the following argument. When $\lambda(k)$ is maximal, we have that $\lambda(k) \geqslant \lambda(k+1)$. This is equivalent to $\lambda(k) \leqslant \psi_{(k+1)}(\theta)$, since the following steps are equivalent

$$0 \geqslant \lambda(k+1) - \lambda(k) \tag{60}$$
$$0 \geqslant \frac{(lk + u(n-k))\lambda(k) + (l-u)\psi_{(k+1)}(\theta)}{l(k+1) + u(n-k-1)} - \lambda(k) \tag{61}$$
$$0 \geqslant (lk + u(n-k))\lambda(k) + (l-u)\psi_{(k+1)}(\theta) - \lambda(k)(l(k+1) + u(n-k-1)) \tag{62}$$
$$0 \geqslant (lk + u(n-k))\lambda(k) + (l-u)\psi_{(k+1)}(\theta) - \lambda(k)(lk + u(n-k)) - \lambda(k)(l+u) \tag{63}$$
$$0 \geqslant (l-u)\psi_{(k+1)}(\theta) - \lambda(k)(l-u) \tag{64}$$
$$\lambda(k) \leqslant \psi_{(k+1)}(\theta), \tag{65}$$

where the last inequality switches because we divide by $l - u$ which is negative. Next, we show that if $\lambda(k) \geqslant \lambda(k+1)$, then $\lambda(k+1) \geqslant \lambda(k+2)$.

$$\lambda(k+1) \tag{66}$$
$$= \frac{(lk + u(n-k))\lambda(k) + (l-u)\psi_{(k+1)}(\theta)}{l(k+1) + u(n-k-1)} \tag{67}$$
$$\leqslant \frac{(lk + u(n-k))\psi_{(k+1)}(\theta) + (l-u)\psi_{(k+1)}(\theta)}{l(k+1) + u(n-k-1)} \tag{68}$$
$$= \psi_{(k+1)}(\theta) \leqslant \psi_{(k+2)}(\theta), \tag{69}$$

and, since we showed above that $\lambda(k) \geqslant \lambda(k+1)$ is equivalent to $\lambda(k) \leqslant \psi_{(k+1)}(\theta)$, we have that $\lambda(k+1) \leqslant \psi_{(k+2)}(\theta)$ is equivalent to $\lambda(k+1) \geqslant \lambda(k+2)$. Thus, $k^* = \inf\{k : \lambda(k) \leqslant \psi_{(k+1)}(\theta)\}$. Hence, the solution of the dual problem is $\tilde{R}_{(i)} = \frac{l\mathbb{1}((i) \leqslant k^*) + u\mathbb{1}((i) > k^*)}{k^*\, l + (n-k^*)\, u}$. Then, the solution of the primal problem can be recovered by $R = \frac{1}{t}\tilde{R}$, where $t = 1/(k^*\, l + (n-k^*)\, u)$. □

A.4 PROOF OF LEMMA 1

For the direct method, we have

$$\psi_i^{\mathrm{DM}}(\theta) = \pi(X_i, \theta)\mu_1(X_i) + (1 - \pi(X_i, \theta))\mu_0(X_i) \tag{70}$$
$$= (\tilde{g}(X_i, \theta) - \tilde{h}(X_i, \theta))\mu_1(X_i) + (1 - \tilde{g}(X_i, \theta) + \tilde{h}(X_i, \theta))\mu_0(X_i). \tag{71}$$

To derive g_i and h_i, we proceed with a case distinction.

Case 1: $\mu_0(X_i) \geqslant 0$ and $\mu_1(X_i) \geqslant 0$
In this case, we have

$$\psi_i^{\mathrm{DM}}(\theta) = (\tilde{g}(X_i, \theta)\mu_1(X_i) + \tilde{h}(X_i, \theta)\mu_0(X_i) + \mu_0(X_i)) \tag{72}$$
$$- (\tilde{h}(X_i, \theta)\mu_1(X_i) + \tilde{g}(X_i, \theta)\mu_0(X_i)), \tag{73}$$

and, hence, the claim follows with

$$g_i(\theta) = \tilde{g}(X_i, \theta)\mu_1(X_i) + \tilde{h}(X_i, \theta)\mu_0(X_i) + \mu_0(X_i) \tag{74}$$
$$h_i(\theta) = \tilde{h}(X_i, \theta)\mu_1(X_i) + \tilde{g}(X_i, \theta)\mu_0(X_i). \tag{75}$$

Case 2: $\mu_0(X_i) < 0$ and $\mu_1(X_i) \geqslant 0$
In this case, we have

$$\psi_i^{\mathrm{DM}}(\theta) = (\tilde{g}(X_i, \theta)\mu_1(X_i) + \tilde{g}(X_i, \theta)|\mu_0(X_i)| - |\mu_0(X_i)|) \tag{76}$$
$$- (\tilde{h}(X_i, \theta)\mu_1(X_i) + \tilde{h}(X_i, \theta)|\mu_0(X_i)|), \tag{77}$$

and, hence, the claim follows with

$$g_i(\theta) = \tilde{g}(X_i,\theta)\mu_1(X_i) + \tilde{g}(X_i,\theta)|\mu_0(X_i)| - |\mu_0(X_i)| \tag{78}$$
$$h_i(\theta) = \tilde{h}(X_i,\theta)(\mu_1(X_i) + |\mu_0(X_i)|). \tag{79}$$

Case 3: $\mu_0(X_i) \geqslant 0$ and $\mu_1(X_i) < 0$
In this case, we have

$$\psi_i^{\mathrm{DM}}(\theta) = (\tilde{h}(X_i,\theta)|\mu_1(X_i)| + \tilde{h}(X_i,\theta)\mu_0(X_i) + \mu_0(X_i)) \tag{80}$$
$$- (\tilde{g}(X_i,\theta)|\mu_1(X_i)| + \tilde{g}(X_i,\theta)\mu_0(X_i)), \tag{81}$$

and, hence, the claim follows with

$$g_i(\theta) = \tilde{h}(X_i,\theta)|\mu_1(X_i)| + \tilde{h}(X_i,\theta)\mu_0(X_i) + \mu_0(X_i) \tag{82}$$
$$h_i(\theta) = \tilde{g}(X_i,\theta)(|\mu_1(X_i)| + \mu_0(X_i)). \tag{83}$$

Case 4: $\mu_0(X_i) < 0$ and $\mu_1(X_i) < 0$
In this case, we have

$$\psi_i^{\mathrm{DM}}(\theta) = (\tilde{h}(X_i,\theta)|\mu_1(X_i)| + \tilde{g}(X_i,\theta)|\mu_0(X_i)| - |\mu_0(X_i)|) \tag{84}$$
$$- (\tilde{g}(X_i,\theta)|\mu_1(X_i)| + \tilde{h}(X_i,\theta)|\mu_0(X_i)|), \tag{85}$$

and, hence, the claim follows with

$$g_i(\theta) = \tilde{h}(X_i,\theta)|\mu_1(X_i)| + \tilde{g}(X_i,\theta)|\mu_0(X_i)| - |\mu_0(X_i)| \tag{86}$$
$$h_i(\theta) = \tilde{g}(X_i,\theta)|\mu_1(X_i)| + \tilde{h}(X_i,\theta)|\mu_0(X_i)|. \tag{87}$$

For the normalized inverse propensity weights method, we have

$$\psi_i^{\mathrm{NIPW}}(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}(1-2T_i)(1-T_i-\pi(X_i,\theta))Y_i. \tag{88}$$

Again, by a case distinction, we yield for $T_i = 1$:

$$\psi_i^{\mathrm{NIPW}}(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\pi(X_i,\theta)Y_i \tag{89}$$
$$= \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\tilde{g}(X_i,\theta)Y_i - \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\tilde{h}(X_i,\theta)Y_i, \tag{90}$$

and, hence, the claim follows with

$$g_i(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\tilde{g}(X_i,\theta)Y_i \tag{91}$$
$$h_i(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\tilde{h}(X_i,\theta)Y_i. \tag{92}$$

For $T_i = 0$, we derive,

$$\psi_i^{\mathrm{NIPW}}(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}(1-\pi(X_i,\theta))Y_i \tag{93}$$
$$= \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}(1-\tilde{g}(X_i,\theta)+\tilde{h}(X_i,\theta))Y_i, \tag{94}$$

and, hence, the claim follows with

$$g_i(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}(\tilde{h}(X_i,\theta)+1)Y_i \tag{95}$$

$$h_i(\theta) = \frac{2W_i^{\mathrm{IPW}}}{\frac{1}{n}\sum_{j=1}^n W_j^{\mathrm{IPW}}}\tilde{g}(X_i,\theta)Y_i. \tag{96}$$

For the doubly robust method, we can use the decomposition of the direct method. By defining

$$\nu_i = (1-2T_i)(Y_i - \mu_{T_i}(X_i)), \tag{97}$$

and rewriting

$$W_i^{\mathrm{IPW}}(1-2T_i)(1-T_i-\pi(X_i,\theta))(Y_i-\mu_{T_i}(X_i)) = W_i^{\mathrm{IPW}}(1-T_i-\pi(X_i,\theta))\nu_i, \tag{98}$$

we proceed again by a case distinction for the rest. For $\nu_i \geqslant 0$ we have

$$W_i^{\mathrm{IPW}}(1-T_i-\pi(X_i,\theta))\nu_i = W_i^{\mathrm{IPW}}(1-T_i)\nu_i + W_i^{\mathrm{IPW}}\nu_i\tilde{h}(X_i,\theta) \tag{99}$$

$$- W_i^{\mathrm{IPW}}\nu_i\tilde{g}(X_i,\theta). \tag{100}$$

For $\nu_i < 0$, we have that

$$W_i^{\mathrm{IPW}}(1-T_i-\pi(X_i,\theta))\nu_i = W_i^{\mathrm{IPW}}(1-T_i)\nu_i + W_i^{\mathrm{IPW}}|\nu_i|\tilde{g}(X_i,\theta) \tag{101}$$

$$- W_i^{\mathrm{IPW}}|\nu_i|\tilde{h}(X_i,\theta). \tag{102}$$

and, hence, the claim follows. □

A.5 PROOF OF THEOREM 3

By Theorem 2, we know that

$$\max_{R\in\mathcal{R}} \frac{\sum_{i=1}^n R_i\psi_i(\theta)}{\sum_{i=1}^n R_i} = \max_{R\in\mathcal{S}\subseteq\mathcal{R}} \frac{\sum_{i=1}^n R_i\psi_i(\theta)}{\sum_{i=1}^n R_i}, \tag{103}$$

with $|\mathcal{S}| = n+1 < \infty$. Hence, we can write the inner maximum as

$$\max_{j\in\mathcal{J}} \frac{\sum_{i=1}^n R_i^j\psi_i(\theta)}{\sum_{i=1}^n R_i^j}, \tag{104}$$

where R^j for $j \in \mathcal{J} = \{0,\ldots,n\}$ denotes one of the $n+1$ possible assignments of l and u, i. e., for $j=0$, it is the vector with all entries equal to l; for $j=1$, it is the vector with all entries equal to l except for the first one being u and so on. By defining the convex functions

$$G^j(\theta) = \frac{\sum_{i=1}^n R_i^j g_i(\theta)}{\sum_{i=1}^n R_i^j}, \tag{105}$$

$$H^j(\theta) = \frac{\sum_{i=1}^n R_i^j h_i(\theta)}{\sum_{i=1}^n R_i^j}, \tag{106}$$

we have

$$\frac{\sum_{i=1}^n R_i^j\psi_i(\theta)}{\sum_{i=1}^n R_i^j} = G^j(\theta) - H^j(\theta), \tag{107}$$

and, hence,

$$\max_{j\in\mathcal{J}} \frac{\sum_{i=1}^n R_i^j \psi_i(\theta)}{\sum_{i=1}^n R_i^j} = \max_{j\in\mathcal{J}}\{G^j - H^j\} \tag{108}$$

$$= \max_{j\in\mathcal{J}}\{G^j + \sum_{\substack{k=1\\k\neq j}}^n H^k - \sum_{k=1}^n H^k\} \tag{109}$$

$$= \underbrace{\max_{j\in\mathcal{J}}\{G^j + \sum_{\substack{k=1\\k\neq j}}^n H^k\}}_{=:g} - \underbrace{\sum_{k=1}^n H^k}_{=:h}. \tag{110}$$

Note that g and h are convex as the sum of convex functions is convex and the maximum of convex functions is convex. Now, g can be rewritten as follows

$$g(\theta) = \max_{j\in\mathcal{J}}\{G^j + \sum_{\substack{k=1\\k\neq j}}^n H^k\} = \max_{j\in\mathcal{J}}\{G^j - H^j + \sum_{k=1}^n H^k\} \tag{111}$$

$$= \max_{j\in\mathcal{J}}\{G^j - H^j\} + \sum_{k=1}^n H^k = \max_{R\in\mathcal{R}}\left\{\frac{\sum_{i=1}^n R_i\psi_i(\theta)}{\sum_{i=1}^n R_i}\right\} + \sum_{k=1}^n H^k \tag{112}$$

$$= \max_{R\in\mathcal{R}}\left\{\frac{\sum_{i=1}^n R_i\psi_i(\theta)}{\sum_{i=1}^n R_i}\right\} + h. \tag{113}$$

Furthermore, we can use the special structure of the worst case policy solutions to rewrite h as

$$h = \sum_{k=1}^n H^k = \sum_{k=1}^n \frac{\sum_{i=1}^n R_i^k h_i(\theta)}{\sum_{i=1}^n R_i^k} = \sum_{k=1}^n\sum_{i=1}^n \frac{R_i^k}{\sum_{i=1}^n R_i^k} h_i(\theta) \tag{114}$$

$$= \sum_{i=1}^n h_i(\theta) \underbrace{\sum_{k=1}^n \frac{R_i^k}{\sum_{i=1}^n R_i^k}}_{=:c_i} = \sum_{i=1}^n h_i(\theta)c_i, \tag{115}$$

where c_i can be calculated as

$$c_i = l\left(\sum_{k=1}^{i} \frac{1}{(n-k+1)l + (k-1)u}\right) + u\left(\sum_{k=i+1}^{n} \frac{1}{(n-k+1)l + (k-1)u}\right), \tag{116}$$

for all i by combinatorial arguments.

A.6 PROOF OF THEOREM 4

The convergence analysis of MMCCP follows from the convergence analysis of the DC-algorithm (DCA) [Tao and An, 1997]. More precisely, DCA for minimizing a function $f = g - h$ reduces to the convex-concave procedure in case that the function h is differentiable [Phan et al., 2018, Sriperumbudur and Lanckriet, 2009]. This is exactly what we have in our case, as by our assumption on $\tilde{g}$ and $\tilde{h}$ we have that each h_i (as a linear combination of differentiable functions) is differentiable and, hence, h is differentiable.

Now, 1. in Theorem 4 directly follows from (i) of Theorem 3 in Tao and An [1997]. For 2. in Theorem 4, we have to proof the following:

1. $\inf_{\theta\in\Theta}\max_{R\in\mathcal{R}} \frac{\sum_{i=1}^n R_i\psi_i(\theta)}{\sum_{i=1}^n R_i}$ is finite.
2. It holds $\rho(g) + \rho(h) > 0$.
3. $(\theta^k)_{k\in\mathbb{N}}$ is bounded.

Ad Item 1: Since $|Y| \leqslant C$, we have that $|\mu_t(X_i)| \leqslant C$. Also, the rest of the terms involved in each of the three cases for ψ_i are bounded constants, and $l \leqslant R_i \leqslant u$ for all $i \in \{1, \dots, n\}$. Hence, since $\pi(\cdot, \theta) \in [0, 1]$, we have that Item 1 holds true.

Ad Item 2: For all $i \in \{1, \dots, n\}$, we have in each of the three cases for ψ_i, that h_i is, up to a constant, a linear combination of $\tilde{g}$ and $\tilde{h}$ with positive weights. By our assumptions, we have that $\rho(\tilde{g}) > 0$ and $\rho(\tilde{h}) > 0$ and, hence, $\rho(h_i) > 0$. By Theorem 3, we have that $h = \sum_{i=1}^n h_i c_i$ with non-negative weights c_i, which yields $\rho(h) > 0$. Item 2 follows by observing that $\rho(g) \geqslant 0$.

Ad Item 3: Follows directly from Assumption 1.

Then, 2. in Theorem 4 follows by (iii) and (iv) of Theorem 3 in Tao and An [1997]. □

B DETAILS ON COVARIATES IN THE ACTG 175 STUDY

The ACTG 175 study assigned four treatments randomly to 2,139 subjects with human immunodeficiency virus (HIV) type 1, whose CD4 counts were 200–500 cells/mm^3. The four treatments that were compared are the zidovudine (ZDV) monotherapy, the didanosine (ddI) monotherapy, the ZDV combined with ddI, and the ZDV combined with zalcitabine (ZAL).

There are 5 continuous covariates: age (year), weight (kg, coded as wtkg), CD4 count (cells/mm^3) at baseline, Karnofsky score (scale of 0-100, coded as karnof), CD8 count (mm^3) at baseline. They are centered and scaled before further analysis. In addition, there are 7 binary variables: gender ($1 =$ male, $0 =$ female), homosexual activity (homo, $1 =$ yes, $0 =$ no), race ($1 =$ nonwhite, $0 =$ white), history of intravenous drug use (drug, $1 =$ yes, $0 =$ no), symptomatic status (symptom, $1 =$ symptomatic, $0 =$ asymptomatic), antiretroviral history (str2, $1 =$ experienced, $0 =$ naive) and hemophilia (hemo, $1 =$ yes, $0 =$ no).

C ASSUMPTION 1 FOR LINEAR POLICIES

Linear policies are defined by $\pi(X, \theta) = \sigma(\theta^\intercal X)$, where $\sigma(z) = \min(1, \max(z, 0))$. A DC-representation for $\sigma(z)$, with $z = \theta^\intercal X$, is given by

$$\tilde{g}_{\text{Lin}}(z) = \max(z, 0), \tag{117}$$

$$\tilde{h}_{\text{Lin}}(z) = \max(\max(z, 0) - 1, 0). \tag{118}$$

It is straightforward to check that both functions are convex. Again, they can be made strongly convex by adding $\frac{\lambda}{2} z^2$ to both functions. Note however, that $\tilde{g}_{\text{Lin}}$ is not differentiable in 0 and $\tilde{h}_{\text{Lin}}$ is not differentiable in $\{0, 1\}$. As a remedy, one can set $\Theta_i^\epsilon = \{\theta \in \mathbb{R}^d : \epsilon \leqslant \theta^\intercal X_i \leqslant 1 - \epsilon\}$ for an $\epsilon > 0$ and define $\Theta^\epsilon = \bigcap_{i=1}^n \Theta_i^\epsilon$. The intersection has to be nonempty to make this approach work.

D IMPLEMENTATION DETAILS

Our code is available at `github.com/tobhatt/GeneralOPL`. For our experiments, we used the policy class of logistic policies as introduced in the main paper. To fulfill Assumption 1, we choose Θ to be a hypercube with large bounds to ensure a large enough search space, i. e., $\Theta = [-10,000; 10,000]^d$. In order to solve the subproblems in MMCCP, we draw upon the L-BFGS-B algorithm implemented in the open-source Python library SciPy. At this point, we note that the subproblems are convex but not necessarily differentiable, as the point-wise maximum of differentiable functions is not necessarily differentiable. However, logistic policies are continuously differentiable and the above choice for Θ is compact. Hence, the functions $\psi_i(\theta)$ are Lipschitz and, thus,

$$\max_{R \in \mathcal{R}} \frac{\sum_{i=1}^n R_i \psi_i(\theta)}{\sum_{i=1}^n R_i} \tag{119}$$

is Lipschitz as the point-wise maximum of Lipschitz functions. By Rademacher's theorem, (119) is therefore almost everywhere differentiable. The points θ where (119) is not differentiable are given by the points in which the maximizing argument R changes. Due to this fact, we find empirically that L-BFGS-B can efficiently solve these subproblems. The rest of the parameters are set as follows. The parameter for the stopping criterion is set to δ_{tol} to 10^{-4}. In order to make

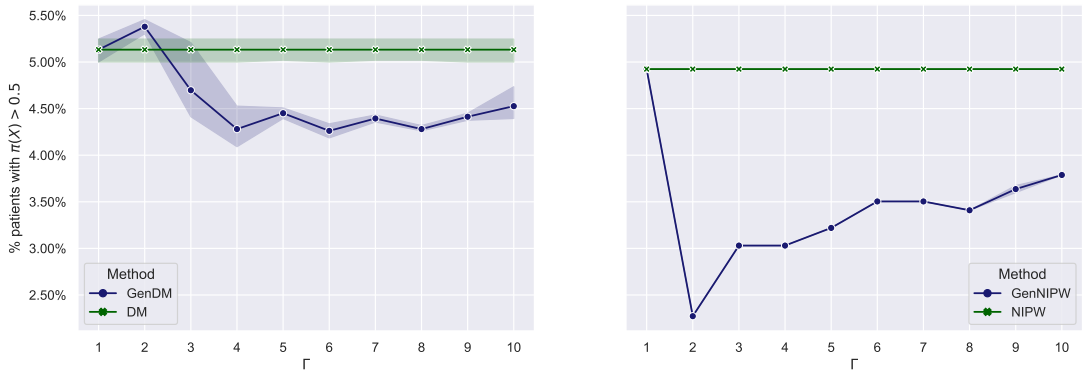

Figure 1: Percentage of patients with $\pi(X) > 0.5$ for our GenDM and GenNIPW policy method. Fewer patients are treated for increasing Γ.

$\tilde{g}$ and $\tilde{h}$ strongly convex, λ is set to 10^{-3}. In every run, the starting points are initialized via a normal distribution, i. e., $\theta^0 \sim \mathcal{N}_d(\mathbf{0}_d, 0.1 \cdot \mathbf{I}_d)$. For each method, we ran our algorithm 5 times on the datasets.

We run all of our experiments on a server with two 16 Core Intel Xeon Gold 6242 processors each with 2.8GHz, and 192GB of RAM.

E RESULTS FOR GENDM AND GENNIPW ON ACTG 175 STUDY

We present the results on the ACTG 175 study for our method GenDM, which uses $\psi_i^{\mathrm{DM}}(\pi)$ from (8) and our method GenNIPW, which uses $\psi_i^{\mathrm{NIPW}}(\pi)$ from (9). Analogously to Section 5.2, we study the percentage of patients that are treated (i. e., $\pi(X) > 0.5$) for varying Γ. The results are presented in Figure 1. Similar to the results for GenDR in Section 5.2, we find that compared to the baseline policy, our policy treats fewer patients for increasing Γ. GenNIPW shows little variance across several runs on the dataset. For each run, GenNIPW obtains different, but similar θ. However, the percentage of patients treated remains consistent across different runs.

References

S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3-4): 181–186, 1962.

N. Kallus and A. Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, 2018.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 2013.

D. N. Phan, H. M. Le, and H. A. Le Thi. Accelerated difference of convex functions algorithm and its application to sparse binary logistic regression. In *International Joint Conference on Artificial Intelligence*, 2018.

B. K. Sriperumbudur and G. R. Lanckriet. On the convergence of the concave-convex procedure. In *Advances in Neural Information Processing Systems*, 2009.

P. D. Tao and L. T. H. An. Convex analysis approach to dc programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.