Supporting Information: Multi-Task Neural Networks for Distributional Treatment Effects

August 10, 2025

Contents

1 Summary of Notation **2**

2 Proofs **2**

3 Multiplier Bootstrap Procedure **6**

4 Simulation Study **7**
4.1 Data Generating Process (DGP) 7
4.2 Model Implementation and Experiment Environment 7
4.3 Pointwise Estimation Result 7

5 Nudges to Reduce Water Consumption **10**

6 ABEMA Content Promotion **10**

7 Validity of Sub-linear Assumption **11**
7.1 Computational Analysis of Linear Regression 11
7.2 Empirical Computation Complexity Analysis of Neural Networks 12

Supporting Information

The supporting information is structured as follows. Section 1 summarizes the notations used in the main text and the supporting Information. Section 2 provides the proofs of the theorems. Section 3 outlines the multiplier bootstrap procedure used for obtaining the confidence intervals in this paper. Section 4 describes the simulation settings, Section 5 presents the additional details of the field experiment to reduce water consumption, and Section 6 offers the supplemental information about the content promotion experiment on ABEMA.

1 Summary of Notation

The notations used in the main text are summarized in Table A1.

Notation	Definition
X_i	pre-treatment covariates
W_i	treatment variable
Y_i	outcome variable
Z_i	observed samples, $Z_i := (X_i, W_i, Y_i)$
$Y_i(w)$	potential outcome for treatment group w
$\tilde{\mathcal{Y}}$	a set of scalar-valued locations to estimate the treatment effect for
π_w	treatment assignment probability for treatment group w
n_w	number of observations in treatment group w
n	number of observations
$\widehat{\pi}_w$	n_w/n, estimated treatment assignment probability for treatment group w
$F_{Y(w)}(y)$	$E[\mathbf{1}_{\{Y(w)\leq y\}}]$, potential outcome distribution function
$\gamma_y^{(w)}(x)$	$E[\mathbf{1}_{\{Y(w)\leq y\}}\|X=x]$, conditional distribution function
$\widehat{\gamma}_y^{(w)}(x)$	estimator of conditional distribution function $\gamma_y^{(w)}(x)$
$\mathbf{\Gamma}^{(w)}$	a set of conditional distribution functions $\gamma_y^{(w)}(\cdot)$ for $y \in \tilde{\mathcal{Y}}$
$\widehat{\mathbf{\Gamma}}^{(w)}$	estimator of $\mathbf{\Gamma}^{(w)}$

Table A1: Summary of Notation

2 Proofs

Proof of Theorem 4.1
We follow the approach used in [2] to prove the efficiency gain of the adjusted estimation. For completeness of the proof, we first present a variant of Lagrange's identity and Bergström's inequality in the below lemma, which is useful to prove the efficiency gain of the regression adjustment.

Lemma 1. *For any* $(a_1, \ldots, a_K) \in \mathbb{R}^K$ *and* $(b_1, \ldots, b_K) \in \mathbb{R}^K$ *with* $b_k > 0$ *for all* $k = 1, \ldots, K$, *we can show that*

$$\sum_{k=1}^{K} \frac{a_k^2}{b_k} - \frac{\left(\sum_{k=1}^{K} a_k\right)^2}{\sum_{k=1}^{K} b_k} = \frac{1}{\sum_{k=1}^{K} b_k} \cdot \frac{1}{2} \sum_{k=1}^{K} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} \frac{(a_k b_\ell - a_\ell b_k)^2}{b_k b_\ell},$$

which implies Bergström's inequality, given by

$$\sum_{k=1}^{K} \frac{a_k^2}{b_k} \geq \frac{\left(\sum_{k=1}^{K} a_k\right)^2}{\sum_{k=1}^{K} b_k}.$$

Proof. Lagrange's identity is that, for any $(c_1, \ldots, c_K) \in \mathbb{R}^K$ and $(d_1, \ldots, d_K) \in \mathbb{R}^K$,

$$\left(\sum_{k=1}^K c_k^2\right)\left(\sum_{k=1}^K d_k^2\right) - \left(\sum_{k=1}^K c_k d_k\right)^2 = \frac{1}{2}\sum_{k=1}^K \sum_{\substack{\ell=1 \\ \ell \neq k}}^K (c_k d_\ell - c_\ell d_k)^2. \tag{1}$$

Fix arbitrary $(a_1, \ldots, a_K) \in \mathbb{R}^K$ and $(b_1, \ldots, b_K) \in \mathbb{R}^K$ with $b_k > 0$ for all $k = 1, \ldots, K$. Then, taking $c_k = a_k/\sqrt{b_k}$ and $d_k = \sqrt{b_k}$ for all $k = 1, \ldots, K$ in (1), we can show that

$$\begin{aligned}\left(\sum_{k=1}^K \frac{a_k^2}{b_k}\right)\left(\sum_{k=1}^K b_k\right) - \left(\sum_{k=1}^K a_k\right)^2 &= \frac{1}{2}\sum_{k=1}^K \sum_{\substack{\ell=1 \\ \ell \neq k}}^K \left(\frac{a_k}{\sqrt{b_k}}\sqrt{b_\ell} - \frac{a_\ell}{\sqrt{b_\ell}}\sqrt{b_k}\right)^2 \\ &= \frac{1}{2}\sum_{k=1}^K \sum_{\substack{\ell=1 \\ \ell \neq k}}^K \frac{(a_k b_\ell - a_\ell b_k)^2}{b_k b_\ell},\end{aligned}$$

which leads to the desired equality. Also, the last expression in the math display above is non-negative, which leads to Bergström's inequality. □

To establish Theorem 4.1, we first introduce additional notation. Specifically, we define the empirical probability measures of X as

$$\widehat{\mathbb{F}}_X := \frac{1}{n}\sum_{i=1}^n \delta_{X_i} \quad \text{and} \quad \widehat{\mathbb{F}}_X^{(w)} := \frac{1}{n_w}\sum_{i=1}^n \mathbf{1}_{\{W_i = w\}} \cdot \delta_{X_i},$$

for all observations and observations in the treatment group $w \in \mathcal{W}$, respectively. Here, δ_x is the measure that assigns mass 1 at $x \in \mathcal{X}$ and thus $\widehat{\mathbb{F}}_X$ and $\widehat{\mathbb{F}}_X^{(w)}$ can be interpreted as the random discrete probability measures, which put mass $1/n$ and $1/n_w$ at each of the n and n_w points $\{X_i\}_{i=1}^n$ and $\{X_i : W_i = w\}_{i=1}^n$, respectively. Given a real-valued function $f : \mathcal{X} \to \mathbb{R}$, we denote by

$$\widehat{\mathbb{F}}_X f = \int f d\widehat{\mathbb{F}}_X = \frac{1}{n}\sum_{i=1}^n f(X_i)$$

and

$$\widehat{\mathbb{F}}_X^{(w)} f = \int f d\widehat{\mathbb{F}}_X^{(w)} = \frac{1}{n_w}\sum_{i=1}^n \mathbf{1}_{\{W_i = w\}} \cdot f(X_i).$$

Given that the true conditional distribution $\gamma_y^{(w)}(X) \equiv F_{Y(w)|X}(y|X)$, the infeasible version of regression-adjusted distribution function for treatment $w \in \mathcal{W}$ is written as

$$\widetilde{\mathbf{F}}_{Y(w)}(y) = \widehat{\mathbf{F}}_{Y(w)}^{empirical}(y) - (\widehat{\mathbb{F}}_X^{(w)} - \widehat{\mathbb{F}}_X)\gamma_y^{(w)}.$$

Proof of Theorem 4.1. **Part (a)** Choose any arbitrary $w \in \mathcal{W}$ and $y \in \mathcal{Y}$. Applying the quadratic expansion for $\widetilde{\mathbf{F}}_{Y(w)}(y) = \widehat{\mathbf{F}}_{Y(w)}^{empirical}(y) - (\widehat{\mathbb{F}}_X^{(w)} - \widehat{\mathbb{F}}_X)\gamma_y^{(w)}$, we can show that

$$\mathrm{Var}\big(\widetilde{\mathbf{F}}_{Y(w)}(y)\big) = \mathrm{Var}\big(\widehat{\mathbf{F}}_{Y(w)}^{empirical}(y)\big) - 2\mathrm{Cov}\Big(\widehat{\mathbf{F}}_{Y(1)}^{empirical}(y), (\widehat{\mathbb{F}}_X^{(w)} - \widehat{\mathbb{F}}_X)\gamma_y^{(w)}\Big) + \mathrm{Var}\Big((\widehat{\mathbb{F}}_X^{(w)} - \widehat{\mathbb{F}}_X)\gamma_y^{(w)}\Big). \tag{2}$$

We can write $\widehat{\mathbb{F}}_X = \sum_{w' \in \mathcal{W}} \hat{\pi}_{w'} \widehat{\mathbb{F}}_X^{(w')}$. It is assumed that observations are a random sample and $n_{w'}/n = \pi_{w'} + o(1)$ for every $w' \in \mathcal{W}$ as $n \to \infty$. Furthermore, all unconditional and conditional functions are bounded. By applying the dominated convergence theorem, we can show

$$\begin{aligned} n\mathrm{Cov}\Big(\widehat{\mathbf{F}}_{Y(w)}^{empirical}(y), (\widehat{\mathbb{F}}_X^{(w)} - \widehat{\mathbb{F}}_X)\gamma_y^{(w)}\Big) &= n\mathrm{Cov}\Big(\widehat{\mathbf{F}}_{Y(w)}^{empirical}(y), (1-\hat{\pi}_w)\widehat{\mathbb{F}}_X^{(w)}\gamma_y^{(w)}(X)\Big) \\ &= \frac{1-\pi_w}{\pi_w}\mathrm{Cov}\big(\mathbf{1}_{\{Y(w)\le y\}}, \gamma_y^{(w)}(X)\big) + o(1). \end{aligned} \tag{3}$$

Similarly, we can show that

$$\begin{aligned} n\mathrm{Var}\big((\widehat{\mathbb{F}}_X^{(w)} - \widehat{\mathbb{F}}_X)\gamma_y^{(w)}\big) &= n\mathrm{Var}\big((1-\hat{\pi}_w)\widehat{\mathbb{F}}_X^{(w)}\gamma_y^{(w)}\big) + n\sum_{w':w'\neq w}\mathrm{Var}\big(\hat{\pi}_{w'}\widehat{\mathbb{F}}_X^{(w')}\gamma_y^{(w)}\big) \\ &= \frac{(1-\pi_w)^2}{\pi_w}\mathrm{Var}\big(\gamma_y^{(w)}(X)\big) + \sum_{w':w'\neq w}\frac{\pi_{w'}^2}{\pi_{w'}}\mathrm{Var}\big(\gamma_y^{(w)}(X)\big) + o(1) \\ &= \left(\frac{(1-\pi_w)^2}{\pi_w} + \sum_{w':w'\neq w}\pi_{w'}\right)\mathrm{Var}\big(\gamma_y^{(w)}(X)\big) + o(1) \\ &= \frac{1-\pi_w}{\pi_w}\mathrm{Var}\big(\gamma_y^{(w)}(X)\big) + o(1). \end{aligned} \tag{4}$$

It follows from (2)-(4) that

$$n\big\{\mathrm{Var}\big(\widehat{\mathbf{F}}_{Y(w)}^{empirical}(y)\big) - \mathrm{Var}\big(\widetilde{\mathbf{F}}_{Y(w)}(y)\big)\big\} = \frac{1-\pi_w}{\pi_w}\big\{2\mathrm{Cov}\big(\mathbf{1}_{\{Y(w)\le y\}}, \gamma_y^{(w)}(X)\big) - \mathrm{Var}\big(\gamma_y^{(w)}(X)\big)\big\} + o(1). \tag{5}$$

An application of the law of iterated expectation yields

$$\mathrm{Cov}\big(\mathbf{1}_{\{Y(w)\le y\}}, \gamma_y^{(w)}(X)\big) = \mathrm{Var}\big(E[\mathbf{1}_{\{Y(w)\le y\}}|X]\big),$$

which together with (5) shows

$$n\big\{\mathrm{Var}\big(\widehat{\mathbf{F}}_{Y(w)}^{empirical}(y)\big) - \mathrm{Var}\big(\widetilde{\mathbf{F}}_{Y(w)}(y)\big)\big\} = \frac{1-\pi_w}{\pi_w}\mathrm{Var}\big(\gamma_y^{(w)}(X)\big) + o(1).$$

Since $\pi_w \in (0,1)$ and $\mathrm{Var}\big(\gamma_y^{(w)}(X)\big) \ge 0$, it follows that $\mathrm{Var}\big(\widehat{\mathbf{F}}_{Y(w)}^{empirical}(y)\big) \ge \mathrm{Var}\big(\widetilde{\mathbf{F}}_{Y(w)}(y)\big) + o(n^{-1})$. Here, the equality hold only when $F_{Y(w)|X}(y) = F_{Y(w)}(y)$ or X has no predictive power for the event $\mathbf{1}_{\{Y(w)\le y\}}$.

Part (b) Choose any arbitrary $y \in \mathcal{Y}$. First, we shall show that, for any $w, w' \in \mathcal{W}$,

$$n\mathrm{Cov}\big(\widehat{\mathbf{F}}_{Y(w)}(y), \widetilde{\mathbf{F}}_{Y(w')}(y)\big) = \mathrm{Cov}\big(\gamma_y^{(w)}(X), \gamma_y^{(w')}(X)\big). \tag{6}$$

Fix any two distinct treatment statuses $w, w' \in \mathcal{W}$. We can write $\widetilde{\mathbf{F}}_{Y(w)}(y) = \big(\widehat{\mathbf{F}}_{Y(w)}^{empirical}(y) - \widehat{\mathbb{F}}_X^{(w)}\gamma_y^{(w)}\big) + \widehat{\mathbb{F}}_X\gamma_y^{(w)}$ and also $\widehat{\mathbb{F}}_X\gamma_y^{(w)} = \sum_{v\in\mathcal{W}}\hat{\pi}_v\widehat{\mathbb{F}}_X^{(v)}\gamma_y^{(w)}$. Given random sample and the bi-linear property of the covariance function, we can show that

$$\begin{aligned} \mathrm{Cov}\Big(\widetilde{\mathbf{F}}_{Y(w)}(y), \widetilde{\mathbf{F}}_{Y(w')}(y)\Big) =& \mathrm{Cov}\Big(\widehat{\mathbf{F}}_{Y(w)}^{empirical}(y) - \widehat{\mathbb{F}}_X^{(w)}\gamma_y^{(w)}, \hat{\pi}_{w'}\widehat{\mathbb{F}}_X^{(w')}\gamma_y^{(w')}\Big) \\ &+ \mathrm{Cov}\big(\hat{\pi}_w\widehat{\mathbb{F}}_X^{(w)}\gamma_y^{(w)}, \widehat{\mathbf{F}}_{Y(w')}^{empirical} - \widehat{\mathbb{F}}_X^{(w')}\gamma_y^{(w')}\big) \\ &+ \mathrm{Cov}\big(\widehat{\mathbb{F}}_X\gamma_y^{(w)}, \widehat{\mathbb{F}}_X\gamma_y^{(w')}\big), \end{aligned}$$

where it can be shown that the first and second terms on the right-hand side are equal zero, due to the fact that $E\big[\widehat{\mathbf{F}}^{empirical}_{Y(w)}(y) - \widehat{\mathbb{F}}^{(w)}_X \gamma^{(w)}_y | X_1, \ldots, X_n\big] = 0$. Furthermore, under the random sample assumption, we can show that

$$\mathrm{Cov}\big(\widehat{\mathbb{F}}_X \gamma^{(w)}_y, \widehat{\mathbb{F}}_X \gamma^{(w')}_y\big) = n^{-1}\mathrm{Cov}\big(\gamma^{(w)}_y(X), \gamma^{(w')}_y(X)\big)$$

. Thus, we can prove the equality in (6).

Next, we compare the variance-covariance matrices of the simple and regression-adjusted estimators. By applying the result from part (a) of this theorem and the one in (6), we are able to show that

$$n\big\{\mathrm{Var}\big(\widehat{\theta}^{empirical}_y\big) - \mathrm{Var}\big(\widetilde{\theta}_y\big)\big\} = \begin{bmatrix} \frac{1-\pi_1}{\pi_1}\mathrm{Var}\big(\gamma^{(1)}_y(X)\big), & -\mathrm{Cov}\big(\gamma^{(1)}_y(X), \gamma^{(2)}_y(X)\big), & \ldots, & -\mathrm{Cov}\big(\gamma^{(1)}_y(X), \gamma^{(K)}_y(X)\big) \\ -\mathrm{Cov}\big(\gamma^{(2)}_y(X), \gamma^{(1)}_y(X)\big), & \frac{1-\pi_2}{\pi_2}\mathrm{Var}\big(\gamma^{(2)}_y(X)\big), & \ldots, & -\mathrm{Cov}\big(F_{Y(2)|X}, F_{Y(K)|X}\big) \\ \vdots & \vdots & \ddots & \vdots \\ -\mathrm{Cov}\big(\gamma^{(K)}_y(X), \gamma^{(1)}_y(X)\big), & -\mathrm{Cov}\big(\gamma^{(K)}_y(X), \gamma^{(2)}_y(X)\big), & \ldots, & \frac{1-\pi_K}{\pi_K}\mathrm{Var}\big(\gamma^{(K)}_y(X)\big) \end{bmatrix} + o(1),$$

which can be written as

$$n\big\{\mathrm{Var}\big(\widehat{\theta}^{empirical}_y\big) - \mathrm{Var}\big(\widetilde{\theta}_y\big)\big\} = E\Big[\big(\gamma_y(X) - E[\gamma_y(X)]\big)A\big(\gamma_y(X) - E[\gamma_y(X)]\big)^\top\Big] + o(1),$$

where $\gamma_y(X) = [\gamma^{(1)}_y(X), \ldots, \gamma^{(K)}_y(X)]^\top$ and

$$A := \begin{bmatrix} \pi_1^{-1} - 1, & -1, & \ldots, & -1 \\ -1, & \pi_2^{-1} - 1, & \ldots, & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1, & -1, & \ldots, & \pi_K^{-1} - 1 \end{bmatrix}.$$

The variant of Lagrange's identity in Lemma 1 with $\sum_{w\in\mathcal{W}} \pi_w = 1$ shows that, for an arbitrary vector $v := (v_1, \ldots, v_K)^\top \in \mathbb{R}^k$,

$$\begin{aligned} & v^\top\big(\gamma_y(X) - E[\gamma_y(X)]\big)A\big(\gamma_y(X) - E[\gamma_y(X)]\big)^\top v \\ &= \sum_{w\in\mathcal{W}} \frac{v_w^2\big(\gamma^{(w)}_y(X) - E[\gamma^{(w)}_y(X)]\big)^2}{\pi_w} - \Big(\sum_{w\in\mathcal{W}} v_w\big(\gamma^{(w)}_y(X) - E[\gamma^{(w)}_y(X)]\big)\Big)^2 \\ &= \frac{1}{2}\sum_{w\in\mathcal{W}}\sum_{\substack{w'\in\mathcal{W} \\ w'\neq w}} \frac{\big\{v_w\big(\gamma^{(w)}_y(X) - E[\gamma^{(w)}_y(X)]\big)\pi_{w'} - v_{w'}\big(\gamma^{(w')}_y(X) - E[\gamma^{(w')}_y(X)]\big)\pi_w\big\}^2}{\pi_w\pi_{w'}}. \end{aligned}$$

It follows that

$$v^\top\big\{\mathrm{Var}\big(\widehat{\theta}^{empirical}_y\big) - \mathrm{Var}\big(\widetilde{\theta}_y\big)\big\}v = \frac{1}{2}\sum_{w\in\mathcal{W}}\sum_{\substack{w'\in\mathcal{W} \\ w'\neq w}} \frac{\mathrm{Var}\Big(v_w\gamma^{(w)}_y(X)\pi_{w'} - v_{w'}\gamma^{(w')}_y(X)\pi_w\Big)}{\pi_w\pi_{w'}} + o(n^{-1}).$$

The above equality implies the desired positive semi-definiteness result, because $\mathrm{Var}\big(v_w\gamma^{(w)}_y(X)\pi_{w'} - v_{w'}\gamma^{(w')}_y(X)\pi_w\big) \geq 0$ for any $w, w' \in \mathcal{W}$ with $w \neq w'$.

Furthermore, the positive definite result holds when $\mathrm{Var}\big(v_w\gamma^{(w)}_y(X)\pi_{w'} - v_{w'}\gamma^{(w')}_y(X)\pi_w\big) > 0$ for any $v \in \mathbb{R}^k$ with $v \neq 0$ and for any $w, w' \in \mathcal{W}$ with $w \neq w'$. Because $v \in \mathbb{R}^k$ is chosen arbitrarily except $v \neq 0$ and $\pi_w \in (0, 1)$ for all $w \in \mathcal{W}$, the condition for the positive definiteness can be written as $\mathrm{Var}\big(\gamma^{(w)}_y(X) - r \cdot \gamma^{(w')}_y(X)\big) > 0$ for any $r \in \mathbb{R}$ and for any $w, w' \in \mathcal{W}$ with $w \neq w'$. □

3 Multiplier Bootstrap Procedure

We obtain pointwise confidence bands for distributional parameters, which are functionals of distribution functions, using multiplier bootstrap following [3] and [1]. We outline the procedure to obtain pointwise confidence bands in Algorithm 1 where $\phi((\theta_y)_{y\in\mathcal{Y}})$ is some functional of $(\theta_y)_{y\in\mathcal{Y}}$.

As a prerequisite, letting $\pi := (\pi_1, \ldots, \pi_K)^\top$, define moment functions

$$\psi_y(Z;\theta_y,\gamma_y,\pi) := \left(\psi_y^{(1)}(Z;\theta_y,\gamma_y,\pi_1),\ldots,\psi_y^{(K)}(Z;\theta_y,\gamma_y,\pi_K)\right)^\top,$$

where, for each $w \in \mathcal{W}$,

$$\psi_y^{(w)}(Z;\theta_y,\gamma_y,\pi_w) := \frac{\mathbf{1}_{\{W=w\}}\cdot(\mathbf{1}_{\{Y\le y\}} - \gamma_y^{(w)}(X))}{\pi_w} + \gamma_y^{(w)}(X) - \theta_y^{(w)}. \tag{7}$$

Algorithm 1 Multiplier bootstrap procedure to obtain pointwise confidence bands

Input: Data $\{(X_i, W_i, Y_i)\}_{i=1}^n$; point estimates $\hat{\theta}_y$; influence functions $\hat{\psi}_y(Z_i) := \psi_y(Z_i;\hat{\theta}_y,\hat{\gamma}_y,\hat{\pi})$

(1) Draw multipliers $\{\xi_i\}_{i=1}^n = \{m_{1,i}/\sqrt{2} + ((m_{2,i})^2 - 1)/2\}_{i=1}^n$ independently from the data $\{Z_i\}_{i=1}^n$, where $m_{1,i}$ and $m_{2,i}$ are i.i.d. draws from two independent standard normal random variables.

(2) For each $y \in \mathcal{Y}$, obtain the bootstrap draws $\phi^b(\hat{\theta}_y)$ of $\phi(\hat{\theta}_y)$ as

$$\phi^b(\hat{\theta}_y) = \phi(\hat{\theta}_y^b) \text{ where } \hat{\theta}_y^b = \hat{\theta}_y + \frac{1}{n}\sum_{i=1}^n \xi_i \hat{\psi}_y(Z_i).$$

(3) Repeat (1)-(2) B times and index the bootstrap draws by $b = 1, \ldots, B$.

(4) Obtain bootstrap standard error estimates for $\phi(\hat{\theta}_y)$ for each $y \in \mathcal{Y}$ as

$$\hat{\Sigma}(y) = \sum_{b=1}^B \frac{(\phi^b(\hat{\theta}_y) - \bar{\phi}(\hat{\theta}_y))^2}{B-1},$$

where $\bar{\phi}(\hat{\theta}_y) = \sum_{b=1}^B \frac{\phi^b(\hat{\theta}_y)}{B}$.

(5) Construct $(1-\alpha)\times 100\%$ pointwise confidence band for $\phi((\theta_y)_{y\in\mathcal{Y}})$ as

$$I^{1-\alpha} := \{[\phi(\hat{\theta}_y) \pm \hat{z}_{1-\alpha} \times \hat{\Sigma}(y)] : y \in \mathcal{Y}\}.$$

Result: $(1-\alpha)\times 100\%$ confidence band $I^{1-\alpha}$ for $\phi((\theta_y)_{y\in\mathcal{Y}})$

4 Simulation Study

Here, we describe the parameters used in the simulation experiment and present additional experimental results not included in the main text. The first subsection describes the data generating process, the second subsection states the parameter details of the ML models and experiment environment, and the third subsection shows the numerical results of the simulation experiment.

4.1 Data Generating Process (DGP)

For the simulation experiment, we used the following DGP. We fix the number of covariates d_x as $d_x = 20$ and the sample size n to be $n = 1000$. For each $i = 1, \ldots, n$, we generate $X_i = (X_{1i}, \ldots, X_{20i})$ from $U_{20}((0,1)^{20})$, a multivariate uniform distribution on $(0,1)$. Binary treatment variable W_i follows a Bernoulli distribution with a success probability of $\rho = 0.5$. A continuous outcome variable Y_i is then generated from the outcome equation $Y_i = f(X_i, W_i) + U_i$, where the error term $U_i \sim N(0,1)$. We consider the functional form of

$$f(X_i, W_i) = \sum_{j=1}^{20} \sum_{k=1}^{20} \beta_j \beta_k X_{ji} X_{ki} \tag{8}$$

so that the outcome includes the interactions of covariates. For coefficients β_j, we set

$$\beta_j = \begin{cases} 1 & \text{for } j \in \{1, \ldots, 18\} \\ W_i & \text{for } j \in \{19, 20\} \end{cases}$$

Because β_{19} and β_{20} depend on the treatment variable, the records with $W_i = 1$ are more likely to take higher outcome values. We used quantiles $q \in \{0.05, 0.1, \ldots, 0.95\}$ for the locations, and observed a negative DTE across all of them. The absolute size of the DTE takes the maximum around the median of the outcome distribution.

4.2 Model Implementation and Experiment Environment

The model used for the simulation study follows has the following parameters in Table A2. Here, h_i denotes the number of neurons in the ith hidden layer, Optimizer is the optimization algorithm for training neural networks, and Folds L is the number of folds used for cross-fitting. The models are trained with binary cross-entry loss and Adam Optimizer. We implemented the experiment in Python and used the PyTorch [7] framework to build the models. Experiments are run on a Macbook Pro with 36 GB memory and the Apple M3 Pro chip. All models are trained on the CPU, and the same environment was used for all experiments.

4.3 Pointwise Estimation Result

The pointwise RMSE reduction (%) is summarized in Table A3 and its raw values for each y are available in Table A4. In Table A3, the minimum, 25 percentile, median, 75 percentile and the maximum of pointwise SE reduction are reported. As shown in the result, the multi-task adjustment with the monotonic constraint NN achieves the highest RMSE reduction in most locations, while the multi-task adjustment without the constraint also outperforms the other two methods. We also applied BART [4], DNet [8], and Jiang et al. [6] to our DGP and compared our DTE RMSE reduction (%) with their QTE RMSE reduction (%) as reported below. BART and DNet are not designed to estimate unconditional QTE precisely, while Jiang et al. proposed a regression-adjustment method like ours, designed to decrease the variance of QTE estimations compared to unadjusted QTE. The result shows that the proposed multi-task NN adjustment achieves a higher variance reduction than BART and Jiang et al. and improves the variance across all quantiles, unlike DNet.

Parameter	Simulation	Water Consumption	ABEMA
Input size d_x	20	12	10
Number of hidden layers	3	3	3
h_1	128	128	16
h_2	64	64	16
h_3	19	200	51
g	exp	$ReLU$	exp
$f(x)$	$arctan(x)/(\pi/2)$	$\frac{1-exp(-x)}{1+exp(-x)}$	$arctan(x)/(\pi/2)$
σ	ReLU	ReLU	ReLU
Learning Rate	0.01	0.001	0.001
Batch Size	16	64	128
Folds L	2	2	2

Table A2: Model Parameters for Empirical Studies
Notes: All neural network models use identical parameters for each experiment, including the number of folds and batch size except for the size of the last hidden layer.

Method	min	p25	p50	p75	max
Single Task (Linear)	7.9	23.9	30.0	36.2	39.7
Single Task (NN)	4.8	23.7	30.3	34.9	36.4
Multi Task (NN)	26.6	37.3	42.0	45.0	49.2
Multi Task (Monotonic NN)	**28.0**	**38.1**	**43.9**	**47.6**	**51.1**

Table A3: RMSE Reduction in DTE Estimation: Simulation Study
Notes: Percentage reduction in RMSE compared to empirical DTE across models. Sample size n=1,000 with S=500 simulation iterations.

Quantile	Single-Task (Linear) [2] - DTE	Single-Task (NN) [2] - DTE	Multi-Task (NN) - DTE	Multi-Task (Monotonic NN) - DTE	DNet [8] - QTE	Jiang et al. [6] - QTE	BART [4] - QTE
0.05	7.85	4.80	27.76	28.02	-134.38	9.71	-16.38
0.10	15.39	15.35	37.34	36.57	-97.73	15.75	0.09
0.15	23.84	24.02	42.99	41.05	-81.55	20.37	-1.00
0.20	28.02	30.93	43.51	45.66	-55.88	20.80	6.09
0.25	29.97	30.55	43.92	46.51	-36.35	26.09	3.18
0.30	30.95	30.26	41.67	43.75	-13.21	30.04	6.64
0.35	36.68	32.10	44.29	47.55	21.71	29.67	5.50
0.40	37.26	31.40	45.18	48.45	43.48	31.42	3.26
0.45	39.68	35.04	48.55	51.15	58.24	31.23	5.69
0.50	38.22	34.37	49.23	50.69	67.80	33.60	3.81
0.55	36.15	36.42	47.16	48.69	56.98	35.06	6.09
0.60	35.75	30.80	47.15	47.84	37.77	31.96	3.06
0.65	33.74	31.52	42.38	44.00	18.12	32.89	6.82
0.70	32.48	28.83	39.48	41.30	5.36	30.08	4.57
0.75	34.48	27.18	40.77	45.26	-20.38	27.30	6.58
0.80	29.45	23.22	37.47	40.44	-34.22	26.47	3.02
0.85	23.88	21.75	37.00	38.15	-61.08	20.25	5.97
0.90	18.44	17.41	33.25	33.57	-65.64	19.39	4.08
0.95	8.03	8.12	26.60	32.26	-59.62	12.99	-1.08

Table A4: Pointwise RMSE reduction (%) relative to the empirical DTE or QTE across different estimation methods in the simulation study. n=1,000, S=500.

Model	**Accuracy ↑**	**Precision ↑**	**Recall ↑**
Single-Task (Linear)	0.933	0.937	0.928
Single-Task (NN)	0.891	0.889	0.892
Multi-Task (NN)	0.954	0.946	**0.963**
Multi-Task (Monotonic NN)	**0.959**	**0.970**	0.945

Table A5: Comparison of prediction performance on the simulation data.

5 Nudges to Reduce Water Consumption

Here, we describe the parameters used in Nudges to Reduce Water Consumption experiment and present additional experimental results that were not included in the main text. The dataset from the randomized experiment can be downloaded at https://doi.org/10.7910/DVN1/22633 [5]. The covariates used for this experiment are monthly water consumption during the year prior to the experiment. The locations used for this experiment are $\tilde{\mathcal{Y}} = (1, \ldots, 199, 200)^{\mathrm{T}}$ and Figure A1 shows the histogram of outcomes. As described in the main text, most records are distributed from 0 to 100 and the range 100 to 200 contains a limited number of records. The parameters used for this experiment are described in Table A2. The numerical results in Table A6 align with the simulation experiment, showing that the multi-task neural network outperforms the other two methods, with the monotonic constraint further enhancing precision.

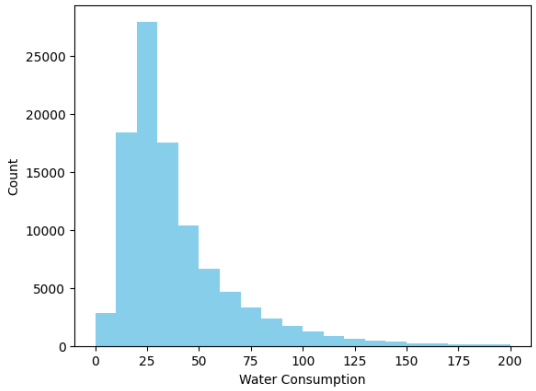

Figure A1: Water Consumption Distribution
Notes: Values in thousands of gallons. Most observations range between 0 and 100 thousand gallons.

Method	min	p25	p50	p75	max
Single-Task (Linear)	-0.35	7.39	12.6	20.5	26.5
Single-Task (NN)	-2.11	12.6	18.2	28.7	33.0
Multi-Task (NN)	**6.78**	17.4	19.1	29.5	32.9
Multi-Task (Monotonic NN)	3.08	**17.5**	**20.3**	**29.5**	**33.3**

Table A6: Summary of SE Reduction in DTE Estimation (%): Water Conservation Experiment
Notes: Percentage reduction in standard errors compared to the empirical DTE across models. Sample size n=78,500 with B=5,000 multiplier bootstrap iterations.

6 ABEMA Content Promotion

Here, we describe the parameters used in the ABEMA content promotion experiment and present additional experimental results that were not included in the main text. The locations used for this experiment are $\tilde{\mathcal{Y}} = (0, 1, \ldots, 50)^{\mathrm{T}}$. Table A7 presents the numerical values of the SE reduction rate. As in the previous two experiments, the multi-task NN and monotonic multi-task NN adjustments outperform the other two adjustment methods. Unlike the other two experiments, no clear improvement was seen by the monotonicity constraint. We hypothesize that the large sample size in this experiment eliminates the need for neural network models to benefit from the monotonic constraint. The parameters used for this experiment are described

in Table A2. We limit the size of neurons in the NN in this experiment due to its large sample size.

Method	min	p25	p50	p75	max
Single-Task (Linear)	**0.406**	5.10	6.31	7.16	14.3
Single-Task (NN)	0.249	5.49	6.45	7.86	15.3
Multi-Task (NN)	0.305	**5.65**	**7.02**	**7.95**	**15.7**
Multi-Task (Monotonic NN)	0.310	**5.65**	7.01	**7.95**	**15.7**

Table A7: Summary of SE Reduction in DTE Estimation (%): ABEMA Content Promotion

Notes: SE reduction (%) is calculated over the empirical DTE across models. Sample size n=4,311,905 with B=5,000 multiplier bootstrap iterations.

7 Validity of Sub-linear Assumption

7.1 Computational Analysis of Linear Regression

Consider the following notation:

- $X \in \mathbb{R}^{n \times d}$: input matrix
- $Y \in \mathbb{R}^{n \times p}$: output matrix with p target dimensions
- $B \in \mathbb{R}^{d \times p}$: parameter matrix

We assume $n \gg d, p$.

Joint Multivariate Linear Regression. The closed-form solution for joint multivariate linear regression is given by

$$B = (X^\top X)^{-1} X^\top Y.$$

The computational complexity consists of the following components:

- $\mathcal{O}(nd^2)$ for computing $X^\top X$
- $\mathcal{O}(d^3)$ for matrix inversion
- $\mathcal{O}(ndp)$ for computing $X^\top Y$
- $\mathcal{O}(d^2 p)$ for the final matrix multiplication

Therefore, the total computational cost of joint multivariate linear regression is

$$\mathcal{O}(nd^2 + d^3 + ndp + d^2 p). \tag{9}$$

Separate Univariate Linear Regression. For training univariate linear regression separately, we compute the regression coefficient for each output dimension $y^{(i)} \in \mathbb{R}^n$:

$$\beta^{(i)} = (X^\top X)^{-1} X^\top y^{(i)}.$$

The computational cost per output dimension $y^{(i)}$ is $\mathcal{O}(nd^2 + d^3)$. Consequently, the total computational cost over all p outputs is

$$\mathcal{O}(pnd^2 + pd^3). \tag{10}$$

Efficiency Comparison. Given that $n \gg d, p$, we observe that

$$\mathcal{O}(nd^2 + d^3 + ndp + d^2p) \ll \mathcal{O}(pnd^2 + pd^3), \tag{11}$$

which demonstrates that linear regression satisfies the sub-linear computational scaling property outlined in Assumption 3, confirming that joint estimation provides significant computational advantages over separate univariate approaches.

7.2 Empirical Computation Complexity Analysis of Neural Networks

We hypothesize that deeper architectures, such as ResNet can achieve substantially higher computational cost reduction under our proposed framework. To empirically validate this hypothesis, we analyze the computational complexity of several neural network architectures by measuring multiply-accumulate (MAC) operations for models with single versus multiple outputs using the `ptflop` Python library.

Model	1 output	20 outputs
Dense NN used in our simulation	11.2 K	12.4 K
BERT	852.23 M	852.24 M
ResNet-18	1.82 G	1.82 G

Table A8: Computational complexity comparison across different neural network architectures measured in MAC operations.

These results demonstrate that complex models satisfy the sub-linear scaling property outlined in Assumption 3. Notably, the computational complexity of both BERT and ResNet-18 remains virtually unchanged when scaling from single to multiple outputs, confirming near-constant computational overhead. This behavior arises because modern deep architectures such as convolutional neural networks (CNNs) and Transformers typically employ deep encoding layers that extract shared representations, making them particularly well-suited for our multi-task learning approach.

References

[1] Alexandre Belloni, Victor Chernozhukov, Ivan Fernandez-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

[2] Undral Byambadalai, Tatsushi Oka, and Shota Yasui. Estimating distributional treatment effects in randomized experiments: Machine learning for variance reduction. 2024.

[3] Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

[4] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), March 2010.

[5] Paul J. Ferraro and Michael K. Price. Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *The review of economics and statistics*, 95:64–73, 2013.

[6] Liang Jiang, Peter CB Phillips, Yubo Tao, and Yichong Zhang. Regression-adjusted estimation of quantile treatment effects under covariate-adaptive randomizations. *Journal of Econometrics*, 234(2):758–776, 2023.

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[8] Guojun Wu, Ge Song, Xiaoxiang Lv, Shikai Luo, Chengchun Shi, and Hongtu Zhu. Dnet: Distributional network for distributional individualized treatment effects. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5215–5224, New York, NY, USA, 2023. Association for Computing Machinery.