Adaptive Design of Multiple Stage Experiments using the Propensity Score*

Jinyong Hahn
UCLA†

Keisuke Hirano
University of Arizona‡

Dean Karlan
Yale University§

1 October 2007

Abstract

Many social experiments are run in multiple waves. In principle, the sampling design of such experiments can be modified in later stages to allow for more efficient estimation of causal effects. We consider the design of a two-stage experiment for estimating an average treatment effect, when covariate information is available for experimental subjects. We use data from the first stage to choose a conditional treatment assignmen rule for units in the second stage of the experiment. This amounts to choosing the "propensity score," the conditional probability of treatment given covariates. We propose to select the propensity score in a way that minimizes the asymptotic variance bound for estimating the average treatment effect, show how to implement this numerically using standard statistical software, and derive large-sample properties of our procedure.

1 Introduction

Social experiments have become increasingly important for the evaluation of social policies and the testing of economic theories. Random assignment of individuals to different treatments makes it possible to conduct valid counterfactual comparisons without strong auxiliary assumptions. On the other hand, social experiments can be costly, especially when they involve policy-relevant treatments and a large number of individuals. Thus, it is important to design experiments carefully to maximize the information gained from them. In this paper, we consider social experiments run

*We thank David Reiley, Don Rubin, Dylan Small, and seminar participants at USC, Ohio State, Chicago GSB, Princeton, Singapore Management University, Xiamen University, Academic Sinica, Uppsala University, UC Irvine, and Cornell for their comments and suggestions.

†Department of Economics, University of California, Los Angeles, Box 951477, Los Angeles, CA 90095-1466 (hahn@econ.ucla.edu)

‡Department of Economics, University of Arizona, Tucson, AZ 85721 (hirano@u.arizona.edu)

§Department of Economics, Yale University, PO Box 208209, New Haven, CT 06520-8209 (dean.karlan@yale.edu)

in multiple stages, and examine the possibility of using initial results from the first stage of an experiment to modify the design of the second stage, in order to estimate the average treatment effect more precisely. We suppose that in the second stage, assignment to different treatments can be randomized *conditional* on some observed characteristics of the individual. We show that data from the first wave can reveal potential efficiency gains from altering conditional treatment assignment probabilities, and suggest a procedure for using the first-stage data to construct second-stage assignment probabilities. In general, the treatment effect can be estimated with a lower variance than under pure random sampling using our sequential procedure.

Many social experiments have a pilot phase or some more general multi-stage or group-sequential structure. For instance, Johnson and Simester (2006) conduct repeated experiments with the same retailers to study price sensitivities. Karlan and Zinman (2006) conduct repeated experiments with a microfinance lender in South Africa to study interest rate sensitivities. In addition, get-out-the-vote experiments, although they are run with different organizations and campaigns, are often similar enough to be thought of as replications (see for example Green and Gerber, 2004).

Randomizing treatment conditional on covariates amounts to choosing the *propensity score*—the conditional treatment probability. Rosenbaum and Rubin (1983) proposed to use the propensity score as a tool for estimating treatment effects in observational studies of treatments under the assumption of unconfoundedness. Some studies have used propensity score methods as a way to improve precision in pure randomized experiments (for example, see Flores-Lagunes, Gonzalez, and Neumann, 2006). When treatment is random conditional on covariates, the semiparametric variance bound for estimating the average treatment effect depends on the propensity score and the conditional variance of outcomes given treatment and covariates. We propose to use data from the first stage to estimate the conditional variance. Then we *choose* the propensity score in the second stage in order to minimize the asymptotic variance for estimating the average treatment effect. Finally, after data from both stages has been collected, we pool the data and construct an overall estimate of the average treatment effect. If both stages have a large number of observations, the estimation error in the first-stage preliminary estimates does not affect the asymptotic distribution of the final, pooled estimate of the treatment effect. Our procedure is "adaptive" in the sense that the design uses an intermediate estimate of the conditional variance structure, and does as well asymptotically as an infeasible procedure that uses knowledge of the conditional variances. We illustrate our approach using data from two recent field experiments, showing how a hypothetical second stage of the experiment should assign treatments based on the data from the first stage.

There is an extensive literature on sequential experimentation and experimental design, but much of the classic work focuses on stopping rules for sequential sampling of individuals, or on "play-the-winner" rules which increase the probability of treatments which appear to be better based on past data. Bayesian methods have also been developed for sequential experimental de-

sign; for a recent review of Bayesian experimental design, see Chaloner and Verdinell (1995). Unlike some recent work taking a simulation-based Bayesian approach, our approach is very simple and does not require extensive computations.[1] However, our analysis is based on asymptotic approximations where the sample size in each stage of the experiment is taken as large. Thus, our formal results would apply best to large-scale social experiments, rather than the very small experiments sometimes conducted in laboratory economic experiments.

Our approach is also closely related to the Neyman allocation formula (Neyman, 1934) for optimal stratified sampling. Some authors, such as Sukhatme (1935), have considered the problem of estimating the optimal strata sizes using preliminary samples, but in a finite-population setting where it is difficult to obtain sharp results on optimal procedures. A review of this literature is given in Solomon and Zacks (1970). Our asymptotic analysis lead to a simple adaptive rule which has attractive large-sample properties.

2 Adaptive Design Algorithm and Asymptotic Theory

2.1 Two-Stage Design Problem

We consider a two-stage social experiment comparing two treatments. In each stage, we draw a random sample from the population. We assume that the population of interest remains the same across the two stages of experimentation. For each individual, we observe some background variables X, and assign the individual to one of two treatments. We will use "treatment" and "control" and "1", "0" to denote the two treatments. Let n_1 denote the number of observations in the first stage, and let n_2 denote the number of observations in the second stage, and let $n = n_1 + n_2$.

In order to develop the formal results below, we assume that the covariate X_i has finite support. If X_i is continuously distributed, we can always discretize it. Further, since we will be making treatment assignment probabilities depend on X_i, it often would be sensible to work with discretized covariates for operational purposes. All of our results to follow will still hold under discretization, although discretizing too coarsely may sacrifice some precision in estimating treatment effects.

In the first stage, individuals are assigned to treatment 1 with probability π_1, which does not depend on their observed covariates. Before the second stage, the outcomes from the first stage are realized, and observed by the experimental designer. In the second stage, the designer can make treatment assignment probabilities depend on the individual's covariate X. Let $\hat{\pi}_2(x)$ denote the probability that a second-stage individual with $X_i = x$ receives treatment 1. We use the "hat" to indicate the these probabilities can depend on all the data from the first stage. The goal is to estimate the population average treatment effect with low mean-squared error.

[1] We have written simple programs in R and Stata to implement our procedures, which are available on request.

Formally, for individuals $i = 1, 2, \ldots, (n_1+n_2)$, let (X_i, Y_{0i}, Y_{1i}) be IID from a joint distribution. We interpret X_i as the (always observed) vector of covariates, and Y_{ti} as the potential outcome under treatment $t = 0, 1$. We are interested in estimation of the average treatment effect

$$\beta := E[Y_{1i} - Y_{0i}].$$

Individuals $i = 1, \ldots, n_1$, drawn in the first stage, are assigned treatment D_i equal to 1 with probability π_1 and 0 with probability $1 - \pi_1$. The experimental planner then observes (X_i, D_i, Y_i), where

$$Y_i := D_i Y_{1i} + (1 - D_i) Y_{0i}.$$

Similarly, for $i = n_1 + 1, \ldots, n_1 + n_2$, we assign individual to treatments according to $P(D_i = 1|X_i = x) = \hat{\pi}_2(x)$, and we observe (X_i, D_i, Y_i).

We can also consider the experimental design problem, when there is an overall budget constraint, that the overall probability of treatment is equal to a fixed number p. In this case, the assignment rule $\hat{\pi}_2(\cdot)$ must satisfy

$$p = \frac{n_1}{n}\pi_1 + \frac{n_2}{n} E[\hat{\pi}_2(X_i)],$$

where $n = n_1 + n_2$ and the expectation is with respect to the marginal distribution of X. In the sequel, we will consider both the unconstrained and constrained design problems. It would also be straightforward to extend the analysis to cases where there is an upper or lower bound on the the overall treatment probability.

2.2 One-Stage Problem and Optimal Propensity Score

Before giving our proposal for an adaptive experimental design rule, it is useful to consider the simpler problem of estimating the average treatment effect under a fixed treatment assignment rule.

Suppose that $(X_i, Y_{0i}, Y_{1i}, D_i)$ are IID from a population for $i = 1, \ldots, n$, and that the treatment assignment rule depends only on X_i:

$$D_i \perp (Y_{0i}, Y_{1i})|X_i.$$

Let

$$p(x) := Pr(D_i = 1|X_i = x).$$

The function $p(x)$ is often called the propensity score (Rosenbaum and Rubin, 1984). We also

require that for all possible values of X, $0 < p(x) < 1$.[2]

As before, the average treatment effect $\beta = E[Y(1) - Y(0)]$ is the object of interest. Typically, there will exist estimators $\hat{\beta}$ that $\hat{\beta} \xrightarrow{p} \beta$ and

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V).$$

We wish to find an estimator with minimal asymptotic variance V. The following result, due to Hahn (1998), provides a lower bound for the variance of regular[3] estimators:

Proposition 1 *(Hahn, 1998) Let*

$$\begin{aligned} \beta(x) &:= E[Y_{1i} - Y_{0i}|X_i = x], \\ \sigma_0^2(x) &:= V[Y_{0i}|X_i = x], \\ \sigma_1^2(x) &:= V[Y_{1i}|X_i = x] \end{aligned}$$

Then any regular estimator $\hat{\beta}$ for β has asymptotic variance

$$V \geq E\left[\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{1 - p(X_i)} + (\beta(X_i) - \beta)^2\right].$$

Estimators that achieve this bound have been constructed by Hahn (1998), Hirano, Imbens, and Ridder (2003) (hereafter HIR), and others. Consider the following two-step estimator proposed by HIR. Let $\hat{p}(x)$ be a nonparametric regression estimate of $E[D_i|X_i = x]$. The HIR estimator is

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\hat{p}(X_i)} - \frac{(1 - D_i)Y_i}{1 - \hat{p}(X_i)}\right).$$

HIR consider the case where X is continuous, and $\hat{p}$ is a estimated using a sieve estimator, and they show that the estimator achieves the semiparametric efficiency bound. In the case that X is finitely supported, it is natural to estimate the propensity score as

$$\hat{p}(x) = \frac{\sum_{i=1}^{n} D_i 1(X_i = x)}{\sum_{i=1}^{n} 1(X_i = x)}.$$

This is simply the empirical probability of treatment for observations with $X_i = x$.

[2] In a randomized experiment, this overlap condition can be guaranteed by design.

[3] See Chamberlain (1986), for a discussion of regularity and semiparametric variance bounds.

An alternative estimator suggested by Hahn (1998) is:

$$\tilde{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{r}_1\left(X_i\right) - \widehat{r}_0\left(X_i\right)\right)$$

where $\widehat{r}_1\left(X_i\right)$ and $\widehat{r}_0\left(X_i\right)$ are nonparametric analogs of

$$r_1\left(X_i\right) = \frac{E\left[D_i Y_i \middle| X_i = x\right]}{E\left[D_i \middle| X_i = x\right]}, \quad r_0\left(X_i\right) = \frac{E\left[\left(1 - D_i\right) Y_i \middle| X_i = x\right]}{E\left[1 - D_i \middle| X_i = x\right]}$$

In the case we consider here, where the covariate X has finite support, the two estimators are equal. The proof of the following proposition is straightforward and is omitted.

Proposition 2 *$\widehat{\beta} = \tilde{\beta}$ when X is multinomial with finite support.*

Now suppose that the researcher can choose the propensity score $p(x)$. The researcher would like to solve

$$\min_{p(\cdot)} E\left[\frac{\sigma_1^2\left(X_i\right)}{p\left(X_i\right)} + \frac{\sigma_0^2\left(X_i\right)}{1 - p\left(X_i\right)} + \left(\beta\left(X_i\right) - \beta\right)^2\right] \tag{1}$$

If there is a constraint on the overall treatment probability, this minimization is subject to the constraint

$$E\left[p\left(X_i\right)\right] = p$$

In the constrained case, an interior solution $p\left(\cdot\right)$ will satisfy

$$-\frac{\sigma_1^2\left(x\right)}{p\left(x\right)^2} + \frac{\sigma_0^2\left(x\right)}{\left(1 - p\left(x\right)\right)^2} = \lambda \tag{2}$$

for all x in the support of X, where λ denotes the Lagrange multiplier.

Thus, the solution depends on the conditional variances $\sigma_0(x)$ and $\sigma_1(x)$. Intuitively, if the data exhibit large differences in conditional variances by x, then allowing for different treatment probabilities for different x may permit more precise estimation of the treatment effect. In essence, heteroskedasticity drives the possibility for improved precision.

2.3 Two-Stage Adaptive Design and Estimator

The optimization problem (1) implicitly assumes that the conditional variance functions $\sigma_1^2\left(X_i\right)$ and $\sigma_0^2\left(X_i\right)$ are known to the researcher, and therefore is not feasible in a one-stage setting. However, if the experiment is run in two stages, one can use the first stage to estimate the unknown variance functions. We propose to use the first stage results to estimate $\sigma_1^2\left(X_i\right)$ and $\sigma_0^2\left(X_i\right)$, and then use these estimates to modify the treatment assignment probabilities in the second stage. We show

that if the sample sizes in both stages are large, the overall design is "adaptive" — we achieve the same overall efficiency as the infeasible optimum. Our overall design and estimation procedure is implemented in the following steps:

1. In Stage 1, we assign individuals $i = 1, \ldots, n_1$ to treatment 1 with probability π_1, irrespective of their covariate values. We collect data (D_i, X_i, Y_i) for these individuals.

2. Using data from Stage 1, we estimate the conditional variances: $\widehat{\sigma}_0^2(x)$ and $\widehat{\sigma}_1^2(x)$ by their empirical analogs: $\widehat{\sigma}_0^2(x)$ is the sample variance of Y for first-stage observations with $D = 0$ and $X = x$, and $\hat{\sigma}_1(x)$ is the sample variance of Y for first-stage observations with $D = 1$ and $X = x$. We then choose $\hat{\pi}_2(x)$ to minimize:

$$E\left[\frac{\widehat{\sigma}_1^2(X_i)}{\pi(X_i)} + \frac{\widehat{\sigma}_0^2(X_i)}{1-\pi(X_i)} + (\beta(X_i) - \beta)^2\right]$$

   where

$$\pi(x) = \kappa\pi_1 + (1-\kappa)\hat{\pi}_2(x).$$

   As before, if there is a constraint that the overall treatment probability is equal to p, then the minimization is subject to:

$$E[\pi(X_i)] = p.$$

   Here, all of the expectations are with respect to the marginal distribution of X_i. Note that the solution does not depend on $(\beta(X_i) - \beta)^2$, so we can drop this term from the objective function when solving the minimization problem.

3. We assign individuals $i = n_1 + 1, \ldots, n_1 + n_2$ to treatment 1 with probabilities $\hat{\pi}_2(X_i)$. We collect data (D_i, X_i, Y_i) from the second stage individuals, and estimate the average treatment effect β using the Hahn/HIR estimator

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\hat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}\right).$$

   Note that this estimator involves estimating a propensity score. Although the propensity score is known (because it is controlled by the researcher), the estimator does not use the true propensity score.[4]

In the second step of our procedure, it is possible to have a corner solution, because the first stage randomization restricts the set of possible propensity scores achievable over the two stages. In

[4]The efficiency gain from using an estimate of the propensity score rather than the true propensity score is discussed in HIR.

particular, for any x, the overall conditional probability $\pi(x)$ cannot be less than $\kappa\pi_1$, and cannot be greater than $\kappa\pi_1 + (1-\kappa) = 1 - \kappa(1-\pi_1)$. However, our results to follow do not require an interior solution.

2.4 Asymptotic Theory

Our asymptotic theory is based on the regularity conditions stated below as Assumption 1:

Assumption 1 *(i) $n_1 \to \infty$ and $n_2 \to \infty$ such that $n_1/(n_1+n_2) \to \kappa$; (ii) X_i has a multinomial distribution with finite support; (iii) $\pi_2^*(\cdot)$ depend smoothly on the vectors $\sigma_0^2(\cdot)$ and $\sigma_1^2(\cdot)$; (iv) the estimators $\widehat{\sigma}_0^2(x)$ and $\widehat{\sigma}_1^2(x)$ are $\sqrt{n}$-consistent for the true variances $\sigma_0^2(x)$ and $\sigma_1^2(x)$.*

The most notable aspect of Assumption 1 is the double asymptotics, in which n_1 and n_2 go to infinity at the same rate. The assumption that $\pi_2^*(\cdot)$ depends smoothly on the vectors $\sigma_0^2(\cdot)$ and $\sigma_1^2(\cdot)$ is innocuous when the X_i has a multinomial distribution with finite support. The assumption that $\widehat{\sigma}_0^2(x)$ and $\widehat{\sigma}_1^2(x)$ are $\sqrt{n_1}$-consistent is also harmless under the multinomial assumption. Because $n_1 = O(n)$, it follows that $\widehat{\sigma}_0^2(x)$ and $\widehat{\sigma}_1^2(x)$ are $\sqrt{n}$-consistent.

Since the estimators $\widehat{\sigma}_0^2(x)$ and $\widehat{\sigma}_1^2(x)$ are $\sqrt{n}$-consistent for the true variances, it follows that the second stage assignment probabilities $\hat{\pi}_2(x)$ are $\sqrt{n}$-consistent for $\pi_2^*(x)$. We also use $\pi^*(x)$ to denote the target overall propensity scores, defined as

$$\pi^*(x) := \kappa\pi_1 + (1-\kappa)\pi_2^*(x).$$

Because the assignment probabilities in the second stage depend on the realization of the first-stage data, we do not have classic IID sampling. To develop the formal results, we must take into account the dependence of the second-stage DGP on the first stage data. We do this by viewing the treatment indicators as being generated by IID uniform random variables. In the first stage,

$$D_i = 1(U_i \leq \pi_1),$$

where U_i are IID Uniform[0,1] random variables, independent of the X and Y variables. For individuals $i = n_1+1, \ldots, n_1+n_2$, drawn in the second stage, treatment is determined according to an assignment rule as $\hat{\pi}_2(X_i)$, where the "hat" indicates that the rule can depend on first-stage data. Treatment is defined as

$$D_i = 1(U_i \leq \hat{\pi}_2(X_i)),$$

and we observe (X_i, D_i, Y_i) where Y_i is defined as before. There is no loss of generality in defining treatment randomization this way, and it permits us to define empirical processes based on the U_i in the proof of the main theorem below.

The following result shows that the two-stage design procedure, combined with the Hahn/HIR estimator, is "adaptive": the estimator has asymptotic variance equal to the variance that would obtain had we used $\pi^*(x)$ to assign individuals to treatment.

Theorem 1 *Let (i)* $\pi_2^*(x) := plim\ \hat{\pi}_2(x)$*; and (ii)* $\pi^*(x) := \kappa\pi_1 + (1-\kappa)\pi_2^*(x)$*. Assume that* $\widehat{\pi}_2(x) = \pi_2^*(x) + O_p\left(\frac{1}{\sqrt{n}}\right)$*. Further assume that* $0 < \pi^*(x) < 1$*. We then have*

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, E\left[\frac{1}{\pi^*(X_i)}\sigma_1^2(X_i) + \frac{1}{1-\pi^*(X_i)}\sigma_0^2(X_i) + (\beta(X_i) - \beta)^2\right]\right)$$

Proof: See Appendix A □

3 Examples

In this section we give two simple numerical examples of our adaptive design algorithm, using data from recently conducted field experiments. Both of these applications were single-stage experiments. For the purpose of illustration, we suppose that the researcher has the ability to carry out a second round of the same experiment. We use our adaptive algorithm, along with the data from the "first" round, to determine how the second stage should be carried out. In the first example, a charitable fundraising experiment, we find significant efficiency gains from employing our adaptive treatment assignment rule. In the second example, a media experiment, we find that the data are very close to homoskedastic, so that simple random treatment assignment in the second round is close to optimal.

3.1 Direct Mail Fundraising Experiment

For the first example, we use data from a direct mail fundraising experiment reported in Karlan and List (2007). In this experiment, a charitable organization mailed 50,083 direct mail solicitations to prior donors to their organization. Of the 50,083, two-thirds (33,396) received a matching grant offer, and one-third (16,687) received the same solicitation but without mention of a matching grant. The matching grant test included several sub-features (i.e., the ratio of the match, the ceiling of the match, and the example amount provided), but for the sake of simplicity we will only consider the main treatment of receiving the matching grant offer. We now ask the question: in a second wave of an experiment with this organization, how should we allocate treatments, conditional on covariates?

There are various covariates available to us, but to keep the analysis simple, we focus on a single binary covariate, an indicator for whether the individual's prior giving amount was greater than the median amount. Thus, $X_i = 0, 1$ with equal probability in our setting. The outcome of interest

is the individual's donation amount after receiving the direct mail solicitation. We fix the overall fraction of individuals treated at 2/3, and set $\kappa = 1/2$, so that the second round will be the same size as the first round.

Using the original data as our "first" stage, we estimate the following conditional variances:

$$\begin{aligned}
\hat{\sigma}_0^2(0) &= 8.73 \\
\hat{\sigma}_0^2(1) &= 117.62 \\
\hat{\sigma}_1^2(0) &= 19.90 \\
\hat{\sigma}_1^2(1) &= 133.35
\end{aligned}$$

Notice that, for $X = 0$ (donors with low prior giving), the variance under treatment 1 is more than double the variance under treatment zero. This suggests that the low donors should be treated more, because it is difficult to learn the expected outcome under treatment for this subpopulation.

We applied our algorithm without the constraint that the overall treatment probability be 2/3. Table 1 gives the overall optimal treatment assignment probabilities, and the implied second-stage treatment probabilities.

Table 1: Optimal Treatment Assignment Probabilities, Unconstrained

X	Overall	Second Stage
0	0.60	0.54
1	0.52	0.36

Using our adaptive rule would lead to a normalized asymptotic variance of 292, compared with 320 from 2/3 random sampling in the second stage. This is a 8.7% gain in efficiency, implying that we could achieve the same precision as 2/3 random sampling with 4354 fewer observations.

We also considered adaptive treatment assignment under the constraint that the overall treatment probability be 2/3. These are given in Table 2.

Table 2: Optimal Treatment Assignment Probabilities, Constrained

X	Overall	Second Stage
0	0.79	0.92
1	0.56	0.45

Thus, nearly all the second round individuals with $X = 0$ should be treated. Using our adaptive rule would lead to a normalized asymptotic variance of approximately 300; this is a 6.4% gain in

efficiency, and would permit the same precision as 2/3 random treatment assignment with 3190 fewer observations.

3.2 Media Experiment

For the second example, we use data from a media experiment reported in Gerber, Karlan and Bergan (2007). This experiment aimed to measure the impact on political opinions, attitudes and behaviors from exposure to the media. In this experiment, a sample frame of individuals from a county in northern Virginia were surveyed to identify households that did not currently receive a newspaper. These households were then assigned to either receive a free two-month subscription to the Washington Post (29% probability), the Washington Times (28% probability). The Washington Times is widely viewed as being a more politically conservative newspaper than the Post. The remaining 43% of the individuals formed a control group and received neither treatment. We discard the control group, and focus on the comparison between the individuals who receive the Post and the Times.

After receiving one of the two newspapers for 2 months, individuals were contacted and surveyed about their political opinions. We focus on one of the outcome measures, "ConservativeS," which is a normalized composite measure of political leaning. We use an indicator for female as our covariate, and take $p = 1/2$ and $\kappa = 1/2$.

When we calculate conditional variances, we find that all are close to 1, indicating that the data are essentially homoskedastic. As a consequence, the optimal treatment assignment is to simply assign individuals one of the two treatments with equal probability. We have also considered other covariates and combinations of covariates, and find similar results.

4 Conclusion

In this paper, we considered the optimal design of a two-stage experiment for estimating an average treatment effect. We propose to choose the propensity score in the second stage based on the data from the first stage, in order to minimize an estimated version of the asymptotic variance bound. We argue, using a double asymptotic approximation, that our proposal leads to an adaptive estimation procedure for the average treatment effect. Using this double asymptotics leads to a very simple, intuitive procedure that is easily implemented in practice, and has good theoretical properties. Extending our approach to more than two time periods is straightforward.

Throughout this paper, we have assumed that the population of interest, the treatments, and the effects of the treatments are stable across periods, so that it is meaningful to combine the data from both stages. In some cases, the second stage might be substantially different from the first stage, for example if the treatments under consideration are modified in later time periods. Then,

the idea of using earlier experiments to inform experimental design could still be fruitful, but this would require additional modeling assumptions to link the data across time periods.

A Proof of Theorem 1

We can write

$$\begin{aligned}\widehat{\beta}-\beta &= \frac{1}{n}\sum_{i=1}^{n}\left(r_1\left(X_i\right)-r_0\left(X_i\right)-\beta\right)\\ &+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i\left(Y_i-r_1\left(X_i\right)\right)}{\widehat{\pi}\left(X_i\right)}-\frac{\left(1-D_i\right)\left(Y_i-r_0\left(X_i\right)\right)}{1-\widehat{\pi}\left(X_i\right)}\right)\\ &+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i r_1\left(X_i\right)}{\widehat{\pi}\left(X_i\right)}-\frac{\left(1-D_i\right)r_0\left(X_i\right)}{1-\widehat{\pi}\left(X_i\right)}-\left(r_1\left(X_i\right)-r_0\left(X_i\right)\right)\right)\\ &+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_iY_i}{\widehat{p}\left(X_i\right)}-\frac{D_iY_i}{\widehat{\pi}\left(X_i\right)}\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left(1-D_i\right)Y_i}{1-\widehat{p}\left(X_i\right)}-\frac{\left(1-D_i\right)Y_i}{1-\widehat{\pi}\left(X_i\right)}\right) \qquad (3)\end{aligned}$$

Note that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i r_1\left(X_i\right)}{\widehat{\pi}\left(X_i\right)}-r_1\left(X_i\right)\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_iY_i}{\widehat{p}\left(X_i\right)}-\frac{D_iY_i}{\widehat{\pi}\left(X_i\right)}\right)\\ &= \sum_{x}\left(r_1\left(x\right)-\widehat{r}_1\left(x\right)\right)\left(\frac{\widehat{p}\left(x\right)-\widehat{\pi}\left(x\right)}{\widehat{\pi}\left(x\right)}\right)\left(\frac{1}{n}\sum_{i=1}^{n}1\left(X_i=x\right)\right) \qquad (4)\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left(1-D_i\right)r_0\left(X_i\right)}{1-\widehat{\pi}\left(X_i\right)}-r_0\left(X_i\right)\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left(1-D_i\right)Y_i}{1-\widehat{p}\left(X_i\right)}-\frac{\left(1-D_i\right)Y_i}{1-\widehat{\pi}\left(X_i\right)}\right)\\ &= -\sum_{x}\left(r_0\left(x\right)-\widehat{r}_0\left(x\right)\right)\left(\frac{\widehat{p}\left(x\right)-\widehat{\pi}\left(x\right)}{1-\widehat{\pi}\left(x\right)}\right)\left(\frac{1}{n}\sum_{i=1}^{n}1\left(X_i=x\right)\right) \qquad (5)\end{aligned}$$

Furthermore, Lemmas 1 and 2 in Appendix B show that

$$\widehat{p}\left(x\right)-\widehat{\pi}\left(x\right)=O_p\left(n^{-1/2}\right),\ r_1\left(x\right)-\widehat{r}_1\left(x\right)=O_p\left(n^{-1/2}\right),\ r_0\left(x\right)-\widehat{r}_0\left(x\right)=O_p\left(n^{-1/2}\right),$$

which implies that (4) and (5) are $o_p\left(n^{-1/2}\right)$. We therefore obtain the following approximation for (3):

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}-\beta\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(r_1\left(X_i\right)-r_0\left(X_i\right)-\beta\right)\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{D_i\left(Y_i-r_1\left(X_i\right)\right)}{\widehat{\pi}\left(X_i\right)}-\frac{\left(1-D_i\right)\left(Y_i-r_0\left(X_i\right)\right)}{1-\widehat{\pi}\left(X_i\right)}\right)+o_p\left(1\right) \qquad (6)\end{aligned}$$

By Lemma 4 in Appendix B, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{D_i\left(Y_i-r_1\left(X_i\right)\right)}{\widehat{\pi}\left(X_i\right)}-\frac{\left(1-D_i\right)\left(Y_i-r_0\left(X_i\right)\right)}{1-\widehat{\pi}\left(X_i\right)}\right)$$
$$=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{D_i^*\left(Y_{1i}-r_1\left(X_i\right)\right)}{\pi^*\left(X_i\right)}-\frac{\left(1-D_i^*\right)\left(Y_{0i}-r_0\left(X_i\right)\right)}{1-\pi^*\left(X_i\right)}\right)+o_p\left(1\right)$$

where $D_i^*:=1\left(U_i\leq\pi_1\right)$ for the first sample, and $D_i^*:=1\left(U_i\leq\pi_2^*\left(X_i\right)\right)$ for the second sample.

Therefore, we can write

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}-\beta\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\beta\left(X_i\right)-\beta\right)\\ &\quad+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{D_i^*\left(Y_{1i}-r_1\left(X_i\right)\right)}{\pi^*\left(X_i\right)}-\frac{\left(1-D_i^*\right)\left(Y_{0i}-r_0\left(X_i\right)\right)}{1-\pi^*\left(X_i\right)}\right)+o_p\left(1\right)\end{aligned}$$

or

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)=\frac{\sqrt{n_1}}{\sqrt{n}}\times\left(I\right)+\frac{\sqrt{n_2}}{\sqrt{n}}\times\left(II\right)+o_p\left(1\right)$$

where

$$\begin{aligned}\left(I\right) &:= \frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left(\beta\left(X_i\right)-\beta+\frac{D_i^*\left(Y_i-r_1\left(X_i\right)\right)}{\pi^*\left(X_i\right)}-\frac{\left(1-D_i^*\right)\left(Y_{0i}-r_0\left(X_i\right)\right)}{1-\pi^*\left(X_i\right)}\right)\\ \left(II\right) &:= \frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\left(\beta\left(X_i\right)-\beta+\frac{D_i^*\left(Y_i-r_1\left(X_i\right)\right)}{\pi^*\left(X_i\right)}-\frac{\left(1-D_i^*\right)\left(Y_{0i}-r_0\left(X_i\right)\right)}{1-\pi^*\left(X_i\right)}\right)\end{aligned}$$

By the Central Limit Theorem (CLT), we obtain that

$$\begin{aligned}\left(I\right) &\xrightarrow{d} N\left(0,E\left[\frac{\pi_1}{\pi^*\left(X_i\right)^2}\sigma_1^2\left(X_i\right)+\frac{1-\pi_1}{\left(1-\pi^*\left(X_i\right)\right)^2}\sigma_0^2\left(X_i\right)+\left(\beta\left(X_i\right)-\beta\right)^2\right]\right)\\ \left(II\right) &\xrightarrow{d} N\left(0,E\left[\frac{\pi_2^*\left(X_i\right)}{\pi^*\left(X_i\right)^2}\sigma_1^2\left(X_i\right)+\frac{1-\pi_2^*\left(X_i\right)}{\left(1-\pi^*\left(X_i\right)\right)^2}\sigma_0^2\left(X_i\right)\right]\right)\end{aligned}$$

Noting that $\left(I\right)$ and $\left(II\right)$ are independent of each other, and that $\kappa\pi_1+\left(1-\kappa\right)\pi_2^*\left(X_i\right)=\pi^*\left(X_i\right)$ by definition, we obtain that $\sqrt{n}\left(\widehat{\beta}-\beta\right)$ converges weakly to a normal distribution with mean

zero and variance equal to

$$E\left[\frac{\kappa\pi_1+(1-\kappa)\,\pi_2^*(X_i)}{\pi^*(X_i)^2}\sigma_1^2(X_i)+\frac{1-(\kappa\pi_1+(1-\kappa)\,\pi_2^*(X_i))}{(1-\pi^*(X_i))^2}\sigma_0^2(X_i)+(\beta(X_i)-\beta)^2\right]$$
$$=E\left[\frac{1}{\pi^*(X_i)}\sigma_1^2(X_i)+\frac{1}{1-\pi^*(X_i)}\sigma_0^2(X_i)+(\beta(X_i)-\beta)^2\right]$$

which proves the theorem.

B Auxiliary Results

Lemma 1 $\widehat{p}(x)-\widehat{\pi}(x)=O_p\left(n^{-1/2}\right)$

Proof: We will write

$$\begin{aligned}
\widehat{p}(x) &= \frac{\sum_{i=1}^n D_i 1(X_i=x)}{\sum_{i=1}^n 1(X_i=x)} \\
&= \frac{\sum_{i=1}^{n_1} D_i 1(X_i=x)+\sum_{i=n_1+1}^n D_i 1(X_i=x)}{\sum_{i=1}^n 1(X_i=x)} \\
&= \frac{\sum_{i=1}^{n_1} 1(U_i\le\pi_1)1(X_i=x)+\sum_{i=n_1+1}^n 1(U_i\le\widehat{\pi}_2(x))1(X_i=x)}{\sum_{i=1}^n 1(X_i=x)} \\
&= \frac{n_1}{n}\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}1(U_i\le\pi_1)1(X_i=x)}{\frac{1}{n}\sum_{i=1}^n 1(X_i=x)}+\frac{n_2}{n}\frac{\frac{1}{n_2}\sum_{i=n_1+1}^{n}1(U_i\le\widehat{\pi}_2(x))1(X_i=x)}{\frac{1}{n}\sum_{i=1}^n 1(X_i=x)} \\
&= \kappa\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}1(U_i\le\pi_1)1(X_i=x)}{\frac{1}{n}\sum_{i=1}^n 1(X_i=x)} \\
&\quad +(1-\kappa)\frac{\frac{1}{n_2}\sum_{i=n_1+1}^{n}1(U_i\le\widehat{\pi}_2(x))1(X_i=x)}{\frac{1}{n}\sum_{i=1}^n 1(X_i=x)} \qquad (7)
\end{aligned}$$

By the law of large numbers and central limit theorem, we would have

$$\begin{aligned}
\frac{1}{n_1}\sum_{i=1}^{n_1}1(U_i\le\pi_1)1(X_i=x) &= E[1(U_i\le\pi_1)1(X_i=x)]+O_p\left(\frac{1}{\sqrt{n_1}}\right) \\
&= \pi_1\Pr(X_i=x)+O_p\left(\frac{1}{\sqrt{n}}\right) \qquad (8)
\end{aligned}$$

In order to deal with the second component on the far RHS of (7), we define the empirical process

$$\xi_2(\cdot,\pi_2):=\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\left(1(U_i\le\pi_2(x))1(X_i=x)-E[1(U_i\le\pi_2(x))1(X_i=x)]\right)$$

The set of functions $\{1(U_i \leq \pi_2(x))\, 1(X_i = x)\}$indexed by $\pi_2(x)$ is Euclidean, and satisfy stochastic equicontinuity. We therefore have $\xi_2(\cdot, \widehat{\pi}_2) = \xi_2(\cdot, \pi_2^*) + o_p(1)$, or

$$\begin{aligned}\frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} 1(U_i \leq \widehat{\pi}_2(x))\, 1(X_i = x) &= \frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} 1(U_i \leq \pi_2^*(x))\, 1(X_i = x) \\ &\quad + G_2 \sqrt{n_2}\,(\widehat{\pi}_2(x) - \pi_2^*(x)) \end{aligned} \tag{9}$$

where

$$G_2 := \frac{\partial}{\partial \pi_2} E\left[1(U_i \leq \pi_2(x))\, 1(X_i = x)\right]\Big|_{\pi_2(x)=\pi_2^*(x)}$$

Because $E[1(U_i \leq \pi_2(x))\, 1(X_i = x)] = \pi_2(x) \Pr(X_i = x)$, we have $G_2 = \Pr(X_i = x)$, and hence,

$$G_2 \sqrt{n_2}\,(\widehat{\pi}_2(x) - \pi_2^*(x)) = O_p(1) \tag{10}$$

as long as $\widehat{\pi}_2(x)$ is chosen to be the $\sqrt{n}$-consistent estimator of $\pi_2^*(x)$. We also have

$$\frac{1}{n_2} \sum_{i=n_1+1}^{n} 1(U_i \leq \pi_2^*(x))\, 1(X_i = x) = \pi_2^*(x)\, E[1(X_i = x)] + O_p\left(\frac{1}{\sqrt{n}}\right) \tag{11}$$

by the law of large numbers and CLT. Combining (9), (10), and (11), we obtain

$$\frac{1}{n_2} \sum_{i=n_1+1}^{n} 1(U_i \leq \widehat{\pi}_2(x))\, 1(X_i = x) = \pi_2^*(x) \Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right) \tag{12}$$

Now note that, by the law of large numbers and CLT, we have

$$\frac{1}{n} \sum_{i=1}^{n} 1(X_i = x) = \Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right) \tag{13}$$

Combining (7), (8), (12), and (13), we obtain

$$\begin{aligned}\widehat{p}(x) &= \kappa \frac{\pi_1 \Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)}{\Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)} + (1-\kappa) \frac{\pi_2^*(x) \Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)}{\Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right)} \\ &= \kappa\pi_1 + (1-\kappa)\,\pi_2^*(x) + O_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

Therefore, as long as $\widehat{\pi}_2(x)$ is chosen to be the $\sqrt{n}$-consistent estimator of $\pi_2^*(x)$, we will have $\widehat{p}(x) = \widehat{\pi}_2(x) + O_p(1/\sqrt{n})$. □

Lemma 2 $r_1(x) - \widehat{r}_1(x) = O_p\left(n^{-1/2}\right)$, $r_0(x) - \widehat{r}_0(x) = O_p\left(n^{-1/2}\right)$

Proof: We only prove that $r_1(x) - \widehat{r}_1(x) = O_p\left(n^{-1/2}\right)$. The proof of the other equality is similar, and omitted. Our proof is based on the equality

$$\begin{aligned}
\widehat{r}_1(x) &= \frac{\sum_{i=1}^n D_i Y_i 1(X_i = x)}{\sum_{i=1}^n D_i 1(X_i = x)} \\
&= \frac{\sum_{i=1}^{n_1} D_i Y_i 1(X_i = x) + \sum_{i=n_1+1}^n D_i Y_i 1(X_i = x)}{\sum_{i=1}^{n_1} D_i 1(X_i = x) + \sum_{i=n_1+1}^n D_i 1(X_i = x)} \\
&= \frac{\frac{n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) Y_i 1(X_i = x) + \frac{n_2}{n}\frac{1}{n_2}\sum_{i=n_1+1}^n 1(U_i \leq \widehat{\pi}_2(x)) Y_i 1(X_i = x)}{\frac{n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) 1(X_i = x) + \frac{n_2}{n}\frac{1}{n_2}\sum_{i=n_1+1}^n 1(U_i \leq \widehat{\pi}_2(x)) 1(X_i = x)} \\
&= \frac{\kappa\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) Y_i 1(X_i = x) + (1-\kappa)\frac{1}{n_2}\sum_{i=n_1+1}^n 1(U_i \leq \widehat{\pi}_2(x)) Y_i 1(X_i = x)}{\kappa\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) 1(X_i = x) + (1-\kappa)\frac{1}{n_2}\sum_{i=n_1+1}^n 1(U_i \leq \widehat{\pi}_2(x)) 1(X_i = x)}
\end{aligned}$$

We take care of the numerator first. We note that

$$\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) Y_i 1(X_i = x) = E\left[1(U_i \leq \pi_1) Y_i 1(X_i = x)\right] + O_p\left(\frac{1}{\sqrt{n}}\right)$$

by the law of large numbers and central limit theorem. Because

$$\begin{aligned}
E\left[1(U_i \leq \pi_1) Y_i 1(X_i = x)\right] &= \pi_1 E\left[Y_{1i} | X_i = x\right] \Pr(X_i = x) \\
&= \pi_1 r_1(x) \Pr(X_i = x),
\end{aligned}$$

we obtain

$$\frac{1}{n_1}\sum_{i=1}^{n_1} 1(U_i \leq \pi_1) Y_i 1(X_i = x) = \pi_1 r_1(x) \Pr(X_i = x) + O_p\left(\frac{1}{\sqrt{n}}\right). \tag{14}$$

In order to deal with $\frac{1}{n_2}\sum_{i=n_1+1}^n 1(U_i \leq \widehat{\pi}_2(x)) Y_i 1(X_i = x)$, we note that the set of functions $\{1(U_i \leq \pi_2(x)) Y_i 1(X_i = x)\}$ indexed by $\pi_2(x)$ is Euclidean, and satisfy stochastic equicontinuity. We therefore have

$$\begin{aligned}
\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^n 1(U_i \leq \widehat{\pi}_2(x)) Y_i 1(X_i = x) &= \frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^n 1(U_i \leq \pi_2^*(x)) Y_i 1(X_i = x) \\
&\quad + G_3\sqrt{n_2}\left(\widehat{\pi}_2(x) - \pi_2^*(x)\right)
\end{aligned}$$

where

$$G_3 := \frac{\partial}{\partial \pi_2} E\left[1(U_i \leq \pi_2(x)) Y_i 1(X_i = x)\right]\Big|_{\pi_2(x) = \pi_2^*(x)}$$

Because

$$\begin{aligned} E\left[1\left(U_i \leq \pi_2\left(x\right)\right) Y_i 1\left(X_i = x\right)\right] &= \pi_2\left(x\right) E\left[Y_{1i} \middle| X_i = x\right] \Pr\left(X_i = x\right) \\ &= \pi_2\left(x\right) r_1\left(x\right) \Pr\left(X_i = x\right), \end{aligned}$$

we have $G_3 = r_1\left(x\right) \Pr\left(X_i = x\right)$, and hence, $G_3\sqrt{n_2}\left(\widehat{\pi}_2\left(x\right) - \pi_2^*\left(x\right)\right) = O_p\left(1\right)$ as long as $\widehat{\pi}_2\left(x\right)$ is chosen to be a $\sqrt{n}$-consistent estimator of $\pi_2^*\left(x\right)$. We also have

$$\frac{1}{n_2}\sum_{i=n_1+1}^{n} 1\left(U_i \leq \pi_2^*\left(x\right)\right) Y_i 1\left(X_i = x\right) = \pi_2^*\left(x\right) r_1\left(x\right) \Pr\left(X_i = x\right) + O_p\left(\frac{1}{\sqrt{n}}\right)$$

by the law of large numbers and the central limit theorem. We may therefore conclude that

$$\frac{1}{n_2}\sum_{i=n_1+1}^{n} 1\left(U_i \leq \widehat{\pi}_2\left(x\right)\right) Y_i 1\left(X_i = x\right) = \pi_2^*\left(x\right) r_1\left(x\right) \Pr\left(X_i = x\right) + O_p\left(\frac{1}{\sqrt{n}}\right) \tag{15}$$

Combining (14) and (15), we obtain

$$\begin{aligned} &\kappa\frac{1}{n_1}\sum_{i=1}^{n_1} 1\left(U_i \leq \pi_1\right) Y_i 1\left(X_i = x\right) + \left(1-\kappa\right)\frac{1}{n_2}\sum_{i=n_1+1}^{n} 1\left(U_i \leq \widehat{\pi}_2\left(x\right)\right) Y_i 1\left(X_i = x\right) \\ &= \kappa\pi r_1\left(x\right) \Pr\left(X_i = x\right) + \left(1-\kappa\right)\pi_2^*\left(x\right) r_1\left(x\right) \Pr\left(X_i = x\right) + O_p\left(\frac{1}{\sqrt{n}}\right) \\ &= \pi^*\left(x\right) r_1\left(x\right) \Pr\left(X_i = x\right) + O_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

We can take care of the denominator in a similar manner, and obtain

$$\begin{aligned} &\kappa\frac{1}{n_1}\sum_{i=1}^{n_1} 1\left(U_i \leq \pi_1\right) 1\left(X_i = x\right) + \left(1-\kappa\right)\frac{1}{n_2}\sum_{i=n_1+1}^{n} 1\left(U_i \leq \widehat{\pi}_2\left(x\right)\right) 1\left(X_i = x\right) \\ = \;& \pi^*\left(x\right) \Pr\left(X_i = x\right) + O_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

and hence, we conclude that

$$\widehat{r}_1\left(x\right) = \frac{\pi^*\left(x\right) r_1\left(x\right) \Pr\left(X_i = x\right) + O_p\left(\frac{1}{\sqrt{n}}\right)}{\pi^*\left(x\right) \Pr\left(X_i = x\right) + O_p\left(\frac{1}{\sqrt{n}}\right)} = r_1\left(x\right) + O_p\left(\frac{1}{\sqrt{n}}\right)$$

□

Lemma 3

$$\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{D_i\left(Y_{1i}-r_1\left(x\right)\right)}{\widehat{\pi}\left(x\right)}1\left(X_i=x\right)=\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{1\left(U_i\leq\pi_1\right)\left(Y_{1i}-r_1\left(x\right)\right)}{\pi^*\left(x\right)}1\left(X_i=x\right)+o_p\left(1\right)$$

$$\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\frac{D_i\left(Y_{1i}-r_1\left(x\right)\right)}{\widehat{\pi}\left(x\right)}1\left(X_i=x\right)=\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\frac{1\left(U_i\leq\pi_2^*\left(x\right)\right)\left(Y_{1i}-r_1\left(x\right)\right)}{\pi^*\left(x\right)}1\left(X_i=x\right)+o_p\left(1\right)$$

$$\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{\left(1-D_i\right)\left(Y_{0i}-r_0\left(x\right)\right)}{1-\widehat{\pi}\left(x\right)}1\left(X_i=x\right)=\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\frac{1\left(U_i>\pi_1\right)\left(Y_{0i}-r_0\left(x\right)\right)}{1-\pi^*\left(x\right)}1\left(X_i=x\right)+o_p\left(1\right)$$

$$\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\frac{\left(1-D_i\right)\left(Y_{0i}-r_0\left(x\right)\right)}{1-\widehat{\pi}\left(x\right)}1\left(X_i=x\right)=\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\frac{1\left(U_i>\pi_2^*\left(x\right)\right)\left(Y_{0i}-r_0\left(x\right)\right)}{1-\pi^*\left(x\right)}1\left(X_i=x\right)+o_p\left(1\right)$$

Proof: We only prove the first two claims. The proof of the last two claims is identical, and omitted.

We first note that

$$\begin{aligned}
&\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\sum_x\frac{D_i\left(Y_{1i}-r_1\left(x\right)\right)}{\widehat{\pi}\left(x\right)}1\left(X_i=x\right)\\
=\;&\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\sum_x\frac{1\left(U_i\leq\pi_1\right)\left(Y_{1i}-r_1\left(x\right)\right)}{\pi^*\left(x\right)+O_p\left(\frac{1}{\sqrt{n}}\right)}1\left(X_i=x\right)\\
=\;&\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\sum_x\frac{1\left(U_i\leq\pi_1\right)\left(Y_{1i}-r_1\left(x\right)\right)}{\pi^*\left(x\right)}1\left(X_i=x\right)+O_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

as long as $\widehat{\pi}_2\left(x\right)$ is chosen to be a $\sqrt{n}$-consistent estimator of $\pi_2^*\left(x\right)$, and the latter is an interior point of $(0,1)$, which proves the first claim.

In order to prove the second claim, we define the empirical process

$$\nu_2\left(\cdot,\pi_2\right):=\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n}\left(\frac{D_i\left(Y_{1i}-r_1\left(x\right)\right)}{\pi\left(x\right)}1\left(X_i=x\right)-E\left[\frac{D_i\left(Y_{1i}-r_1\left(x\right)\right)}{\pi\left(x\right)}1\left(X_i=x\right)\right]\right)$$

where $\pi\left(x\right)=\kappa\pi_1+\left(1-\kappa\right)\pi_2\left(x\right)$. Recall that $D_i=1\left(U_i\leq\pi_1\right)$ for the first sample, and $D_i=1\left(U_i\leq\widehat{\pi}_2\left(X_i\right)\right)$ for the second sample. Because the sets of functions

$$\left\{\frac{1\left(U_i\leq\pi_2\left(x\right)\right)D_i\left(Y_{1i}-r_1\left(x\right)\right)}{\kappa\pi_1+\left(1-\kappa\right)\pi_2\left(x\right)}1\left(X_i=x\right)\right\}$$

indexed by $\pi_2\left(x\right)$ is Euclidean, we can use stochastic equicontinuity, and conclude that $\nu_2\left(\cdot,\widehat{\pi}_2\right)=$

$\nu_2(\cdot, \pi_2^*) + o_p(1)$, or

$$\frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} \frac{1(U_i \leq \widehat{\pi}_2(x))(Y_{1i} - r_1(x))}{\widehat{\pi}(x)} 1(X_i = x)$$
$$= \frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} \frac{1(U_i \leq \pi_2^*(x))(Y_{1i} - r_1(x))}{\pi^*(x)} 1(X_i = x) + F_2 \sqrt{n_2}(\widehat{\pi}_2 - \pi_2^*) + o_p(1)$$

where

$$F_2 = \frac{\partial}{\partial \pi_2} E\left[\frac{1(U_i \leq \pi_2(x))(Y_{1i} - r_1(x))}{\kappa\pi_1 + (1-\kappa)\pi_2(x)} 1(X_i = x)\right]\Big|_{\pi_2 = \pi_2^*}$$

Because U_i is independent of (X_i, Y_{1i}, Y_{0i}), we have

$$E\left[\frac{1(U_i \leq \pi_2(x))(Y_{1i} - r_1(x))}{\kappa\pi_1 + (1-\kappa)\pi_2(x)} 1(X_i = x)\right] = 0$$

regardless of the value of $\pi(x)$. This implies that the derivative F_2 is identically zero, from which the validity of the second claim follows. □

Lemma 4

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{D_i(Y_i - r_1(X_i))}{\widehat{\pi}(X_i)} - \frac{(1-D_i)(Y_i - r_0(X_i))}{1 - \widehat{\pi}(X_i)}\right)$$
$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{D_i^*(Y_{1i} - r_1(X_i))}{\pi^*(X_i)} - \frac{(1-D_i^*)(Y_{0i} - r_0(X_i))}{1 - \pi^*(X_i)}\right) + o_p(1)$$

Proof: Write

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{D_i(Y_i - r_1(X_i))}{\widehat{\pi}(X_i)} = \sum_x \left(\frac{\sqrt{n_1}}{\sqrt{n}} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n} \left(\frac{D_i(Y_{1i} - r_1(x))}{\widehat{\pi}(x)} 1(X_i = x)\right)\right)$$
$$+ \sum_x \left(\frac{\sqrt{n_2}}{\sqrt{n}} \frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} \left(\frac{D_i(Y_{1i} - r_1(x))}{\widehat{\pi}(x)} 1(X_i = x)\right)\right)$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(1-D_i)(Y_i - r_0(X_i))}{1 - \widehat{\pi}(X_i)} = \sum_x \left(\frac{\sqrt{n_1}}{\sqrt{n}} \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n} \left(\frac{(1-D_i)(Y_i - r_0(X_i))}{1 - \widehat{\pi}(X_i)} 1(X_i = x)\right)\right)$$
$$+ \sum_x \left(\frac{\sqrt{n_2}}{\sqrt{n}} \frac{1}{\sqrt{n_2}} \sum_{i=n_1+1}^{n} \left(\frac{(1-D_i)(Y_i - r_0(X_i))}{1 - \widehat{\pi}(X_i)} 1(X_i = x)\right)\right)$$

The conclusion then follows by using Lemma 3. □

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B., 1996, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association* 91(434).

Chaloner, K., and Verdinelli, I., 1995, "Bayesian Experimental Design: A Review," *Statistical Science* 10(3), 273-304.

Chamberlain, G. 1986, "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics* 32, 189-218.

Flores-Lagunes, A., Gonzalez, A., and Neumann, T., 2006, "Learning But Not Earning? The Impact of Job Corps Training for Hispanics," working paper, University of Arizona.

Green, Donald and Gerber, Alan, 2004, "Get Out the Vote! How to Increase Voter Turnout," Washington, DC: Brookings Institution Press.

Hahn, J., 1998, "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66(2), 315-331.

Hirano, K., Imbens, G. W., and Ridder, G., 2003, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71(4), 1161-1189.

Imbens, G. W., and Angrist, J. D., 1994, "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62(2): 467-475.

Johnson and Simester, 2006, "Dynamic Catalog Mailing Policies," *Management Science* 52(5): 683-696.

Karlan, D., and List, J., 2007, "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment," forthcoming, *American Economic Review*.

Karlan, D. S., and Zinman, J., 2006, "Credit Elasticities in Less-Developed Economies: Implications for Microfinance," working paper, Yale University.

Neyman, J., 1934, "On the Two Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, Series A, 97, 558-625.

Rosenbaum, P. R., and Rubin, D. B., 1983, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70(1): 41-55.

Solomon, H., and Zacks, S., 1970, "Optimal Design of Sampling from Finite Populations: A Critical Review and Indication of New Research Areas," *Journal of the American Statistical Association*, 65(330), 653-677.

Sukhatme, P. V., 1935, "Contributions to the Theory of the Representative Method," *Journal of the Royal Statistical Society*, Supplement 2, 253-268.