A. Theorems and Lemmas

Theorem A.1. *(Chernoff Bound for unbounded sub-Gaussian random variables) Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with parameter σ. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. For all $\varepsilon > 0$,*

$$\mathbb{P}\left[\left|\overline{X}\right| \geq \varepsilon\right] \leq \exp\left\{\frac{-n\varepsilon^2}{2\sigma^2}\right\}.$$

Corollary A.2. *(High probability bound on the sum of unbounded sub-Gaussian random variables) For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\overline{X}\right| < \sigma\sqrt{\frac{2\log(1/\delta)}{n}}$$

Theorem A.3. *(Chernoff/Hoeffding's inequality) Let $X_1, \ldots, X_n$ be independent and bounded random variables such that $a \leq X_i \leq b$ for all i. Then*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \geq \varepsilon\right] \leq \exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)$$

Corollary A.4. *(High probability upper bound on the sum of bounded random variables) For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^{n} X_i \leq (b-a)\sqrt{\frac{\log(1/\delta)}{2n}},$$

where $X_i \in [a,b]$ for all i from 1 to n.

Lemma A.5. *(Cauchy-Schwarz Inequality) For any n-dimensional vectors $u, v \in \mathbb{R}^n$, the L^2-norm of the inner product of u and v is less than or equal to the L^2-norm of u times the L^2-norm of v, i.e.*

$$\left\|\langle u, v\rangle\right\|_2 \leq \left\|u\right\|_2 \cdot \left\|v\right\|_2.$$

Alternatively, for any $m \times n$-dimensional matrices $A \in \mathbb{R}^{m\times n}$ and n-dimensional vector $v \in \mathbb{R}^n$, the L^2-norm of the dot product of A and v is less than or equal to the spectral norm of A times the L^2-norm of v, i.e.

$$\left\|Av\right\|_2 \leq \left\|A\right\|_2 \cdot \left\|v\right\|_2.$$

Theorem A.6. *(Matrix Chernoff) Consider a finite sequence X_k of independent, random, self-adjoint matrices with common dimension d. Assume that:*

$$0 \leq \lambda_{\min}(X_k) \quad \textit{and} \quad \lambda_{\max}(X_k) \leq \omega \quad \textit{for each index k.}$$

Introduce the random matrix $Y = \sum_k X_k$. Define the minimum eigenvalue μ_{min} and maximum eigenvalue μ_{max} of the expectation $\mathbb{E}[Y]$.

$$\mu_{min} = \lambda_{\min}\left\{\mathbb{E}[Y]\right\} = \lambda_{\min}\left\{\sum_k \mathbb{E}[X_k]\right\}, \quad \textit{and}$$

$$\mu_{\max} = \lambda_{\max}\left\{\mathbb{E}[Y]\right\} = \lambda_{\max}\left\{\sum_k \mathbb{E}[X_k]\right\}$$

Then, for $\theta > 0$,

$$\mathbb{E}[\lambda_{\min}(Y)] \geq \frac{1-e^{-\theta}}{\theta}\mu_{\min} - \frac{1}{\theta}L\log d, \quad \textit{and}$$

$$\mathbb{E}[\lambda_{\max}(Y)] \leq \frac{e^{\theta}-1}{\theta}\mu_{\max} + \frac{1}{\theta}L\log d$$

Furthermore,

$$\mathbb{P}[\lambda_{\min}(Y) \leq (1-\varepsilon)\mu_{\min}] \leq d\left[\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right]^{\mu_{\min}/\omega} \quad \textit{for } \varepsilon \in [0,1)$$

$$\mathbb{P}[\lambda_{\max}(Y) \leq (1+\varepsilon)\mu_{\max}] \leq d\left[\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right]^{\mu_{\max}/\omega} \quad \textit{for } \varepsilon \geq 0$$

Theorem A.7. *(Union bound): For a countable set of events $A_1, A_2, \ldots$, we have*

$$\mathbb{P}\left[\bigcup_i A_i\right] \leq \sum_i \mathbb{P}(A_i)$$

B. IV Estimator Proof for Control-Treatment Setting

Recall that our reward model can be stated as the following equation:

$$y_i = \theta x_i + g_i^{(u_i)} \tag{13}$$

To analyze the Wald estimator, we introduce two conditional probabilities that an agent chooses the treatment given a recommendation $\hat{\gamma}_0$ and $\hat{\gamma}_1$, given as proportions over a set of n samples $(x_i, z_i)_{i=1}^n$ and formally defined as

$$\hat{\gamma}_0 = \hat{\mathbb{P}}_{(x_i,z_i)_{i=1}^n}[x_i = 1 | z_i = 0] = \frac{\sum_{i=1}^n x_i(1-z_i)}{\sum_{i=1}^n (1-z_i)^2} \quad \text{and} \quad \hat{\gamma}_1 = \hat{\mathbb{P}}_{(x_i,z_i)_{i=1}^n}[x_i = 1 | z_i = 1] = \frac{\sum_{i=1}^n x_i z_i}{\sum_{i=1}^n z_i^2}$$

Then, we can write the action choice x_i as such:

$$\begin{aligned} x_i &= \hat{\gamma}_1 z_i + \hat{\gamma}_0(1-z_i) + \eta_i \\ &= \hat{\gamma} z_i + \hat{\gamma}_0 + \eta_i \end{aligned}$$

where $\eta_i = x_i - \hat{\gamma}_1 z_i - \hat{\gamma}_0(1-z_i)$ and $\hat{\gamma} = \hat{\gamma}_1 - \hat{\gamma}_0$ is the in-sample *compliance coefficient*. Now, we can rewrite the reward y_i as

$$\begin{aligned} y_i &= \theta\left(\hat{\gamma} z_i + \hat{\gamma}_0 + \eta_i\right) + g_i^{(u_i)} \\ &= \underbrace{\theta\,\hat{\gamma}}_{\beta} z_i + \theta\hat{\gamma}_0 + \theta\eta_i + g_i^{(u_i)} \end{aligned}$$

Let operator $\bar{\cdot}$ denote the sample mean, e.g. $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{g} := \frac{1}{n}\sum_{i=1}^n g_i^{(u_i)}$. $\bar{\eta} = \frac{1}{n}\sum_{i=1}^n \eta_i = 0$, by definition.

Then,

$$\bar{y} = \beta\bar{z} + \theta\hat{\gamma}_0 + \theta\bar{\eta} + \bar{g} + \bar{\varepsilon}$$

Thus, the centered reward and treatment choice at round i are given as:

$$\begin{cases} y_i - \bar{y} = \theta(x_i - \bar{x}) + g_i^{(u_i)} - \bar{g} \\ y_i - \bar{y} = \beta(z_i - \bar{z}) + \theta(\eta_i - \bar{\eta}) + g_i^{(u_i)} - \bar{g} \\ x_i - \bar{x}_i = \hat{\gamma}(z_i - \bar{z}) + \eta_i - \bar{\eta} \end{cases} \tag{14}$$

This formulation of the centered reward $y_i - \bar{y}$ allows us to express and bound the error between the treatment effect θ and its instrumental variable estimate $\hat{\theta}_S$, which we show in the following Theorem 2.1.

Theorem 2.1 (Finite-sample error bound for Wald estimator)**.** *Let $z_1, z_2, \ldots, z_n \in \{0,1\}$ be a sequence of instruments. Suppose there is a sequence of n agents such that each agent i has their private type u_i drawn independently from $\mathcal{U}$, selects action x_i under instrument z_i, and receives reward y_i. Let sample set $S = (x_i, y_i, z_i)_{i=1}^n$. Let $A : (\{0,1\}^n \times \{0,1\}^n \times \mathbb{R}^n) \to \mathbb{R}$ denote the approximation bound for set S, such that*

$$A(S, \delta) := \frac{2\sigma_g\sqrt{2n\log(2/\delta)}}{|\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})|}$$

and the Wald estimator given by (3) satisfies

$$\left|\hat{\theta}_S - \theta\right| \leq A(S, \delta)$$

with probability at least $1 - \delta$, for any $\delta \in (0,1)$.

Proof. Given a sample set $S = (x_i, y_i, z_i)_{i=1}^n$ of size n, we form an estimate of the treatment effect $\hat{\theta}_S$ via a Two-Stage Least Squares (2SLS). In the first stage, we regress $y_i - \bar{y}$ onto $z_i - \bar{z}$ to get the empirical estimate $\hat{\beta}_S$ and $x_i - \bar{x}$ onto $z_i - \bar{z}$ to get $\hat{\gamma}_S$ as such:

$$\hat{\beta}_S := \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2} \qquad \text{and} \qquad \hat{\gamma}_S := \frac{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2} \tag{15}$$

In the second stage, we take the quotient of these two empirical estimates as the predicted treatment effect $\hat{\theta}_S$, i.e.

$$\begin{aligned} \hat{\theta}_S = \frac{\hat{\beta}_S}{\hat{\gamma}_S} &= \left(\frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2} \right) \left(\frac{\sum_{i=1}^n (z_i - \bar{z})^2}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} \right) \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} \end{aligned} \tag{16}$$

Next, we can express the absolute value of the difference between the true treatment effect θ and the IV estimate of the treatment effect $\hat{\theta}_S$ given a sample set S of size n as such:

$$\begin{aligned} \left| \hat{\theta}_S - \theta \right| &= \left| \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} - \theta \right| \\ &= \left| \frac{\sum_{i=1}^n \left(\theta(x_i - \bar{x}) + g_i^{(u_i)} - \bar{g} \right)(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} - \theta \right| && \text{(by Equation (14))} \\ &= \left| \theta + \frac{\sum_{i=1}^n \left(g_i^{(u_i)} - \bar{g} \right)(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} - \theta \right| \\ &= \frac{\left| \sum_{i=1}^n \left(g_i^{(u_i)} - \bar{g} \right)(z_i - \bar{z}) \right|}{\left| \sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z}) \right|} \end{aligned} \tag{17}$$

In order to complete our proof, we demonstrate an upper bound on the numerator $\left| \sum_{i=1}^n \left(g_i^{(u_i)} - \bar{g} \right)(z_i - \bar{z}) \right|$ of Equation (17) in the last line above. We do so in Lemma B.1. □

Lemma B.1. *For all $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\left| \sum_{i=1}^n \left(g_i^{(u_i)} - \bar{g} \right)(z_i - \bar{z}) \right| \leq 2\sigma_g \sqrt{2n \log(2/\delta)} \tag{18}$$

if the set of $g_i^{(u_i)}$ are i.i.d. sub-Gaussian random variables with sub-Gaussian norm σ_g.

Proof. We can rewrite the left hand side as follows

$$\begin{aligned}
&\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\bar{g}\right)(z_i-\bar{z})\right| \\
&=\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]+\mathbb{E}[g^{(u)}]-\bar{g}\right)(z_i-\bar{z})\right| \\
&=\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)z_i-\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)\bar{z}+\sum_{i=1}^{n}\left(\mathbb{E}[g^{(u)}]-\bar{g}\right)(z_i-\bar{z})\right| \\
&=\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)z_i-\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)\bar{z}\right| && \text{(since } \textstyle\sum_{i=1}^{n}(z_i-\bar{z})=0\text{)} \\
&\leq\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)z_i\right|+\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)\right| && \text{(by the triangle inequality and } |\bar{z}|\leq 1\text{)}
\end{aligned}$$

Now, if $g_i^{(u_i)}$ is sub-Gaussian, then the last line in the system of inequalities above is given as:

$$\begin{aligned}
&\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)z_i\right|+\left|\sum_{i=1}^{n}\left(g_i^{(u_i)}-\mathbb{E}[g^{(u)}]\right)\right| \\
&\leq\left|\sigma_g\sqrt{2n_1\log(1/\delta_1)}\right|+\left|\sigma_g\sqrt{2n\log(1/\delta_2)}\right| && \text{(by Corollary A.2, where } n_1:=\textstyle\sum_{i=1}^{n}z_i\text{)} \\
&\leq 2\sigma_g\sqrt{2n\log(2/\delta)} && \text{(since } n_1\leq n \text{ and by Theorem A.7, where } \delta_1=\delta_2=\delta/2\text{)}
\end{aligned}$$

This recovers the stated bound and finishes the proof for Theorem 2.1. □

Next, we demonstrate a lower bound on the denominator of Theorem 2.1, in terms of the level of compliance at each phase of Algorithms 1 and 2.

Theorem B.2 (Lower bound on $|\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})|$ for a type 0 compliant sample set)**.** *Let $S=(x_i,y_i,z_i)_{i=1}^{n}$ denote a sample set which satisfies the conditions of Theorem 2.1. Furthermore, assume that there are p_c fraction of agents in the population who would be compliant. Recall that $\bar{z}=\frac{1}{n}\sum_{i=1}^{n}z_i$ and $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$. Then, the denominator of the approximation bound $A(S,\delta)$ (from Theorem 2.1) is lower bounded as such:*

$$\left|\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})\right|\geq\begin{cases}n\bar{z}(1-\bar{z}) & \text{if } p_c=1 \text{ (i.e. if all agents are compliant)};\\ n\bar{z}(1-\bar{z})p_c-(3-\bar{z})\sqrt{\frac{n\bar{z}\log(3/\delta)}{2(1-\bar{z})}} & \text{with probability at least } 1-\delta \text{ for any } \delta\in(0,1) \text{ otherwise}.\end{cases}$$

Proof. In this theorem, we formulate the denominator of the approximation bound in Theorem 2.1 in terms of $\bar{z}$, since $\bar{z}$ is determined by the social planner. For any type u, let $u\in U_c$ denote that agents of type u comply; let $u\in U_0$ denote that agents of type u are never-takers (agents which prefer control, according to their prior); and let $u\in U_1$ denote that agents of type u are always-takers (agents which prefer treatment, according to their prior). Let p_0 and p_1 be the fractions of never-takers and always-takers, respectively.

Next, we expand the binomial in the denominator and arrive at the following simplified form:

$$\left|\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})\right|=\left|\sum_{i=1}^{n}x_iz_i-\bar{z}\sum_{i=1}^{n}x_i-\bar{x}\sum_{i=1}^{n}z_i+\bar{z}\bar{x}\right|=\left|\sum_{i=1}^{n}x_iz_i-n\bar{z}\bar{x}\right| \tag{19}$$

First, observe that at any round i, the product $x_i=1$ only when agent i is a non-compliant always-taker or when $z_i=1$ and agent i is compliant. Formally, for any agent i, action choice $x_i=1$ is equivalent to the following:

$$x_i=1\equiv(z_i=1\wedge u_i\in U_c)\vee(u_i\in U_1\wedge u_i\notin U_c) \tag{20}$$

Then, the sum $\sum_{i=1}^{n} x_i$ can be expressed as follows:

$$\begin{aligned}\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} \mathbb{1}\left[(z_i = 1 \wedge u_i \in U_c) \vee (u_i \in U_1 \wedge u_i \notin U_c)\right] \\ &= \sum_{i=1}^{n} \mathbb{1}\left[(z_i = 1 \wedge u_i \in U_c\right] + \sum_{i=1}^{n} \mathbb{1}\left[u_i \in U_1 \wedge u_i \notin U_c\right] \\ &= \left(\sum_{i=1}^{n} z_i\right)\left(\hat{p}_{c:z_i=1} + n\hat{p}_{nc1}\right) \\ &= n\left(\bar{z}\hat{p}_{c:z_i=1} + \hat{p}_{nc1}\right) \end{aligned} \tag{21}$$

where we define $\hat{p}_{c:z_i=1}$ as the empirical proportion of agents with types in U_c when the recommendation $z = 1$ and $\hat{p}_{nc1}$ as the empirical proportion of non-compliant always-takers. Formally, $\hat{p}_{c:z_i=1} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[u_i \in U_c, z_i = 1]$ and $\hat{p}_{nc1} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[u_i \in U_1 \wedge u_i \notin U_c]$. Define p_{nc1} to be the proportion of non-compliant always-takers in the population of agents. Then, in expectation over the randomness of how agents arrive, $\mathbb{E}[\hat{p}_{c:z_i=1}] = p_c$ and $\mathbb{E}[\hat{p}_{nc1}] = p_{nc1}$.

Next, we rewrite the sum $\sum_{i=1}^{n} x_i z_i$ in terms of $\bar{z}$ and some population constants. Observe that at any round i, the product $x_i z_i = 1$ only when both $x_i = 1$ and $z_i = 1$. Thus, by Equation (20), for any agent i, the event $x_i z_i = 1$ is equivalent to the following:

$$\begin{aligned} x_i z_i = 1 &\equiv z_i = 1 \wedge ((z_i = 1 \wedge u_i \in U_c) \vee (u_i \in U_1 \wedge u_i \notin U_c)) \\ &\equiv z_i = 1 \wedge (u_i \in U_c \vee (u_i \in U_1 \wedge u_i \notin U_c)) \end{aligned}$$

Then, the sum $\sum_{i=1}^{n} x_i z_i$ can be expressed as follows:

$$\begin{aligned}\sum_{i=1}^{n} x_i z_i &= \sum_{i=1}^{n} \mathbb{1}\left[z_i = 1 \wedge (u_i \in U_c \vee (u_i \in U_1 \wedge u_i \notin U_c))\right] \\ &= \left(\sum_{i=1}^{n} \mathbb{1}[z_i = 1]\right)\left(\hat{p}_c|_{z_i=1} + \hat{p}_{nc1}|_{z_i=1}\right) \\ &= n\bar{z}\left(\hat{p}_{c:z_i=1} + \hat{p}_{nc1:z_i=1}\right) \end{aligned} \tag{22}$$

where we define $\hat{p}_{nc1:z_i=1}$ as the empirical proportions of non-compliant always-takers who arrive when $z_i = 1$ — i.e. $\hat{p}_{nc1:z_i=1} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[u_i \in U_1 \wedge u_i \notin U_c, z_i = 1]$. In expectation over the randomness of how agents arrive, $\mathbb{E}[\hat{p}_{nc1:z_i=1}] = p_{nc1}$.

Finally, by Equations (19), (21) and (22), we can provide a high probability lower bound on the denominator as such:

$$\begin{aligned}
\left|\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})\right| &= \left|\sum_{i=1}^{n}x_iz_i - n\bar{z}\bar{x}\right| && \text{(by Equation (19))}\\
&= \left|\sum_{i=1}^{n}x_iz_i - n\bar{z}\bar{x}\right| && \text{(by Equation (19))}\\
&= |n\bar{z}\left(\hat{p}_{c:z_i=1}+\hat{p}_{nc1:z_i=1}\right) - n\bar{z}\left(\bar{z}\hat{p}_{c:z_i=1}+\hat{p}_{nc1}\right)| && \text{(by Equations (21) and (22))}\\
&= |n\bar{z}\left((1-\bar{z})\hat{p}_{c:z_i=1}+\hat{p}_{nc1:z_i=1}-\hat{p}_{nc1}\right)|\\
&\geq \left|n\bar{z}\left((1-\bar{z})\left(p_c-\sqrt{\frac{\log(1/\delta_1)}{2n\bar{z}}}\right)+p_{nc1}-\sqrt{\frac{\log(1/\delta_2)}{2n\bar{z}}}-\left(p_{nc1}+\sqrt{\frac{\log(1/\delta_3)}{2n}}\right)\right)\right| \\
& && \text{(by Theorem A.3)}\\
&\geq \left|n\bar{z}\left((1-\bar{z})p_c-(1-\bar{z})\sqrt{\frac{\log(3/\delta)}{2n\bar{z}}}-\sqrt{\frac{\log(3/\delta)}{2n\bar{z}}}-\sqrt{\frac{\log(3/\delta)}{2n}}\right)\right| \\
& && \text{(by Theorem A.7 where } \delta_1=\delta_2=\delta_3=\delta/3)\\
&\geq \left|n\bar{z}\left((1-\bar{z})p_c-(3-\bar{z})\sqrt{\frac{\log(3/\delta)}{2n\bar{z}}}\right)\right|\\
&= n\bar{z}(1-\bar{z})p_c-(3-\bar{z})\sqrt{\frac{n\bar{z}\log(3/\delta)}{2(1-\bar{z})}}
\end{aligned}$$

with probability at least $1-\delta$ for any $\delta\in(0,1)$. □

C. Missing Proofs for Section 3

Claim C.1. *For any agent t at round t with recommendation policy π_t with a positive probability of recommending either control or treatment, according to the prior $\mathcal{P}^{(u_t)}$, i.e. $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=0]>0$ and $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=1]>0$. Furthermore, $a^{(u)}$ and $b^{(u)}$ denote the initially preferred and unpreferred actions for any type u, i.e. $a^{(u)}:=\mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta]\geq 0]$ and $b^{(u)}:=\mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta]<0]$. Formally, the following holds:*

$$\left\{(-1)^{a^{(u_t)}}\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t=b^{(u_t)}]\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=b^{(u_t)}]\geq 0\right\}\Rightarrow\left\{(-1)^{b^{(u_t)}}\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t=a^{(u_t)}]\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=a^{(u_t)}]<0\right\}$$

Proof. Note that $a^{(u)}$ is defined in such a way that $(-1)^{a^{(u)}}\mathbb{E}_{\mathcal{P}^{(u)}}[\theta]<0$ always: if agents of type u prefer initially control, then $a^{(u)}=0$ and $(-1)^{a^{(u)}}\mathbb{E}_{\mathcal{P}^{(u)}}[\theta]=\mathbb{E}_{\mathcal{P}^{(u)}}[\theta]<0$; if agents of type u initially prefer treatment, then $a^{(u)}=1$ and $(-1)^{a^{(u)}}\mathbb{E}_{\mathcal{P}^{(u)}}[\theta]=-\mathbb{E}_{\mathcal{P}^{(u)}}[\theta]<0$. Then,

Recall that we assume that type 0 agents prefer the control, i.e. the expected treatment effect $\mathop{\mathbb{E}}_{\mathcal{P}^{(0)}}[\theta]<0$. Then:

$$\begin{aligned}
&(-1)^{a^{(u_t)}}\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t=b^{(u_t)}]\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=b^{(u_t)}]-(-1)^{b^{(u_t)}}\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t=a^{(u_t)}]\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=a^{(u_t)}]\\
&=(-1)^{a^{(u_t)}}\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t=b^{(u_t)}]\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=b^{(u_t)}]+(-1)^{a^{(u_t)}}\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t=a^{(u_t)}]\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=a^{(u_t)}]\\
&=(-1)^{a^{(u_t)}}\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta]<0.
\end{aligned}$$

Therefore, given that both $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=0]>0$ and $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=1]>0$ and, by assumption, $(-1)^{a^{(u_t)}}\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t=b^{(u_t)}]\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=b^{(u_t)}]\geq 0$, then it must be that $(-1)^{b^{(u_t)}}\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t=a^{(u_t)}]\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t=a^{(u_t)}]<0$. □

C.1. Algorithm 1 Proofs and Extension 1

Lemma 3.2 (Type 0 compliance with Algorithm 1)**.** *Under Assumption 3.1, any type 0 agent who arrives in the last ℓ rounds of Algorithm 1 is compliant with any recommendation, as long as the exploration probability ρ satisfies*

$$\rho \leq 1 + \frac{4\mu^{(0)}}{\mathbb{P}_{\mathcal{P}^{(0)}}[\xi] - 4\mu^{(0)}} \tag{5}$$

where the event ξ is defined above in Equation (4).

Proof. Let the event $\xi = \xi^{(0)}$ (as given by Definition C.2). By Lemma C.3, if ρ satisfies the following condition, then any type 0 agent will comply with any recommendation of the last ℓ rounds of Algorithm 1:

$$\rho \leq 1 + \frac{4\mu^{(0)}}{\mathbb{P}_{\mathcal{P}^{(0)}}[\xi^{(0)}] - 4\mu^{(0)}} \tag{23}$$

□

Definition C.2 (Extension 1 of Algorithm 1)**.** Here, we formalize the recommendation policy of Extension 1 in Section 3.1, which modifies Algorithm 1 in two ways:

1. We redefine event ξ as $\xi^{(u)}$ such that it is relative to any type u, defined as follows:

$$\xi^{(u)} = \left\{ \bar{y}^1 > \bar{y}^0 + \sigma_g \left(\sqrt{\frac{2\log(2/\delta)}{\ell_0}} + \sqrt{\frac{2\log(2/\delta)}{\ell_1}} \right) + G^{(u_t)} + \frac{1}{2} \right\}, \tag{24}$$

   where $G^{(u_t)}$ is an upper bound on the difference between the prior mean of the treatment versus the control according to type u, i.e. $G^{(u_t)} > \mathbb{E}_{\mathcal{P}^{(u)}}[g^1 - g^0]$, and where $\mathbb{E}_{\mathcal{P}^{(u)}}[g^0]$ and $\mathbb{E}_{\mathcal{P}^{(u)}}[g^1]$ are the expected baseline rewards for initial never-takers and always-takers.

2. If we are trying to incentivize compliance for always-takers, then those agents in the exploration set E are recommended control (rather than treatment, as described in the pseudocode for Algorithm 1).

Lemma C.3 (Arbitrary Type Compliance with Extension 1 of Algorithm 1)**.** *Under Assumption 3.1, any type u_t agent who arrives at round t in the last ℓ rounds of Extension 1 of Algorithm 1 (given in Definition C.2) is compliant with any recommendation z_t, as long as the exploration probability ρ satisfies:*

$$\rho \leq 1 + \frac{4\mu^{(u_t)}}{\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - 4\mu^{(u_t)}} \tag{25}$$

where the event $\xi^{(u_t)}$ is defined in Definition C.2.

Proof. This proof follows a similar structure to the Sampling Stage BIC proof in (Mansour et al., 2015).

We will prove compliance for any type u in the more general Extension 1 of Algorithm 1, as given in Definition C.2, which admits arbitrarily many types and the option to incentivize initial always-takers, instead of initial never-takers, to comply.

Let recommendation policy π be that described in Definition C.2, i.e. Extension 1 of Algorithm 1 which admits arbitrarily many types and allows for the exploration recommendations to be given in order to incentivize initial always-takers, instead of initial never-takers, to comply. Throughout this proof, we will assume that the exploration set E is defined relative to the initial preference of any agent of type u_t, who we are proving compliance for.

According to the selection function in Equation (2), if any agent t expects the treatment effect θ to be positive, they will select the treatment $x_t = 1$. Conversely, if they expect the treatment effect θ to be negative, they will select control $x_t = 0$. Thus, for any agent of type u_t at round t, proving compliance entails the expected treatment effect θ over the prior of type u_t and policy π_t is positive given that the recommendation $z_t = 1$ and negative given that the recommendation $z_t = 0$, i.e.

$$\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = 1] \geq 0 \qquad \text{and} \qquad \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = 0] < 0.$$

Next, we show that we can reduce our proof to demonstrating only one of the above statements, depending on the prior preference of type u. Let $a^{(u)}$ and $b^{(u)}$ denote the prior preferred and unpreferred actions for any type u, i.e. $a^{(u)} := \mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] \geq 0]$ and $b^{(u)} := \mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0]$. Because policy π (Algorithm 1 extension) is designed in a such way that at any round t in the last ℓ rounds, treatment or control is recommended each with positive probability —i.e. $\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = 1] > 0$ and $\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = 0] > 0$,— Claim C.1 applies and the following holds:

$$\left\{ (-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \geq 0 \right\} \Rightarrow \left\{ (-1)^{b^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = a^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = a^{(u_t)}] < 0 \right\}.$$

Thus, at round t, in order to prove compliance for agents of type u_t with prior preferred and unpreferred actions $a^{(u_t)}$ and $b^{(u_t)}$, respectively, it suffices to demonstrate that $(-1)^{a^{(u_t)}} \mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}] \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \geq 0$. The remainder of the proof is devoted to demonstrating this.

We first rewrite $\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}]$ in terms of the event $\xi^{(u_t)}$:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \\
&= \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)} \,\&\, t \notin E] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = b^{(u_t)} \,\&\, t \notin E] + \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)} \,\&\, t \in E] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = b^{(u_t)} \,\&\, t \in E] \\
&\qquad \text{(for explore set } E \text{ defined to recommend action } b^{(u_t)}) \\
&= \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi \,\&\, t \notin E] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi \,\&\, t \notin E] + \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)} \,\&\, t \in E] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = b^{(u_t)} \,\&\, t \in E] \\
&\qquad \text{(since the only way } z_t = b^{(u_t)} \text{ when exploiting (i.e. when } t \notin E) \text{ is when event } \xi \text{ occurs)} \\
&= \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi \,\&\, t \notin E] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi \,\&\, t \notin E] + \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | t \in E] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[t \in E], \quad (t \in E \Rightarrow z_t = 1 \text{ by definition of } E) \\
&= \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[t \notin E] + \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[t \in E] \qquad \text{(since } \theta \perp t \in E \text{ and } \xi \perp t \notin E) \\
&= (1 - \rho) \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi] + \rho \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta] \qquad \text{(since agent } t \in E \text{ with probability } \rho) \\
&= (1 - \rho) \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi] + \rho \mu^{(u)} \qquad \text{(by definition, } \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta] = \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}}[\theta] = \mu^{(u)}) \qquad (26)
\end{aligned}$$

Now, we can rewrite our compliance condition as such:

$$(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \geq 0 \equiv (-1)^{a^{(u_t)}} \left((1 - \rho) \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi] + \rho \mu^{(u)} \right) \geq 0.$$

Now, we rewrite this compliance condition strictly in terms of the exploration probability ρ and relative to a number of constants which depend on the prior $\mathcal{P}^{(u_t)}$. Thus, if we set ρ to satisfy the following condition (in Equation (27)), then all

agents of type u will comply with recommendations from policy π (Algorithm 1 extension):

$$(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \geq 0$$

$$(-1)^{a^{(u_t)}} \left((1-\rho) \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}] + \rho \mu^{(u_t)} \right) \geq 0 \quad \text{(by Equation (26))}$$

$$(-1)^{a^{(u_t)}} \left(\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \rho \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}] + \rho \mu^{(u_t)} \right) \geq 0$$

$$(-1)^{a^{(u_t)+1}} \rho \left(\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \mu^{(u_t)} \right) \geq (-1)^{a^{(u_t)+1}} \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}]$$

$$\rho \leq \frac{\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}]}{\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \mu^{(0)}}$$

(since $(-1)^{a^{(u_t)+1}} \rho \left(\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \mu^{(u_t)} \right) < 0$ for any u_t[17])

$$\rho \leq 1 + \frac{\mu^{(u_t)}}{\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \mu^{(u_t)}} \tag{27}$$

Finally, we can further simplify the upper bound on ρ given in Equation (27) above by showing that $\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}]$ satisfies some constant lower bound. This will complete our proof.

For any type u, the baseline reward $g^{(u)}$ is a random variable independently distributed according to a sub-Gaussian distribution with variance $\sigma^{(u)}$ which is bounded above by σ_g, i.e. $\sigma^{(u)} < \sigma_g$ for any u. Furthermore, recall that $G^{(u_t)} > \mathbb{E}_{\mathcal{P}^{u_t}}[g^1 - g^0]$, where $\mathbb{E}_{\mathcal{P}^{u_t}}[g^1]$ and $\mathbb{E}_{\mathcal{P}^{u_t}}[g^0]$ are the expected value of the baseline rewards of always-takers and never-takers over the prior of type u_t, respectively.

Now, we define 3 clean events: $\mathcal{C}_0$ and $\mathcal{C}_1$ pertain to these baseline reward random variables, and $\mathcal{C}_2$ occurs when the first stage of Algorithm 1 generates at least ℓ_0 control samples and at least ℓ_1 treatment samples:

$$\mathcal{C}_0 := \left\{ \bar{y}^0 = \frac{1}{\sum_{t=1}^{\ell} \mathbb{1}[\mu^{(u_t)} < 0]} \sum_{t=1}^{\ell} g^{(u_t)} \mathbb{1}[\mu^{(u_t)} < 0] \leq \sigma_g \sqrt{\frac{2\log(1/\delta_0)}{\ell_0}} - \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[g^0] \right\} \tag{28}$$

$$\mathcal{C}_1 := \left\{ \bar{y}^1 = \frac{1}{\sum_{t=1}^{\ell} \mathbb{1}[\mu^{(u_t)} > 0]} \sum_{t=1}^{\ell} g^{(u_t)} \mathbb{1}[\mu^{(u_t)} > 0] \geq -\sigma_g \sqrt{\frac{2\log(1/\delta_1)}{\ell_1}} - \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[g^1] \right\} \tag{29}$$

$$\mathcal{C}_2 := \left\{ \ell_1 \leq \sum_{i=1}^{\ell'} x_i \leq \ell' - \ell_0 \right\} \tag{30}$$

where $\ell' = 2\max(\ell_0/p_0, \ell_1/p_1)$ is the number of rounds in the first stage of Algorithm 1. Let $\delta_0 = \delta_1 = \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}]/24$. Furthermore, event $\mathcal{C}_2$ occurs when the binomial random variable with success $u_t = x_t = 1$ (since $x_t = u_t$ in the first stage of Algorithm 1) and success probability p_1 is lower bounded by ℓ_1 and upper bounded by $\ell' - \ell_0$. For $\ell' = 2\max(\ell_0/p_0, \ell_1/p_1)$ total trials, the probability of this event is less than $\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}]/24$.

Now, define another clean event $\mathcal{C}$ where all $\mathcal{C}_0$, $\mathcal{C}_1$, and $\mathcal{C}_2$ happen simultaneously. Letting $\delta = \delta_0 + \delta_1 + \delta_2$, the event $\mathcal{C}$ occurs with probability at least $1 - \delta$ where $\delta < \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}]/8$. We can now rewrite $\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\xi^{(u_t)}]$ in

[17] This point is not entirely obvious: If $a^{(u_t)} = 0$, then $(-1)^{a^{(u_t)+1}} < 0$ and $\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] - \mu^{(u_t)} > 0$, since $\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] > 0$ and $\mu^{(u_t)} < 0$. If $a^{(u_t)} = 1$, then $(-1)^{a^{(u_t)+1}} > 0$ and $\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] - \mu^{(u_t)} < 0$, since $\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | \xi^{(u_t)}] < 0$ and $\mu^{(u_t)} > 0$.

terms of event $\mathcal{C}$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] &= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)},\mathcal{C}] + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\neg\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)},\neg\mathcal{C}] \\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)},\mathcal{C}] - \delta \\
&\qquad\qquad \text{(since } \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\neg\mathcal{C}] < \delta \text{ and } \theta \geq -1 \text{ by definition)} \\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \left(\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\neg\mathcal{C}] \right) - \delta \\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \left(\mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \delta \right) - \delta \\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - \delta \left(1 + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \right) \\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - 2\delta \qquad \left(\text{since } \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \leq 1\right) \qquad (31)
\end{aligned}$$

This comes down to finding a lower bound on the denominator of the expression above. We can reduce the dependency of the denominator to a single prior-dependent constant $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]$ if we lower bound the prior-dependent expected value $\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}]$. That way, assuming we know the prior and can calculate the probability of event $\xi^{(u_t)}$, we can pick an appropriate exploration probability ρ to satisfy the compliance condition for all agents of type 0. Then:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}\left[\theta \middle| \bar{y}^1 > \bar{y}^0 + \sigma_g \left(\sqrt{\frac{2\log(1/\delta)}{\ell_0}} + \sqrt{\frac{2\log(1/\delta)}{\ell_1}} \right) + G^{(u_t)} + \frac{1}{2}, \mathcal{C} \right] \\
&\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}\left[\theta \middle| \theta > \frac{1}{\ell_0}\sum_{t=1}^{\ell_0} g^0 - \frac{1}{\ell_1}\sum_{t=1}^{\ell_1}\theta + \sigma_g \left(\sqrt{\frac{2\log(1/\delta)}{\ell_0}} + \sqrt{\frac{2\log(1/\delta)}{\ell_1}} \right) + G^{(u_t)} + \frac{1}{2}, \mathcal{C} \right] \\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}\left[\theta \middle| \theta > -\left(\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[g^0] - \sigma_g\sqrt{\frac{2\log(1/\delta_1)}{\ell_0}} \right) + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[g^1] - \sigma_g\sqrt{\frac{2\log(1/\delta_2)}{\ell_1}} \right. \\
&\qquad\qquad \left. + \sigma_g \left(\sqrt{\frac{2\log(1/\delta)}{\ell_0}} + \sqrt{\frac{2\log(1/\delta)}{\ell_1}} \right) + G^{(u_t)} + \frac{1}{2}, \mathcal{C} \right] \qquad \text{(by event } \mathcal{C}) \\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}\left[\theta \middle| \theta > -\sigma_g \left(\sqrt{\frac{2\log(2/\delta)}{\ell_0}} - \sqrt{\frac{2\log(2/\delta)}{\ell_1}} \right) + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[g^1 - g^0] + \sigma_g \left(\sqrt{\frac{2\log(2/\delta)}{\ell_0}} + \sqrt{\frac{2\log(2/\delta)}{\ell_1}} \right) + G^{(u_t)} + \frac{1}{2}, \mathcal{C} \right] \\
&\qquad\qquad \text{(by Theorem A.7, where } \delta_1 = \delta_2 = \delta/2) \\
&> \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}\left[\theta \middle| \theta > \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[g^1 - g^0] - \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[g^1 - g^0] + \frac{1}{2} \right] \qquad \text{(by definition of } G^{(u_t)}) \\
&> \frac{1}{2} \qquad (32)
\end{aligned}$$

Hence, the term $\mathbb{E}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}]\,\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]$ satisfies the following lower bound:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] &\geq \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|\xi^{(u_t)},\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - 2\delta && \text{(by Equation (31))}\\
&> \frac{1}{2} \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - 2\delta && \text{(by Equation (32))}\\
&= \frac{\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]}{4} + \frac{\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]}{4} - 2\delta\\
&> \frac{\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]}{4} && \text{(since } \delta < \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}]/8)
\end{aligned}$$

Substituting this into Equation (27), we arrive at a lower bound to set the exploration probability ρ for the agent any round t with type u_t to comply with recommendation policy π_t (extension of Algorithm 1):

$$\rho \leq 1 + \frac{4\mu^{(u_t)}}{\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\xi^{(u_t)}] - 4\mu^{(u_t)}}$$

□

Theorem 3.3 (Treatment Effect Confidence Interval after Algorithm 1)**.** *With sample set $S_\ell = (x_i, y_i, z_i)_{i=1}^{\ell}$ of ℓ samples collected from the second stage of Algorithm 1 —run with exploration probability ρ small enough so that type 0 agents are compliant (see Lemma 3.2),— approximation bound $A(S_\ell, \delta)$ satisfies the following, with probability at least $1-\delta$:*

$$A(S_\ell, \delta) \leq \frac{2\sigma_g\sqrt{2\log(5/\delta)}}{\rho(1-\rho)p_0\sqrt{\ell} - (3-\rho)\sqrt{\frac{\rho\log(5/\delta)}{2(1-\rho)}}}$$

for any $\delta \in (0,1)$. Recall σ_g is the variance of $g^{(u_i)}$, p_0 is the fraction of compliant never-takers in the population of agents,[18] and $A(S_\ell, \delta)$ is defined as in Theorem 2.1.

Proof. First, Theorem 2.1 demonstrates, for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$ that the approximation bound

$$|\theta - \hat{\theta}_{S_\ell}| \leq A(S_\ell, \delta) = \frac{2\sigma_g\sqrt{2\ell\log(2/\delta_1)}}{\left|\sum_{i=1}^{\ell}(x_i - \bar{x})(z_i - \bar{z})\right|}. \tag{33}$$

Next, recall that the mean recommendation $\bar{z} = \rho$ for exploration probability ρ in the second stage of Algorithm 1. We assume Algorithm 1 to be initialized with parameters (see Lemma 3.2 for details) such that its recommendations are compliant for agents of type 0. In the worst case, only type 0 agents are compliant. Therefore, Theorem B.2 implies that, for any $\delta_2 \in (0,1)$, with probability at least $1-\delta_2$ that

$$\left|\sum_{i=1}^{\ell}(x_i - \bar{x})(z_i - \bar{z})\right| \geq \rho\ell\left(p_0(1-\rho) - \sqrt{\frac{(1-\rho)\log(1/\delta_2)}{2\ell}}\right). \tag{34}$$

With a union bound over Equations (33) and (34) while letting $\delta_1 = \delta_2 = \frac{\delta}{3}$ for any $\delta \in (0,1)$, we conclude: with probability at least $1-\delta$,

$$A(S_\ell, \delta) \leq \frac{2\sigma_g\sqrt{2\ell\log(3/\delta)}}{\rho\ell\left(p_0(1-\rho) - \sqrt{\frac{(1-\rho)\log(3/\delta)}{2\ell}}\right)} = \frac{2\sigma_g\sqrt{2\log(3/\delta)}}{\rho\left(p_0(1-\rho)\sqrt{\ell} - \sqrt{\frac{(1-\rho)\log(3/\delta)}{2}}\right)}$$

□

D. Missing Proofs for Section 4

D.1. Algorithm 2 Proofs

Lemma 4.2 (Algorithm 2 Partial Compliance)**.** *Recall that Algorithm 2 is initialized with input samples $S_0 = (x_i, y_i, z_i)_{i=1}^{|S_0|}$. For any type u with the following prior preference (control or treatment), if S_0 satisfies the following condition, with probability at least $1-\delta$, then all agents of type u will comply with recommendations of Algorithm 2:*

$$A(S_0, \delta) \leq \begin{cases} \tau\, \mathbb{P}_{\mathcal{P}^{(u)}}[\theta > \tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0; \\ \tau\, \mathbb{P}_{\mathcal{P}^{(u)}}[\theta < -\tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}^{(u)}}[\theta] \geq 0, \end{cases}$$

for some $\tau \in (0,1)$, where $A(S_0, \delta)$ is the approximation bound for S_0 and any $\delta \in (0,1)$ (see Theorem 2.1).

Proof. Just as in the proof for Lemma C.3, let $a^{(u)}$ and $b^{(u)}$ denote the prior preferred and unpreferred actions for agents of any type u, i.e. $a^{(u)} := \mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] \geq 0]$ and $b^{(u)} := \mathbb{1}[\mathbb{E}_{\mathcal{P}^{(u)}}[\theta] < 0]$. Let π denote the recommendation policy defined by Algorithm 2. At any round t of Algorithm 2, recommendation policy π_t has a positive probability of recommending either control or treatment, according to the prior $\mathcal{P}^{(u_t)}$ for type u_t, i.e. $\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = 0] > 0$ and $\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = 1] > 0$. Thus, by Claim C.1, the following holds:

$$\left\{ (-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}] \geq 0 \right\} \Rightarrow \left\{ (-1)^{b^{(u_t)}} \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta | z_t = a^{(u_t)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = a^{(u_t)}] < 0 \right\}$$

and it suffices to prove the premise $(-1)^{a^{(u_t)}} \mathbb{E}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}]\, \mathbb{P}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}] \geq 0$ in order to prove that agent t of type u_t complies with recommendation z_t.

Recall that the sample set S_q^{BEST} is made up of the best samples up until phase q of Algorithm 2, i.e. the samples which produce the smallest approximation bound A_q. The treatment effect estimate derived from set S_q^{BEST} is denoted $\hat{\theta}_q$. We define the event $\mathcal{C}$ as the event that the treatment effect estimate $\hat{\theta}_q$ satisfies the approximation bound A_q at every phase q throughout Algorithm 2:

$$\mathcal{C} := \left\{ \forall q \geq 0 : |\theta - \hat{\theta}_q| < A_q \right\}. \tag{35}$$

By Theorem 2.1, for event $\mathcal{C}$, the failure probability $\mathbb{P}[\neg\mathcal{C}] \leq \delta$. Furthermore, we assume here that

$$\delta \leq \begin{cases} \frac{\tau\, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta \geq \tau]}{2(\tau\, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta \geq \tau] + 1)} & \text{if } \mu^{(u_t)} < 0; \\ \frac{\tau\, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta < -\tau]}{2(\tau\, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta < -\tau] + 1)} & \text{if } \mu^{(u_t)} \geq 0. \end{cases}$$

Therefore, since $|\theta| \leq 1$, we have:

$$\begin{aligned}
&(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}] \\
&= (-1)^{a^{(u_t)}} \left(\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}, \mathcal{C}] + \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}, \neg\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}, \neg\mathcal{C}] \right) \\
&\geq (-1)^{a^{(u_t)}} \left(\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}, \mathcal{C}] - (-1)^{a^{(u_t)}} \delta \right) \\
&\geq (-1)^{a^{(u_t)}} \left(\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}, \mathcal{C}] \right) - \frac{\tau\, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta \geq \tau]}{2\tau\, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta \geq \tau] + 2}
\end{aligned}$$

In order to lower bound the last line above, we marginalize $\mathbb{E}_{\mathcal{P}^{(u_t)}, \pi_t}[\theta | z_t = b^{(u_t)}, \mathcal{C}]\, \mathbb{P}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = b^{(u_t)}, \mathcal{C}]$ based on four

possible ranges which θ lies on:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[\theta|z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = b^{(u_t)}, \mathcal{C}] \\
&= \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] \\
&+ \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}, \mathcal{C}, 0 \leq (-1)^{a^{(u_t)}}\theta < \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, 0 \leq (-1)^{a^{(u_t)}}\theta < \tau] \\
&+ \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}, \mathcal{C}, -2A_q < (-1)^{a^{(u_t)}}\theta < 0] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, -2A_q < (-1)^{a^{(u_t)}}\theta < 0] \\
&+ \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta|z_t = b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta \leq -2A_q] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta \leq -2A_q]
\end{aligned} \tag{36}
$$

Because A_q is the smallest approximation bound derived from samples collected over any phase q of Algorithm 2 (including the initial sample set S_0), the following holds:

$$
\begin{aligned}
2A_q &\leq 2A(S_0, \delta) \\
&\leq \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2} && \text{(by assumption } A(S_0,\delta) \leq \tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]/4) \\
&\leq \tau
\end{aligned}
$$

Conditional on $\mathcal{C}$, $|\theta - \hat{\theta}_q| < A_q$. Thus, if $(-1)^{a^{(u_t)}}\theta \geq \tau \geq 2A_q$, then $(-1)^{a^{(u_t)}}\hat{\theta}_q \geq \tau - A_q \geq A_q$, which invokes the stopping criterion for the while loop in Algorithm 2. Thus, type u_t's preferred action $a^{(u_t)}$ must have been eliminated from the race before phase $q = 1$ and the unpreferred action $b^{(u_t)}$ is recommended almost surely throughout Algorithm 2, i.e. $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] = \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau]$. Similarly, if $(-1)^{a^{(u_t)}}\theta \leq -2A_q$, then $(-1)^{a^{(u_t)}}\hat{\theta}_q \leq -A_q$ by phase $q = 1$ and the unpreferred action $b^{(u_t)}$ is recommended almost never, i.e. $\mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t =$

$b^{(u_t)}, \mathcal{C}, (-1)^{a^{(u_t)}}\theta < -2A_q] = 0$. Substituting in these probabilities, we proceed:

$$
\begin{aligned}
&(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[\theta | z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = b^{(u_t)}, \mathcal{C}] \\
&= (-1)^{a^{(u_t)}} \Bigg(\mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta | \mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] \\
&\quad + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}, \mathcal{C}, 0 \leq (-1)^{a^{(u_t)}}\theta < \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, 0 \leq (-1)^{a^{(u_t)}}\theta < \tau] \\
&\quad + \mathop{\mathbb{E}}_{\pi_t,\mathcal{P}^{(u_t)}}[\theta | z_t = b^{(u_t)}, \mathcal{C}, -2A_q < (-1)^{a^{(u_t)}}\theta < 0] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, -2A_q < (-1)^{a^{(u_t)}}\theta < 0] \Bigg) \\
&\geq (-1)^{a^{(u_t)}} \Bigg((-1)^{a^{(u_t)}} \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] + 0 \cdot \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, 0 \leq (-1)^{a^{(u_t)}}\theta < \tau] \\
&\qquad\qquad - (-1)^{a^{(u_t)}} 2A_q \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[z_t = b^{(u_t)}, \mathcal{C}, -2A_q < (-1)^{a^{(u_t)}}\theta < 0] \Bigg) \\
&\geq (-1)^{a^{(u_t)}} \left((-1)^{a^{(u_t)}} \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] - (-1)^{a^{(u_t)}} 2A_q \right) \\
&\geq \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C}, (-1)^{a^{(u_t)}}\theta \geq \tau] - \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2} \\
&\geq \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[\mathcal{C} | (-1)^{a^{(u_t)}}\theta \geq \tau] \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] - \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2} \\
&\geq (1-\delta)\tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] - \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2} \\
&= \left(\frac{1}{2} - \delta \right) \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] \\
&\geq \left(\frac{1}{2} - \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] + 2} \right) \tau \mathop{\mathbb{P}}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] \\
&= \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] + 2}
\end{aligned}
$$

Putting everything together, we get that

$$
\begin{aligned}
&(-1)^{a^{(u_t)}} \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[\theta | z_t = b^{(u_t)}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = b^{(u_t)}] \\
&\geq (-1)^{a^{(u_t)}} \left(\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[\theta | z_t = b^{(u_t)}, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = b^{(u_t)}, \mathcal{C}] + \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)},\pi_t}[\theta | z_t = b^{(u_t)}, \neg\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)},\pi_t}[z_t = b^{(u_t)}, \neg\mathcal{C}] \right) \\
&\geq \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] + 2} - \frac{\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau] + 2} \\
&= 0
\end{aligned}
$$

Therefore, so long as $A(S_0, \delta) \leq \tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]/4$ and $\delta < \tau \, \mathbb{P}_{\pi_t,\mathcal{P}^{(u_t)}}[(-1)^{a^{(u_t)}}\theta \geq \tau]/2$, any agent of type u_t will comply with recommendations from Algorithm 2. □

D.2. Lemma D.1 and Theorem D.2: Full Compliance and Subsequent Estimation Bound

Lemma D.1 (Algorithm 2 Full Compliance)**.** *Suppose that some fraction $p_c > 0$ of agents is compliant from the beginning of Algorithm 2 and assume that $p_c < 1$. Of all types u which were not compliant from the beginning, let type u^* agents be the most resistant to compliance. Suppose that phase q satisfies one of the following bounds (depending on whether type u^* agents prefer control or treatment):*

$$q \geq \begin{cases} \left(\frac{1}{2hp_c} \left(\frac{(32\sigma_g\sqrt{2\log(5/\delta)}}{\tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta > \tau]} + \sqrt{50\log(5/\delta)} \right) \right)^2 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] < 0 \\ \left(\frac{1}{2hp_c} \left(\frac{(32\sigma_g\sqrt{2\log(5/\delta)}}{\tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta < -\tau]} + \sqrt{50\log(5/\delta)} \right) \right)^2 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] \geq 0, \end{cases}$$

for some $\tau \in (0,1)$ and any $\delta \in (0,1)$. Then, with probability at least $1-\delta$, for any phase q greater or equal to the following lower bound all agents will comply with recommendations from Algorithm 2.

Proof. First, recall that the set S_q is made up of the input samples S_0 plus samples collected following Algorithm 2 over all phases up to q. Let $S_{q-0} = (x_i, y_i, z_i)_{i=1}^{2hq}$ denote S_q sans S_0 (i.e. just samples collected following Algorithm 2 up to phase q). Note that for any $\delta \in (0,1)$, the approximation bound $A_q \leq A(S_{q-0}, \delta)$.

We want to prove that type u^* is compliant by and beyond phase q. By Lemma 4.2, it suffices to prove that the approximation bound A_q satisfies the following upper bound with probability at least $1-\delta$:[19]

$$\begin{cases} A_q \leq \tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta > \tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] < 0; \\ A_q \leq \tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta < -\tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] \geq 0, \end{cases}$$

for any $\delta \in (0,1)$ and some $\tau \in (0,1)$.

In order to prove this, recall that each phase q of Algorithm 2 is $2hq$ rounds long and the mean recommendation $\bar{z} = \frac{1}{2}$. By assumption, p_c proportion of agents are compliant. Thus, by Theorem B.2, with probability $1-\delta$ for any $\delta \in (0,1)$, the set S_{q-0} satisfies:

$$A(S_{q-0}, \delta) \leq \frac{8\sigma_g\sqrt{2\log(3/\delta)}}{p_c\sqrt{|S_{q-0}|} - \sqrt{\log(3/\delta)}}$$

By assumption, q satisfies the following lower bound for some $\tau \in (0,1)$:

$$q \geq \begin{cases} \left(\frac{1}{2hp_c} \left(\frac{(32\sigma_g\sqrt{2\log(5/\delta)}}{\tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta > \tau]} + \sqrt{50\log(5/\delta)} \right) \right)^2 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] < 0 \\ \left(\frac{1}{2hp_c} \left(\frac{(32\sigma_g\sqrt{2\log(5/\delta)}}{\tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta < -\tau]} + \sqrt{50\log(5/\delta)} \right) \right)^2 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] \geq 0, \end{cases}$$

Substituting these lower bound values for q in $A(S_{q-0}, \delta)$, we get that the approximation bound A_q satisfies the following inequalities (since $A_q \leq A(S_{q-0}, \delta)$):

$$A_q \leq A(S_{q-0}, \delta) \leq \begin{cases} \tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta > \tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] < 0; \\ \tau\,\mathbb{P}_{\mathcal{P}(u^*)}[\theta < -\tau]/4 & \text{if } \mathbb{E}_{\mathcal{P}(u^*)}[\theta] \geq 0. \end{cases}$$

□

Finally, after Algorithm 2 has become compliant for both types of agents, we achieve the following accuracy guarantee for the final treatment estimate $\hat{\theta}_S$.

[19] Lemma 4.2 doesn't exactly state this: it states that any type u will be compliant if the input samples S_0 satisfy the above bounds. Yet, we can simply imagine that phase q is . Proving compliance starting from any phase $q > 0$ is just the same as proving compliance from phase 0. Intuitively, you can imagine we simply run Algorithm 2 starting at phase q initialized with the samples collected up until phase q.

Theorem D.2 (Treatment Effect Confidence Interval from Algorithm 2 with Full Compliance)**.** *Suppose sample set $S = (x_i, y_i, z_i)_{i=1}^{|S|}$ is collected from Algorithm 2 during $|S|$ rounds when all agents comply. We form estimate $\hat{\theta}_S$ of the treatment effect θ. For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\hat{\theta}_S - \theta\right| \leq 8\sigma_g \sqrt{\frac{2\log(2/\delta)}{|S|}}$$

Proof. We assume Algorithm 2 is initialized and allowed to run long enough such that both types 0 and 1 become compliant at some point. From samples $S = (x_i, y_i, z_i)_{i=1}^{|S|}$ collected during these rounds (from i to $|S|$)), we form an estimate $\hat{\theta}_S$ of the treatment effect θ. By Theorem 2.1, this estimate satisfies the following bound with probability at least $1-\delta$ for any $\delta \in (0,1)$:

$$|\hat{\theta}_S - \theta| \leq A(S,\delta) = \frac{2\sigma_g\sqrt{2|S|\log(2/\delta)}}{\left|\sum_{i=1}^{|S|}(x_i - \bar{x})(z_i - \bar{z})\right|}. \tag{37}$$

Recall that $\bar{z} = \frac{1}{2}$ throughout Algorithm 2. Then, by Theorem B.2, the denominator of the bound in Equation (37) above satisfies the following bound:

$$\left|\sum_{i=1}^{|S|}(x_i - \bar{x})(z_i - \bar{z})\right| \geq \frac{|S|}{4}. \tag{38}$$

Therefore, by Equations (37) and (38), the confidence interval $|\hat{\theta}_S - \theta|$ satisfies the following upper bound:

$$|\hat{\theta}_S - \theta| \leq 8\sigma_g \sqrt{\frac{2\log(2/\delta)}{|S|}}$$

□

D.3. Proof of Lemma 5.2

Lemma 5.2 (Lower bound on ℓ for Type u Compliance in Algorithm 2)**.** *Recall that S_ℓ denotes the samples collected from the second stage of Algorithm 1. Let S_ℓ be the input samples S_0 in Algorithm 2. Assume that p_{c_1} proportion of agents in the population are compliant with recommendations of Algorithm 1 and length ℓ satisfies:*

$$\ell \geq \begin{cases} \left(\frac{\kappa_1}{\tau\,\mathbb{P}_{\mathcal{P}(u)}[\theta > \tau]} + \kappa_2\right)^2 & \text{if } \mathop{\mathbb{E}}_{\mathcal{P}(u)}[\theta] < 0 \\ \left(\frac{\kappa_1}{\tau\,\mathbb{P}_{\mathcal{P}(u)}[\theta < -\tau]} + \kappa_2\right)^2 & \text{if } \mathop{\mathbb{E}}_{\mathcal{P}(u)}[\theta] \geq 0 \end{cases} \tag{6}$$

for some $\tau \in (0,1)$ and where $\kappa_1 := \frac{8\sigma_g\sqrt{2\log(5/\delta)}}{p_{c_1}\rho(1-\rho)}$ and $\kappa_2 := (3-\rho)\sqrt{\frac{\rho\log(5/\delta)}{2(1-\rho)}}$ for any $\delta \in (0,1)$. Then any agent of type u will comply with recommendations of Algorithm 2.

Proof. By assumption, Algorithm 1 is initialized so that agents of type 0 comply and we collect S_ℓ samples from the second stage. Then, for any $\delta \in (0,1)$ and some $\tau \in (0,1)$, approximation bound $A(S_\ell, \delta)$ satisfies:

$$\begin{aligned} A(S_\ell,\delta) &\leq \frac{2\sigma_g\sqrt{2\log(3/\delta)}}{\rho\left(p_0(1-\rho)\sqrt{\ell} - \sqrt{\frac{(1-\rho)\log(3/\delta)}{2}}\right)} && \text{(by Theorem 3.3)} \\ &\leq \frac{2\sigma_g\sqrt{2\log(3/\delta)}}{\rho\left(p_0(1-\rho)\left(\frac{8\sigma_g\sqrt{2\log(3/\delta)}}{p_0\rho(1-\rho)\tau\,\mathbb{P}_{\mathcal{P}(0)}[\theta>\tau]} + \frac{\sqrt{(1-\rho)\log(3/\delta)}}{2p_0(1-\rho)}\right) - \sqrt{\frac{(1-\rho)\log(3/\delta)}{2}}\right)} && \text{(by Equation (6))} \\ &\leq \frac{\tau\,\mathbb{P}_{\mathcal{P}(0)}[\theta > \tau]}{4} \end{aligned}$$

Thus, by Lemma 4.2, if we let the samples S_ℓ collected from the second stage of Algorithm 1 be the input samples S_0 in Algorithm 2, i.e. $S_0 = S_\ell$, then that agents of type 0 will comply with recommendations of Algorithm 2.

□

D.3.1. RACING STAGE FIRST PART ESTIMATION BOUND

Theorem 4.3 (Treatment Effect Confidence Interval from Algorithm 2 with Partial Compliance)**.** *With set $S = (x_i, y_i, z_i)_{i=1}^{|S|}$ of $|S|$ samples collected from Algorithm 2 where p_c is the fraction of compliant agents in the population, we form an estimate $\hat{\theta}_S$ of the treatment effect θ. With probability at least $1-\delta$,*

$$\left|\hat{\theta}_S - \theta\right| \leq \frac{8\sigma_g\sqrt{2\log(5/\delta)}}{p_c\sqrt{|S|} - \sqrt{50\log(5/\delta)}}$$

for any $\delta \in (0,1)$, where σ_g is the variance of $g^{(u_i)}$.

Proof. First, Theorem 2.1 demonstrates, for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$ that the approximation bound

$$|\theta - \hat{\theta}_S| \leq A(S,\delta) = \frac{2\sigma_g\sqrt{2|S|\log(2/\delta_1)}}{\left|\sum_{i=1}^{\ell}(x_i - \bar{x})(z_i - \bar{z})\right|}. \tag{39}$$

Next, recall that the mean recommendation $\bar{z} = \frac{1}{2}$ throughout Algorithm 2. We assume Algorithm 2 to be initialized with parameters such that its recommendations are compliant for agents of type 0. In the worst case, only type 0 agents are compliant. Therefore, Theorem B.2 implies that, for any $\delta_2 \in (0,1)$, with probability at least $1-\delta_2$ that

$$\left|\sum_{i=1}^{|S|}(x_i - \bar{x})(z_i - \bar{z})\right| \geq \frac{|S|}{4}\left(p_0 - \sqrt{\frac{\log(1/\delta_2)}{|S|}}\right). \tag{40}$$

With a union bound over Equations (39) and (40) while letting $\delta_1 = \delta_2 = \frac{\delta}{3}$ for any $\delta \in (0,1)$, we conclude: with probability at least $1-\delta$,

$$A(S,\delta) \leq \frac{8\sigma_g\sqrt{2\log(3/\delta)}}{p_0\sqrt{|S|} - \sqrt{\log(3/\delta)}}$$

□

E. Missing Regret Proofs for Section 5

E.1. Pseudo-regret Proof

Lemma 5.4 (Pseudo-regret)**.** *The pseudo-regret accumulated from policy π_c is bounded for any $\theta \in [-1,1]$ as follows, with probability at least $1-\delta$ for any $\delta \in (0,1)$:*

$$R_\theta(T) \leq L_1 + O(\sqrt{T\log(T/\delta)}) \tag{8}$$

for sufficiently large time horizon T, where the length of Algorithm 1 is $L_1 = \ell + 2\max\left(\frac{\ell_0}{p_0}, \frac{\ell_1}{p_1}\right)$.

Proof. Recall that the clean event $\mathcal{C}$, as defined in the proof of Lemma 4.2, entails that the approximation bound over all rounds. If event $\mathcal{C}$ fails (i.e. $\neg\mathcal{C}$ holds), then the pseudo-regret may only be bounded by the maximum possible value, which is at most $T|\theta|$.

Assume now that $\mathcal{C}$ holds for every round L_2 in Algorithm 2. Then, the absolute value of the treatment $|\theta| \leq |\hat{\theta}_{S_{L_2}}| + A(S_{L_2},\delta)$ where $A(S_{L_2},\delta)$ is the approximation bound based on the samples S_{L_2} collected from L_2 rounds of Algorithm 2. Before the stopping criterion of Algorithm 2 is invoked, we also have $|\hat{\theta}_{S_{L_2}}| \leq A(S_{L_2},\delta)$. Hence, the treatment effect absolute value satisfies the following inequalities:

$$|\theta| \leq 2A(S_{L_2},\delta) \leq \frac{16\sigma_g\sqrt{2\log(5T/\delta)}}{p_{c_2}\sqrt{L_2} - \sqrt{50\log(5T/\delta)}}$$

Assuming that $L_2 \geq \frac{200\log(5T/\delta)}{p_{c_2}^2}$, then $p_{c_2}\sqrt{L_2} - \sqrt{50\log(5T/\delta)} \geq \frac{p_{c_2}\sqrt{L_2}}{2}$ and, to carry on:

$$|\theta| \leq \frac{32\sigma_g\sqrt{2\log(5T/\delta)}}{p_{c_2}\sqrt{L_2}}$$
$$\Rightarrow L_2 \leq \frac{2048\sigma_g^2\log(5T/\delta)}{p_{c_2}^2|\theta|^2}$$

Therefore, a winner is declared —i.e. the treatment effect is definitively either positive or not— by the following round of Algorithm 2:

$$L_2^* = \frac{2048\sigma_g^2\log(5T/\delta)}{p_{c_2}^2|\theta|^2}$$

After this, the winner is recommended for the remainder of the rounds and (because we assume event $\mathcal{C}$ holds) no regret is accumulated for the remaining rounds (until time horizon T).

Note that, by the length the assumption that $L_2 \geq \frac{200\log(5T/\delta)}{p_{c_2}^2}$ holds trivially for L_2^* if $|\theta| \leq \frac{16\sigma_g}{5}$. Otherwise, we simply assume that $L_2 \geq \frac{200\log(5T/\delta)}{p_{c_2}^2}$ holds. We may incorporate this bound into a necessary lower bound on the time horizon T, such that we assume $T \geq L_1 + L_2 \geq L_1 + \frac{200\log(5T/\delta)}{p_{c_2}^2}$.

Now, we demonstrate the amount of regret accumulated during these n^* rounds of Algorithm 2 before a winner is declared. During each phase Algorithm 2, each control and treatment each get recommended $n^*/2$ times.

Furthermore, recall that p_{c_2} fraction of agents are compliant throughout all rounds of Algorithm 2. Without loss of generality, assume that these agents initially prefer control and the rest of the $1 - p_{c_2}$ fraction of agents prefer treatment (and do not comply). Then, if the treatment $\theta < 0$, then in expectation over the randomness of the arrival of agents, the regret is on average $(1 - p_{c_2}/2)|\theta|$. Then, the total accumulated regret throughout Algorithm 2 in policy π_c is given as such:

$$R_2(T) \leq \frac{2048(1 - p_{c_2}/2)\sigma_g^2\log(5T/\delta)}{p_{c_2}^2|\theta|} \tag{41}$$

On the other hand, if treatment $\theta \geq 0$, then on average, the regret for each phase is $p_{c_2}|\theta|/2$, then the total accumulated regret for Algorithm 2 is given:

$$R_2(T) \leq \frac{1024\sigma_g^2\log(5T/\delta)}{p_{c_2}|\theta|} \tag{42}$$

Observe that the pseudo-regret for each round t of the policy π_c over the entire T rounds is at most that of Algorithm 2 plus $|\theta|$ per round of Algorithm 1. Recall that, by policy π_c, there are L_1 total rounds of Algorithm 1. Alternatively, we can also upper bound the total pseudo-regret by $|\theta|$ per each round. Therefore, the total accumulated pseudo-regret for policy π_c once we reach the time horizon T is bounded as follows.

If the treatment effect $\theta \geq 0$, then the total regret of policy π_c satisfies the following bound with probability at least $1 - \delta$, for any $\delta \in (0, 1)$ which satisfies the condition that $L_1 \geq \frac{2\sqrt{\log(5T/\delta)}}{p_{c_2}}$:

$$R(T) \leq \min\left\{L_1\rho|\theta| + \frac{1024\sigma_g^2\log(5T/\delta)}{p_{c_2}|\theta|}, T|\theta|\right\}$$

We can solve for $|\theta|$ in terms of T at the point when $T|\theta|$ is the better regret:

$$\begin{aligned} T|\theta| &= \frac{1024\sigma_g^2 \log(5T/\delta)}{p_{c_2}|\theta|} \\ \Rightarrow |\theta|^2 &= \frac{1024\sigma_g^2 \log(5T/\delta)}{p_{c_2} T} \\ \Rightarrow |\theta| &= 32\sigma_g \sqrt{\frac{\log(5T/\delta)}{p_{c_2} T}} \end{aligned}$$

Substituting this expression for $|\theta|$ back into our expression for the total pseudo-regret, we get the following:

$$\begin{aligned} R(T) &\leq \min\left\{ L_1\rho|\theta| + \frac{1024\sigma_g^2 \log(5T/\delta)}{32\sigma_g p_{c_2}\sqrt{\frac{\log(5T/\delta)}{p_{c_2}T}}}, 32\sigma_g T\sqrt{\frac{\log(5T/\delta)}{p_{c_2}T}} \right\} \\ &= \min\left\{ L_1\rho|\theta| + 32\sigma_g\sqrt{\frac{T\log(5T/\delta)}{p_{c_2}}}, 32\sigma_g\sqrt{\frac{T\log(5T/\delta)}{p_{c_2}}} \right\} \\ &\leq L_1\rho + O(\sqrt{T\log(T/\delta)}) \end{aligned}$$

The $L_1\rho$ regret for Algorithm 1 in the last line above is given because $|\theta| \leq 1$.

Following a similar analysis, if the treatment effect $\theta < 0$, then the total pseudo-regret accumulated following policy π_c satisfies the following bound with probability at least $1-\delta$ for any $\delta \in (0,1)$ which satisfies the condition that $L_1 \geq \frac{2\sqrt{\log(5T/\delta)}}{p_{c_2}}$:

$$R(T) \leq \min\left\{ L_1|\theta| + \frac{2048(1-p_{c_2}/2)\sigma_g^2\log(5T/\delta)}{p_{c_2}^2|\theta|}, T|\theta| \right\} \tag{43}$$

$$\leq L_1 + O(\sqrt{T\log(T/\delta)}) \tag{44}$$

Note that (as stated above) this regret holds only if the time horizon T is sufficiently large such that $T \geq L_1 + \frac{200\log(5T/\delta)}{p_{c_2}^2}$. □

E.2. Regret Proof

Lemma 5.5 (Regret)**.** *Policy π_c achieves regret as follows:*

$$\mathbb{E}[R(T)] = O(\sqrt{T\log(T)}) \tag{9}$$

for sufficiently large time horizon T.

Proof. We can set parameters $\delta, \ell_0, \ell_1, \ell$, and ρ in terms of the time horizon T, in order to both guarantee compliance throughout policy π_c and to obtain sublinear (expected) regret bound relative to T.

First, to guarantee sublinear expected regret, we must guarantee that $\delta = 1/T^2$. To meet our compliance conditions for Algorithm 2, we must set

$$\delta \leq \frac{\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[|\theta| \geq \tau]}{2\left(\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[|\theta| \geq \tau] + 1\right)},$$

for some τ. These may be expressed as conditions on the time horizon T: for any $\delta \in (0,1)$ which satisfies the above compliance conditions, we set T sufficiently large to satisfy the following condition:

$$T \geq \frac{1}{\sqrt{\delta}} \geq \sqrt{\frac{2\left(\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[|\theta| \geq \tau] + 1\right)}{\tau\,\mathbb{P}_{\mathcal{P}^{(u)}}[|\theta| \geq \tau]}} \tag{45}$$

Second, recall that p_0 and p_1 denote the fractions in the population of agents who are never-takers and always-takers, respectively. Furthermore, recall that p_{c_1} and p_{c_2} denote the fractions of agents who comply with Algorithm 1 and Algorithm 2, respectively. Assume that the length of the first stage of Algorithm 1 is non-zero and the exploration probability ρ is set to be small enough in order to guarantee compliance throughout Algorithm 1. The length L_1 of Algorithm 1 must be sufficiently large so that p_c fraction of agents comply in Algorithm 2, as well. However, in order to guarantee sublinear regret, we also need that

$$T \geq L_1^2 = (2\max(\ell_0/p_0, \ell_1/p_1) + \ell)^2 \tag{46}$$

Recall that the clean event $\mathcal{C}$, as defined in the proof of Lemma 4.2, entails that the approximation bound over all rounds. This event $\mathcal{C}$ holds with probability at least $1-\delta$ for any $\delta \in (0,1)$. Conditional on the failure event $\neg\mathcal{C}$, policy π_c accumulates at most linear pseudo-regret in terms of T, i.e. $T|\theta|$. Thus, in expectation it accumulates at most $T|\theta|\delta$ regret.

Then, with the above assumptions on T in mind, the expected regret of policy π_c is:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{P}^{(u)}}[R(T)] &= \mathbb{E}[R(T)|\neg\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u)}}[\neg\mathcal{C}] + \mathbb{E}[R(T)|\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u)}}[\mathcal{C}] \\
&\leq T\delta + \left(L_1 + O\left(\sqrt{T\log(T/\delta)}\right)\right) \\
&= \frac{1}{T} + \left(\sqrt{T} + O\left(\sqrt{T\log(T^3)}\right)\right) \\
&= \frac{1}{T} + O\left(\sqrt{T\log(T)}\right) \\
&= O\left(\sqrt{T\log(T)}\right)
\end{aligned}$$

Therefore, assuming that all hyperparameters $\delta, \ell_0, \ell_1, \ell$, and ρ are set to incentivize compliance of some nonzero proportion of agents throughout π_c and assuming that T is sufficiently large so as to satisfy both Equations (45) and (46) above, policy π_c achieves sublinear regret. □

F. General Setting: Many Arms & Many Types

F.1. Model

We now consider a general setting for the sequential game between a social planner and a sequence of agents over T rounds, as first mentioned in Section 2. In this setting, there are k treatments of interest, each with unknown treatment effect. In each round t, a new agent indexed by t arrives with their private type u_t drawn independently from a distribution $\mathcal{U}$ over the set of all private types U. Each agent t has k actions to choose from, numbered 1 to k. Let $x_t \in \mathbb{R}^k$ be a one-hot encoding of the action choice at round t, i.e. a k-dimensional unit vector in the direction of the action. For example, if the agent at round t chooses action 2, then $x_t = \mathbf{e}_2 = (0, 1, 0, \cdots, 0) \in \mathbb{R}^k$. Additionally, agent t receives an action recommendation $z_t \in \mathbb{R}^k$ from the planner upon arrival. After selecting action $x_t \in \mathbb{R}^k$, agent t receives a reward $y_t \in \mathbb{R}$, given by

$$y_t = \langle \theta, x_t \rangle + g_t^{(u_t)} \tag{47}$$

where $g_t^{(u_t)}$ denotes the confounding *baseline reward* which depends on the agent's private type u_t. Each $g_t^{(u_t)}$ is drawn from a sub-Gaussian distribution with a sub-Gaussian norm of σ_g. The social planner's goal is to estimate the treatment effect vector $\theta \in \mathbb{R}^k$ and maximize the total expected reward of all T agents.

History, beliefs, and action choice. As in the body of the paper, the history H_t is made up of all tuples (z_i, x_i, y_i) over all rounds from $i = 1$ to t. Additionally, before the game starts, the social planner commits to recommendation policy π, which is known to all agents. Each agent also knows the number of the round t when they arrive. Their private type u_t maps to their prior belief $\mathcal{P}^{(u_t)}$, which is a joint distribution over the treatment effect θ and noisy error term $g^{(u)}$. With all this information, the agent t selects the action x_t which they expect to produce the most reward:

$$x_t := \mathbf{e}_{a_t} \qquad \text{where} \qquad a_t := \operatorname*{argmax}_{1 \le j \le k} \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t} \left[\theta^j \mid z_t, t \right]. \tag{48}$$

F.2. Instrumental Variable Estimate and Finite Sample Approximation Bound

As in the body of the paper, we view the planner's recommendations as instruments and perform *instrumental variable (IV) regression* to estimate θ.

IV Estimator for $k > 1$ Treatments Our mechanism periodically solves the following IV regression problem: given a set S of n observations $(x_i, y_i, z_i)_{i=1}^n$, compute an estimate $\hat{\theta}_S$ of θ. We consider the following two-stage least square (2SLS) estimator:

$$\hat{\theta}_S = \left(\sum_{i=1}^n z_i x_i^\intercal \right)^{-1} \sum_{i=1}^n z_i y_i, \tag{49}$$

where $(\cdot)^{-1}$ denotes the pseudoinverse.

To analyze the 2SLS estimator, we introduce a *compliance matrix* of conditional probabilities that an agent chooses some treatment given a recommendation $\hat{\Gamma}$, given as proportions over a set of n samples $S = (x_i, z_i)_{i=1}^n$, where any entry in $\hat{\Gamma}$ is given as such:

$$\hat{\Gamma}_{ab}(S) = \hat{\mathbb{P}}_S[x = \mathbf{e}_a | z = \mathbf{e}_b] = \frac{\sum_{i=1}^n \mathbb{1}[x = \mathbf{e}_a, z = \mathbf{e}_b]}{\sum_{i=1}^n \mathbb{1}[z = \mathbf{e}_b]} \tag{50}$$

Then, we can write the action choice x_i as such:

$$x_i = \hat{\Gamma} z_i + \eta_i, \tag{51}$$

where $\eta_i = x_i - \hat{\Gamma} z_i$. Now, we can rewrite the reward y_i at round i as such:

$$\begin{aligned} y_i &= \left\langle \theta, (\hat{\Gamma} z_i + \eta_i) \right\rangle + g_i^{(u_i)} \\ &= \langle \underbrace{\theta \hat{\Gamma}}_{\beta}, z_i \rangle + \langle \theta, \eta_i \rangle + g_i^{(u_i)} \\ &= \langle \beta, z_i \rangle + \langle \theta, \eta_i \rangle + g_i^{(u_i)}. \end{aligned}$$

This formulation allows us to express and bound the error between the treatment effect θ and its IV estimate $\hat{\theta}$ in Theorem 6.1.

F.3. Proof of Theorem 6.1

Theorem 6.1 (Many Treatments Effect Approximation Bound)**.** *Let $z_1, \ldots, z_n \in \{0,1\}^k$ be a sequence of instruments. Suppose there is a sequence of n agents such that each agent i has private type u_i drawn independently from $\mathcal{U}$, selects x_i under instrument z_i and receives reward y_i. Let sample set $S = (x_i, y_i, z_i)_{i=1}^n$. The approximation bound $A(S,\delta)$ is given as such:*[20]

$$A(S,\delta) = \frac{\sigma_g \sqrt{2nk\log(k/\delta)}}{\sigma_{\min}\left(\sum_{i=1}^n z_i x_i^\intercal\right)},$$

and the IV estimator given by Equation (11) satisfies

$$\left\|\hat{\theta}_S - \theta\right\|_2 \leq A(S,\delta)$$

with probability at least $1-\delta$ for any $\delta \in (0,1)$.

Proof. Given a sample set $S = (x_i, y_i, z_i)_{i=1}^n$ of size n, we form an estimate of the treatment effect $\hat{\theta}_S$ via Two-Stage Least Squares regression (2SLS). In the first stage, we regress y_i onto z_i to get the empirical estimate $\hat{\beta}_S$ and x_i onto z_i to get $\hat{\Gamma}_S$ as such:

$$\hat{\beta}_S := \left(\sum_{i=1}^n z_i z_i^\intercal\right)^{-1}\left(\sum_{i=1}^n z_i y_i\right) \qquad \text{and} \qquad \hat{\Gamma}_S := \left(\sum_{i=1}^n z_i z_i^\intercal\right)^{-1}\left(\sum_{i=1}^n z_i x_i^\intercal\right) \tag{52}$$

Now, note that by definition $\theta = (\hat{\Gamma})^{-1}\beta$. In the second stage, we take the inverse of $\hat{\Gamma}$ times $\hat{\beta}$ as the predicted causal effect vector $\hat{\theta}_S$, i.e.

$$\begin{aligned}
\hat{\theta}_S &= \hat{\Gamma}_S^{-1}\hat{\beta}_S \\
&= \left(\sum_{i=1}^n z_i x_i^\intercal\right)^{-1}\left(\sum_{i=1}^n z_i z_i^\intercal\right)\left(\sum_{i=1}^n z_i z_i^\intercal\right)^{-1}\sum_{i=1}^n z_i y_i \\
&= \left(\sum_{i=1}^n z_i x_i^\intercal\right)^{-1}\sum_{i=1}^n z_i y_i
\end{aligned}$$

Hence, the L2-norm of the difference between θ and $\hat{\theta}_S$ is given as:

$$\begin{aligned}
\left\|\hat{\theta}_S - \theta\right\|_2 &= \left\|\left(\sum_{i=1}^n z_i x_i^\intercal\right)^{-1}\sum_{i=1}^n z_i y_i - \theta\right\|_2 \\
&= \left\|\left(\sum_{i=1}^n z_i x_i^\intercal\right)^{-1}\sum_{i=1}^n z_i\left(\langle\theta, x_i\rangle + g_i^{(u_i)}\right)^\intercal - \theta\right\|_2 \\
&= \left\|\left(\sum_{i=1}^n z_i x_i^\intercal\right)^{-1}\left(\sum_{i=1}^n z_i x_i^\intercal\theta + \sum_{i=1}^n z_i g_i^{(u_i)}\right) - \theta\right\|_2 \\
&= \left\|\theta + \left(\sum_{i=1}^n z_i x_i^\intercal\right)^{-1}\sum_{i=1}^n z_i g_i^{(u_i)} - \theta\right\|_2 \\
&= \left\|\left(\sum_{i=1}^n z_i x_i^\intercal\right)^{-1}\sum_{i=1}^n z_i g_i^{(u_i)}\right\|_2 \\
&\leq \left\|\left(\sum_{i=1}^n z_i x_i^\intercal\right)^{-1}\right\|_2 \left\|\sum_{i=1}^n z_i g_i^{(u_i)}\right\|_2 && \text{(by Lemma A.5)} \\
&= \frac{\left\|\sum_{i=1}^n z_i g_i^{(u_i)}\right\|_2}{\sigma_{\min}\left(\sum_{i=1}^n z_i x_i^\intercal\right)}
\end{aligned}$$

[20] The operator $\sigma_{\min}(\cdot)$ denotes the smallest singular value.

Finally, we may bound $\left\|\hat{\theta}_S - \theta\right\|_2$ by upper bounding $\left\|\sum_{i=1}^n z_i g_i^{(u_i)}\right\|_2$ in the following lemma F.1.

Lemma F.1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left\|\sum_{i=1}^n z_i g_i^{(u_i)}\right\|_2 \leq \sigma_g \sqrt{2nk \log(k/\delta)} \tag{53}$$

Proof. Recall that the baseline reward $g^{(u)}$ is an independently distributed random variable which, by assumption, has a mean of zero, i.e. $\mathbb{E}[g^{(u)}] = 0$. Because of these properties of $g^{(u)}$, with probability at least $1-\delta$ for any $\delta \in (0,1)$, the numerator above satisfies the following upper bound:

$$\begin{aligned}
\left\|\sum_{i=1}^n g_i^{(u_i)} z_i\right\|_2 &= \sqrt{\sum_{j=1}^k \left(\sum_{i=1}^n g_i^{(u_i)} \mathbb{1}[z_i = \mathbf{e}_j]\right)^2} \\
&= \sqrt{\sum_{j=1}^k \left(\sum_{i=1}^{n_j} g_i^{(u_i)}\right)^2} && \text{(where } n_j = \textstyle\sum_{i=1}^n \mathbb{1}[z_i = \mathbf{e}_j]\text{)} \\
&\leq \sqrt{\sum_{j=1}^k \left(\sigma_g \sqrt{2n_j \log(1/\delta_j)}\right)^2} && \text{(by Corollary A.2 and, by assumption, } \mathbb{E}[g^{(u)}] = 0\text{)} \\
&\leq \sqrt{\sum_{j=1}^k \left(\sigma_g \sqrt{2n_j \log(k/\delta)}\right)^2} && \text{(by a Theorem A.7 where } \delta_j = \tfrac{\delta}{k} \text{ for all } j\text{)} \\
&\leq \sqrt{k \left(\sigma_g \sqrt{2n \log(k/\delta)}\right)^2} && \text{(since } n_j \leq n \text{ for all } j\text{)} \\
&= \sigma_g \sqrt{2nk \log(k/\delta)}
\end{aligned}$$

□

This recovers the stated bound and finishes the proof for Theorem 6.1. □

Next, we demonstrate a lower bound which the denominator of the approximation bound $A(S,\delta)$ in Theorem 6.1 equals $O(1/\sqrt{|S|})$, where $|S|$ is the size of sample set S.

Theorem F.2 (Treatment Effect Confidence Interval for General k Treatments)**.** *Let $z_1, \ldots, z_n \in \{0,1\}^k$ be a sequence of instruments. Suppose there is a sequence of n agents such that each agent i has private type u_i drawn independently from $\mathcal{U}$, selects x_i under instrument z_i and receives reward y_i. Assume that each agent initially prefers treatment 1, i.e. $x = \mathbf{e}_1$. Let sample set $S = (x_i, y_i, z_i)_{i=1}^n$. Let r be the proportion of recommendations for each treatment $j > 1$ and let $1-(k-1)r$ be the proportion of recommendations for treatment 1. Let p_c fraction of agents in the population of agents be compliant over the rounds from which S is collected. For any $\delta \in (0,1)$, if $n \geq \frac{rp_c^2}{\log(k/\delta)}$, then the approximation bound $A(S,\delta)$ is given as such:*

$$A(S,\delta) \leq \frac{\sigma_g \sqrt{2k \log(k/\delta)}}{\alpha \sqrt{n}} = O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

and the IV estimator given by Equation (11) satisfies

$$\left\|\hat{\theta}_S - \theta\right\|_2 \leq A(S,\delta)$$

with probability at least $1-\delta$, where $\alpha > 0$ is a constant of proportionality given in Claim F.3 below.

Proof. Note that Theorem 6.1 holds in this case and it suffices to demonstrate that the denominator is bounded by $\frac{1}{\alpha n}$.

Claim F.3 (Proportionality of the Denominator of the Approximation Bound for k Treatments)**.** *Given all assumptions in Theorem F.2 above, the denominator $\sigma_{\min}\left(\sum_{i=1}^{n} z_i x_i^\intercal\right)$ of the approximation bound $A(S,\delta)$ is positive and increases proportionally to n. Formally, $\sigma_{\min}\left(\sum_{i=1}^{n} z_i x_i^\intercal\right) = \Omega(n)$.*

Proof. Recall that we assume that every agent initially prefers treatment 1. Thus, whenever agent is recommended any treatment greater than 1 and does not comply, the agent takes treatment 1. At any round i, if agent i is always compliant, then $x_i = z_i$; if not, then $x_i = \mathbf{e}_1$. (If $z_i = \mathbf{e}_1$, then $x_i = z_i = \mathbf{e}_1$ always.) Furthermore, at any round i when $x_i = z_i$, the outer product $z_i x_i^\intercal = diag(z_i)$, i.e. a diagonal matrix where the diagonal equals z_i. If $x_i = \mathbf{e}_1$, then the outer product

$$z_i x_i^\intercal = \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ z_i & \mathbf{0} & \cdots & \mathbf{0} \\ \downarrow & \downarrow & & \downarrow \end{pmatrix},$$

which is a $k \times k$ matrix where the first column is z_i and all other entries are 0. Thus, as long as we have at least one sample of each treatment, i.e. at least one round i where $x_i = z_i = \mathbf{e}_j$ for all $1 \le j \le k$, then the sum $\sum_{i=1}^{n} z_i x_i^\intercal$ is a lower triangular matrix with all positive entries in the diagonal. To illustrate this, let $\mathbf{A}$ denote the expected mean values of the sum $\sum_{i=1}^{n} z_i x_i^\intercal$, such that

$$\mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\intercal\right] = n \begin{pmatrix} 1-rk & 0 & \cdots & & \cdots & 0 \\ r(1-p_c) & rp_c & 0 & \cdots & \cdots & \vdots \\ r(1-p_c) & 0 & rp_c & 0 & \cdots & \\ \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ r(1-p_c) & 0 & \cdots & \cdots & 0 & rp_c \end{pmatrix} = n\mathbf{A}.$$

Note that

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} z_i x_i^\intercal\right)^\intercal \left(\sum_{i=1}^{n} z_i x_i^\intercal\right)\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} z_i x_i^\intercal\right)\right]^\intercal \mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\intercal\right]$$

$$= n^2 \begin{pmatrix} (1-rk)^2 + (k-1)r^2(1-p_c)^2 & r^2p_c(1-p_c) & \cdots & & \cdots & r^2p_c(1-p_c) \\ r^2p_c(1-p_c) & r^2p_c^2 & 0 & \cdots & \cdots & 0 \\ r^2p_c(1-p_c) & 0 & r^2p_c^2 & 0 & \cdots & \vdots \\ \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ r^2p_c(1-p_c) & 0 & \cdots & \cdots & 0 & r^2p_c^2 \end{pmatrix}.$$

Furthermore, let $\hat{\mathbf{A}}$ denote the empirical approximation of $\mathbf{A}$ over our n samples, given as such:

$$\sum_{i=1}^{n} z_i x_i^\intercal = n \begin{pmatrix} 1-rk & 0 & \cdots & & \cdots & 0 \\ r(1-\hat{p}_{c,2}) & r\hat{p}_{c,2} & 0 & \cdots & \cdots & \vdots \\ r(1-\hat{p}_{c,3}) & 0 & r\hat{p}_{c,3} & 0 & \cdots & \vdots \\ \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ r(1-\hat{p}_{c,k}) & 0 & \cdots & \cdots & 0 & r\hat{p}_{c,k} \end{pmatrix} = n\hat{\mathbf{A}},$$

where for any $j \ge 2$, the empirical proportion of agents who comply with the recommended treatment j is denoted as $\hat{p}_{c,j}$. Note that, since $\mathbb{E}[\hat{p}_{c,j}] = p_c$ for all $j \ge 2$, the expected value $\mathbb{E}\left[\hat{\mathbf{A}}\right] = \mathbf{A}$. We may bound the difference between $\hat{p}_{c,j}$ and p_c with high probability, based on the number of times each treatment j is recommended, which is rn. Over n samples, with probability at least $1 - \delta_j$ for any $\delta_j \in (0,1)$ for any treatment j, the proportion $\hat{p}_{c,j}$ satisfies the following:

$\hat{p}_{c,j} \geq p_c - \sqrt{\frac{\log(1/\delta_j)}{2rn}}$. In order for this bound to hold for all $j \geq 2$, let $\delta_2 = \delta_3 = \cdots = \delta_k = \delta/k$. Then, by a union bound, with probability $1-\delta$ for any $\delta \in (0,1)$, the bound $\hat{p}_{c,j} \geq p_c - \sqrt{\frac{\log(k/\delta)}{2rn}}$ holds simultaneously for all $2 \leq j \leq k$. Thus, for any $\delta \in (0,1)$ and $n \geq \frac{rp_c^2}{\log(k/\delta)}$, each entry in the diagonal of $\hat{\mathbf{A}}$ is positive. Thus, (since it is a triangular matrix) the eigenvalues of $\hat{\mathbf{A}}$ equal the entries in the diagonal and are all positive. Furthermore, because $\text{rank}(\hat{\mathbf{A}}) = \text{rank}(\hat{\mathbf{A}}^\intercal \hat{\mathbf{A}})$, the singular values of $\hat{\mathbf{A}}$ are all positive, as well.

Thus, for $n \geq \frac{rp_c^2}{\log(k/\delta)}$, the minimum singular value

$$\sigma_{\min}\left(\sum_{i=1}^{n} z_i x_i^\intercal\right) = n\sigma_{\min}\{\hat{\mathbf{A}}\} = n\alpha = \Omega(n),$$

where $\alpha = \sigma_{\min}(\hat{\mathbf{A}}) > 0$ is some (possibly small) constant of proportionality. □

Thus, by Claim F.3 and Theorem 6.1, the approximation bound

$$A(S,\delta) \leq \frac{\sigma_g\sqrt{2nk\log(k/\delta)}}{n\alpha} = O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

□

Corollary F.4 (Treatment Effect Confidence Interval for General k Treatments)**.** *Given all assumptions in Theorem F.2, plus the assumptions that the minimum compliance rate for any arm is at least $1/k$ and the minimum proportion of treatment 1 recommendations is at least $1/k$, for any $\delta \in (0,1)$, with a large enough sample size n, the approximation bound $A(S,\delta)$ is given as such:*

$$A(S,\delta) = O\left(k\sqrt{\frac{k\log(1/\delta)}{n}}\right) \tag{54}$$

Proof. Note that Claim F.3 holds in this case and it suffices to demonstrate that the α is bounded by $\frac{1}{k}$. We focus on the denominator $\sigma_{\min}\left(\sum_{i=1}^{n} z_i x_i^\intercal\right)$ of the approximation bound $A(S,\delta)$. Note that since z_i and x_i are one-hot encoded vectors, we have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\intercal\right] &= \sum_{j=1}^{k}\sum_{i\in S_j} x_i\left(\sum_{i\in S_j} x_i\right)^\intercal && \text{(where } S_j = \{i : z_i = \mathbf{e}_j\}\text{)} \\
&= n\sum_{j=1}^{k} v_j v_j^\intercal && \text{(where vector } v_j = (v_{j1}, 0, \cdots 0, v_{jj}, 0, \cdots 0) \in \mathbb{R}^k\text{)}
\end{aligned}$$

where $\forall j : v_{j1} = \frac{r}{k}(1-p_j)$ is the probability of getting a treatment 1 sample when the recommendation is $j > 1$, the term $v_{11} = 1 - r$ is the probability of recommending treatment 1 (since we assume agents always comply with treatment 1 recommendations), and the term $v_{jj} = \frac{r}{k}p_j$ is the probability of getting a treatment j sample when the recommendation is $j > 1$. By definition, we can write the denominator term squared as:

$$\begin{aligned}
\sigma_{\min}\left(\mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\intercal\right]\right)^2 &= \min_{a:\|a\|=1} a^\intercal\left(n\sum_{j=1}^{k} v_j v_j^\intercal\right)a \\
&= n\left[\min_{a:\|a\|=1}(a_1 v_{11})^2 + (a_1 v_{21} + a_2 v_{22})^2 + \cdots + (a_1 v_{k1} + a_k v_{kk})^2\right].
\end{aligned}$$

Also, without loss of generality, assume that $a_1 > 0$ and $\forall j > 1 : a_j \leq 0$.
Substituting the expression above with algorithm-specific variables, we have:

$$\begin{aligned}
&(a_1 v_{11})^2 + (a_1 v_{21} + a_2 v_{22})^2 + \cdots + (a_1 v_{k1} + a_k v_{kk})^2 \\
&= a_1^2(1-r)^2 + \frac{r^2}{k^2} a_1^2 \sum_{j=2}^{k} (1-p_j)^2 + \sum_{j=2}^{k} a_j^2 p_j^2 + \frac{2r^2}{k^2} a_1 \sum_{j=2}^{k} a_j p_j (1-p_j) \\
&\geq a_1^2(1-r)^2 + \frac{r^2}{k^2} a_1^2 \sum_{j=2}^{k} (1-p_j)^2 + p_{\min}^2 \sum_{j=2}^{k} a_j^2 + \frac{2r^2}{k^2} a_1 \sum_{j=2}^{k} a_j \frac{1}{4} \\
&\geq a_1^2(1-r)^2 + \frac{r^2}{k^2} a_1^2 \sum_{j=2}^{k} (1-p_j)^2 + p_{\min}^2 (1-a_1^2) - \frac{r^2 a_1}{2k^2} \sqrt{(k-1)(1-a_1^2)}
\end{aligned}$$

where the second line is direct substitution, the third line comes from lower bounding all p_j^2 terms with the minimum compliance rate $p_{\min}^2$ and lower bounding $p_j(1-p_j)$ by $1/4$. The last line comes from the fact that $\|a\| = 1$ and from applying Lemma A.5 on the last term. Since we assume that the probability of recommending treatment 1 is $1 - r \geq \frac{1}{k}$, we have:

$$\begin{aligned}
&\sigma_{\min}\left(\mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\intercal\right]\right)^2 \\
&\geq n \left[\min_{a:\|a\|=1} \frac{a^2}{k^2} + \frac{\left(1-\frac{1}{k}\right)^2}{k^2} a_1^2 \sum_{j=2}^{k} (1-p_j)^2 + p_{\min}^2 (1-a_1)^2 - \frac{\left(1-\frac{1}{k}\right)^2}{2k^2} a_1 \sqrt{(k-1)(1-a_1^2)}\right] \\
&\geq n \left[\min_{a:\|a\|=1} \frac{a_1^2}{k^2} + p_{\min}^2 (1-a_1)^2 - \frac{(k-1)^2}{2k^4} a_1 \sqrt{(k-1)(1-a_1^2)}\right] \\
&\geq n \left[\min_{a:\|a\|=1} \frac{a_1^2}{k^2} + p_{\min}^2 (1-a_1)^2 - \frac{a_1 \sqrt{(k-1)(1-a_1^2)}}{2k^2}\right] \\
&= n \left[\left(a_1 \sqrt{\frac{1}{k^2} - p_{\min}^2} - \frac{1}{4k^2} \sqrt{\frac{(k-1)(1-a_1^2)}{\frac{1}{k^2} - p_{\min}^2}}\right)^2 + p_{\min}^2 - \frac{(k-1)(1-a_1^2)}{16k^4\left(\frac{1}{k^2} - p_{\min}^2\right)}\right] \\
&\geq n \left[p_{\min}^2 - \frac{(k-1)(1-a_1^2)}{16k^4\left(\frac{1}{k^2} - p_{\min}^2\right)}\right] \\
&\geq n \left[p_{\min}^2 - \frac{(1-a_1^2)}{16k - 16k^3 p_{\min}^2}\right]
\end{aligned}$$

Since we assume that the minimum compliance rate $p_{\min} \geq \frac{1}{k}$, we have:

$$p_{\min}^2 \geq \frac{1}{k^2} \Rightarrow 16k - 16k^3 p_{\min}^2 \leq 0$$

Therefore, we have $\alpha = \frac{1}{k}$ and

$$\sigma_{\min}\left(\mathbb{E}\left[\sum_{i=1}^{n} z_i x_i^\intercal\right]\right)^2 \geq \frac{n}{k^2}$$

We apply Theorem A.6 to this matrix to get that, with probability at least $1 - \delta$, for $\delta \in (0, 1)$:

$$\sigma_{\min}\left(\sum_{i=1}^{n} z_i x_i^\intercal\right) \geq \sqrt{\frac{n}{k^2} - \log(k/\delta)}$$

Hence, we have the approximation bound $A(S,\delta)$ for Algorithm 4 is given as

$$A(S,\delta) \leq \frac{\sigma_g\sqrt{2nk\log(k/\delta)}}{\sqrt{\frac{n}{k^2}-\log(k/\delta)}\sqrt{n}} = O\left(k\sqrt{\frac{k\log(1/\delta)}{n}}\right)$$

F.4. Extensions of Algorithms 1 and 2 and Recommendation Policy π_c to k Treatments

We assume that every agent —regardless of type— shares the same prior ordering of the treatments, such that all agents prior expected value for treatment 1 is greater than their prior expected value for treatment 2 and so on. First, Algorithm 3 is a generalization of Algorithm 1 which serves the same purpose: to overcome complete non-compliance and incentivize some agents to comply eventually. The incentivization mechanism works the same as in Algorithm 1, where we begin by allowing all agents to choose their preferred treatment —treatment 1— for the first ℓ rounds. Based on the ℓ samples collected from the first stage, we then define a number of events $\xi_j^{(u)}$ —which are similar to event ξ from Algorithm 1— that each treatment $j \geq 2$ has the largest expected reward of any treatment and treatment 1 has the smallest, according to the prior of type u:

$$\xi_i^{(u)} := \left(\bar{y}_\ell^1 + C \leq \min_{1<j<i} \bar{y}_\ell^j - C \text{ and } \max_{1<j<i} \bar{y}_\ell^j + C \leq \mu_i^{(u)}\right), \tag{55}$$

where $C = \sigma_g\sqrt{\frac{2\log(3/\delta)}{\ell}} + \frac{1}{4}$ for any $\delta \in (0,1)$ and where $\bar{y}_\ell^1$ denotes the mean reward for treatment 1 over the ℓ samples of the first stage of Algorithm 3. Thus, if we set the exploration probability ρ small enough, then some subset of agents will comply with all recommendations in the second stage of Algorithm 3.

Algorithm 3 Overcoming complete non-compliance for k treatments

Input: exploration probability $\rho \in (0,1)$, minimum number of samples of any treatment $\ell \in \mathbb{N}$ (assume w.l.o.g. $(\ell/\rho) \in \mathbb{N}$), failure probability $\delta \in (0,1)$, compliant type u
1st stage: The first ℓ agents are given no recommendation (they choose treatment 1)
for each treatment $i > 1$ in increasing lexicographic order **do**
  if $\xi_i^{(u)}$ holds, based on the ℓ samples from the first phase and any samples of treatment $2 \leq j < i$ collected thus far **then**
    $a_i^* = i$
  else
    $a_i^* = 1$
  end if
  From the next ℓ/ρ agents, pick ℓ agents uniformly at random to be in the explore set E[21]
  for the next ℓ rounds **do**
    if agent t is in explore set E **then**
      $z_t = 1$
    else
      $z_t = a^*$
    end if
  end for
end for

Second, Algorithm 4 is a generalization of Algorithm 2, which is required to start with at least partial compliance and more rapidly and incentivizes more agents to comply eventually. The incentivization mechanism works the same as in Algorithm 1, where we begin by allowing all agents to choose their preferred treatment —treatment 1— for the first ℓ rounds. Based on the ℓ samples collected from the first stage, we then define a number of events —which are similar to event ξ from Algorithm 1— that each treatment $j \geq 2$ has the largest expected reward of any treatment and treatment 1 has the smallest. Thus, if we set the exploration probability ρ small enough, then some subset of agents will comply with all recommendations in the second stage of Algorithm 3.

Definition F.5 (General recommendation policy π_c for k treatments)**.** Recommendation policy π_c over T rounds is given as such:

[21] We set the length of each phase i of the second stage to be ℓ/ρ so that we get ℓ samples of each treatment i and the exploration probability is ρ.

Algorithm 4 Overcoming partial compliance for k treatments

Input: samples $S_0 := (x_i, z_i, y_i)_{i=1}^{|S_0|}$ which meet Theorem 6.1 conditions and produce IV estimate $\hat{\theta}_{S_0}$, time horizon T, number of recommendations of each action per phase h, failure probability $\delta \in (0,1)$
Split the remaining rounds (up to T) into consecutive phases of h rounds each, starting with $q = 1$;
Let $\hat{\theta}_0 = \hat{\theta}_{S_0}$ and $A_0 = A(S_0, \delta)$
Initialize set of active treatments: $B = \{\text{all treatments}\}$.
while $|B| > 1$ **do**
  Let $\hat{\theta}^*_{q-1} = \max_{i \in B} \hat{\theta}^i_{q-1}$ be the largest entry i in $\hat{\theta}_{q-1}$
  Recompute $B = \left\{\text{treatments } i : \hat{\theta}^*_{q-1} - \hat{\theta}^i_{q-1} \leq A_{q-1}\right\}$;
  The next $|B|$ agents are recommended each treatment $i \in B$ sequentially in lexicographic order;
  Let S_q^{BEST} be the sample set with the smallest approximation bound so far, i.e. $S_q^{\text{BEST}} = \operatorname{argmin}_{S_r, 0 \leq r \leq q} A(S_r, \delta)$;
  Define $\hat{\theta}_q = \hat{\theta}_{S_q^{\text{BEST}}}$ and $A_q = A(S_q^{\text{BEST}}, \delta)$;
  $q = q + 1$
end while
For all remaining agents, recommend a^* that remains in B.

1) Run Algorithm 3 with exploration probability ρ set to incentivize at least $p_{c_1} > 0$ fraction of agents of the population to comply in Algorithm 3. Let S_0 be the sample set given from Algorithm 3. By Corollary F.8 and Theorem F.9, we can use S_0 as the initial samples in Algorithm 4 to incentivize compliance for any arm a if the approximation bound $A(S_0, \delta)$ given by S_0 is small enough (see Theorem F.9). Thus, we run Algorithm 3 long enough (i.e. we set ℓ large enough) so that the approximation bound is small enough and at least $p_{c_2} > 1/k$ fraction of agents comply with recommendations of every treatment in Algorithm 4.

2) Initialize Algorithm 4 with samples S_0 from Algorithm 3. At least $p_{c_2} > 1/k$ fraction of agents comply with recommendations from Algorithm 4 from the beginning and until time horizon T.

Lemma F.6 (Algorithm 3 compliance)**.** *Let event $\xi^{(u)}$ be defined such that $\mathbb{P}[\xi^{(u)}] = \min_i \mathbb{P}[\xi_i^{(u)}]$. In Algorithm 3, any type u agent who arrives in the last ℓ/ρ rounds of Algorithm 3 is compliant with any recommendation if $\mu_i^{(u)} > 0$ for all $1 < i \leq k$, and the exploration probability ρ satisfies:*

$$\rho \leq 1 + \frac{8\left(\mu_j^{(u)} - \mu_i^{(u)}\right)}{\mathbb{P}_{\pi_c, \mathcal{P}^{(u)}}[\xi^{(u)}]} \tag{56}$$

Proof. Let the recommendation policy π here be Algorithm 3.

Part I (Compliance with recommendation for treatment $i > 1$): We first argue that an agent t of type u who is recommended treatment i will not switch to any other treatment j. For treatments $j > i$, there is no information about treatments i or j collected by the algorithm and by assumption, we have $\mu_i^{(u)} \geq \mu_j^{(u)}$. Hence, it suffices to consider when $j < i$. We want to show that

$$\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u)}}[\theta^j - \theta^1 | z_t = \mathbf{e}_i] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u)}}[z_t = \mathbf{e}_i] \geq 0. \tag{57}$$

Part II (Recommendation for treatment 1): When agent t is recommended treatment 1, they know that they are not in the explore group E. Therefore, they know that the event $\neg\xi_i^{(u)}$ occurred. Thus, in order to prove that Algorithm 3 is BIC for an agent of type u, we need to show the following for any treatment $j > 1$:

$$\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u)}}[\theta^1 - \theta^j | z_t = \mathbf{e}_1] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u)}}[z_t = \mathbf{e}_1] = \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u)}}[\theta^1 - \theta^j | \neg\xi_i^{(u)}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u)}}[\neg\xi_i^{(u)}] \geq 0 \tag{58}$$

We omit the remainder of this proof due to its similarity with the proof of Lemma 3.2

□

Lemma F.7 (Algorithm 4 Partial Compliance)**.** *Recall that Algorithm 4 is initialized with input samples $S_0 = (x_i, y_i, z_i)_{i=1}^{|S_0|}$. For any type u, if S_0 satisfies the following condition, then with probability at least $1-\delta$ all agents of type u will comply with recommendations of Algorithm 4:*

$$A(S_0, \delta) \leq \tau \mathop{\mathbb{P}}_{\mathcal{P}^{(u)}}[\min_{a,b} \left(|\theta^a - \theta^b|\right) > \tau]/4$$

for some $\tau \in (0,1)$, where $A(S_0, \delta)$ is the approximation bound for S_0 and any $\delta \in (0,1)$ (see Theorem 6.1).

Proof. Let the recommendation policy π here be Algorithm 4. We want to show that for any agent at time t with a type $i < u$ in the racing stage and for any two treatments $a, b \in B$:

$$\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u)}}[\theta^a - \theta^b | z_t = \mathbf{e}_a] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u)}}[z_t = \mathbf{e}_a] \geq 0$$

We omit the remainder of this proof due to its similarity with Lemma 4.2.

□

□

Corollary F.8. *(Pairwise Treatment Effect Confidence Interval for General k Treatments) Given all assumptions in Corollary F.4, the pairwise approximation bound between any two particular arms a, b is given as*

$$|(\theta^a - \theta^b) - (\hat{\theta}^a - \hat{\theta}^b)| = A^{ab}(S, \delta) \leq \sqrt{2} A(S, \delta)$$

where $\hat{\theta}^a$ and $\hat{\theta}^b$ are the IV estimate for the treatment effect of arm a and arm b, respectively.

Proof. We have:

$$\begin{aligned}
|(\theta^a - \theta^b) - (\hat{\theta}^a - \hat{\theta}^b)| &= |(\hat{\theta}^a - \theta^a) + (\hat{\theta}^b - \theta^b)| \\
&\leq |\hat{\theta}^a - \theta^a| + |\hat{\theta}^b - \theta^b| && \text{(by Triangle Inequality)} \\
&\leq \sqrt{2}\sqrt{(\hat{\theta}^a - \theta^a)^2 + (\hat{\theta}^b - \theta^b)^2} && \text{(by Lemma A.5)} \\
&\leq \sqrt{2}\sqrt{\sum_{i=1}^{k} (\hat{\theta}^i - \theta^i)^2} \\
&\leq \sqrt{2} A(S, \delta)
\end{aligned}$$

This recovers the stated bound and we only pay a small constant ($\sqrt{2}$) to obtain a pairwise approximation bound from our IV estimator. □

Theorem F.9. *Let G^{vw} denote the gap between the causal effects of any arms v and w, i.e. $G^{vw} := \theta^v - \theta^w$ and let G^v denote the smallest gap for arm v, i.e. $G^v := \theta^v - \max_{w \neq v} \theta^w = \min_{w \neq v} \theta^v - \theta^w$.*

Let $A^{vw}(S, \delta)$ denote a high-probability upper bound (with probability at least $1-\delta$) on the difference between the true gap G^{vw} (for causal effects θ^v and θ^w for arms v and w) and its estimate $\widehat{G}^{vw}$ based on the sample set S, i.e.

$$\left|G^{vw} - \widehat{G}^{vw}\right| = \left|\theta^v - \theta^w - \left(\hat{\theta}^v - \hat{\theta}^w\right)\right| < A^{vw}(S, \delta).$$

Furthermore, let $A^v(S, \delta)$ denote a high-probability upper bound on the difference between the true minimum gap G^v for arm v and its empirical estimate $\widehat{G}^v$ based on sample set S, i.e.

$$\left|G^v - \widehat{G}^v\right| = \left|\theta^v - \theta^{w_{\min}} - \left(\hat{\theta}^v - \hat{\theta}^{w_{\min}}\right)\right| < A^v(S, \delta),$$

where $w_{\min} = \operatorname{argmin}_{w \neq v} \theta^v - \theta^w$. For shorthand, let A_q^v denote the best (i.e. smallest) approximation bound $A^v(S_q^{BEST}, \delta)$ by phase q.

Recall that Algorithm 4 is initialized with samples $S_0 = (x_i, y_i, z_i)_{i=1}^{|S_0|}$. Any agent at time t with type u_t will comply with recommendation $z_t = \mathbf{e}_v$ for arm v from policy π_t according to Algorithm 4, if the following holds for some $\tau \in (0, 1)$:

$$A^v(S_0, \delta) \leq \tau \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]/4.$$

Proof. Any agent at time t with type u_t will comply with an arm v recommendation $z_t = \mathbf{e}_v$ from policy π_t following Algorithm 4, if the following holds: For any two treatments $v, w \in B$,

$$\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta^v - \theta^w | z_t = \mathbf{e}_v] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v] \geq 0.$$

We will prove a stronger statement:

$$\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta^v - \max_{w \neq v} \theta^w | z_t = \mathbf{e}_v] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v] \geq 0.$$

We can prove this in largely the same way as we proved Lemma 4.2 in Appendix D.1: we simply replace θ and A_q^v in the proof for Lemma 4.2 with G^v and A_q^v, respectively.

The clean event C is given as:

$$\mathcal{C} := \left(\forall q \geq 0 : |G^v - \widehat{G}^v| < A_q^v\right).$$

By Corollary F.4, for event $\mathcal{C}$, the failure probability $\mathbb{P}[\neg\mathcal{C}] \leq \delta$. We assume that

$$\delta \leq \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau] + 2}.$$

First, we marginalize $\mathbb{E}_{\pi_t, \mathcal{P}^{(u_t)}}[\theta^v - \max_{w \neq v} \theta^w | z_t = \mathbf{e}_v] \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v]$ based on the clean event $\mathcal{C}$, such that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v] \\
&= \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}] + \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \neg\mathcal{C}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \neg\mathcal{C}] \\
&\geq \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}] - \delta \\
&\geq \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}] - \delta \\
&\geq \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}] - \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau] + 2}.
\end{aligned}$$

Next, we marginalize $\mathbb{E}_{\mathcal{P}^{(u_t)}, \pi_t}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \, \mathbb{P}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = \mathbf{e}_v, \mathcal{C}]$ based on four possible ranges which G^v lies on:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = \mathbf{e}_v, \mathcal{C} \\
&= \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}, G^v \geq \tau] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, G^v \geq \tau] \\
&+ \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}, 0 \leq G^v < \tau] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, 0 \leq G^v < \tau] \\
&+ \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}, -2A_q^v < G^v < 0] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, -2A_q^v < G^v < 0] \\
&+ \mathop{\mathbb{E}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v | z_t = \mathbf{e}_v, \mathcal{C}, G^v \leq -2A_q^v] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, G^v \leq -2A_q^v]
\end{aligned} \tag{59}$$

Because A_q^v is the smallest approximation bound derived from samples collected over any phase q of Algorithm 4 (including the initial sample set S_0), the following holds:

$$\begin{aligned}
2A_q^v &\leq 2A^v(S_0, \delta) \\
&\leq \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2} && \text{(by assumption } A^v(S_0, \delta) \leq \tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]/4) \\
&\leq \tau
\end{aligned}$$

Conditional on $\mathcal{C}$, $|G^v - \widehat{G}_q^v| < A_q^v$. Thus, if $G^v \geq \tau \geq 2A_q^v$, then $\widehat{G}_q^v \geq \tau - A_q^v \geq A_q^v$, which invokes the stopping criterion for the while loop in Algorithm 4. Thus, all other arms must have been eliminated from the race before phase $q = 1$ and arm v is recommended almost surely throughout Algorithm 4, i.e. $\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, G^v \geq \tau] = \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[\mathcal{C}, G^v \geq \tau]$. Similarly, if $G^v \leq -2A_q^v$, then $\widehat{G}_q^v \leq -A_q^v$ by phase $q = 1$ and arm v is recommended almost never, i.e. $\mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, G^v < -2A_q^v] = 0$. Substituting in these probabilities (and substituting minimum possible expected values), we proceed:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = \mathbf{e}_v, \mathcal{C}] \\
&\geq \tau \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\mathcal{C}, G^v \geq \tau] - 2A_q^v \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[z_t = \mathbf{e}_v, \mathcal{C}, -2A_q^v < G^v < 0] \\
&\geq \tau \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\mathcal{C}, G^v \geq \tau] - \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2} \\
&\geq \tau \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[\mathcal{C} | G^v \geq \tau] \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau] - \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2} \\
&\geq (1 - \delta)\tau \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau] - \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2} \\
&\geq \left(\frac{1}{2} - \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau] + 2} \right) \tau \mathop{\mathbb{P}}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau] \\
&= \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau] + 2}.
\end{aligned}
$$

Putting everything together, we get that

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[G^v | z_t = \mathbf{e}_v] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = \mathbf{e}_v] \\
&= \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[G^v | z_t = \mathbf{e}_v, \mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = \mathbf{e}_v, \mathcal{C}] + \mathop{\mathbb{E}}_{\mathcal{P}^{(u_t)}, \pi_t}[G^v | z_t = \mathbf{e}_v, \neg\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u_t)}, \pi_t}[z_t = \mathbf{e}_v, \neg\mathcal{C}] \\
&\geq \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau] + 2} - \frac{\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau]}{2\tau \, \mathbb{P}_{\pi_t, \mathcal{P}^{(u_t)}}[G^v \geq \tau] + 2} \\
&= 0.
\end{aligned}
$$

Thus, as long as $A^v(S_0, \delta) \leq \tau \, \mathbb{P}_{\pi, \mathcal{P}^{(u)}}[G^v \geq \tau]/4$, any agent of type u will comply with a recommendation of arm v from recommendation policy π according to Algorithm 4. □

Finally, we present the (expected) regret from the k treatment extension of policy π_c given in Definition F.5.

Lemma 6.3 (Regret of Policy π_c for k Treatments)**.** *An extension of policy π_c achieves (expected) regret as follows:*

$$
\mathbb{E}[R(T)] = O\left(k\sqrt{kT\log(kT)}\right) \tag{12}
$$

for sufficiently large time horizon T.

Proof. Let θ^* be the best treatment effect overall and the gap between θ^* and any treatment effect θ^i be $\Delta_i = |\theta^* - \theta_i|$. Recall that the clean event $\mathcal{C}$ entails that the approximation bound holds for all rounds. If event $\mathcal{C}$ fails, then we can only bound the pseudo-regret by the maximum value, which is at most $T \min_i \Delta_i$.

For the rest of this proof, assume that the event $\mathcal{C}$ holds for every round of Algorithm 4. This proof follows the standard technique from (Even-Dar et al., 2006). Since $\mathcal{C}$ holds, we have $\Delta_i \leq A(S_{L_2}, \delta) + |\hat{\theta}^* - \hat{\theta}^i|$ for any treatment i, where $A(S_{L_2}, \delta)$ is the approximation bound based on S_{L_2} samples of Algorithm 4. Before the stopping criteria is invoked, we also have $|\hat{\theta}^* - \hat{\theta}^i| \leq A(S_{L_2,\delta})$. Hence, the gap between the best treatment effect and any other treatment effect is:

$$
\Delta_i \leq 2A(S_{L_2}, \delta) \leq \frac{2\sigma_g\sqrt{2k\log(2kT/\delta)}}{\sqrt{\frac{L_2}{k^2} - \log(k/\delta)}},
$$

where σ_g is the variance parameter for the baseline reward $g^{(u)}$ and α_2 is defined as in Claim F.3 relative to the proportion $r_2 = 1/|B|$ of recommendations for each treatment during Algorithm 4 and the proportion of compliant agents p_{c_2}. Hence, we must have eliminated treatment i by round

$$L_2 = \frac{8k\sigma_g^2 \log(2kT/\delta)}{\Delta_i^2 \left(\frac{1}{k^2} - \log(k/\delta)\right)},$$

assuming that $L_2 \geq \frac{p_c^2}{k \log(k/\delta)}$ (in order to satisfy the criterion for Theorem F.2). During Algorithm 4, the social planner gives out $|B|$ recommendations for each treatment $i \in B$ sequentially. Hence, the contribution of each treatment i for each phase is Δ. Conditioned on event $\mathcal{C}$, the treatment a^* at the end of Algorithm 4 is the best treatment overall; so, no more regret is collected after Algorithm 4 is finished.

If treatment 1 is not the winner, then we accumulate $R_1(T) = \Delta_i \left((1 - k\rho)L_1 + L_2/k\right)$ regret for treatment 1. If some other treatment $i > 1$ is not the winner, then we accumulate $R_i(T) = \Delta_i \left(\rho L_1 + L_2/k\right)$ regret for treatment i. Hence, the total regret accumulated in Algorithm 4 is:

$$R(T) \leq \Delta_i \left((1-\rho)L_1 + \left(\frac{k-1}{k}\right) L_2\right) \leq (1-\rho)L_1\Delta_i + \frac{8(k-1)\sigma_g^2 \log(2kT/\delta)}{\Delta_i \left(\frac{1}{k^2} - \log(k/\delta)\right)}$$

Observe that the pseudo-regret of the combined recommendation policy is at most that of Algorithm 4 plus $\Delta = \min_i \Delta_i$ per each round of Algorithm 3. Alternatively, we can also upper bound the regret by Δ per each round of the combined recommendation policy. Following the same argument as Lemma 5.4, we can derive the pseudo-regret of the policy π_c for k treatments:

$$R(T) \leq \min\left(L_1(1-\rho)\Delta_i + \frac{8(k-1)\sigma_g^2 \log(2kT/\delta)}{\Delta_i \left(\frac{1}{k^2} - \log\left(k/\delta\right)\right)}, T\Delta\right) \leq L_1 + O(k\sqrt{kT\log(kT/\delta)}).$$

For the expected regret, we can set the parameters δ and L_1 in terms of the time horizon T, in order to both guarantee compliance throughout policy π_c and to obtain sublinear expected regret bound relative to T.

First, we must guarantee that the failure probability δ in Algorithm 4 is small, i.e. $\delta = 1/T^2$. To meet our compliance condition for Algorithm 4, we must set

$$\delta \leq \frac{\tau\, \mathbb{P}_{\mathcal{P}^{(u)}}[\theta \geq \tau]}{2(\tau\, \mathbb{P}_{\mathcal{P}^{(u)}}[\theta \geq \tau] + 1)}$$

for some constant $\tau \in (0, 1)$. Hence, we can set T sufficiently large such that, for any $\delta in (0, 1)$, we have

$$T \geq \frac{1}{\sqrt{\delta}} \geq \sqrt{\frac{2(\tau\, \mathbb{P}_{\mathcal{P}^{(u)}}[\theta \geq \tau] + 1)}{\tau\, \mathbb{P}_{\mathcal{P}^{(u)}}[\theta \geq \tau]}}$$

We also recall that the length L_1 of Algorithm 3 needs to be sufficiently large so that p_{c_2} fraction of agents comply in Algorithm 4. Moreover, we accumulate linear regret in each round of Algorithm 3. Hence, in order to guarantee sublinear regret, we also require that T satisfies the following:

$$T \geq L_1^2 = (\ell + \ell/\rho)^2$$

Finally, recall that the clean event $\mathcal{C}$ in Algorithm 4 holds with probability at least $1 - \delta$ for any $\delta \in (0, 1)$. Conditioned on the failure event $\neg\mathcal{C}$, policy π_c accumulates at most linear pseudo-regret in terms of T. Thus, in expectation, it accumulates at most $T \max_{i,j} |\theta^i - \theta^j| \delta$ regret

Therefore, we can derive the expected regret of k treatment recommendation policy π_c as:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{P}^{(u)}}[R(T)] &= \mathbb{E}[R(T)|\neg\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u)}}[\neg\mathcal{C}] + \mathbb{E}[R(T)|\mathcal{C}] \mathop{\mathbb{P}}_{\mathcal{P}^{(u)}}[\mathcal{C}] \\
&\leq T\delta + (L_1 + O(\sqrt{kT\log(kT/\delta)})) \\
&= \frac{1}{T} + (\sqrt{T} + O(k\sqrt{kT\log(kT)})) \\
&= \frac{1}{T} + O(k\sqrt{kT\log(kT)}) \\
&= O(k\sqrt{kT\log(kT)})
\end{aligned}$$

Therefore, assuming that all hyperparameters δ, L_1 are set to incentivize compliance for some nonzero proportion of agents throughout π_c and assuming that T is sufficiently large so as to satisfy the conditions above, policy π_c (for k treatments) achieves sublinear regret. □

G. Experiments Omitted from Section 7

In this section, we present additional experiments to evaluate Algorithm 1 and Algorithm 2, which were previously omitted from Section 7. Our code is available here: https://github.com/DanielNgo207/Incentivizing-Compliance-with-Algorithmic-Instruments. We are interested in (1) the effect of different prior choices on the exploration probability ρ, (2) comparing the approximation bound in Algorithm 1 to that of Algorithm 2 and (3) the total regret accumulated by the combined recommendation policy. Firstly, we observed that the exploration probability ρ in Figure 1 is small, leading to slow improvement in accuracy of Algorithm 1. Since ρ depends on the event ξ (as defined in Equation (5)), we want to investigate whether changes in the agents' priors would increase the exploration probability. Secondly, we claimed earlier in the paper that the estimation accuracy increases much quicker in Algorithm 2 compared to Algorithm 1. This improvement motivates the social planner to move to Algorithm 2, granted there is a large enough portion of agents that comply with the recommendations. Finally, while we provide a regret guarantee in Lemma 5.4, it is not immediately clear how the magnitude of Algorithm 1 length L_1 would affect the overall regret. There is a tradeoff: if we run Algorithm 1 for a small number of rounds, then it would not affect the regret by a significant amount, but a portion of the agents in Algorithm 2 may not comply. For our combined recommendation policy, we run Algorithm 1 until it is guaranteed that type 0 agents will comply in Algorithm 2.

Experimental Description For Algorithm 1, we consider a setting with two types of agents: type 0 who are initially never-takers, and type 1 who are initially always takers. For Algorithm 2, we consider a setting with two types of agents: type 0 who are compliant, and type 1 who are initially always-takers. We let each agent's prior on the treatment be a truncated Gaussian distribution between -1 and 1. The noisy baseline reward $g_t^{(u_t)}$ for each type u of agents is drawn from a Gaussian distribution $\mathcal{N}(\mu_{g^{(u)}}, 1)$, with its mean $\mu_{g^{(u)}}$ also drawn from a Gaussian prior. We let each type of agents have equal proportion in the population, i.e. $p_0 = p_1 = 0.5$.

For the first experiment, we are interested in finding the correlation between the exploration probability ρ and different prior parameters, namely the difference between mean baseline rewards $\mu_{g^{(1)}} - \mu_{g^{(0)}}$ and the variance of Gaussian prior on the treatment effect θ. Similar to the experiment in Section 7, we use Monte Carlo simulation by running the first stage of Algorithm 1 with varying choices of the two prior parameters above. From these initial samples, we calculate the probability of event ξ, and subsequently the exploration probability ρ. For the second experiment, we are interested in finding when agents of type 1 also comply with the recommendations. This shift in compliance depends on a constant τ (as defined in Lemma 4.2). We find two values of the constant τ that minimizes the number of samples needed to guarantee that agents of type 0 and type 1 are compliant in Algorithm 2 (as defined in Lemma 5.2). After this, Algorithm 2 is run for increasing number of rounds. Similar to the Algorithm 1 experiment, we repeated calculate the IV estimate of the treatment effect and compare it to the naive OLS estimate over the same samples as a benchmark. On a separate attempt, we evaluate the combined recommendation policy by running Algorithm 1 and Algorithm 2 successively using the priors above. We calculate the accumulated regret of this combined policy using the pseudo-regret notion (as defined in Definition 5.3).

Results In Table 1 and Table 2, we calculate the exploration probability ρ with different initialization of the agents' priors. In Table 1, we let the mean baseline reward of type 1 $\mu_{g^{(1)}}$ be drawn from $\mathcal{N}(0.5, 1)$ and the mean baseline reward of type 0 $\mu_{g^{(0)}}$ be drawn from $\mathcal{N}(c, 1)$ with $c \in [0, 1]$. The gap between these priors is defined as $\mathbb{E}[\mu_{g^{(1)}}] - \mathbb{E}[\mu_{g^{(0)}}]$. We observe that

Gap between $\mathbb{E}[\mu_{g^{(1)}}]$ and $\mathbb{E}[\mu_{g^{(0)}}]$	ρ
-0.5	0.004480
-0.4	0.004975
-0.3	0.007936
-0.2	0.003984
-0.1	0.003488
0.1	0.003984
0.2	0.004480
0.3	0.003488
0.4	0.004480
0.5	0.003488

Table 1. ρ with different gaps between $\mathbb{E}[\mu_{g^{(1)}}]$ and $\mathbb{E}[\mu_{g^{(0)}}]$

Variance in prior over treatment effect	ρ
0.1	0.002561
0.2	0.003112
0.3	0.002561
0.4	0.002561
0.5	0.003982
0.6	0.004643
0.7	0.005422
0.8	0.002790
0.9	0.001389
1	0.003488

Table 2. ρ with different variances in the prior over treatment effect θ

the exploration probability does not change monotonically with increasing gap between mean baseline reward. In Table 2, we calculate the exploration probability ρ with different variance in prior over treatment effect θ. Similarly, in Table 1, we observe that the exploration probability ρ does not change monotonically with increasing variance in prior over θ. In both tables, ρ value lies between $[0.001, 0.008]$, which implies infrequent exploration by Algorithm 1. This slow rate of exploration is also reflected in Figure 1, which motivates the social planner to transition to Algorithm 2.

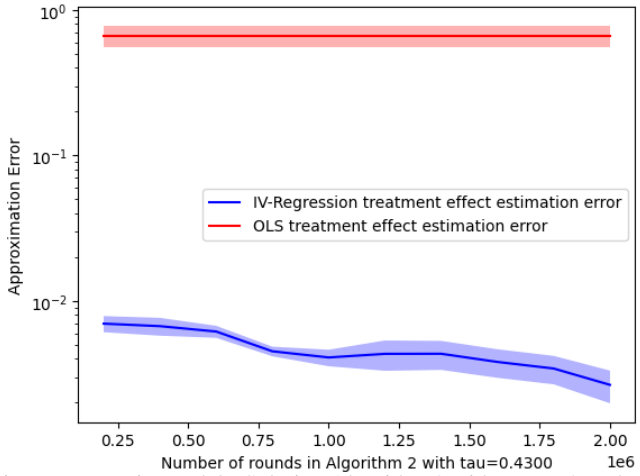

Figure 2. Approximation bound using IV regression and OLS during Algorithm 2 with $\tau = 0.43$. The y-axis uses a log scale. Results are averaged over 5 runs; error bars represent one standard error.

In Figure 2, we compare the approximation bound on $|\theta - \hat{\theta}|$ between the IV estimate $\hat{\theta}$ and the naive estimate for Algorithm 2. In our experiments, the constant τ generally lies within $[0.4, 0.6]$. Similar to the experiment in Section 7, we let the hidden treatment effect $\theta = 0.5$, type 0 and type 1 agents' priors on the treatment effect be $\mathcal{N}(-0.5, 1)$ and $\mathcal{N}(0.9, 1)$ — each truncated into $[-1, 1]$ — respectively. We also let the mean baseline reward for type 0 and type 1 agents be $\mu_{g^{(0)}} \sim \mathcal{N}(0, 1)$ and $\mu_{g^{(1)}} \sim \mathcal{N}(0.1, 1)$, respectively. With these priors, we have found a suitable value of $\tau = 0.43$ for Algorithm 2. Instead of using the theoretical bound on ℓ in Lemma 5.2, we compare the approximation bound $|\theta - \hat{\theta}|$ with the conditions in Lemma 4.2. In Figure 2, the IV estimate consistently outperform the naive estimate for any number of rounds. Furthermore, we observe that the scale of the IV estimate approximation bound in Figure 2 is much smaller than that of Figure 1. This difference shows the improvement of Algorithm 2 over Algorithm 1 on estimating the treatment effect θ. It takes Algorithm 2 a small number of rounds to get a better estimate than Algorithm 1 due to the small exploration probability ρ.