**Research Article**

Samir Khan* and Johan Ugander

# Adaptive normalization for IPW estimation

https://doi.org/10.1515/jci-2022-0019
received March 17, 2022; accepted October 28, 2022

**Abstract:** Inverse probability weighting (IPW) is a general tool in survey sampling and causal inference, used in both Horvitz–Thompson estimators, which normalize by the sample size, and Hájek/self-normalized estimators, which normalize by the sum of the inverse probability weights. In this work, we study a family of IPW estimators, first proposed by Trotter and Tukey in the context of Monte Carlo problems, that are normalized by an affine combination of the sample size and a sum of inverse weights. We show how selecting an estimator from this family in a data-dependent way to minimize asymptotic variance leads to an iterative procedure that converges to an estimator with connections to regression control methods. We refer to such estimators as adaptively normalized estimators. For mean estimation in survey sampling, the adaptively normalized estimator has asymptotic variance that is never worse than the Horvitz–Thompson and Hájek estimators. Going further, we show that adaptive normalization can be used to propose improvements of the augmented IPW (AIPW) estimator, average treatment effect (ATE) estimators, and policy learning objectives. Appealingly, these proposals preserve both the asymptotic efficiency of AIPW and the regret bounds for policy learning with IPW objectives, and deliver consistent finite sample improvements in simulations for all three of mean estimation, ATE estimation, and policy learning.

**Keywords:** inverse probability weighting, Horvitz–Thompson, Hájek, ATE estimation, survey sampling

**MSC 2020:** 62F12

## 1 Introduction

Consider the problem of estimating a mean from samples that are observed with nonuniform probabilities. Formally, we want to estimate the mean $\mu$ of a set of responses $Y_1, \ldots, Y_n$, but only observe $Y_1 I_1, \ldots, Y_n I_n$, where $I_k$ is an indicator of whether unit $k$ was observed. This problem is a fundamental primitive of many problems in survey sampling, causal inference, and beyond. We focus on the case where the $I_k$ are independent and distributed as $I_k \sim \mathrm{Ber}(p_k)$, which can correspond to a randomized experiment under a Bernoulli design with known $p_k$ or to certain observational contexts.

The standard estimators for $\mu$ are the Horvitz–Thompson and Hájek estimators [1,2], both of which are based on the idea of inverse probability weighting (IPW). To introduce these estimators, we first define

$$\hat{S} = \sum_{k=1}^{n} \frac{Y_k I_k}{p_k} \quad \text{and} \quad \hat{n} = \sum_{k=1}^{n} \frac{I_k}{p_k}$$

as estimates of the population total and sample size. Then, the Horvitz–Thompson and Hájek estimators are

$$\hat{\mu}_{\mathrm{HT}} = \hat{S}/n \quad \text{and} \quad \hat{\mu}_{\mathrm{Hájek}} = \hat{S}/\hat{n}.$$

*** Corresponding author: Samir Khan,** Department of Statistics, Stanford University, Stanford, CA 94305, United States, e-mail: samirk@stanford.edu
**Johan Ugander:** Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, United States, e-mail: jugander@stanfod.edu

These estimators have several desirable properties: $\hat{\mu}_{\text{HT}}$ is unbiased and admissible in the class of all unbiased estimators [3], while $\hat{\mu}_{\text{Hájek}}$ is approximately unbiased and often has lower variance than $\hat{\mu}_{\text{HT}}$ [4].

The idea of IPW also figures prominently in the Monte Carlo literature on importance sampling (IS), where the Horvitz–Thompson and Hájek estimators are known as the IS and self-normalized importance sampling estimators, respectively. In 1954, working in this context, Trotter and Tukey [5] briefly entertained the idea of a family of estimators that generalizes $\hat{\mu}_{\text{HT}}$ and $\hat{\mu}_{\text{Hájek}}$, namely,

$$\hat{\mu}_\lambda = \frac{\hat{S}}{(1-\lambda)n + \lambda\hat{n}}, \tag{1}$$

for $\lambda \in \mathbb{R}$. This family contains Horvitz–Thompson estimator as the special case $\lambda = 0$ and Hájek as the special case $\lambda = 1$. Curiously, Trotter and Tukey emphasized that $\lambda$ need not be constrained to $[0, 1]$ and that values outside that range might sometimes be useful [5].

## 1.1 Main contributions

Although this proposal first appeared nearly 70 years ago, it has not received any significant attention in either the Monte Carlo literature or the causal inference literature. In this article, we consider this proposal in detail. In particular:

- We propose a method for iteratively selecting data-dependent values of $\lambda$ in (1) to minimize asymptotic variance. We refer to this as *adaptive normalization* and show in Section 3 how our proposal leads to an estimator that improves on both the Horvitz–Thompson and Hájek estimators in terms of asymptotic variance. We also study our estimator when propensities are estimated rather than known exactly and demonstrate connections between our estimator and other ideas from the causal inference literature.
- As applications, we consider a series of problems where the Horvitz–Thompson and Hájek estimators are traditionally used as a primitive, and evaluate the use of adaptively normalized estimators instead. Specifically, we develop a novel extension of the augmented IPW (AIPW) estimator of ref. [6], new estimators for average treatment effect (ATE) estimation, and a new objective for policy learning. We show that our proposals consistently preserve desirable theoretical guarantees while improving performance in simulations.

## 1.2 Motivation for adaptive normalization

Why would we want to use values of $\lambda$ learned from the data in (1) rather than $\lambda = 0$ or $\lambda = 1$, especially values of $\lambda < 0$ or $\lambda > 1$? As a starting point, consider $\lambda = 0$, the Horvitz–Thompson estimator, and $\lambda = 1$, the Hájek estimator. As mentioned earlier, the Hájek estimator often has lower variance than the Horvitz–Thompson estimator. One reason for this variance reduction is that the numerator and denominator of the Hájek estimator are positively correlated, while the numerator and denominator of the Horvitz–Thompson estimator are uncorrelated.

To see the role that this correlation plays, consider a toy example with $n = 10$ units with responses and response probabilities

$$Y_1 = Y_2 = \cdots = Y_{10} = 1, \quad p_1 = 10^{-5}, \quad p_2 = \cdots = p_{10} = 0.5.$$

Then, we claim that the Horvitz–Thompson and Hájek estimators have very different behaviors on the event that $Y_1$ is observed. For our illustration, assume units 2–5 are observed but 6–10 are not.

For the Horvitz–Thompson estimator, if $Y_1$ is observed, then the numerator of the estimator increases from 8 to $8 + 1/10^{-5} \approx 10^5$, while the denominator, $n$, remains fixed at 10, so our estimate increases from $8/10$ to $(8 + 10^5)/10 \approx 10^4$. For the Hájek estimator, if $Y_1$ is observed, the numerator similarly increases by $10^5$, but now the denominator also increases by $10^5$, and so our estimate is less affected. The positive correlation

between the numerator and denominator of the Hájek estimator provides a kind of shrinkage that serves to reduce variance.

Since the covariance between numerator and denominator in the Hájek estimator is one way that it reduces variance, we would expect that we can further reduce variance by introducing more covariance, which motivates taking $\lambda > 1$ in (1). By increasing the weight on the random factor of $\hat{n}$ in the denominator of (1), we increase the covariance between the numerator and the denominator, and thus reduce variance. Put crudely, if going from $\lambda = 0$ to $\lambda = 1$ reduces variance, couldn't going from $\lambda = 1$ to $\lambda = 2$ reduce it even more?

Of course, there is a bias-variance trade-off. Increasing $\lambda$ decreases the variance of $\hat{\mu}_\lambda$, but increasing $\lambda$ to be arbitrarily large also shrinks our estimates toward 0 and increases bias. Thus, the problem becomes one of choosing a value of $\lambda$ that provides the correct amount of shrinkage for a given problem. This trade-off is visualized in Figure 1, which shows the mean squared error (MSE) of (1) for a simulated problem at a range of values of $\lambda$. Ideally, we would choose $\lambda$ to minimize the curve shown in the left panel of this figure. Unfortunately, we cannot compute the exact MSE of $\hat{\mu}_\lambda$, and even if we could, choosing $\lambda$ to minimize the MSE would lead to an asymptotically biased estimator, complicating issues of inference. (This is analogous to the problem of constructing confidence intervals in nonparametric regression after choosing an MSE-optimal bandwidth [7,8].) In this work, we instead take the approach of showing that $\hat{\mu}_\lambda$ is asymptotically normal and then choose $\lambda$ to iteratively minimize estimates of the asymptotic variance.

This approach will also nearly minimize the MSE of $\hat{\mu}_\lambda$ as long the bias of $\hat{\mu}_\lambda$ is not too large, in which the case the variance and MSE will be almost equal; this is indeed the case in the setting of Figure 1. In general, the bias of $\hat{\mu}_\lambda$ is of order $1/n$ and should not be a concern in reasonably large sample sizes. However, if the sample size is small, there is a possibility that choosing $\lambda$ according to our scheme will be unfavorable in an root-mean squared error (RMSE) sense because it does not account for bias. A second-order Taylor expansion of $\hat{\mu}_\lambda$ shows that the bias is controlled by the parameters $T$ and $\pi$ defined in (2), and these parameters are large when $p_k$ is likely to be close to zero. Thus, in a setting with small sample sizes and many $p_k$ close to zero (i.e., with poor overlap), practitioners should be aware that our proposed estimators may offer improvements only in variance, and not necessarily in MSE. We present a further simulation study of the bias of $\hat{\mu}_\lambda$ in Appendix B.

Finally, we note that other procedures for selecting $\lambda$ based on different criteria (e.g., minimizing upper bounds of the MSE) would lead to other estimators, suggesting a general class of adaptively normalized estimators that may be worthy of further study.

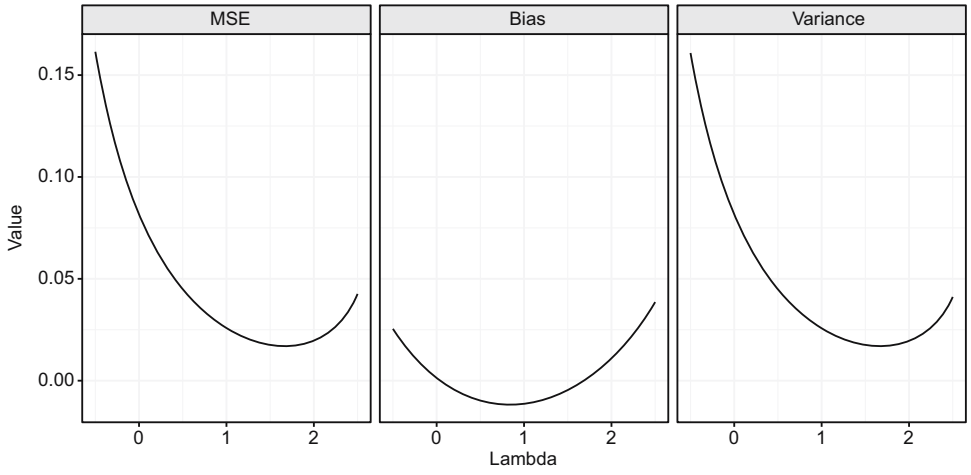


**Figure 1:** The MSE, variance, and bias of $\hat{\mu}_\lambda$ as a function of $\lambda$ in an instance of the normal model described in (18) with $n = 250$, $\mu = 1$, and $\rho = 0.2$. Note that neither the Horvitz–Thompson estimator ($\lambda = 0$) nor Hájek ($\lambda = 1$) estimator minimizes the MSE, and that the MSE-optimal choice of $\lambda$ is greater than 1. The variance shows essentially the same behavior as the MSE because the squared-bias is much smaller then the variance, and so a variance-minimizing choice of $\lambda$ is nearly the same as an MSE-minimizing choice of $\lambda$ in this example.

### 1.3 Related work

The present work is most closely related to the literature on variance reduction in IS. Since the proposal of IS in ref. [9], there have been a variety of proposals for variance reduction methods, many of which are surveyed in ref. [10]. The one most relevant to our work is the method of control variates, which is also known as difference estimation in the survey sampling community [4]. We show in Section 3 that the estimator we derive is algebraically equivalent to a particular kind of control variate estimator, and although there are multiple known interpretations of control variates [11], this adaptive normalized interpretation appears to be new. A slightly modified form of this estimator has also been briefly discussed by ref. [12] and studied in Monte Carlo simulations by ref. [13]; our work differs from theirs in focusing on the missing data model and in the development of new causal inference applications. In addition, we study this estimator in the case where the propensities $p_k$ are estimated rather than known, an issue that is unique to the causal inference context, relative earlier Monte Carlo studies.

Our proposed modifications of the AIPW estimator stand alongside many other proposals: for example, refs [14,15] use an empirical likelihood approach, while [16] uses a weighting approach, among others. We do not, however, know of any attempt or alternative derivation that is equivalent to the estimator we derive in this context. Similarly, our policy learning proposals were build directly based on the work of ref. [17], which is itself closely related to the work on counterfactual risk minimization in the bandit setting [18], and was extended by ref. [19] using ideas from AIPW estimation.

Finally, our work is also related to the literature on limited overlap in two ways. First, one strand of the limited overlap literature aims to reduce variance in the presence of small propensities, often by trimming units with extreme propensities [20,21]. Our proposed methods can offer an alternative form of variance reduction by exploiting correlations between these extreme propensities and the responses, and this approach also has the advantage of not changing the estimand as trimming-based approaches do. Second, another line of work in the limited overlap literature studies IPW estimators in a regime where the propensities are allowed to converge to 0 or 1 asymptotically [22,23]. We do not consider such a regime in this work, but extending our results to this setting is an interesting future direction.

## 2 Problem formulation and notation

In this section, we formally specify our model and introduce notation. We assume that pairs $(Y_1, p_1), \ldots, (Y_n, p_n)$ are drawn i.i.d. from a super-population distribution $\mathcal{D}$ on $\mathbb{R} \times [0, 1]$ and that we observe both $p_1, \ldots, p_n$ and $Y_1 I_1, \ldots, Y_n I_n$, where $I_k$ are independent and $I_k | p_k \sim \mathrm{Ber}(p_k)$.

This model differs from previous design-based work on asymptotics of the Horvitz–Thompson and Hájek estimators in assuming a super-population model for the $Y_k$ rather than modeling $Y_k$ as a sequence of fixed finite populations [24,25]. However, we believe that our results are still true in a finite population model, with slightly modified proofs, and we choose to work in a super-population model mainly to avoid making many cumbersome assumptions about the existence of limiting moments of the $Y_k$. In particular, although we are working in a super-population model, we make no parametric assumptions on $\mathcal{D}$, such as assuming a regression model.

Our assumption that the $p_k$ are random as well is best understood in the context of a more general model that we consider in Sections 4.1 and 4.3. There, we consider pairs $(Y_1, X_1), \ldots, (Y_n, X_n)$ drawn i.i.d. from a super-population distribution $\mathcal{D}$ on $\mathbb{R} \times \mathcal{X}$, where the $Y_k$ are responses and the $X_k$ are known covariates. Then, we can think of the treatment probability $p_k$ as a function $p_k = p(X_k)$ of the observed covariates.

We assume for several of our theoretical results in Section 3 that the $p_k$ (or equivalently the map $p(\cdot)$) are known exactly or estimated from a well-specified logistic regression model. This is mainly to facilitate asymptotic comparisons, since the asymptotics of IPW estimators with estimated propensities are highly dependent on the method of estimating the propensity. A very specific choice of propensity score estimator has been shown to lead to efficient estimators [26], but we choose to study more general methods of

propensity estimation and efficiency under appropriate rate and consistency conditions in the context of AIPW estimation in Section 4.

We now specify our assumptions on $\mathcal{D}$.

**Assumption 1.** (Boundedness and overlap) There exist constants $M, \delta > 0$ such that $|Y_k| \le M$ and $\delta \le p_k \le 1 - \delta$ almost surely.

Under Assumption 1, all of the moments

$$\mu = \mathbb{E}[Y_1], \quad \pi = \mathbb{E}\left[\frac{1-p_1}{p_1}\right], \quad T = \mathbb{E}\left[Y_1\frac{1-p_1}{p_1}\right] \tag{2}$$

of $\mathcal{D}$ are all finite.

Throughout the next section, our goal is to estimate $\mu = \mathbb{E}[Y_k]$ in the model presented here. In Section 4, we consider more diverse models and goals.

# 3 Adaptive normalization

In this section, we propose and analyze a novel procedure for selecting a data-dependent value of $\lambda$ and describe properties of the adaptively normalized IPW estimator that follows. We establish that the resulting estimator has smaller asymptotic variance than either the Horvitz–Thompson estimator or Hájek estimator.

## 3.1 The optimal choice of $\lambda$

Before we can understand how to choose $\lambda$ from the data, we must first understand the behavior of $\hat{\mu}_\lambda$ for a fixed $\lambda$, as a function of $\lambda$ and the problem parameters. To this end, we contribute the following central limit theorem, which is a straightforward generalization of known results on the bias and variance of the Hájek estimator. We defer the proof this result, as well as all other results in this section, to Appendix A.

**Theorem 1.** *Suppose Assumption* 1 *holds. Then, for any fixed* $\lambda \in \mathbb{R}$, *we have the CLT*

$$\sqrt{n}(\hat{\mu}_\lambda - \mu) \xrightarrow{d} N(0, \sigma^2(\lambda)), \quad \sigma^2(\lambda) = \operatorname{var}(Y_1) + \mathbb{E}\left[\frac{1-p_1}{p_1}(Y_1 - \lambda\mu)^2\right]. \tag{3}$$

With this result in hand, we now choose $\lambda$ to minimize $\sigma^2(\lambda)$. Moving forward, in Sections 3.1 and 3.2, we assume that $\mu \neq 0$, since if $\mu = 0$, then $\sigma(\lambda)$ does not actually depend on $\lambda$, and minimizing over $\lambda$ is no longer meaningful. However, in Section 3.3 onward, when we study the properties of our estimator, we do not require $\mu \neq 0$.

Under the assumption that $\mu \neq 0$, there is a unique value of $\lambda$ that minimizes $\sigma(\lambda)^2$ in (3), given by

$$\lambda^* = \frac{\mathbb{E}\left[\frac{1-p_1}{p_1}Y_1\right]}{\mu\mathbb{E}\left[\frac{1-p_1}{p_1}\right]} = \frac{T}{\pi\mu}, \tag{4}$$

where $\pi$ and $T$ are the moments defined in (2). Assumption 1 precludes the possibility that $\pi = 0$.

We can interpret (4) to shed light on the role $\lambda$ plays. If $Y_k$ and $p_k$ are positively correlated, then $Y_k$ and $\frac{1-p_k}{p_k}$ are negatively correlated, so $T < \mu\pi$ and $\lambda^* < 1$. Similarly, if the $Y_k$ and $p_k$ are negatively correlated, we obtain $\lambda^* > 1$. This interpretation of (4) extends the conventional wisdom that Hájek estimator is preferable to Horvitz–Thompson estimator when $Y_k$ and $p_k$ are negatively correlated [4].

## 3.2 Estimating $\lambda^*$ from the data

On the basis of the aforementioned asymptotic analysis, we would seem to prefer $\hat{\mu}_{\lambda^*}$ over $\hat{\mu}_{\mathrm{HT}}$ and $\hat{\mu}_{\mathrm{Hájek}}$. However, we cannot use $\hat{\mu}_{\lambda^*}$ directly as an estimator of $\mu$ because the prescribed choice of $\lambda^*$ depends on unknown moments of $\mathcal{D}$, including the very mean $\mu$ we are trying to estimate. What happens if we estimate $\lambda^*$ from data?

Our expression of $\lambda^*$ depends on three moments of $\mathcal{D}$, namely, $T$, $\pi$, $\mu$. The first two of these can readily be estimated from the data by the IPW-style estimators

$$\hat{T} = \frac{1}{n}\sum_{k=1}^{n} \frac{1-p_k}{p_k} Y_k \frac{I_k}{p_k}, \quad \hat{\pi} = \frac{1}{n}\sum_{k=1}^{n} \frac{1-p_k}{p_k} \frac{I_k}{p_k}, \tag{5}$$

and we already have the estimator $\hat{\mu}_{\mathrm{HT}}$ of $\mu$.

Before moving forward, we address two possible questions that might be raised by (5): Can we also use an adaptively normalized estimator when estimating $T$ and $\pi$, rather than a standard IPW estimator? And why do we inverse propensity weight our estimate of $\pi$, rather than using the sample mean of the $(1-p_k)/p_k$? For the first question, any consistent estimators of $T$ and $\pi$ will suffice for our theoretical results, and we focus on $\hat{T}$ and $\hat{\pi}$ as defined in (5) for simplicity. In practice, for some estimators of $T$ and $\pi$ with lower asymptotic variance, the asymptotic variance is not a good approximation of the finite sample behavior unless $n$ is extremely large, thus worsening finite sample performance at moderate sample sizes. Similarly, we could use other estimators of $\mu$ that concentrate around $\mu$, and we focus on $\hat{\mu}_{\mathrm{HT}}$ for simplicity. For the second, the reason we estimate $\pi$ with an IPW estimator is that, to estimate $\lambda^*$, we will use expressions of the form $\hat{T}/\hat{\pi}$, and IPW estimating $\pi$ introduces correlation between the numerator and denominator of this expression, thus reducing the variance of our estimate of $\lambda^*$.

Now, the estimators in (5) lead us to estimate $\lambda^*$ and $\mu$ by

$$\hat{\lambda}^* = \frac{\hat{T}}{\hat{\pi}\hat{\mu}_{\mathrm{HT}}}, \quad \hat{\mu}_{\hat{\lambda}^*} = \frac{\hat{S}}{(1-\hat{\lambda}^*)n + \hat{\lambda}^*\hat{n}}. \tag{6}$$

At this point, however, we can make useful observation: We expect that $\hat{\mu}_{\hat{\lambda}^*}$ is a better estimator of $\mu$ than $\hat{\mu}_{\mathrm{HT}}$, so should we not use $\hat{\mu}_{\hat{\lambda}^*}$ rather than $\hat{\mu}_{\mathrm{HT}}$ when estimating $\lambda$? In fact, every time we obtain a better estimate of $\mu$, we can use this to obtain a better estimate of $\lambda^*$. On the other hand, a better estimate of $\lambda^*$ will, because $\lambda^*$ is the optimal amount of normalization, lead to a better estimate of $\mu$. Combining these ideas leads to the following alternating scheme.

Formally, we construct a sequence of estimators $(\hat{\lambda}^{(t)}, \hat{\mu}^{(t)})$ initialized at $(\hat{\lambda}^{(0)}, \hat{\mu}^{(0)}) = (0, \hat{\mu}_{\mathrm{HT}})$ and defined for $t > 0$ by the recursions

$$\hat{\lambda}^{(t)} = \frac{\hat{T}}{\hat{\pi}\hat{\mu}^{(t-1)}}, \quad \hat{\mu}^{(t)} = \frac{\hat{S}}{(1-\hat{\lambda}^{(t)})n + \hat{\lambda}^{(t)}\hat{n}}. \tag{7}$$

The first equation in (7) corresponds to estimating $\lambda^*$ using $\hat{\mu}^{(t-1)}$ as an estimate of $\mu$, while the second corresponds to estimating $\mu$ using $\hat{\lambda}^{(t)}$ as an estimate of $\lambda^*$. There are two possible stable limiting behaviors for this sequence: the first is to trivially have $\hat{\mu}^{(t)} \to 0$ and $\hat{\lambda}^{(t)} \to \infty$, while the second is to converge to a fixed point at a pair $(\hat{\mu}_{\mathrm{AN}}, \hat{\lambda}_{\mathrm{AN}})$ satisfying

$$\hat{\lambda}_{\mathrm{AN}} = \frac{\hat{T}}{\hat{\pi}\hat{\mu}_{\mathrm{AN}}}, \quad \hat{\mu}_{\mathrm{AN}} = \frac{\hat{S}}{(1-\hat{\lambda}_{\mathrm{AN}})n + \hat{\lambda}_{\mathrm{AN}}\hat{n}}.$$

This system of equations has the unique solution

$$\hat{\mu}_{\mathrm{AN}} = \frac{\hat{S}}{n} + \frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right). \tag{8}$$

The following theorem formally establishes that the iterations (7) converge to the nontrivial solution, the fixed point at (8).

**Theorem 2.** *Suppose Assumption* 1 *holds, we have* $\mu \neq 0$, *and consider the sequence of estimators* $(\hat{\lambda}^{(t)}, \hat{\mu}^{(t)})$ *initialized at* $\hat{\lambda}^{(0)} = 0$, $\hat{\mu}^{(0)} = \hat{\mu}_{\mathrm{HT}}$ *and defined for* $t > 0$ *by the recursion* (7). *Then*

(*i*) *the sequence* $\hat{\mu}^{(t)}$ *converges as* $t \to \infty$ *to an estimator* $\hat{\mu}_{\lim}$;

(*ii*) *the estimator* $\hat{\mu}_{\lim}$ *satisfies*

$$\lim_{n\to\infty} \mathbb{P}(\hat{\mu}_{\lim} = \hat{\mu}_{\mathrm{AN}}) = 1,$$

*so that* $\hat{\mu}_{\lim} - \hat{\mu}_{\mathrm{AN}}$ *converges in probability to* 0.

Thus, our attempts to develop an estimator of $\mu$ by estimating the optimal normalization parameter $\lambda^*$ culminate in the estimator $\hat{\mu}_{\mathrm{AN}}$, which we refer to as the adaptively normalized IPW estimator. The role of Theorem 2 and the iterative scheme we have presented is to make explicit the connection between adaptive normalization and the final estimator $\hat{\mu}_{\mathrm{AN}}$. The fixed-point equations alone do not characterize this connection, since, based on those equations alone, we may be led to take $\hat{\mu} = 0$ and think of this as using an infinite value of $\lambda$. Thus, this theorem fully develops the proposal of ref. [5] into a usable estimator, providing the first complete analysis of IPW estimators with normalizations other than $n$ or $\hat{n}$.

Before proceeding, we note a very attractive property of $\hat{\mu}_{\mathrm{AN}}$ in (8): its simplicity. It is, for practical purposes, the Horvitz–Thompson estimator with a correction term and can replace the Horvitz–Thompson and Hájek estimators in essentially any application where they are used, with minimal additional computation. Beyond this simplicity, we will see in Section 3.3 that $\hat{\mu}_{\mathrm{AN}}$ typically has lower asymptotic variance than either the Horvitz–Thompson estimator or the Hájek estimator, and in Section 3.4 that it is closely related to several existing ideas in the literature.

### 3.2.1 An optimization perspective

It may feel unsettled to minimize the asymptotic variance and then commence an iteration scheme. We can also derive $\hat{\mu}_{\mathrm{AN}}$ more directly by framing the problem of selecting $\lambda$ as a joint minimization, over $\mu$ and $\lambda$, of the asymptotic variance. Minimizing the asymptotic variance $\sigma_\lambda^2$ in (3) is equivalent to minimizing

$$-2\lambda\mu\mathbb{E}\left[\frac{1-p_k}{p_k}Y_k\right] + \lambda^2\mu^2\mathbb{E}\left[\frac{1-p_k}{p_k}\right] = -2\lambda\mu T + \lambda^2\mu^2\pi.$$

As mentioned earlier, we use $\hat{T}$ and $\hat{\pi}$ defined in (5) to estimate $T$ and $\pi$, but now we estimate $\mu$ by $\hat{\mu}_\lambda$ directly. This leads to the nonconvex optimization problem:

$$\min_\lambda \; -2\lambda\hat{\mu}_\lambda\hat{T} + \lambda^2\hat{\mu}_\lambda^2\hat{\pi}.$$

Rather than directly solving this optimization problem, we consider the equivalent problem

$$\begin{aligned} &\min_{\lambda,\hat{\mu}} \; -2\lambda\hat{\mu}\hat{T} + \lambda^2\hat{\mu}^2\hat{\pi}, \\ &\text{s.t.} \;\; \hat{\mu} = \frac{\hat{S}}{(1-\lambda)\hat{n} + \lambda n}. \end{aligned} \tag{9}$$

This problem is also nonconvex, but it is more amenable to analysis. In particular, we claim that, despite its nonconvexity, (9) can be easily solved analytically to find a unique optimum.

To solve (9), note that the objective is a quadratic function of $\lambda\hat{\mu}$, and so checking first-order stationarity conditions shows that the unconstrained minimum is achieved for any pair $(\lambda, \hat{\mu})$ satisfying $\lambda\hat{\mu} = \hat{T}/\hat{\pi}$. Then, direct algebra yields that there is a unique pair $(\hat{\lambda}_{\mathrm{opt}}, \hat{\mu}_{\mathrm{opt}})$ satisfying both $\hat{\lambda}_{\mathrm{opt}}\hat{\mu}_{\mathrm{opt}} = \hat{T}/\hat{\pi}$ as well as the constraint in (9), and the resulting $\hat{\mu}_{\mathrm{opt}}$ is precisely $\hat{\mu}_{\mathrm{AN}}$.

Finally, we note that it is possible to relax the constraint in (9) to merely require that $\hat{\mu}$ is unbiased, rather than that it has a particular functional form. In this case, the resulting problem would be nonconvex, but solving a relaxation or upper bound may lead to other interesting estimators.

## 3.3 Properties of adaptive normalization

We now study the asymptotic behavior of $\hat{\mu}_{\text{AN}}$, and show that its asymptotic variance generically improves on the asymptotic variance of the Hájek and Horvitz–Thompson estimators.

### 3.3.1 Asymptotic variance

To understand the asymptotics of

$$\hat{\mu}_{\text{AN}} = \frac{\hat{S}}{n} + \frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right),$$

we use the fact that $\hat{T}/\hat{\pi}$ is consistent for $T/\pi$ and then apply a CLT to the remaining terms. This argument produces the following result.

**Theorem 3.** *Under Assumption* 1, *and regardless of whether* $\mu = 0$, *the estimator* $\hat{\mu}_{\text{AN}}$ *satisfies the CLT*

$$\sqrt{n}(\hat{\mu}_{\text{AN}} - \mu) \xrightarrow{d} N(0, \sigma^2(\lambda^*)), \tag{10}$$

*where* $\sigma^2(\cdot)$ *is the asymptotic variance of* $\hat{\mu}_\lambda$ *from* (3). *Furthermore, the asymptotic variance in* (10) *is always smaller than the asymptotic variances of* $\hat{\mu}_{\text{HT}}$ *and* $\hat{\mu}_{\text{Hájek}}$ *and is strictly smaller except if* $T = \mu\pi$ *or* $T = 0$.

We defer the proof of the CLT to Appendix A, but note here that the variance comparison follows from the fact that $\lambda^*$ is the minimizer of $\sigma^2(\lambda)$. In particular, unless $\lambda^* = 1$, which corresponds to $T = \mu\pi$, or $\lambda^* = 0$, which corresponds to $T = 0$, $\sigma^2(\lambda^*)$ is smaller than both of $\sigma^2(1)$ and $\sigma^2(0)$. Thus, $\hat{\mu}_{\text{AN}}$ improves on $\hat{\mu}_{\text{HT}}$ and $\hat{\mu}_{\text{Hájek}}$ unless they were already using the optimal value of $\lambda$.

We also note that, since Theorem 3 does not require $\mu \neq 0$, there should be no concerns about using the adaptively normalized estimator in problems where the true mean may be zero. This assumption was required for connecting the iterations (7) to $\hat{\mu}_{\text{AN}}$, but the favorable variance properties of $\hat{\mu}_{\text{AN}}$ do not rely on it. Thus, even in problems where the true mean may be zero, we recommend using $\hat{\mu}_{\text{AN}}$, since it will be equivalent to other IPW estimators asymptotically if the true mean is zero, and lower variance asymptotically if the true mean is not zero.

### 3.3.2 Finite sample variance

Examining the form of $\hat{\mu}_{\text{AN}}$ directly, we can see that if it has lower variance than $\hat{\mu}_{\text{HT}}$ in finite samples, it must be because the correction term $\frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right)$ is negatively correlated with the first term, $\hat{S}/n = \hat{\mu}_{\text{HT}}$. However, by using the iterative scheme introduced in (7), we can give an alternative explanation for why the finite sample variance of $\hat{\mu}_{\text{AN}}$ is smaller than that of $\hat{\mu}_{\text{HT}}$, one that builds on the result of Theorem 2.

To make this argument, we first combine the two update equations in (7) into the single equation

$$\hat{\mu}^{(t)} = \frac{\hat{S}}{\left(1 - \frac{\hat{T}}{\hat{\pi}\hat{\mu}^{(t-1)}}\right)n + \frac{\hat{T}}{\hat{\pi}\hat{\mu}^{(t-1)}}\hat{n}} = \frac{\hat{S}/n}{1 - \frac{\hat{T}}{\hat{\pi}\hat{\mu}^{(t-1)}}\left(1 - \frac{\hat{n}}{n}\right)}.$$

We can write this more succinctly as $\hat{\mu}^{(t)} = f(\hat{\mu}^{(t-1)})$, where

$$f(x) = \frac{ax}{x - b}, \quad a = \frac{\hat{S}}{n}, \quad b = \frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right).$$

In this notation, the fixed point of $f$ is at $x = a + b = \hat{\mu}_{\mathrm{AN}}$. Rather than working with $f$ directly, we will work with the two-step map $g(x) = f(f(x)) = \frac{a^2 x}{x(a-b)+b^2}$ and the sub-sequence $x^{(0)}, x^{(2)}, \ldots$.

We will now argue informally that the iterations of $g$ should be variance reducing. If it were the case that $|g'(x)| \leq 1$ for all $x$, then applying $g$ would certainly reduce variance, and we would conclude that $\hat{\mu}_{\mathrm{AN}}$ has smaller variance than $\hat{\mu}_{\mathrm{HT}}$. But this is not the case: $g(x)$ approaches $\infty$ as its denominator approaches 0, and so $|g'(x)|$ can be arbitrarily large . However, the actual sequence of iterates $\hat{\mu}^{(2t)}$ lies, with high probability, in an interval where $|g'(x)|$ is bounded by 1.

In what follows, we assume again that $\mu \neq 0$ and that $|a| > 2|b|$, which we show holds with high probability by the same arguments as those in the proof of Lemma A3 of Appendix A. We also assume, without the loss of generality, that both $a$ and $b$ are positive, but this is only for clarity. Now, in the proof of Lemma A2, we show that under these conditions, the entire sequence of iterates $\hat{\mu}^{(0)}, \hat{\mu}^{(2)}, \ldots$ lies in the interval $[a, a + 2b]$. By standard concentration arguments, $a$ is concentrated around $\mu$ and $b$ is concentrated around 0, so the interval between $[a, a + 2b]$ lies with high probability in a small interval $\mathcal{I}$ centered at $\mu$. On the other hand, we can check by direct computation that $|g'(x)| \leq 1$ for $x$ outside the interval $(-3b, b)$. Again by concentration arguments, for large $n$, this interval will be concentrated around zero and thus be disjoint from the interval $\mathcal{I}$ centered at $\mu$.

To summarize, there exists an interval $\mathcal{I}$ centered around $\mu$ such that, with high probability, $x^{(2t)} \in \mathcal{I}$ for all $t > 0$ and $|g'(x)| \leq 1$ for all $x \in \mathcal{I}$. If the function $g$ were fixed, this would be enough to conclude that each application of $g$ reduces variance, and thus that $\hat{\mu}_{\mathrm{AN}}$ has smaller variance than $\hat{\mu}_{\mathrm{HT}}$. Unfortunately, because the function $g$ is random, this is not a rigorous argument, only an intuitive one. However, it is clear from simulations, the results of which are shown in Figure 2, that each successive $\hat{\mu}^{(2t)}$ does in fact have lower variance.

We note that an argument similar to ours appears in Section 5 of ref. [27]. In that context, when studying a generalized method of moments estimator, the authors iterate a random data-dependent estimate of an underlying true function. The underlying true function is a contraction mapping and thus reduces variance when applied to a random variable, so they argue heuristically that the iterations of the data-dependent estimated function should reduce variance as well.

In addition, the following theorem verifies the finite-sample variance reduction properties of $\hat{\mu}_{\mathrm{AN}}$ in the special case where $p_k$ is constant.

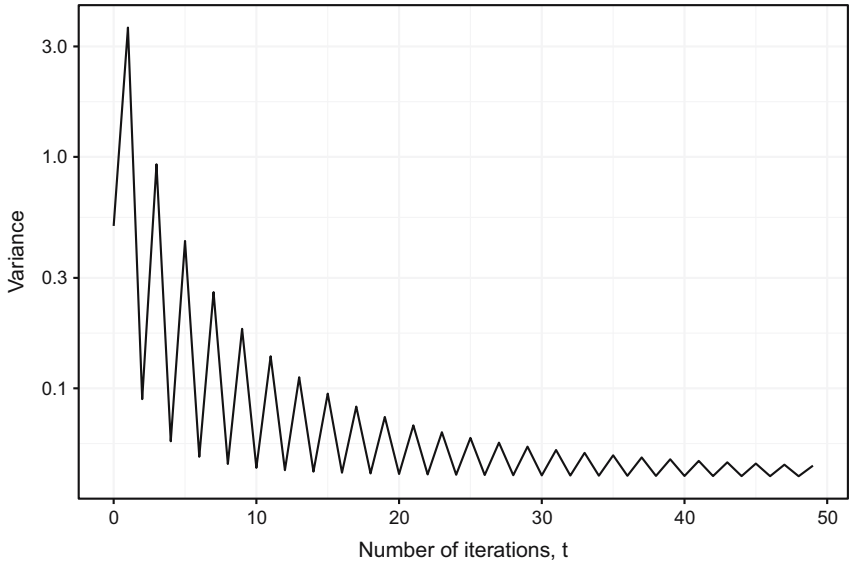


**Figure 2:** The variance of $\hat{\mu}^{(t)}$ as a function of $t$. Although a single iteration may increase the variance, we observe that, in this simulation, every two iterations reduce variance. The data here are generated from the normal model (18) in Section 5 with $n = 100$, $\rho = 0.1$, and $\mu = -1$.

**Theorem 4.** *Under Assumption* 1, *if* $p_k = p$ *is constant, then* $\mathrm{var}(\hat{\mu}_{\mathrm{AN}}) \leq \mathrm{var}(\hat{\mu}_{\mathrm{HT}})$.

We make two important comments on this result. First, in this special case, $\hat{\mu}_{\mathrm{AN}}$ is equivalent to both the sample mean of $Y_1 I_1, \ldots, Y_n I_n$ and $\hat{\mu}_{\mathrm{Hájek}}$. As such, this case is a simple model that does not fully showcase all of the subtleties of the problem, but still presents evidence for the desirable finite-sample variance properties of $\hat{\mu}_{\mathrm{AN}}$. Second, although $\hat{\mu}_{\mathrm{AN}}$ is equivalent to the observed mean in this case, Theorem 4 differs from classical results on the difference-of-means estimators for completely randomized experiments in assuming a Poisson sampled design, which, to the best of our knowledge, has not been previously considered [28].

### 3.3.3 Estimated propensities

We have been assuming until now that $p_k$ is known, or equivalently that the propensity map $p(X_k)$ for covariates $X_k$ is known. In practice, and especially in the context of observational data, $p(\cdot)$ must often be estimated. In such a setting, the results of ref. [29] imply that the semiparametric efficiency bound for estimating $\mu$ is

$$\mathbb{E}\left[\frac{\mathrm{var}(Y_1|X_1)}{p_1} + (\mathbb{E}[Y_1|X_1] - \mu)^2\right] = \mathrm{var}(Y_1) + \mathbb{E}\left[(Y_1 - \mathbb{E}[Y_1|X_1])^2 \frac{1-p_1}{p_1}\right]. \tag{11}$$

Comparing (11) with the result of Theorem 1, we see that the family $\hat{\mu}_\lambda$ cannot achieve the semiparametric lower bound, because the asymptotic variances of $\hat{\mu}_\lambda$ have a $Y_1 - \lambda\mu$ term instead of $Y - \mathbb{E}[Y_1|X_1]$ term.

However, [26] has shown that estimation of the propensities actually reduces the asymptotic variance of IPW estimators. In fact, [26] also shows that when using a linear or logistic sieve estimator for the propensity score, the Horvitz–Thompson and Hájek estimators achieve the semiparametric lower bound in (11). In this work, we do not consider such nonparametric propensity estimators, and instead focus on the case of propensities estimated from a well-specified logistic model; we consider the analysis of $\hat{\mu}_{\mathrm{AN}}$ with propensities estimated from a linear or logistic sieve, or other nonparametric estimators (to which the results of ref. [26] may be generalizable), and a study of efficiency properties like those in ref. [26] to be an interesting future direction. For users who are interested in achieving semiparametric efficiency with our methods, we apply our methods to double machine learning methods in Section 4.1 and give a semiparametrically efficient estimator there.

Now, in the case of propensities estimated from a well-specified logistic model, it is natural to ask how this variance reduction from propensity estimation interacts with the variance reduction of $\hat{\mu}_{\mathrm{AN}}$, a question we answer via the following theorem comparing the asymptotic variances of $\hat{\mu}_{\mathrm{AN}}$ and $\hat{\mu}_{\mathrm{HT}}$ in this case.

**Theorem 5.** *Suppose Assumption* 1 *holds and that the* $p_k$ *are estimated from a well-specified logistic model, i.e.,* $p_k = (1 + \exp(X_k^T\theta^*))^{-1}$, *and we estimate* $\theta^*$ *by the maximum-likelihood estimate* $\hat{\theta}$ *based on the data* $(X_1, I_1), \ldots, (X_n, I_n)$. *Let* $\hat{\mu}_{\mathrm{HT,\,logistic}}$ *and* $\hat{\mu}_{\mathrm{AN,\,logistic}}$ *be estimates of* $\mu$ *using* $\hat{\theta}$. *Then, these estimators satisfy the CLTs*:

$$\sqrt{n}(\hat{\mu}_{\mathrm{HT,\,logistic}} - \mu) \xrightarrow{d} N(0, \sigma^2_{\mathrm{HT,\,logistic}}), \quad \sqrt{n}(\hat{\mu}_{\mathrm{AN,\,logistic}} - \mu) \xrightarrow{d} N(0, \sigma^2_{\mathrm{AN,\,logistic}})$$

*and*

$$\sigma^2_{\mathrm{AN,\,logistic}} - \sigma^2_{\mathrm{HT,\,logistic}} = -\frac{T^2}{\pi} + \frac{2T}{\pi} u_2^T I(\theta^*)^{-1} u_1 - \frac{T^2}{\pi^2} u_2^T I(\theta^*)^{-1} u_2 \tag{12}$$

*where* $u_1 = \mathbb{E}[Y_k(1 - p(X_k^T\theta^*))X_k]$, $u_2 = \mathbb{E}[(1 - p(X_k^T\theta^*))X_k]$, *and* $I(\theta^*) = \mathbb{E}[(1 - p(X_k^T\theta^*))p(X_k^T\theta^*)X_k X_k^T]$ *are the Fisher information.*

The value of Theorem 5 is that it highlights that some caution is necessary when choosing between $\hat{\mu}_{\mathrm{AN}}$ and $\hat{\mu}_{\mathrm{HT}}$ in a problem with estimated propensities, and suggests an approach to making this choice by using

the right-hand side of (12) as a diagnostic quantity. That is, we can select between $\hat{\mu}_{\mathrm{HT}}$ and $\hat{\mu}_{\mathrm{AN}}$ by estimating the right-hand side of (12), and using $\hat{\mu}_{\mathrm{HT}}$ if it is positive, and $\hat{\mu}_{\mathrm{AN}}$ if it is negative. However, practically speaking, when estimating propensities and choosing between $\hat{\mu}_{\mathrm{HT}}$ and $\hat{\mu}_{\mathrm{AN}}$, we should prefer neither – rather, we should use modern doubly robust methods like the AIPW estimator. We discuss these methods, and the role of adaptive normalization in them, in Section 4.1.

## 3.4 Connections to regression/control variate methods

In this section, we show how $\hat{\mu}_{\mathrm{AN}}$ can be understood as a control variate/regression control method. We note briefly that arguments similar to the one's presented here can be used to draw connections between $\hat{\mu}_{\mathrm{AN}}$ and the AIPW estimator of ref. [6], as well as the estimator of ref. [30].

A popular technique for variance reduction in survey sampling is the use of regression controls, and this technique is also known in the Monte Carlo literature, where regression controls are referred to as control variates. We will now show that $\hat{\mu}_{\mathrm{AN}}$ is equivalent to a particular choice of regression control/control variate that is known in the Monte Carlo community, but does not appear to have been widely adopted in the survey sampling and causal inference communities.

We follow the discussion in ref. [10], where the author strongly recommends using the importance weights as control variates whenever they are known, based on the results of ref. [13]. In our setting, the random variables $w_k = I_k / p_k$ play the role of the importance weights, since they have mean 1 and re-weight the observed $Y_k I_k$ to have mean $\mu$. Then, following ref. [10], we define the family of estimators:

$$\hat{\mu}_\beta = \frac{1}{n}\sum_{k=1}^{n} Y_k w_k - \beta\left(\frac{1}{n}\sum_{k=1}^{n} w_k - 1\right). \tag{13}$$

Each estimator in this family is unbiased, so we choose $\beta$ to minimize variance. The variance of $\hat{\mu}_\beta$ is

$$\frac{1}{n}\mathrm{var}(Y_k w_k) - \frac{2\beta}{n}\mathrm{cov}(Y_k w_k, w_k) + \frac{\beta^2}{n}\mathrm{var}(w_k),$$

which is minimized for $\beta^* = \mathrm{cov}(Y_k w_k, w_k)/\mathrm{var}(w_k)$.

Direct computation gives

$$\mathrm{cov}(Y_k w_k, w_k) = \mathbb{E}\left[Y_k \frac{1-p_k}{p_k}\right] = T \quad \text{and} \quad \mathrm{var}(w_k) = \mathbb{E}\left[\frac{1-p_k}{p_k}\right] = \pi,$$

connecting the IS problem to the notation of our survey sampling problem given in (2). Thus, estimating the numerator and denominator of $\beta^*$ separately by $\hat{T}$ and $\hat{\pi}$, respectively, gives the estimator $\hat{\beta} = \hat{T}_0/\hat{\pi}_0$ of $\beta^*$. The resulting estimator of $\mu$ becomes

$$\hat{\mu}_{\hat{\beta}} = \frac{1}{n}\sum_{k=1}^{n} Y_k w_k - \frac{\hat{T}}{\hat{\pi}}\left(\frac{1}{n}\sum_{k=1}^{n} w_k - 1\right) = \hat{\mu}_{\mathrm{HT}} + \frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right),$$

which is algebraically equivalent to the estimator $\hat{\mu}_{\mathrm{AN}}$ in (8) derived via adaptive normalization. We note that the version of this estimator in refs [10,13] estimates $\mathrm{cov}(Y_k, w_k)$ and $\mathrm{var}(w_k)$ directly rather than first simplifying, a minor difference with our version.

### 3.4.1 Value of adaptive normalization as a perspective

While the adaptively normalized estimator we have derived is algebraically equivalent to a known estimator in the Monte Carlo literature, we feel that our derivation, based on the simple idea of combining the denominators of $\hat{\mu}_{\mathrm{HT}}$ and $\hat{\mu}_{\mathrm{Hájek}}$, offers a valuable and unique motivation. Furthermore, our iterative

analysis of $\hat{\mu}_{\text{AN}}$ provides instructive intuition for finite-sample variance reduction. Finally, we have focused on choosing $\lambda$ to minimize asymptotic variance – choosing $\lambda$ by minimizing a different criteria, such as mean-squared error, or choosing $\lambda$ using an alternative approach such as Lepski's method, would lead to other estimators that may be worthy of further study.

Finally, despite these other guises in which the adaptively normalized estimator has appeared, it has not received significant attention in the survey sampling or causal inference community. This inattention is especially surprising when we consider that, in the Monte Carlo setting, the importance weights are often not computable in closed form, and so they cannot be used as control variates. In contrast, in the causal inference setting, the treatment probabilities are often known from the experimental design, and so $\hat{\mu}_{\text{AN}}$ is usually available as an immediate improvement over $\hat{\mu}_{\text{HT}}$ or $\hat{\mu}_{\text{Hájek}}$. Thus, it seems that $\hat{\mu}_{\text{AN}}$ is more widely known in the community where it is of less practical value; we hope our work will remedy this gap by presenting $\hat{\mu}_{\text{AN}}$ in a causal inference context and addressing issues specific to this context, such as the estimation of propensity scores. To this end, we proceed in the next section to identify a variety of template settings in causal inference where $\hat{\mu}_{\text{HT}}$ and $\hat{\mu}_{\text{Hájek}}$ are used by default, and it is desirable to replace them with $\hat{\mu}_{\text{AN}}$.

# 4 Applications

In this section, we consider various causal inference settings in which the Horvitz–Thompson and Hájek estimators are used "by default" and show how adaptively normalized estimators act as a free upgrade.

## 4.1 AIPW estimation

Our first application of adaptive normalization focuses on AIPW estimation. A basic version of AIPW was first introduced for survey sampling in the 1980s [31,32] and then re-discovered and significantly expanded on for ATE estimation by the causal inference community [6,33], where it is also known as the doubly robust estimator. Our approach follows the one developed in ref. [33], although we concentrate on a single group for now and defer ATE estimation until further on.

We discuss AIPW estimation in the more general model where we have access to covariate information. Specifically, we assume the pairs $(Y_1, X_1), \ldots, (Y_n, X_n)$ are i.i.d. from a distribution $\mathcal{D}$ on $\mathbb{R} \times \mathcal{X}$, and we observe all of $X_1, \ldots, X_n$ in addition to $Y_1 I_1, \ldots, Y_n I_n$, where $I_k$ are independently $\text{Ber}(p_k)$ and $p_k = p(X_k)$ is a function of the covariates. Then, the AIPW estimate of $\mu$ is expressed as follows:

$$\hat{\mu}_{\text{AIPW}} = \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_j}\frac{1}{n}\hat{\mu}^{-j}(X_k) + \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_j}\frac{(Y_k - \hat{\mu}^{(-j)}(X_k))I_k}{\hat{p}^{(-j)}(X_k)}, \tag{14}$$

where $\mathcal{I}_1, \ldots, \mathcal{I}_K$ is a partition of the data into $K$ folds, $\hat{\mu}^{(-j)}(X_k)$ is an estimate of $\mu(X_k) = \mathbb{E}[Y_k|X_k]$ based on all the data except $\mathcal{I}_j$, and $\hat{p}(X_k)$ is an estimate of $p(X_k)$ based on all the data except $\mathcal{I}_j$. We denote by $\mathcal{I}_{-j}$ the data on which $\hat{\mu}^{(-j)}$ and $\hat{p}^{(-j)}$ are trained. Thus, our estimators are cross-fit as shown in ref. [33]. We will require that $\hat{\mu}^{(-j)}(\cdot)$ and $\hat{p}^{(-j)}(\cdot)$ satisfy the following standard conditions for each $1 \le j \le K$ [33]:

**Assumption 2.** (Consistency) As $n \to \infty$, $\hat{\mu}^{(-j)}(\cdot)$, and $\hat{p}^{(-j)}(\cdot)$ satisfy

$$\sup_{x\in\mathcal{X}}|\mu(x) - \hat{\mu}^{(-j)}(x)|, \quad \sup_{x\in\mathcal{X}}|\hat{p}^{(-j)}(x) - p(x)| \xrightarrow{\mathbb{P}} 0.$$

**Assumption 3.** (Risk decay) We have that $\hat{\mu}^{(-j)}(\cdot)$, $\hat{p}^{(-j)}(\cdot)$ satisfy

$$\mathbb{E}[(\hat{\mu}^{(-j)}(X_k) - \mu(X_k))^2|\mathcal{I}_{-j}] \times \mathbb{E}[(\hat{p}^{(-j)}(X_k) - p(X_k))^2|\mathcal{I}_{-j}] = o_P(n^{-1}).$$

Under Assumptions 2 and 3, the estimator $\hat{\mu}_{\text{AIPW}}$ is semi-parametrically efficient [29], making it a natural starting point for potential improvements. Of course, this efficiency result means that any potential usage of adaptive normalization will not reduce the asymptotic variance of $\hat{\mu}_{\text{AIPW}}$, but we will see that we can still use the ideas of adaptive normalization to develop an estimator that has the same asymptotic efficiency as $\hat{\mu}_{\text{AIPW}}$, and also has better finite-sample MSE in simulations.

Consider the estimator in (14). The first term is an estimate of $\mu$ based on imputing all of the $Y_k$ by $\hat{\mu}^{(-j)}(X_k)$, while the second term is an inverse probability weighted estimate of the bias of the $\hat{\mu}^{(-j)}(X_k)$. In light of our work in Section 3, we naturally propose replacing the second term with a adaptively normalized estimator, yielding the new estimator

$$\begin{aligned}\hat{\mu}_{\text{AIPW,AN}} = \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_j}\hat{\mu}^{(-j)}(X_k) + \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_j}\frac{(Y_k-\hat{\mu}^{(-j)}(X_k))I_k}{\hat{p}^{(-j)}(X_k)} \\ + \frac{1}{\hat{\pi}}\left(\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_j}(Y_k-\hat{\mu}^{(-j)}(X_k))\frac{1-\hat{p}^{(-j)}(X_k)}{\hat{p}^{(-j)}(X_k)}\frac{I_k}{\hat{p}^{(-j)}(X_k)}\right)\left(1-\frac{\hat{n}}{n}\right),\end{aligned} \tag{15}$$

where now

$$\hat{\pi} = \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in\mathcal{I}_j}\frac{1-\hat{p}^{(-j)}(X_k)}{\hat{p}^{(-j)}(X_k)}\frac{I_k}{\hat{p}^{(-j)}(X_k)}, \quad \hat{n} = \sum_{j=1}^{K}\sum_{i\in\mathcal{I}_j}\frac{I_k}{\hat{p}^{(-j)}(X_k)}$$

are functions of the estimated propensities, rather than the true treatment probabilities as shown in Section 3.

The following theorem shows that this correction is asymptotically negligible; the proof is presented in Appendix A.4.

**Theorem 6.** *Suppose Assumptions* 1–3 *hold. Then* $\sqrt{n}(\hat{\mu}_{\text{AIPW}} - \hat{\mu}_{\text{AIPW,AN}}) \xrightarrow{\mathbb{P}} 0$.

Crucially, Theorem 6 implies that $\hat{\mu}_{\text{AIPW,AN}}$ has the same asymptotic variance as $\hat{\mu}_{\text{AIPW}}$ under the given conditions. These are exactly the conditions required for the efficiency of $\hat{\mu}_{\text{AIPW}}$, so we conclude that $\hat{\mu}_{\text{AIPW,AN}}$ is efficient whenever $\hat{\mu}_{\text{AIPW}}$ is. On the other hand, in finite samples, the additional correction term in $\hat{\mu}_{\text{AIPW,AN}}$ is negatively correlated with the other terms, and reduces variance, a fact we demonstrate empirically in Section 5. Taken together, these observations suggest that $\hat{\mu}_{\text{AIPW,AN}}$ should typically be preferred to $\hat{\mu}_{\text{AIPW}}$.

Our proposal is by no means the only attempt to improve on $\hat{\mu}_{\text{AIPW}}$. However, our proposal differs from existing work in two important ways. First, many other proposals are meant to improve on AIPW in the case when $\hat{\mu}$ is misspecified, which is not our motivation here, although we do explore this setting in simulation and see that $\hat{\mu}_{\text{AIPW,AN}}$ handles misspecification better than $\hat{\mu}_{\text{AIPW}}$. Second, and more importantly, other proposed estimators are significantly more complex, and sometimes requires solving certain estimating equations numerically. In contrast, $\hat{\mu}_{\text{AIPW,AN}}$ has an explicit, simple, closed form, and computing it requires nothing more than what is required to compute $\hat{\mu}_{\text{AIPW}}$.

## 4.2 ATE estimation

Another setting in which IPW estimators play a key role is in the problem of ATE estimation. Our model for ATE estimation is that there are triplets

$$(Y_1(1), Y_1(0), p_1), \ldots, (Y_n(1), Y_n(0), p_n)$$

drawn i.i.d. from a distribution $\mathcal{D}$ on $\mathbb{R} \times \mathbb{R} \times [0, 1]$ satisfying Assumption 1 for both $Y_k(1)$ and $Y_k(0)$. We assume that we observe $Y_1(I_1), \ldots, Y_n(I_n)$, where $I_k|p_k \sim \text{Ber}(p_k)$, and want to estimate the ATE $\tau = \mathbb{E}[Y_k(1) - Y_k(0)]$ from these observations. There are a variety of general approaches to estimating $\tau$ [34]; the one relevant to our work is the approach of using IPW estimators to separately estimate $\mu_1 = \mathbb{E}[Y_k(1)]$ and $\mu_0 = \mathbb{E}[Y_k(0)]$, and then subtracting the two estimates.

Since the problems of estimating $\mu_1$ and $\mu_0$ are survey sampling problems, they are often estimated with Horvitz–Thompson and Hájek estimators. Continuing our general theme, why not use $\hat{\mu}_{\text{AN}}$ instead? Thus, we propose

$$\hat{\tau}_{\text{AN}} = \hat{\mu}_{1,\text{AN}} - \hat{\mu}_{0,\text{AN}}, \tag{16}$$

where $\hat{\mu}_{1,\text{AN}}$ is the adaptively normalized estimator based on $Y_1(1)I_1, \ldots, Y_n(1)I_n$ and $\hat{\mu}_{0,\text{AN}}$ is the same based on $Y_1(0)(1 - I_1), \ldots, Y_n(0)(1 - I_n)$.

This is not the only possible use of adaptive normalization in this setting. Note that $\hat{\mu}_{1,\text{AN}}$ and $\hat{\mu}_{0,\text{AN}}$ here are designed to minimize variance within each group, even though our target estimand is the difference between groups. It is natural to ask what happens if we instead attempt to directly minimize the variance of the estimated difference between the two groups; we discuss this approach in detail in Appendix C and compare it to $\hat{\tau}_{\text{AN}}$ from (16). We conclude that estimating the two groups separately and taking the difference is reasonable and generally advised, except in the presence of strong correlation between the responses and the propensities, in which case the second approach may be preferable.

Similarly, rather than using AIPW estimation in each group for ATE estimation, we can also use adaptively normalized AIPW estimation in each group instead. Indeed, we will show in Section 5 that these estimators improve on both of the usual estimators in simulation.

Finally, we note that, since the adaptively normalized estimator is also a regression control estimator, $\hat{\tau}_{\text{AN}}$ corresponds to a particular form of the interaction estimator of ref. [35]. Appealingly, we derive the same estimator without any reference to an ordinary least squares model, side-stepping issues like those discussed in refs [35,36] of whether the model is correct or can only be used "agnostically."

## 4.3 Policy evaluation

The final setting in which we propose replacing an IPW estimator with an adaptively normalized estimator is policy evaluation. In this context, we wish to learn a statistical rule for assigning treatments to a population that maximizes the total welfare. We follow the regret minimization framework introduced by ref. [37] and build directly on the work of ref. [17], although much of our notation is drawn from [19].

Formally, suppose that individual $k$ has potential outcomes $Y_k(1)$ and $Y_k(0)$ depending on whether they receive a treatment, and we wish to learn a policy $\pi$ that maps known covariates $X_k \in \mathcal{X}$ to a treatment assignment in $\{0, 1\}$. The value of a policy $\pi$ (note that we are no longer using $\pi$ to represent moments of the unknown distribution as in Section 3) is $V(\pi) = \mathbb{E}[Y_i(\pi(X_i))]$, the average outcome of an individual treated using the policy. Assuming we are restricted to a class of policies $\Pi$, the best possible policy is $\pi^* = \arg\max_{\pi' \in \Pi} V(\pi')$, and we evaluate a policy based on its regret $V(\pi^*) - V(\pi)$.

Ideally we would learn a policy by maximizing $V$, but we cannot compute $V$. Instead we estimate $V$ from historical data $(Y_1(I_1), X_1), \ldots, (Y_n(I_n), X_n)$, where $I_k \sim \text{Ber}(p(X_k))$ is an indicator of whether $Y_k$ received the treatment, and we assume that Assumption 1 holds. Under the assumption that the propensity map $p(\cdot)$ is known, ref. [17] proposed

$$\hat{V}_{\text{IPW}}(\pi) = \frac{1}{n} \sum_{k=1}^{n} \frac{\mathbf{1}\{I_k = \pi(X_k)\} Y_k}{\mathbb{P}(I_k = \pi(X_k)|X_k)}$$

as an unbiased estimate of $V(\pi)$.

Of course, at this point, we can sense that a better estimate of $V$ would be the adaptively normalized

$$\hat{V}_{\mathrm{AN}} = \hat{V}_{\mathrm{IPW}}(\pi) + \frac{\sum_{k=1}^{n} Y_k \frac{1-\mathbb{P}(I_k=\pi(X_k)\mid X_k)}{\mathbb{P}(I_k=\pi(X_k)\mid X_k)} \frac{\mathbf{1}\{I_k=\pi(X_k)\}}{\mathbb{P}(I_k=\pi(X_k)\mid X_k)}}{\sum_{k=1}^{n} \frac{1-\mathbb{P}(I_k=\pi(X_k)\mid X_k)}{\mathbb{P}(I_k=\pi(X_k)\mid X_k)} \frac{\mathbf{1}\{I_k=\pi(X_k)\}}{\mathbb{P}(I_k=\pi(X_k)\mid X_k)}} \left(1 - \frac{1}{n}\sum_{k=1}^{n} \frac{\mathbf{1}\{I_k=\pi(X_k)\}}{\mathbb{P}(I_k=\pi(X_k)|X_k)}\right). \tag{17}$$

The main result of ref. [17] showed that maximizing $\hat{V}_{\mathrm{IPW}}(\pi)$ yields a policy whose regret decays at a rate of $1/\sqrt{n}$. We now give an analogous result for $\hat{V}_{\mathrm{AN}}$.

**Theorem 7.** *Let* $\hat{\pi}_{\mathrm{AN}} = \mathrm{argmin}_{\pi\in\Pi}\hat{V}_{\mathrm{AN}}(\pi)$ *and suppose that* $\Pi$ *has finite VC-dimension. Then*

$$\mathbb{E}[V(\pi^*) - V(\hat{\pi}_{\mathrm{AN}})] \leq O\left(\frac{M}{\delta}\sqrt{\frac{\mathrm{VC}(\Pi)}{n}}\right).$$

Thus, $\hat{V}_{\mathrm{AN}}$ preserves the theoretical guarantees associated with $\hat{V}_{\mathrm{IPW}}$, and we verify in the next section that policies learned with $\hat{V}_{\mathrm{AN}}$ are closer to the optimal policy. One drawback of $\hat{V}_{\mathrm{AN}}$, however, is that $\hat{V}_{\mathrm{IPW}}$ can be interpreted as a weighted classification objective [17,19], facilitating optimization, while $\hat{V}_{\mathrm{AN}}$ unfortunately does not have such an interpretation. Finally, we note that [19] introduced the idea of policy learning based on AIPW estimation instead of IPW estimation; applying these ideas in our setting is an exciting opportunity for future work.

# 5 Experiments

In this section, we use a series of experiments to evaluate the empirical performance of adaptive normalization. The goal of our experiments is to validate that applying adaptive normalization for mean estimation as well as in the various settings of Section 4 actually pays the dividends we expect.

Our first set of experiments are focused on survey sampling as discussed in Section 3 and make use of semi-synthetic data to compare $\hat{\mu}_{\mathrm{AN}}$ to $\hat{\mu}_{\mathrm{HT}}$ and $\hat{\mu}_{\mathrm{Hájek}}$. Our second set of experiments are simulations comparing ATE estimation using all of the aforementioned estimators, and also using $\hat{\mu}_{\mathrm{AIPW,AN}}$ and $\hat{\mu}_{\mathrm{AIPW}}$. Our third set of experiments studies the coverage of confidence intervals for these estimators constructed using asymptotic normal approximations. Finally, we use a simulation experiment to compare $\hat{V}_{\mathrm{AN}}$ to $\hat{V}_{\mathrm{IPW}}$ as a policy learning objective. Additional experiments appear in Appendix B.

## 5.1 Survey experiments

We work with a dataset of Swiss municipalities, provided by the R package `sampling` [38] under the GPL-2 license. This data set contains demographic and financial data for 2,896 different municipalities in Switzerland, and we consider two responses: $Y_1$, the wooded area of a municipality, and $Y_2$, the industrial area of a municipality. We assume the standard sampling scheme for studies of this dataset in which the probability $p_k$ of observing $Y_k$ is proportional to the total area of municipality $k$, and set $\sum_k p_k$ (the expected number of observations) to be either 50 or 250. We resample the set of observed municipalities according to these probabilities and compare the three estimators with the true mean of the $Y_k$. The RMSE of the Horvitz–Thompson estimator, Hájek estimator, and adaptively normalized estimator for the four specifications are presented in Table 1.

To better understand Table 1, we remark that $Y_1$ is strongly positively correlated with $p_k$ and $Y_2$ is weakly positively correlated with $p_k$. Thus, in the first two columns of Table 1, the Horvitz–Thompson estimator is much better than the Hájek estimator, but not as good as the adaptively normalized estimator, in accordance with our observations about $\lambda^*$ in Section 3. In the latter two columns, the correlation is weaker, so the Horvitz–Thompson and Hájek estimators are more comparable, but the adaptively normalized estimator still improves on both. These results demonstrate that although it may be unclear whether the

**Table 1:** RMSE of estimators on Swiss municipality data; $Y_1$ is wood area and $Y_2$ is industrial area; probabilities are chosen proportional to total municipality area, which is strongly positively correlated with $Y_1$ and weakly positively correlated with $Y_2$. RMSEs are averaged over 100,000 trials and standard errors are over 10 replications of the 100,000 trials

| | Problem specification | | | |
|---|---|---|---|---|
| | $\sum p_k = 50, Y_1$ | $\sum p_k = 250, Y_1$ | $\sum p_k = 50, Y_2$ | $\sum p_k = 250, Y_2$ |
| $\hat{\mu}_{\text{HT}}$ | 68.4 ± 0.1030 | 27.8 ± 0.0710 | 2.51 ± 0.0051 | 1.07 ± 0.0026 |
| $\hat{\mu}_{\text{Hájek}}$ | 95.3 ± 0.3587 | 39.3 ± 0.1510 | 2.52 ± 0.0076 | 1.06 ± 0.00244 |
| $\hat{\mu}_{\text{AN}}$ | 61.5 ± 0.1035 | 23.1 ± 0.0538 | 2.45 ± 0.0086 | 1.01 ± 0.0028 |

Horvitz–Thompson estimator or Hájek estimator should be used in a particular problem, the adaptively normalized estimator circumvents this question by always improving on both of the other two estimators. For reference, plots of the MSE in each problem as a function of different values of $\lambda$ are shown in Figure 3.

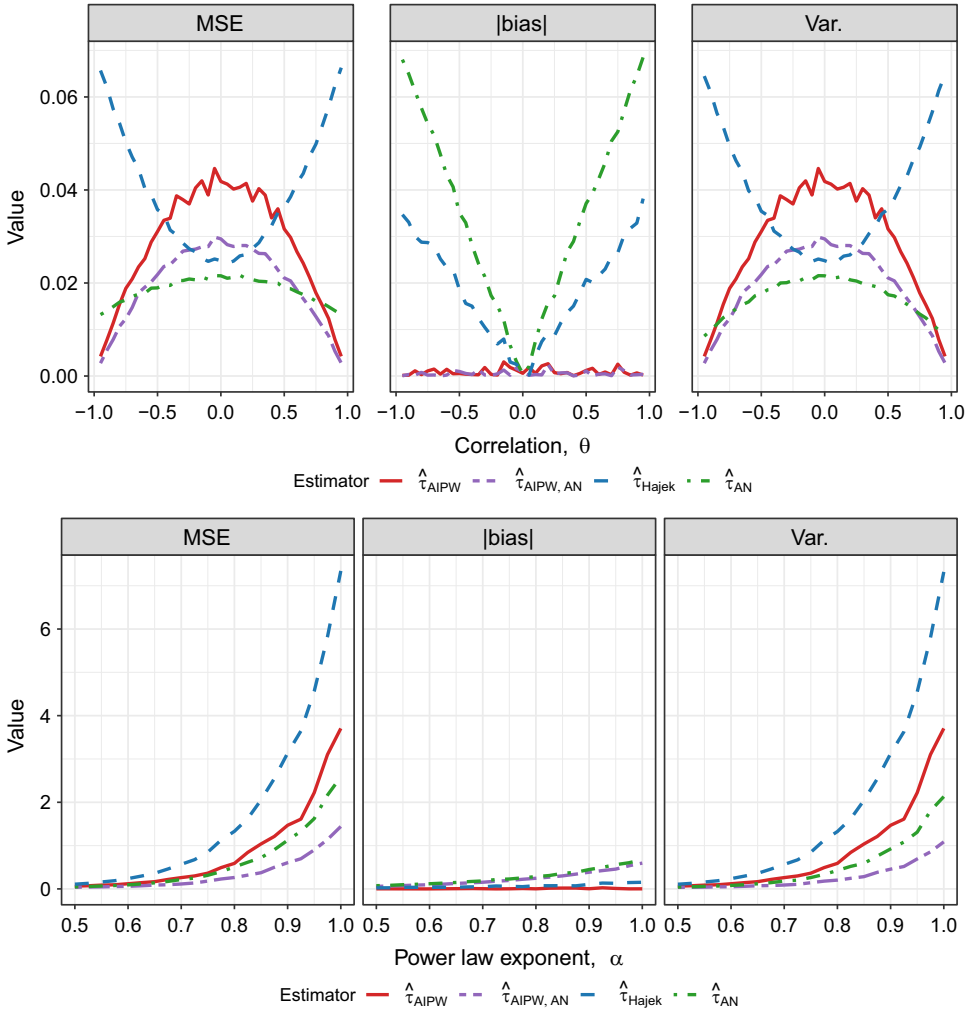


**Figure 3:** Top: the estimated absolute value of bias, variance, and MSE for estimators applied to data from model (18) with $n = 500$, $\mu = 1$, and varying $\theta$. Bottom: the same, but with data from model (19) with $n = 500$ and varying $\alpha$. Results are averaged across 10,000 trials. We omit $\hat{\tau}_{\text{HT}}$ due to extremely large variance. Across the board, adaptively normalized estimators effectively trade off bias and variance to achieve lower MSE than their traditional counterparts.

## 5.2 ATE experiments

For ATE estimation, we consider two models, one of which represents a benign setting in which $Y_k$ and $p_k$ have roughly linear relationship, and one of which represents a more difficult setting in which $Y_k$ and $p_k$ have an extremely strong negative relationship.

The first model, which we refer to as our *normal model*, is a model in which we generate

$$(Y_k(1), X_k) \sim N\left(\begin{bmatrix}\mu\\0\end{bmatrix}, \begin{bmatrix}1 & \theta\\ \theta & 1\end{bmatrix}\right), \tag{18}$$

and $Y_k(0) = Y_k(1) - \tau$, $p_k = (1 + \exp(-2X_k))^{-1}$. In this model, $\theta$ controls the correlation between $Y_k$ and $X_k$, and indirectly the correlation between $Y_k$ and $p_k$. To satisfy Assumption 1, we truncate $p_k$ and $Y_k$ with $M = 50$, $\delta = 0.01$.

The second model we consider is our *power law model*, where we generate

$$p_k \sim \text{Uni}(\varepsilon, 1-\varepsilon), \quad Y_k(1) = p_k^{-\alpha} + N(0, \sigma^2), \tag{19}$$

and $Y_k(0) = Y_k(1) - \tau$, $X_k = \log((1-p_k)/p_k)$. In this model, $\alpha$ controls the strength of the nonlinear negative relationship between $Y_k$ and $p_k$, while $\sigma^2$ is a noise parameter we set to 1. To satisfy Assumption 1, we take $\varepsilon = 10^{-2}$ and truncate $Y_k$ at $M = 10^6$.

In both models, we estimate $p_k$ from logistic regression with covariate $X_k$ and estimate $\mathbb{E}[Y_k|X_k]$ with a generalized additive model. For AIPW estimators, we use two-fold cross-fitting. The treatment effect is set to $\tau = 0.5$. We simulate data from both models, fixing $n = 500$ and $\tau = 0.5$ while varying $\theta$ and $\alpha$, and compare the biases, variance, and MSEs of various ATE estimators. The results are shown in Figure 4, omitting $\hat{\tau}_{\text{HT}}$ for clarity.

We discuss the classical and semi-parametric estimators separately. Comparing $\hat{\tau}_{\text{Hájek}}$ and $\hat{\tau}_{\text{AN}}$, we see in the normal model that $\hat{\tau}_{\text{AN}}$ has larger bias but smaller variance than $\hat{\tau}_{\text{Hájek}}$ across all values of $\theta$. Because

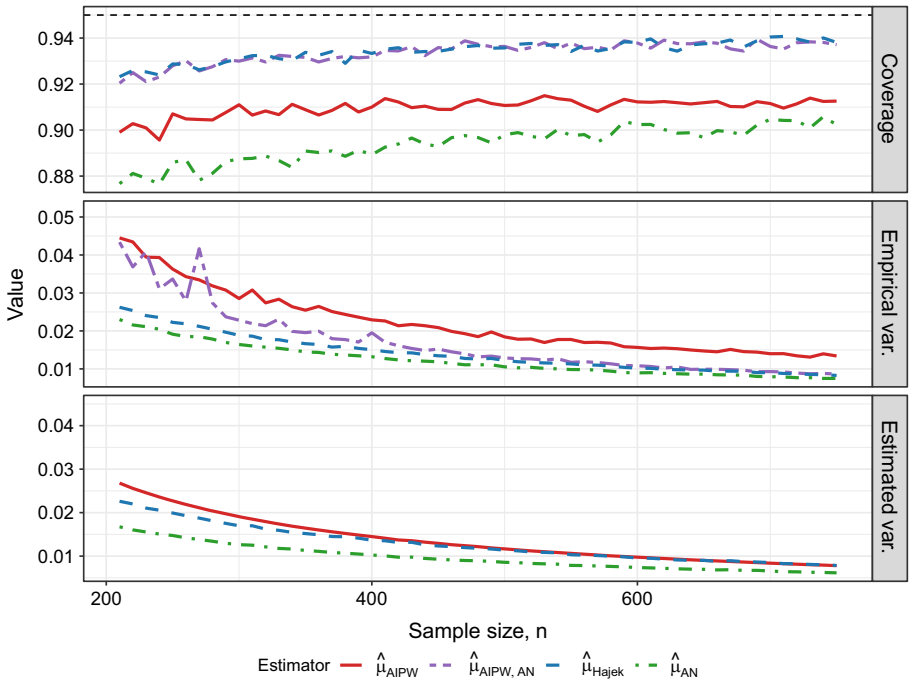


**Figure 4:** Coverage of 95% confidence intervals for different estimators of the mean $\mu$ of $Y_k(1)$ in the normal model (18) when $\mu = 1$, $\theta = 0.1$, and $n$ varies, along with empirical variances and estimated variances based on asymptotics. We see in general that $\hat{\mu}_{\text{AN}}$ and $\hat{\mu}_{\text{AIPW,AN}}$ achieve nearly the target coverage, but that $\hat{\mu}_{\text{AIPW}}$ and $\hat{\mu}_{\text{AN}}$ lag behind. Finite sample coverage issues for $\hat{\mu}_{\text{AIPW}}$ are known in the literature, and the weaker coverage of $\hat{\mu}_{\text{AN}}$ is driven by its larger bias in finite samples (as shown in Figure 4). Thus, some caution is needed when constructing confidence intervals for $\hat{\mu}_{\text{AN}}$.

the MSE is dominated by the variance, $\hat{\tau}_{\text{AN}}$ also has consistently lower MSE than $\hat{\tau}_{\text{Hájek}}$, and the same behavior is found in the power law model. By comparing $\hat{\tau}_{\text{AIPW}}$ and $\hat{\tau}_{\text{AIPW,AN}}$, we see in the normal model that they both have nearly no bias, but that $\hat{\tau}_{\text{AIPW, AN}}$ has slightly lower variance and thus lower MSE across all values of $\theta$. We again see the same behavior in the power law model, except that the difference between $\hat{\tau}_{\text{AIPW}}$ and $\hat{\tau}_{\text{AIPW, AN}}$ is more substantial in this model. The reason for this difference is that estimating $Y_k|X_k$ is more difficult in the power law model, so the more careful estimate of the bias used by $\hat{\tau}_{\text{AIPW,AN}}$ pays larger dividends.

## 5.3 Coverage experiments

We now turn to the issue of inference, and consider the problem of constructing confidence intervals for our estimators. For $\hat{\mu}_{\text{HT}}$, $\hat{\mu}_{\text{Hájek}}$, and $\hat{\mu}_{\text{AN}}$, this can be done using the asymptotic variances in Theorems 1 and 3 and an asymptotic normal approximation; for $\hat{\mu}_{\text{AIPW}}$ and $\hat{\mu}_{\text{AIPW,AN}}$, we use the known asymptotic variance of the AIPW estimator (see, e.g., ref. [29]) and an asymptotic normal approximation. We use the same asymptotic variance for both $\hat{\mu}_{\text{AIPW}}$ and $\hat{\mu}_{\text{AIPW,AN}}$ because, as shown in Theorem 6, these two estimators have the same limiting distributions. We generate data from the model (18), and attempt to construct a 95% confidence interval for the mean $\mu$ of $Y_k(1)$ when $\mu = 1$, $\theta = 0.1$, and $n$ varies. The results are shown in Figure A1.

We see in the top panel of Figure A1 that the coverage of confidence intervals for $\hat{\mu}_{\text{Hájek}}$ and $\hat{\mu}_{\text{AIPW,AN}}$ is fairly close to the target coverage level of 95%. However, the coverage of confidence intervals for $\hat{\mu}_{\text{AIPW}}$ and $\hat{\mu}_{\text{AN}}$ is weaker, especially in smaller samples. Issues with the coverage of confidence intervals for the AIPW estimator are known in the literature [39,40], and we see from the bottom two panels of Figure A1 that these are caused by the empirical variance of the AIPW estimator being larger than our asymptotic approximation of it in the sample sizes considered. Since $\hat{\mu}_{\text{AIPW,AN}}$ has smaller empirical variance than $\hat{\mu}_{\text{AIPW}}$, it also enjoys better coverage in finite samples. On the other hand, the asymptotic approximation of the variance of $\hat{\mu}_{\text{AN}}$ appears relatively accurate, but the confidence intervals are slightly below the target coverage level because of the larger bias of $\hat{\mu}_{\text{AN}}$, as shown in Figure 4. For this reason, practitioners should be aware that confidence intervals for $\hat{\mu}_{\text{AN}}$ may slightly undercover unless the sample size is quite large. This issue could be remedied by using some kind of bias correction, and is an important future direction.

## 5.4 Policy evaluation experiments

Our final simulation compares the two policy learning objectives, $\hat{V}_{\text{IPW}}$ and $\hat{V}_{\text{AN}}$ of the policy evaluation application in Section 4. Our data generating model is inspired by the simulations of ref. [19], and sets (letting $X_{k,i}$ be the $i^{\text{th}}$ entry of $X_k$)

$$X_k \sim N(0, I_{3\times 3}), \quad p(X_k) = \frac{1}{1 + \exp(-X_{k,1})}, \tag{20}$$

with potential outcomes $Y_k(1) = X_{k,1}$, $Y_k(0) = Y_k(1) - \operatorname{sgn}(X_{k,2} + X_{k,3})$. In general, minimizing $\hat{V}_{\text{IPW}}$ or $\hat{V}_{\text{AN}}$ is nonconvex. To obtain a tractable problem, we restrict the class $\Pi$ to be the set of policies of the form $\pi(X_k) = \mathbf{1}\{X_{k,2} > T\}$ for $T \in [-1, 1]$. Note that this is a misspecified setting, in the sense that the optimal policy $\pi(X_k) = \mathbf{1}\{X_{k,2} + X_{k,3} > 0\}$ is not in the class we are optimizing over. For each of $\hat{V}_{\text{IPW}}$ and $\hat{V}_{\text{AN}}$, we generate a sample of size $n$ from (20) and learn a cut-off $T$ that minimizes the corresponding objective by grid search on $T \in [-1, 1]$.

The average threshold learned, for a range of values of $n$, is presented in Table 2. The optimal policy is to threshold at $T = 0$, and so we take this as a point of comparison. We see that the thresholds learned from minimizing $\hat{V}_{\text{AN}}$ are consistently closer to the optimal threshold of zero than those learned by minimizing $\hat{V}_{\text{IPW}}$, and that the gap between the performance of the two objectives is consistent across the range of values of $n$ we consider.

**Table 2:** Thresholds learned by $\hat{V}_{\text{IPW}}$ and $\hat{V}_{\text{AN}}$ for varying $n$ for data generated according to (20). Each entry is an average over 100,000 replications and standard errors are over 10 replications of the 100,000 trials. The optimal policy is to threshold at 0; minimizing $\hat{V}_{\text{AN}}$ consistently learns better thresholds

| Objective | $n = 250$ | $n = 500$ | $n = 750$ | $n = 1{,}000$ |
|---|---|---|---|---|
| $\hat{V}_{\text{IPW}}$ | −0.057 ± 0.0013 | −0.035 ± 0.0011 | −0.026 ± 0.0011 | −0.020 ± 0.0014 |
| $\hat{V}_{\text{AN}}$ | −0.039 ± 0.0018 | −0.015 ± 0.0014 | −0.010 ± 0.0009 | −0.004 ± 0.0009 |

# 6 Discussion

In this article, we study *adaptive normalization* for IPW estimators: rather than normalizing by the sample size or by the sum of the weights, we propose normalizing by a data-dependent affine combination of the two. For mean estimation in survey sampling, our proposed estimator is algebraically equivalent to a control variate method from the Monte Carlo literature that has not been studied in the causal inference literature before. We further develop the adaptive normalization idea in causal inference settings by using it to improve on the AIPW estimators, to propose new estimators for the ATE in randomized experiments, and new objectives for policy learning.

There are several possible future directions for this work. One is to relax the assumption that the treatment indicators $I_k$ are independent, an unrealistic assumption in many observational datasets and also directly violated in certain experimental designs. However, if the correlation structure of the $I_k$ is known or possible to estimate, analogues of our limit theorems and estimators could be developed. If the correlation between the $I_k$ is sufficiently weak, Theorem 1 may hold without modification, and we expect the remainder of our results would go through as well. However, if the correlation between the $I_k$ is strong enough to affect the asymptotic variance in Theorem 1, then the optimal choice of $\lambda$ would change, and these changes would propagate through the analysis. In such a case, we expect that the general form of $\hat{\mu}_{\text{AN}}$ would remains similar, but the $\hat{T}/\hat{\pi}$ term would be replaced by a more complex expression.

In addition, estimands other than the mean, such as quantiles and distributions, are also of interest in causal inference and program evaluation and can be estimated using IPW-style estimators [41–43]. Extending our results to these other estimands is another possible line of further work and will likely give very different results from the ones presented here. IPW estimators for the mean have a simple closed form involving the weights, whereas IPW estimators for other estimands are typically defined as weighted M-estimators and do not have such a closed form [42,43]. Thus, the weights enter into the estimator in a different fashion in such cases, and we would expect the analogue of the adaptively normalized estimator to look different as well. In fact, we are not even aware of any work studying Hájek/self-normalized estimators in these other contexts, and the comparison between the Hájek and Horvitz–Thompson estimators may be different as well.

In a different direction, there are many places within and beyond causal inference where inverse probability weighted estimators are used in context-specific ways, such as off-policy evaluation on networks [44], recommender system evaluation [45], in learning to rank problems [46], and inference from bandits [47,48]. Developing and applying similar ideas in these contexts suggests many promising lines of future work.

**Acknowledgements:** We thank Guillaume Basse, Alex Chin, Dean Eckles, Sharad Goel, Kevin Guo, Ramesh Johari, Brian Karrer, and Fredrik Sävje for helpful discussions and feedback on early versions of this work. This work was partially supported by ARO award 73348-NS-YIP.

**Conflict of interest:** Authors state no conflict of interest.

# References


[1] Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952;47(260):663–85. http://www.jstor.org/stable/2280784.

[2] Basu D. An essay on the logical foundations of survey sampling, Part I. In: Godambe V, Sprott D, editors. Foundations of statistical inferences. Toronto, Canada: Holt, Rinehart and Winston; 1971.

[3] Godambe VP, Joshi VM. Admissibility and bayes estimation in sampling finite populations. I. The annals of mathematical statistics. 1965;36(6):1707–22. http://www.jstor.org/stable/2239112.

[4] Särndal CE, Swensson B, Wretman J. Model assisted survey sampling. Springer Science and Business Media; 2003.

[5] Trotter HF, Tukey JW. Conditional Monte Carlo for normal samples. In: Symposium on Monte Carlo Methods. New York: Wiley; 1954.

[6] Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89(427):846–66. http://www.jstor.org/stable/2290910.

[7] Calonico S, Cattaneo MD, Farrell MH. On the effect of bias estimation on coverage accuracy in nonparametric inference. J Am Stat Assoc. 2018;113(522):767–79.

[8] Hall P. Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. Ann Stat. 1992;20:675–94.

[9] Hammersley JM, Handscomb DC. Monte Carlo methods. New York: Springer; 1964.

[10] Owen AB. Monte Carlo theory, methods and examples; 2013. https://artowen.su.domains/mc/.

[11] Glynn PW, Szechtman R. Some new perspectives on the method of control variates. In: Monte Carlo and Quasi-Monte Carlo methods 2000. Springer; 2002. p. 27–49.

[12] Firth D. On improved estimation for importance sampling. Brazilian J Probab Stat. 2011;25(3):437–43. doi: 10.1214/11-BJPS155.

[13] Hesterberg T.Weighted average importance sampling and defensive mixture distributions. Technometrics. 1995;37(2):185–94. https://amstat.tandfonline.com/doi/abs/10.1080/00401706.1995.10484303.

[14] Chen SX, Leung DHY, Qin J. Improving semiparametric estimation by using surrogate data. J R Stat Soc Ser B (Stat Methodol). 2008;70(4):803–23. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2008.00662.x.

[15] Qin J, Zhang B, Leung DHY. Empirical likelihood in missing data problems. J Am Stat Assoc. 2009;104(488):1492–503. doi: 10.1198/jasa.2009.tm08163.

[16] Rotnitzky A, Lei Q, Sued M, Robins JM. Improved double-robust estimation in missing data and causal inference models. Biometrika. 2012 April;99(2):439–56. doi: 10.1093/biomet/ass013.

[17] Kitagawa T, Tetenov A. Who should be treated? empirical welfare maximization methods for treatment choice. Econometrica. 2018;86(2):591–616.

[18] Swaminathan A, Joachims T. The self-normalized estimator for counterfactual learning. In: Advances in neural information processing systems. New York: Citeseer; 2015. p. 3231–9.

[19] Athey S, Wager S. Policy learning with observational data. Econometrica. 2021;89(1):133–61.

[20] Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009;96(1):187–99.

[21] Yang S, Ding P. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. Biometrika. 2018;105(2):487–93.

[22] Ma X, Wang J. Robust inference using inverse probability weighting. J Am Stat Assoc. 2020;115(532):1851–60.

[23] Hong H, Leung MP, Li J. Inference on finite-population treatment effects under limited overlap. Econometr J. 2020;23(1):32–47.

[24] Delevoye A, Sävje F. Consistency of the Horvitz–Thompson estimator under general sampling and experimental designs. J Stat Planning Inference 2020;207:190–7.

[25] Robinson PM. On the convergence of the Horvitz-Thompson estimator. Australian J Stat. 1982;24(2):234–8. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.1982.tb00829.x.

[26] Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica. 2003;71(4):1161–89.

[27] Hansen BE, Lee S. Inference for iterated GMM under misspecification. Econometrica. 2021;89(3):1419–47.

[28] Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences. Cambridge, England: Cambridge University Press; 2015.

[29] Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica. 1998;66(2):315–31. http://www.jstor.org/stable/2998560.

[30] Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. Biometrika. 2010;97(3):661–82.

[31] Cassel CM, Sarndal CE, Wretman JH. Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika. 1976;63(3):615–20. http://www.jstor.org/stable/2335742.

[32] Little RJA. Estimating a finite population mean from unequal probability samples. J Am Stat Assoc. 1983;78(383):596–604. http://www.jstor.org/stable/2288125.

[33] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. Econom J. 2018 Jan;21(1):C1–C68. doi: 10.1111/ectj.12097.
[34] Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. Review Econ Stat. 2004;86(1):4–29.
[35] Lin W. Agnostic notes on regression adjustments to experimental data: reexamining Freedmanas critique. Annal Appl Stat. 2013;7(1):295–318.
[36] Freedman DA. On regression adjustments to experimental data. Adv Appl Math. 2008;40(2):180–93.
[37] Manski CF. Statistical treatment rules for heterogeneous populations. Econometrica. 2004;72(4):1221–46.
[38] Tillé Y, Matei A. Sampling: survey sampling; 2021. R package version 2.9. https://CRAN.R-project.org/package=sampling.
[39] Hankin M, Chan D, Perry M. A comparison of causal inference methods for estimating sales lift. 2020. https://research.google/pubs/pub49507.
[40] Liu L, Mukherjee R, Robins JM. On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. Stat Sci. 2020;35(3):518–39.
[41] Bitler MP, Gelbach JB, Hoynes HW. What mean impacts miss: distributional effects of welfare reform experiments. Am Econ Rev. 2006;96(4):988–1012.
[42] Firpo S. Efficient semiparametric estimation of quantile treatment effects. Econometrica. 2007;75(1):259–76.
[43] Cattaneo MD. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. J Econ. 2010;155(2):138–54.
[44] Chin A, Eckles D, Ugander J. Evaluating stochastic seeding strategies in networks. Manag Sci. 2021;68:1591–2376.
[45] Schnabel T, Swaminathan A, Singh A, Chandak N, Joachims T. Recommendations as treatments: debiasing learning and evaluation. In: International Conference on Machine Learning. New York: PMLR; 2016. p. 1670–9.
[46] Oosterhuis H, de Rijke M. Unifying online and counterfactual learning to rank: a novel counterfactual estimator that effectively utilizes online interventions. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. New York: ACM Digital Library; 2021. p. 463–71.
[47] Hadad V, Hirshberg DA, Zhan R, Wager S, Athey S. Confidence intervals for policy evaluation in adaptive experiments. 2019. arXiv: http://arXiv.org/abs/arXiv:191102768.
[48] Bibaut A, Chambaz A, Dimakopoulou M, Kallus N, van der Laan M. Post-Contextual-Bandit inference. 2021. arXiv: http://arXiv.org/abs/arXiv:210600418.

# Appendix

# A Technical proofs

## A.1 Proofs of Theorems 1, 3, and 5

In this section, we detail the calculations underlying the CLTs in Section 3. The building block of our results is a joint CLT for the vector $\hat{\beta} = (\hat{S}/n, \hat{n}/n)$, which has mean $\beta = (\mu, 1)$.

**Lemma A1.** *Under Assumption* 1, *we have*

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

*where the entries of* $\Sigma$ *are given by*

$$n\mathrm{var}(\hat{S}/n) = \mathrm{var}(Y_1) + \mathbb{E}\left[Y_k^2 \frac{1-p_k}{p_k}\right], \quad n\mathrm{var}(\hat{n}/n) = \mathbb{E}\left[\frac{1-p_k}{p_k}\right], \quad n\mathrm{cov}(\hat{S}/n, \hat{n}/n) = \mathbb{E}\left[Y_k \frac{1-p_k}{p_k}\right].$$

**Proof.** For any vector $v \in \mathbb{R}^4$, the quantity $\sqrt{n}(v^T\hat{\beta} - v^T\beta)$ is a sum of i.i.d. random variables. These random variables have finite variance by Assumption 1, so the classical Lindeberg CLT applies with a limiting distribution that is $N(0, v^T\Sigma v)$. The lemma then follows from the Cramer-Wold device. □

The CLT for $\hat{\mu}_\lambda$ now follows from an application of the delta method.

**Proof of Theorem 1.** Note that $\hat{\mu}_\lambda = f(\hat{\beta})$ for $f(x, y; \lambda) = \frac{x}{\lambda + (1-\lambda)y}$. The function $f$ is differentiable at the point $\beta$ and has gradient $\nabla_\beta f = (1, -(1-\lambda)\mu)$, and so by the delta method, we conclude that $\hat{\mu}_\lambda$ satisfies a CLT with the asymptotic variance

$$\nabla_\beta f^T \Sigma \nabla_\beta f = \mathrm{var}(Y_1) + \mathbb{E}\left[Y_k^2 \frac{1-p_k}{p_k}\right] - 2(1-\lambda)\mu\mathbb{E}\left[Y_k \frac{1-p_k}{p_k}\right] + (1-\lambda)^2\mu^2\mathbb{E}\left[\frac{1-p_k}{p_k}\right],$$

which rearranges to the variance in (3), as desired. □

**Proof of Theorem 3.** The key step is to replace the factor of $\hat{T}/\hat{\pi}$ in $\hat{\mu}_{\mathrm{AN}}$ with $T/\pi$, so that we can then apply a standard CLT. Thus, we write

$$\hat{\mu}_{\mathrm{AN}} = \frac{\hat{S}}{n} + \frac{T}{\pi}\left(1 - \frac{\hat{n}}{n}\right) + \left(\frac{\hat{T}}{\hat{\pi}} - \frac{T}{\pi}\right)\left(1 - \frac{\hat{n}}{n}\right), \tag{A1}$$

$$= \frac{\hat{S}}{n} + \frac{T}{\pi}\left(1 - \frac{\hat{n}}{n}\right) + o_P(1)O_P(n^{-1/2}), \tag{A2}$$

$$= \frac{\hat{S}}{n} + \frac{T}{\pi}\left(1 - \frac{\hat{n}}{n}\right) + o_P(n^{-1/2}), \tag{A3}$$

where the second equality holds because $\frac{\hat{T}}{\hat{\pi}} - \frac{T}{\pi} = o_P(1)$ by the consistency of $T$ and $\pi$, and $1 - \frac{\hat{n}}{n} = O_P(n^{-1/2})$ by the CLT for i.i.d. sums.

Thus, the limiting distribution of $\sqrt{n}(\hat{\mu}_{\mathrm{AN}} - \mu)$ is the same as the limiting distribution of

$$\sqrt{n}\left(\frac{\hat{S}}{n} + \frac{T}{\pi}\left(1 - \frac{\hat{n}}{n}\right) - \mu\right) = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\left(\frac{Y_k I_k}{p_k} - \mu\right) - \frac{T}{\pi} \cdot \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\left(\frac{I_k}{p_k} - 1\right).$$

This last expression is the difference of two i.i.d. mean zero sums, and so its limiting distribution is asymptotically normal with variance

$$\mathrm{var}(Y_k I_k / p_k) - 2(T/\pi)\mathrm{cov}(Y_k I_k / p_k, I_k / p_k) + (T/\pi)^2 \mathrm{var}(I_k / p_k),$$

which simplifies to $\sigma^2(\lambda^*)$. □

Next we turn to the CLT for the case of estimated propensities. Recall that our set-up here is that $p_k = (1 + \exp(X_k^T \theta^*))^{-1}$, and we estimate $\theta^*$ by the maximum-likelihood estimate $\hat{\theta}$. In what follows, we abuse notation and write $p(X_k^T \theta)$ for $(1 + \exp(X_k^T \theta))^{-1}$.

**Proof of Theorem 5.** Consider the family of estimators

$$\hat{\mu}_\beta(\theta) = \frac{1}{n}\sum_{k=1}^n \frac{Y_k}{I_k} p(X_k^T \theta) + \beta\left(1 - \frac{1}{n}\sum_{k=1}^n \frac{I_k}{p(X_k^T \theta)}\right).$$

Then $\hat{\mu}_{\mathrm{HT}}$ corresponds to $\beta = 0$, and it follows from the consistency and asymptotic normality of $\hat{\theta}$ that we can repeat the arguments of (A3) and show that $\hat{\mu}_{\mathrm{AN}}$ is first-order equivalent to $\beta = T/\pi$. Thus, to establish the theorem, it suffices to characterize the asymptotic variance of $\hat{\mu}_\beta(\hat{\theta})$ and compare the cases $\beta = 0$ and $\beta = T/\pi$.

To understand the asymptotic of $\hat{\mu}_\beta(\hat{\theta})$, we first Taylor expand around $\theta^*$ and obtain

$$\hat{\mu}_\beta(\hat{\theta}) = \hat{\mu}_\beta(\theta^*) + \mathbb{E}\left[\frac{\partial \hat{\mu}_\beta(\theta)}{\partial \theta}\,|_{\theta=\theta^*}\right]^T (\hat{\theta} - \theta^*) + o_P(n^{-1/2}). \tag{A4}$$

Next, it follows from standard results on M-estimators that $\hat{\theta} - \theta^* = I(\theta^*)^{-1} S(\theta^*) + o_P(n^{-1/2})$, where

$$I(\theta^*)^{-1} = \mathbb{E}[p(X_k^T\theta^*)(1 - p(X_k^T\theta^*))X_k X_k^T] \quad \text{and} \quad S(\theta^*) = \frac{1}{n}\sum_{k=1}^n X_k(I_k - p(X_k^T\theta^*)) \tag{A5}$$

are the Fisher information and score. Also, by direct computation,

$$\frac{\partial \hat{\mu}_\beta(\theta)}{\partial \theta} = \frac{1}{n}\sum_{k=1}^n - \frac{Y_k I_k}{p(X_k^T\theta)^2} \cdot p(X_k^T\theta)(1 - p(X_k^T\theta))X_k + \frac{1}{n}\sum_{k=1}^n \frac{\beta I_k}{p(X_k^T\theta)^2} \cdot p(X_k^T\theta)(1 - p(X_k^T\theta)), \tag{A6}$$

which has expectation $-\mathbb{E}[(Y_k - \beta)(1 - p(X_k^T\theta^*))X_k]$ at $\theta^*$.

Combining (A4), (A5), and (A6) gives

$$\begin{aligned}\sqrt{n}(\hat{\mu}_\beta(\hat{\theta}) - \mu) = {} & \frac{1}{\sqrt{n}}\sum_{k=1}^n \frac{Y_k I_k}{p(X_k^T\theta^*)} - \mu + \frac{1}{\sqrt{n}}\sum_{k=1}^n \beta - \frac{\beta I_k}{p(X_k^T\theta^*)} - \mathbb{E}[Y_k(1 - p(X_k^T\theta^*))X_k] I(\theta^*)^{-1} \\ & \times \frac{1}{\sqrt{n}}\sum_{k=1}^n X_k(I_k - p(X_k^T\theta^*)) + \beta\mathbb{E}[(1 - p(X_k^T\theta^*))X_k] I(\theta^*)^{-1} \cdot \frac{1}{\sqrt{n}}\sum_{k=1}^n X_k(I_k - p(X_k^T\theta^*)) \\ & + o_P(n^{-1/2}).\end{aligned} \tag{A7}$$

Since each term of (A7) is a sum of mean-zero i.i.d. terms, we can compute the asymptotic variance directly. In particular, the terms that depend on $\beta$ are

$$\begin{aligned}& \beta^2 \mathrm{var}(I_k / p(X_k^T\theta^*)) - 2\beta\mathrm{cov}\left(\frac{I_k}{p(X_k^T\theta^*)}, \frac{Y_k I_k}{p(X_k^T\theta^*)}\right) + 2\beta\mathrm{cov}\left(\frac{I_k}{p(X_k^T\theta^*)}, u_1^T I(\theta^*)^{-1} X_k(I_k - p(X_k^T\theta^*))\right) \\ & + \beta^2 \mathrm{var}(u_2^T I(\theta^*)^{-1} X_k(I_k - p(X_k^T\theta^*))) + 2\beta\mathrm{cov}\left(u_2^T I(\theta^*)^{-1} X_k(I_k - p(X_k^T\theta^*)), \frac{Y_k I_k}{p(X_k^T\theta^*)}\right) \\ & - 2\beta\mathrm{cov}(u_1^T I(\theta^*)^{-1} X_k(I_k - p(X_k^T\theta^*)), u_2^T I(\theta^*)^{-1} X_k(I_k - p(X_k^T\theta^*))) \\ & - 2\beta^2\mathrm{cov}\left(\frac{I_k}{p(X_k^T\theta^*)}, u_2^T I(\theta^*)^{-1} X_k(I_k - p(X_k^T\theta^*))\right),\end{aligned} \tag{A8}$$

where $u_1 = \mathbb{E}[Y_k(1 - p(X_k^T\theta^*))X_k]$ and $u_2 = \mathbb{E}[(1 - p(X_k^T\theta^*))X_k]$. The first two terms of (A8) are $\beta^2\pi - 2T\beta$. To simplify the remaining terms, we compute

$$\operatorname{cov}\left(\frac{I_k}{p(X_k^T\theta^*)}, u_1^T I(\theta^*)^{-1}X_k(I_k - p(X_k^T\theta^*))\right) = u_1^T I(\theta^*)^{-1}u_2, \tag{A9}$$

$$\operatorname{var}(u_2^T I(\theta^*)^{-1}X_k(I_k - p(X_k^T\theta^*))) = u_2^T I(\theta^*)^{-1}u_2, \tag{A10}$$

$$\operatorname{cov}\left(u_2^T I(\theta^*)^{-1}X_k(I_k - p(X_k^T\theta^*)), \frac{Y_k I_k}{p(X_k^T\theta^*)}\right) = u_2^T I(\theta^*)^{-1}u_1, \tag{A11}$$

$$\operatorname{cov}(u_1^T I(\theta^*)^{-1}X_k(I_k - p(X_k^T\theta^*)), u_2^T I(\theta^*)^{-1}X_k(I_k - p(X_k^T\theta^*))) = u_1^T I(\theta^*)^{-1}u_2, \tag{A12}$$

$$\operatorname{cov}\left(\frac{I_k}{p(X_k^T\theta^*)}, u_2^T I(\theta^*)^{-1}X_k(I_k - p(X_k^T\theta^*))\right) = u_2^T I(\theta^*)^{-1}u_2. \tag{A13}$$

Together, these imply that (A8) can be written as follows:

$$= \beta^2\pi - 2\beta T + 2\beta u_1^T I(\theta^*)^{-1}u_2 + \beta^2 u_2^T I(\theta^*)^{-1}u_2 + 2\beta u_2^T I(\theta^*)^{-1}u_1 - 2\beta u_1^T I(\theta^*)^{-1}u_2 - 2\beta^2 u_2^T I(\theta^*)^{-1}u_2, \tag{A14}$$

$$= \beta^2\pi - 2\beta T + 2\beta(u_2^T I(\theta^*)^{-1}u_1) - \beta^2 u_2^T I(\theta^*)^{-1}u_2. \tag{A15}$$

Finally, recall that $\hat{\mu}_{\mathrm{HT}}$ corresponds to $\beta = 0$ and $\hat{\mu}_{\mathrm{AN}}$ corresponds to $\beta = T/\pi$. Since (A15) evaluated at $\beta = 0$ is 0, the difference in asymptotic variance between $\hat{\mu}_{\mathrm{HT}}$ and $\hat{\mu}_{\mathrm{AN}}$ will be equal to (A15) evaluated at $\beta = T/\pi$, which is $-\frac{T^2}{\pi} + \frac{2T}{\pi}u_2^T A u_1 - \frac{T^2}{\pi^2}u_2^T A u_2$, completing the argument. □

## A.2 Proof of Theorem 2

This subsection contains the proof of Theorem 2 and details for the heuristic argument for variance reduction given in the main text.

As mentioned earlier, we combine the two equations in (7) to obtain

$$\hat{\mu}^{(t)} = \frac{\hat{S}}{\left(1 - \frac{\hat{T}}{\hat{\pi}\hat{\mu}^{(t-1)}}\right)n + \frac{\hat{T}}{\hat{\pi}\hat{\mu}^{(t-1)}}\hat{n}} = \frac{\hat{S}/n}{1 - \frac{\hat{T}}{\hat{\pi}\hat{\mu}^{(t-1)}}\left(1 - \frac{\hat{n}}{n}\right)}$$

and write this as $\hat{\mu}^{(t)} = f(\hat{\mu}^{(t-1)})$, where

$$f(x) = \frac{ax}{x - b}, \quad a = \frac{\hat{S}}{n}, \quad b = \frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right). \tag{A16}$$

In this notation, the fixed point of $f$ is $x = a + b = \hat{\mu}_{\mathrm{AN}}$.

The proof now proceeds in two steps: first, we generalize the problem slightly and consider the discrete dynamical system $x^{(t)}$ initialized at $x^{(0)} = a$ and with iterative map $x^{(t)} = f(x^{(t-1)})$ for arbitrary fixed $a$ and $b$. For this dynamical system, we show that if $|a| > |b|$, then $x^{(t)}$ converges to $a + b$. Second, we show that, with high probability, the particular $a$ and $b$ defined in (A16) satisfy these conditions.

### A.2.1 Analyzing the dynamical system

The function $f$ has two fixed points at $x_1^* = 0$ and at $x_2^* = a + b$. To understand the stability of these fixed points, we compute the derivative

$$|f'(x)| = \left|\frac{ab}{(x - b)^2}\right| \Rightarrow |f'(x_1^*)| = \frac{|a|}{|b|}, \quad |f'(x_2^*)| = \frac{|b|}{|a|}.$$

A fixed point $x^*$ of the map $f$ is stable if and only if $|f'(x^*)| < 1$, so we see that if $|a| > |b|$, then $x_1^*$ is unstable and $x_2^*$ is stable, and if $|a| < |b|$, then $x_1^*$ is stable and $x_2^*$ is unstable. The case $|a| = |b|$ occurs with probability zero, and so we do not consider it.

In light of these stabilities, we should expect $x^{(t)}$ to converge to $a + b$ whenever $|a| > |b|$. The following lemma confirms this.

**Lemma A2.** *If $|a| > |b|$, then the dynamical system with initial point $x_0 = a$ and iterative map $x_{t+1} = f(x_t)$ converges to $x^* = a + b$.*

**Proof.** Our analysis relies on the two-step map

$$f(f(x)) = \frac{a^2 x}{x(a-b) + b^2}.$$

In particular, we will show that the subsequences $x_0, x_2, \dots,$ and $x_1, x_3, \dots,$ both converge to $x^*$, and this will prove the lemma. In both cases, the key observation is that

$$|f(f(x)) - x^*| = \left| \frac{a^2 x}{x(a-b) + b^2} - (a+b) \right| = \left| \frac{b^2}{x(a-b) + b^2} \right| \cdot |x - (a+b)|. \tag{A17}$$

We first consider the subsequence $x_0, x_2, \dots$, which for convenience we denote by $y_t = x_{2t}$. We claim that for any $t \geq 0$, we have

$$|y_{t+1} - x^*| \leq \frac{|b|}{|a|} |y_t - x^*|. \tag{A18}$$

The proof of the claim is by induction. For the base case, which is $t = 0$ and $y_0 = a$, we have that

$$\left| \frac{b^2}{a(a-b) + b^2} \right| \leq \left| \frac{b^2}{ab} \right| = \frac{|b|}{|a|},$$

and substituting this into (A17) gives (A18).

For the inductive step, suppose the result holds for $y_1, \dots, y_{t-1}$. We will show that $y_t$ also satisfies

$$\left| \frac{b^2}{y_t(a-b) + b^2} \right| \leq \frac{|b|}{|a|},$$

and this together with (A17) will prove the claim. The previous display is equivalent to

$$|y_t(a-b) + b^2| \geq |ab|. \tag{A19}$$

This inequality can be established by casework on the signs of $a$ and $b$. We discuss the case $a > 0$, $b > 0$ in detail; the other three cases are analogous.

If $a > 0$ and $b > 0$, then $a + b > a$, and since by the inductive hypothesis, $y_t$ is closer to $a + b$ than $a$ is, we must have $a \leq y_t \leq a + 2b$. Thus,

$$|y_t(a-b) + b^2| \geq \min_{a \leq t \leq a+2b} |t(a-b) + b^2|.$$

Since the function we are minimizing is piecewise linear, the minimum must be attained at an endpoint ($t = a$ or $t = a + 2b$) or where $t(a-b) + b^2 = 0$.

At $t = a$, the objective is $|a^2 - ab + b^2| \geq |ab|$; at $t = a + 2b$, the objective is $|a^2 + ab - b^2|$. Since $|a| > |b|$ and $a, b$ are both positive, this is equal to $a^2 + ab - b^2 \geq ab$ as well. Finally, $t(a-b) + b^2 = 0$ is not possible because this requires $t = \frac{b^2}{b-a}$ and because $a > b$, $\frac{b^2}{b-a} < 0 < a$. Thus, we conclude that (A19) holds, and this completes the induction for the case $a > 0$ and $b > 0$. The other cases are analogous, and combining them establishes (A18).

Now, by using our claim in (A18) repeatedly, we have that for any $t > 0$,

$$|y_t - x^*| \le \frac{|b|}{|a|}|y_{t-1} - x^*| \le \cdots \le \left(\frac{|b|}{|a|}\right)^t |y_0 - x^*|.$$

Since $|b| < |a|$, we thus have $|y_t - x^*| \to 0$ as $t \to \infty$, proving the convergence.

Recalling that $y_t = x_{2t}$, we have shown that the subsequence $x_0, x_2, \ldots$, converges to $x^*$. An analogous argument gives that $x_1, x_3, \ldots$ converges to $x^*$ as well, and these two together imply that $x_t \to x^*$. □

### A.2.2 High-probability guarantees

Our next lemma carries out the proof of the second part of Theorem 2, which is showing that $a$ and $b$ as defined in (A16) satisfy $|a| > |b|$ with high probability.

**Lemma A3.** *Suppose Assumption* 1 *holds and that* $\mu \neq 0$. *Then*,

$$\mathbb{P}\left(\left|\frac{\hat{S}}{n}\right| > \left|\frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right)\right|\right) \ge 1 - 4\exp\left(\frac{-2\mu^2 n}{O(M/\delta)^2}\right).$$

**Proof.** For any $0 \le \varepsilon \le |\mu|$, define the events

$$E_1 = \left\{\left|\frac{\hat{S}}{n}\right| > \left|\frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right)\right|\right\}, \quad E_2 = \left\{\left|\frac{\hat{S}}{n} - \mu\right| \le \varepsilon\right\}, \quad E_3 = \left\{\left|\frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right)\right| \le |\mu| - \varepsilon\right\}.$$

We need to lower bound the probability of $E_1$, which we do by upper bounding the probability of $E_1^c$. By the union bound,

$$\begin{aligned}
\mathbb{P}(E_1^c) &\le \mathbb{P}(E_2^c) + \mathbb{P}(E_3^c), \\
&= \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}\frac{Y_k I_k}{p_k} - \mu\right| > \varepsilon\right) + \mathbb{P}\left(\left|\frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right)\right| > |\mu| - \varepsilon\right), \\
&\le \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}\frac{Y_k I_k}{p_k} - \mu\right| > \varepsilon\right) + \mathbb{P}\left(\left|\left(1 - \frac{\hat{n}}{n}\right)\right| > \frac{|\mu| - \varepsilon}{M}\right), \\
&\le 2\exp\left(-\frac{2\varepsilon^2 n}{(M/\delta)^2}\right) + 2\exp\left(-\frac{2(|\mu| - \varepsilon)^2 n}{(1/\delta)^2}\right).
\end{aligned}$$

The second inequality follows from the bound

$$|\hat{T}| = \left|\frac{1}{\hat{n}}\sum_{k=1}^{n} Y_k \frac{1 - p_k}{p_k} \cdot \frac{I_k}{p_k}\right| \le \frac{1}{\hat{n}}\sum_{k=1}^{n}|Y_k|\frac{1 - p_k}{p_k} \cdot \frac{I_k}{p_k} \le M|\hat{\pi}|, \tag{A20}$$

which implies that $|\hat{T}/\hat{\pi}| \le M$. The third inequality follows from applying Hoeffding's inequality to each term with the bounds $|Y_k I_k/p_k| \le M/\delta$ and $|I_k/p_k| \le 1/\delta$.

Finally, we choose $\varepsilon$ to balance these two terms. The optimal choice is $\varepsilon = \frac{M}{M+1}|\mu|$, and with this value of $\varepsilon$, we conclude that

$$\mathbb{P}(E_1) \ge 1 - 4\exp\left(-\frac{2\mu^2 n}{(M+1)^2/\delta^2}\right),$$

finishing the proof. □

### A.2.3 Combining the lemmas

With these two lemmas, the proof of Theorem 2 is straightforward.

**Proof of Theorem 2.** Let $a$, $b$ be as defined in (A16).

For (i), if $|a| > |b|$, then Lemma A2 implies the result with $\hat{\mu}_{\text{lim}} = \hat{\mu}_{\text{AN}}$. If $|a| < |b|$, then an argument similar to the one in the proof of Lemma A2 establishes that $x^{(t)} \to 0$ as $t \to \infty$, and so the statement holds with $\hat{\mu}_{\text{lim}} = 0$.

For (ii), we have

$$\mathbb{P}(\hat{\mu}_{\text{lim}} \neq \hat{\mu}_{\text{AN}}) \leq \mathbb{P}\left(\left|\frac{\hat{S}}{n}\right| > \left|\frac{\hat{T}}{\hat{n}}\left(1 - \frac{\hat{n}}{n}\right)\right|\right) \leq 4\exp\left(-\frac{2\mu^2 n}{O(M/\delta)^2}\right),$$

where the first inequality is from the contrapositive of Lemma A2 and the second is from that of Lemma A3. This bound goes to zero, implying (ii). □

## A.3 Proof of Theorem 4

Before presenting the proof, we remark that for this result, we must deal explicitly with the possibility that $\sum I_k = 0$, i.e., no units receive treatment. We do not discuss this case elsewhere because our other results are asymptotic, and this event occurs with a probability that is exponentially small in $n$ and so can be ignored, but since Theorem 4 is a finite-sample result, we must account for this possibility.

**Proof.** Recall that

$$\text{var}(\hat{\mu}_{\text{HT}}) = \text{var}(Y_1) + \mathbb{E}[Y_1^2]\frac{1-p}{p}, \tag{A21}$$

so we must upper bound $\text{var}(\hat{\mu}_{\text{AN}})$ by the right-hand side of (A21). Now, under the assumption that $p_k = p$ is constant, $\hat{\mu}_{\text{AN}}$ simplifies to

$$\hat{\mu}_{\text{AN}} = \begin{cases} \dfrac{\sum Y_k I_k}{\sum I_k} & \text{if } \sum I_k \neq 0, \\ 0 & \text{if } \sum I_k = 0. \end{cases} \tag{A22}$$

For convenience, let $N = \sum I_k$. Then we decompose

$$\text{var}(\hat{\mu}_{\text{AN}}) = \mathbb{E}[\text{var}(\hat{\mu}_{\text{AN}}|N)] + \text{var}(\mathbb{E}[\hat{\mu}_{\text{AN}}|N]) \tag{A23}$$

and compute

$$\text{var}(\hat{\mu}_{\text{AN}}|N) = \begin{cases} \dfrac{\text{var}(Y_1)}{N} & \text{if } N \neq 0, \\ 0 & \text{if } N = 0 \end{cases}, \quad \mathbb{E}[\hat{\mu}_{\text{AN}}|N] = \begin{cases} \mathbb{E}[Y_1] & \text{if } N \neq 0, \\ 0 & \text{if } N = 0 \end{cases}. \tag{A24}$$

Substituting (A24) into (A23) yields

$$\text{var}(\hat{\mu}_{\text{AN}}) = \text{var}(Y_1)\mathbb{E}[1/N|N \neq 0] + \mathbb{P}(N \neq 0)(\mathbb{E}[Y_1] - \mathbb{E}[Y_1]\mathbb{P}(N \neq 0))^2 + \mathbb{P}(N = 0)(\mathbb{E}[Y_1]\mathbb{P}(N \neq 0))^2 \tag{A25}$$

$$= \text{var}(Y_1)\mathbb{E}[1/N|N \neq 0] + \mathbb{E}[Y_1^2]\mathbb{P}(N = 0)\mathbb{P}(N \neq 0) \tag{A26}$$

$$\leq \text{var}(Y_1) + \mathbb{E}[Y_1^2]\mathbb{P}(N = 0) \tag{A27}$$

$$\leq \text{var}(Y_1) + \mathbb{E}[Y_1^2]\frac{1-p}{p}, \tag{A28}$$

giving the result. Here, the first equality uses $\mathbb{P}(N = 0) + \mathbb{P}(N \neq 0) = 1$ and algebra, the second inequality uses the fact that $1/N \leq 1$ conditional on $N \neq 0$, and the third inequality uses the fact that $\mathbb{P}(N = 0) = (1-p)^n$ together with the elementary inequality $(1-p)^n \leq \frac{1-p}{p}$ for $p \in (0, 1)$. □

## A.4 Proof of Theorem 6

**Proof.** For the ease of notation, we work within a single fold of the cross-fitted estimator and let $\mathcal{T}_n$ be an auxiliary data set of size $n$ on which $\hat{\mu}(\cdot)$ and $\hat{p}(\cdot)$ are trained. This is equivalent to twofold cross-fitting; the case of more folds is analogous. Then, it is sufficient to show that the correction term we have introduced is $o_P(n^{-1/2})$. That error term is

$$= \frac{1}{\hat{\pi}}\left(\frac{1}{n}\sum_{k=1}^{n}(Y_k - \hat{\mu}(X_k))\frac{1-\hat{p}(X_k)}{\hat{p}(X_k)}\frac{I_k}{\hat{p}(X_k)}\right)\left(1 - \frac{1}{n}\sum_{k=1}^{n}\frac{I_k}{\hat{p}(X_k)}\right), \tag{A29}$$

$$\lesssim \left(\frac{1}{n}\sum_{k=1}^{n}(Y_k - \hat{\mu}(X_k))I_k\right)\left(1 - \frac{1}{n}\sum_{k=1}^{n}\frac{I_k}{\hat{p}(X_k)}\right), \tag{A30}$$

$$= \left(\frac{1}{n}\sum_{k=1}^{n}(Y_k - \mu(X_k))I_k + \frac{1}{n}\sum_{k=1}^{n}(\mu(X_k) - \hat{\mu}(X_k))I_k\right)\left(1 - \frac{1}{n}\sum_{k=1}^{n}\frac{I_k}{p(X_k)} + \frac{1}{n}\sum_{k=1}^{n}\frac{I_k}{p(X_k)} - \frac{I_k}{\hat{p}(X_k)}\right), \tag{A31}$$

$$\leq \left(O_P(n^{-1/2}) + \frac{1}{n}\sum_{k=1}^{n}|(\mu(X_k) - \hat{\mu}(X_k))I_k|\right)\left(O_P(n^{-1/2}) + \frac{1}{n}\sum_{k=1}^{n}\left|\frac{I_k}{p(X_k)} - \frac{I_k}{\hat{p}(X_k)}\right|\right), \tag{A32}$$

$$\begin{aligned} &\leq O_P(n^{-1}) + \frac{O_P(n^{-1/2})}{n}\sum_{k=1}^{n}|\mu(X_k) - \hat{\mu}(X_k)| + \frac{O_P(n^{-1/2})}{n}\sum_{k=1}^{n}\left|\frac{1}{p(X_k)} - \frac{1}{\hat{p}(X_k)}\right| \\ &\quad + \left(\frac{1}{n}\sum_{k=1}^{n}|\mu(X_k) - \hat{\mu}(X_k)|\right)\left(\frac{1}{n}\sum_{k=1}^{n}\left|\frac{1}{p(X_k)} - \frac{1}{\hat{p}(X_k)}\right|\right), \end{aligned} \tag{A33}$$

$$\begin{aligned} &\leq O_P(n^{-1}) + O_P(n^{-1/2})\left(\sup_{x\in\mathcal{X}}|\mu(x) - \hat{\mu}(x)| + \sup_{x\in\mathcal{X}}\left|\frac{1}{p(x)} - \frac{1}{\hat{p}(x)}\right|\right) \\ &\quad + \left(\frac{1}{n}\sum_{k=1}^{n}|\mu(X_k) - \hat{\mu}(X_k)|\right)\left(\frac{1}{n}\sum_{k=1}^{n}\left|\frac{1}{p(X_k)} - \frac{1}{\hat{p}(X_k)}\right|\right), \end{aligned} \tag{A34}$$

$$= O_P(n^{-1}) + O_P(n^{-1/2})o_P(1) + \left(\frac{1}{n}\sum_{k=1}^{n}|\mu(X_k) - \hat{\mu}(X_k)|\right)\left(\frac{1}{n}\sum_{k=1}^{n}\left|\frac{1}{p(X_k)} - \frac{1}{\hat{p}(X_k)}\right|\right), \tag{A35}$$

where the second inequality follows from the fact that, by the consistency of $\hat{p}(\cdot)$, we have $\delta/2 \leq \hat{p}(x) \leq 1 - \delta/2$ for all $x \in \mathcal{X}$ for sufficiently large $n$, the fourth inequality applies the triangle inequality and the CLT for i.i.d. sums, and the final equality uses Assumption 2. (Note that since $p(\cdot)$ is bounded, the consistency of $\hat{p}$ implies the consistency of $1/\hat{p}$ as well.)

Examining (A35), the first two terms are $o_P(n^{-1/2})$ as needed, so it remains to analyze the final term. We thus compute

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n}|\mu(X_k) - \hat{\mu}(X_k)|\right)^2\left(\frac{1}{n}\sum_{k=1}^{n}\left|\frac{1}{p(X_k)} - \frac{1}{\hat{p}(X_k)}\right|\right)^2\right]$$

as

$$\leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n}(\mu(X_k) - \hat{\mu}(X_k))^2\right)\left(\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{p(X_k)} - \frac{1}{\hat{p}(X_k)}\right)^2\right)\right], \tag{A36}$$

$$= \mathbb{E}\left[\frac{1}{n^2}\sum_{k\neq\ell}(\mu(X_k) - \hat{\mu}(X_k))^2\left(\frac{1}{p(X_\ell)} - \frac{1}{\hat{p}(X_\ell)}\right)^2\right] + o(n^{-1}), \tag{A37}$$

$$= \mathbb{E}\left[\frac{1}{n^2}\sum_{k\neq\ell}\mathbb{E}\left[(\mu(X_k)-\hat{\mu}(X_k))^2\left(\frac{1}{p(X_\ell)}-\frac{1}{\hat{p}(X_\ell)}\right)^2|\mathcal{T}_n\right]\right] + o(n^{-1}), \tag{A38}$$

$$= \mathbb{E}\left[\frac{n(n-1)}{n^2}\mathbb{E}\left[(\mu(X_k)-\hat{\mu}(X_k))^2\left(\frac{1}{p(X_\ell)}-\frac{1}{\hat{p}(X_\ell)}\right)^2|\mathcal{T}_n\right]\right] + o(n^{-1}), \tag{A39}$$

$$\leq \mathbb{E}\left[\mathbb{E}[(\mu(X_k)-\hat{\mu}(X_k))^2|\mathcal{T}_n]\mathbb{E}\left[\left(\frac{1}{p(X_\ell)}-\frac{1}{\hat{p}(X_\ell)}\right)^2|\mathcal{T}_n\right]\right] + o(n^{-1}), \tag{A40}$$

where the first inequality is Cauchy-Schwarz, the second equality expands the product and drops the $k = \ell$ terms, the fourth equality uses the fact that the terms of the sum are equal after conditioning on $\mathcal{T}_n$, and the fifth equality uses the fact that the two errors are independent after conditioning on $\mathcal{T}_n$.

By Assumption 3, the product of conditional expectations in (A40) is $o_P(n^{-1})$. Since $\mu(X_k)$ and $p(X_\ell)$ are bounded by Assumption 1, and $\hat{\mu}(\cdot)$ and $\hat{p}(\cdot)$ are sup-norm consistent by Assumption 2, that product of conditional expectations is eventually dominated by a constant, and so the expectation in (A40) is $o(n^{-1})$ as well, completing the proof. □

## A.5 Proof of Theorem 7

Our proof closely follows ideas and tools developed in ref. [17].

**Proof.** Our goal is to control $V(\pi^*) - V(\hat{\pi}_{\mathrm{AN}})$. We do this by first writing $V(\pi^*) - V(\hat{\pi}_{\mathrm{AN}})$ as follows:

$$= V(\pi^*) - \hat{V}_{\mathrm{AN}}(\hat{\pi}_{\mathrm{AN}}) + \hat{V}_{\mathrm{AN}}(\hat{\pi}_{\mathrm{AN}}) - V(\hat{\pi}_{\mathrm{AN}}), \tag{A41}$$

$$\leq V(\pi^*) - \hat{V}_{\mathrm{AN}}(\pi^*) + \sup_{\pi\in\Pi}|\hat{V}_{\mathrm{AN}}(\pi) - V(\pi)|, \tag{A42}$$

$$\leq 2\sup_{\pi\in\Pi}|\hat{V}_{\mathrm{AN}}(\pi) - V(\pi)|, \tag{A43}$$

$$\leq 2\sup_{\pi\in\Pi}|\hat{V}_{\mathrm{AN}}(\pi) - \hat{V}_{\mathrm{IPW}}(\pi)| + 2\sup_{\pi\in\Pi}|\hat{V}_{\mathrm{IPW}}(\pi) - V(\pi)|. \tag{A44}$$

The second term of (A44) is bounded in Theorem 2.1 of ref. [17], and so we are interested in bounding the first term. By the definitions of $\hat{V}_{\mathrm{IPW}}$ and $\hat{V}_{\mathrm{AN}}$, that first term of (A44) is

$$= \sup_{\pi\in\Pi}\left|\frac{\sum_{k=1}^n Y_k\frac{1-\mathbb{P}(I_k=\pi(X_k)\mid X_k)}{\mathbb{P}(I_k=\pi(X_k)\mid X_k)}\frac{\mathbf{1}\{I_k=\pi(X_k)\}}{\mathbb{P}(I_k=\pi(X_k)\mid X_k)}}{\sum_{k=1}^n\frac{1-\mathbb{P}(I_k=\pi(X_k)\mid X_k)}{\mathbb{P}(I_k=\pi(X_k)\mid X_k)}\frac{\mathbf{1}\{I_k=\pi(X_k)\}}{\mathbb{P}(I_k=\pi(X_k)\mid X_k)}}\left(1-\frac{1}{n}\sum_{k=1}^n\frac{\mathbf{1}\{I_k=\pi(X_k)\}}{\mathbb{P}(I_k=\pi(X_k)|X_k)}\right)\right|, \tag{A45}$$

$$\leq M\sup_{\pi\in\Pi}\left|1-\frac{1}{n}\sum_{k=1}^n\frac{\mathbf{1}\{I_k=\pi(X_k)\}}{\mathbb{P}(I_k=\pi(X_k)|X_k)}\right|, \tag{A46}$$

by a calculation analogous to the one in (A20). Combining (A44) and (A46) gives

$$\mathbb{E}[V(\pi^*) - V(\hat{\pi}_{\mathrm{AN}})] \leq M\mathbb{E}\left[\sup_{\pi\in\Pi}\left|1-\frac{1}{n}\sum_{k=1}^n\frac{\mathbf{1}\{I_k=\pi(X_k)\}}{\mathbb{P}(I_k=\pi(X_k)|X_k)}\right|\right] + O\left(\frac{M}{\delta}\sqrt{\frac{\mathrm{VC}(\Pi)}{n}}\right). \tag{A47}$$

Finally, the expectation in the first term of (A47) can be bounded by using standard empirical process tools; in particular, Lemmas A.1 and A.4 of ref. [17] imply that it is $O\left(\frac{1}{\delta}\sqrt{\frac{\mathrm{VC}(\Pi)}{n}}\right)$, finishing the proof. □

# B Further experiments

## B.1 RMSE plots for Swiss data

The plots in Figure 3 show the RMSE of the various problem specifications of our survey sampling simulations in Section 5 as a function of $\lambda$. Interestingly, the optimal choices of $\lambda$ are all between 0 and 1, suggesting that the Hájek estimator is actually "over-normalizing" in this case.

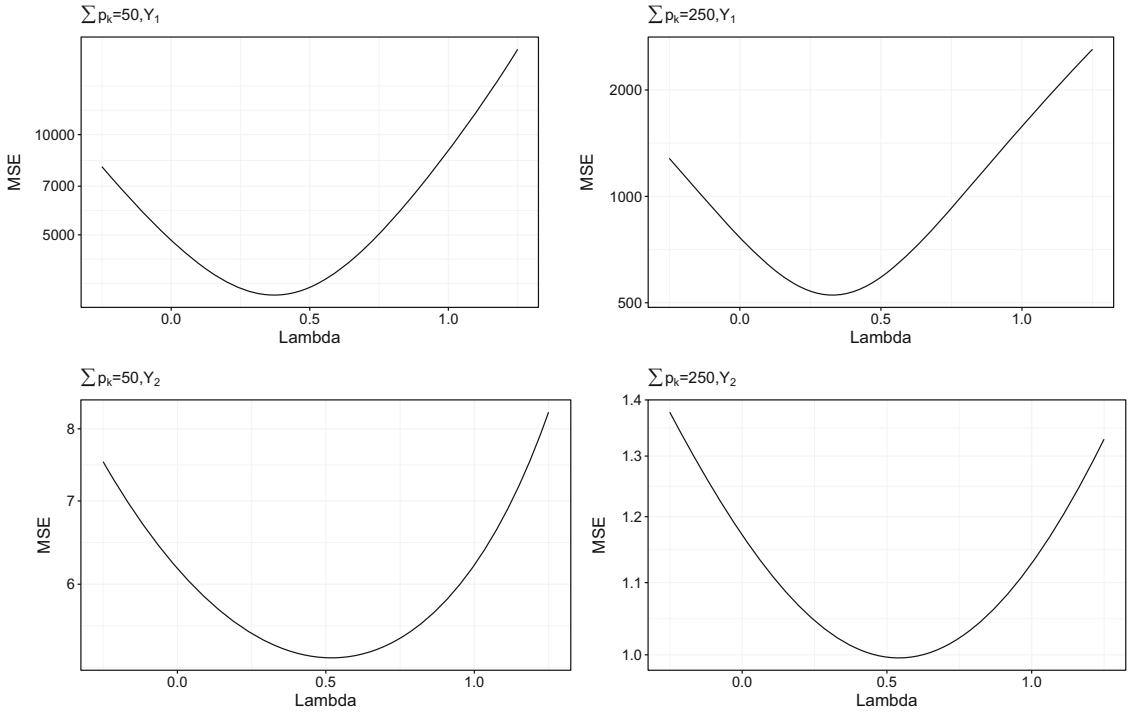


**Figure A1:** MSE of $\hat{\mu}_\lambda$ as a function of $\lambda$ for the four problem specifications of the Swiss municipality data.

## B.2 Overlap and bias

In this section, we further study the bias of $\hat{\mu}_\lambda$ and $\hat{\mu}_{\mathrm{AN}}$ in small samples. To do this, we generate data from the normal model (18) with $\mu = 0.1$ and $\theta = 0.9$, and attempt to estimate the mean $\mu$ of $Y_k(1)$ for varying values of $n$, the sample size, and also for varying values of $\delta$, the cut-off to which the $p_k$ are truncated (i.e., when $\delta$ is smaller, we allow smaller $p_k$, and thus worse overlap). Estimates for the bias of $\hat{\mu}_\lambda$ with $\lambda = 0.25$, 0.5, and 0.75, as well as for $\hat{\mu}_{\mathrm{AN}}$ under these conditions are shown in Figure A2. These experiments confirm the expected behavior of the bias based on theory: the bias of $\hat{\mu}_\lambda$ decreases at approximately a $1/n$ rate, and is larger in the presence of worse overlap. Thus, our proposal to select $\lambda$ by minimizing the asymptotic variance can also be expected to reduce MSE as long as the data have good overlap. The bias of $\hat{\mu}_{\mathrm{AN}}$ shows similar behavior and also appears to decrease more quickly in the presence of better overlap.

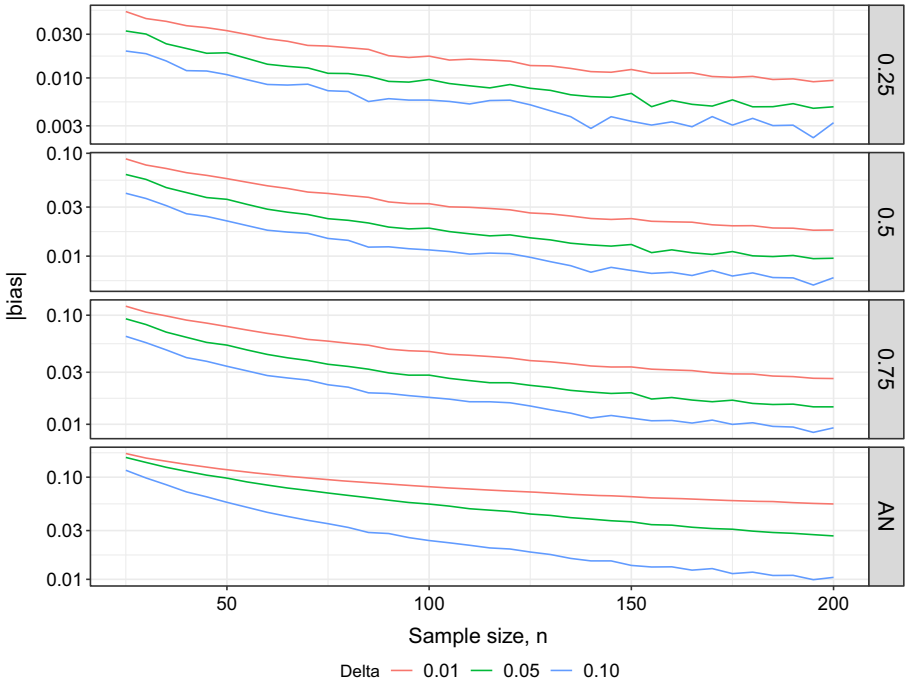


**Figure A2:** The absolute value of the biases of $\hat{\mu}_\lambda$ for $\lambda = 0.25$, 0.5, and 0.75 and $\hat{\mu}_{\text{AN}}$ as $n$, the sample size, and $\delta$, the lower bound on the overlap, vary in the normal model (18). We see that bias decreases with $n$ at roughly a $1/n$ rate, as the theory suggests, and that overlap plays an important role in the bias: worse overlap consistently leads to larger bias.

# C Joint adaptive normalization in ATE estimation

This appendix explores some subtleties of how the adaptive normalizations should be chosen for ATE estimation. In particular, the estimator in (16) is equivalent to selecting $\lambda$ in (1) to separately minimize the asymptotic variance of the two mean estimates. However, this "plug-in" use of adaptive normalization ignores the fact that the asymptotic variance of an adaptively normalized ATE estimator depends not only on the variances of the two mean estimators but also their covariance.

In what follows, we jointly choose the normalizations of each mean estimator to minimize the asymptotic variance of estimating the combined estimand $\tau$. Specifically, we define

$$\hat{\mu}_{1,\lambda_1} = \frac{\hat{S}_1}{(1-\lambda_1)n + \lambda_1 \hat{n}_1}, \quad \hat{\mu}_{0,\lambda_0} = \frac{\hat{S}_0}{(1-\lambda_0)n + \lambda_0 \hat{n}_0},$$

where

$$\hat{S}_1 = \sum_{k=1}^{n} \frac{Y_k(1) I_k}{p_k}, \quad \hat{S}_0 = \sum_{k=1}^{n} \frac{Y_k(0)(1-I_k)}{1-p_k}, \quad \hat{n}_1 = \sum_{k=1}^{n} \frac{I_k}{p_k}, \quad \hat{n}_0 = \sum_{k=1}^{n} \frac{1-I_k}{1-p_k}.$$

Combining these estimators leads to the estimator $\hat{\tau}_{\lambda_1,\lambda_0} = \hat{\mu}_{1,\lambda_1} - \hat{\mu}_{0,\lambda_0}$ of $\tau$. To follow the program of Section 3, we seek to choose $\lambda_1$ and $\lambda_0$ to minimize the following asymptotic variance.

**Theorem A1.** *Assuming that both $Y_k(1)$ and $Y_k(0)$ satisfy Assumption 1, we have*

$$\sqrt{n}\left(\hat{\tau}_{\lambda_1,\lambda_0} - \tau\right) \xrightarrow{d} N(0, \sigma^2),$$

*where*

$$\sigma^2 = \lambda_1^2 \mu_1^2 \pi_1 + \lambda_0^2 \mu_0^2 \pi_0 - 2\lambda_1 \mu_1 (T_1 + \mu_0) + 2\lambda_0 \mu_0 (-\mu_1 - T_0) + 2\lambda_1 \lambda_0 \mu_1 \mu_0 + C,$$

*for*

$$\pi_1 = \mathbb{E}\left[\frac{1-p_k}{p_k}\right], \quad \pi_0 = \mathbb{E}\left[\frac{p_k}{1-p_k}\right], \quad T_1 = \mathbb{E}\left[Y_k(1)\frac{1-p_k}{p_k}\right], \quad T_0 = \mathbb{E}\left[Y_k(0)\frac{p_k}{1-p_k}\right],$$

*and C denotes terms that do not depend on either $\lambda_1$ or $\lambda_0$.*

As with our other CLTs, this is a routine delta method calculation.

**Proof.** Define the vector

$$\hat{\beta} = \left(\frac{1}{n}\sum_{k=1}^{n}\frac{Y_k(1)I_k}{p_k}, \frac{1}{n}\sum_{k=1}^{n}\frac{Y_k(0)(1-I_k)}{1-p_k}, \frac{1}{n}\sum_{k=1}^{n}\frac{I_k}{p_k}, \frac{1}{n}\sum_{k=1}^{n}\frac{1-I_k}{1-p_k}\right)$$

with mean $\beta = (\mu_1, \mu_0, 1, 1)$. By the same arguments as in A1, $\hat{\beta}$ satisfies the usual CLT for the mean of i.i.d. random variables, and so we can apply the delta method with the function $f(x, y, z, w) = \frac{x}{1-\lambda_1+\lambda_1 z} - \frac{y}{1-\lambda_0+\lambda_0 w}$. The relevant gradient is $(1, -1, -\lambda_1\mu_1, \lambda_0\mu_0)$, and so the asymptotic variance of $\hat{\tau}_{\lambda_1,\lambda_0}$ is

$$\lambda_1^2\mu_1^2\mathrm{var}\left(\frac{I_k}{p_k}\right) + \lambda_0^2\mu_0^2\mathrm{var}\left(\frac{1-I_k}{1-p_k}\right) - 2\lambda_1\mu_1\left(\mathrm{cov}\left(\frac{I_k}{p_k}, \frac{Y_k(1)I_k}{p_k}\right) - \mathrm{cov}\left(\frac{I_k}{p_k}, \frac{Y_k(0)(1-I_k)}{1-p_k}\right)\right),$$
$$+2\lambda_0\mu_0\left(\mathrm{cov}\left(\frac{1-I_k}{1-p_k}, \frac{Y_k(1)I_k}{p_k}\right) - \mathrm{cov}\left(\frac{1-I_k}{1-p_k}, \frac{Y_k(0)(1-I_k)}{1-p_k}\right)\right) - 2\lambda_1\lambda_0\mu_1\mu_0\mathrm{cov}\left(\frac{I_k}{p_k}, \frac{1-I_k}{1-p_k}\right) + C,$$

where $C$ denotes terms that do not depend on either $\lambda_1$ or $\lambda_2$.

Finally, we can compute

$$\mathrm{var}\left(\frac{I_k}{p_k}\right) = \pi_1, \quad \mathrm{var}\left(\frac{1-I_k}{1-p_k}\right) = \pi_0$$

and

$$\mathrm{cov}\left(\frac{I_k}{p_k}, \frac{Y_k(1)I_k}{p_k}\right) = T_1, \quad \mathrm{cov}\left(\frac{I_k}{p_k}, \frac{Y_k(0)(1-I_k)}{1-p_k}\right) = -\mu_0,$$

and

$$\mathrm{cov}\left(\frac{1-I_k}{1-p_k}, \frac{Y_k(1)I_k}{p_k}\right) = -\mu_1, \quad \mathrm{cov}\left(\frac{1-I_k}{1-p_k}, \frac{Y_k(0)(1-I_k)}{1-p_k}\right)T_0,$$

and finally, $\mathrm{cov}\left(\frac{I_k}{p_k}, \frac{1-I_k}{1-p_k}\right) = -1$. Substituting these in gives the result. □

We now minimize the asymptotic variance of Theorem A1 in $\lambda_1$ and $\lambda_0$. The first-order stationary conditions tell us that the optimal pair $(\lambda_1^*, \lambda_0^*)$ will satisfy

$$2\lambda_1^*\mu_1^2\pi_1 - 2\mu_1(T_1 + \mu_0) + 2\lambda_0^*\mu_1\mu_0 = 0 \;\Rightarrow\; \lambda_1^* = \frac{T_1 + \mu_0 - \lambda_0^*\mu_0}{\mu_1\pi_1}$$

and

$$2\lambda_0^*\mu_0^2\pi_0 - 2\mu_0(\mu_1 + T_0) + 2\lambda_1^*\mu_1\mu_0 = 0 \;\Rightarrow\; \lambda_0^* = \frac{T_0 + \mu_1 - \lambda_1^*\mu_1}{\mu_0\pi_0}.$$

As mentioned earlier, we replace $T_0$, $T_1$, $\pi_0$, and $\pi_1$ with IPW estimates $\hat{T}_0$, $\hat{T}_1$, $\hat{\pi}_0$, and $\hat{\pi}_1$, and then jointly estimate $\lambda_0$, $\lambda_1$, $\mu_0$, and $\mu_1$ by solving the system of fixed point equations:

$$\hat{\lambda}_{1,\mathrm{AN}} = \frac{\hat{T}_1 + \hat{\mu}_{0,\mathrm{AN}} - \hat{\lambda}_{0,\mathrm{AN}}\hat{\mu}_{0,\mathrm{AN}}}{\hat{\mu}_{1,\mathrm{AN}}\hat{\pi}_1}, \quad \hat{\lambda}_{0,\mathrm{AN}} = \frac{\hat{T}_0 + \hat{\mu}_{1,\mathrm{AN}} - \hat{\lambda}_{1,\mathrm{AN}}\hat{\mu}_{1,\mathrm{AN}}}{\hat{\mu}_{0,\mathrm{AN}}\hat{\pi}_0} \tag{A48}$$

and

$$\hat{\mu}_{1,\mathrm{AN}} = \frac{\hat{S}_1}{(1-\hat{\lambda}_{1,\mathrm{AN}})n + \hat{\lambda}_{1,\mathrm{AN}}\hat{n}_1}, \quad \hat{\mu}_{0,\mathrm{AN}} = \frac{\hat{S}_0}{(1-\hat{\lambda}_{0,\mathrm{AN}})n + \hat{\lambda}_{0,\mathrm{AN}}\hat{n}_0}. \tag{A49}$$

(Note that we have slightly overloaded the notation $\hat{\mu}_{1,\mathrm{AN}}$, which is used differently in Section 4.2. In the entirety of this appendix, the definition mentioned earlier is the one used.)

To actually solve this system, we focus first on (A48), noting that it is linear in $\hat{\lambda}_{1,\mathrm{AN}}\hat{\mu}_{1,\mathrm{AN}}$ and $\hat{\lambda}_{0,\mathrm{AN}}\hat{\mu}_{0,\mathrm{AN}}$ and has solution

$$\hat{\lambda}_{1,\mathrm{AN}}\hat{\mu}_{1,\mathrm{AN}} = \frac{\hat{T}_1\hat{\pi}_0 + \hat{\mu}_{0,\mathrm{AN}}\hat{\pi}_0 - \hat{T}_0 - \hat{\mu}_{1,\mathrm{AN}}}{\hat{\pi}_0\hat{\pi}_1 - 1}, \quad \hat{\lambda}_{0,\mathrm{AN}}\hat{\mu}_{0,\mathrm{AN}} = \frac{\hat{T}_0\hat{\pi}_1 + \hat{\mu}_{1,\mathrm{AN}}\hat{\pi}_1 - \hat{T}_1 - \hat{\mu}_{0,\mathrm{AN}}}{\hat{\pi}_0\hat{\pi}_1 - 1}. \tag{A50}$$

Then, we rewrite (A49) as follows:

$$\hat{\mu}_{1,\mathrm{AN}} = \hat{\mu}_{1,\mathrm{AN}}\hat{\lambda}_{1,\mathrm{AN}}\left(1 - \frac{\hat{n}_1}{n}\right) + \frac{\hat{S}_1}{n}, \quad \hat{\mu}_{0,\mathrm{AN}} = \hat{\mu}_{0,\mathrm{AN}}\hat{\lambda}_{0,\mathrm{AN}}\left(1 - \frac{\hat{n}_0}{n}\right) + \frac{\hat{S}_0}{n}. \tag{A51}$$

By using (A50), we can conclude that $\hat{\mu}_{1,\mathrm{AN}}$ and $\hat{\mu}_{0,\mathrm{AN}}$ must satisfy the system of linear equations:

$$\begin{bmatrix} 1 + \frac{1}{\hat{\pi}_0\hat{\pi}_1 - 1}\left(1 - \frac{\hat{n}_1}{n}\right) & -\frac{\hat{\pi}_0}{\hat{\pi}_0\hat{\pi}_1 - 1}\left(1 - \frac{\hat{n}_1}{n}\right) \\ -\frac{\hat{\pi}_1}{\hat{\pi}_0\hat{\pi}_1 - 1}\left(1 - \frac{\hat{n}_0}{n}\right) & 1 + \frac{1}{\hat{\pi}_0\hat{\pi}_1 - 1}\left(1 - \frac{\hat{n}_0}{n}\right) \end{bmatrix} \begin{bmatrix} \hat{\mu}_{1,\mathrm{AN}} \\ \hat{\mu}_{0,\mathrm{AN}} \end{bmatrix} = \begin{bmatrix} \frac{\hat{S}_1}{n} + \frac{\hat{T}_1\hat{\pi}_0 - \hat{T}_0}{\hat{\pi}_0\hat{\pi}_1 - 1}\left(1 - \frac{\hat{n}_1}{n}\right) \\ \frac{\hat{S}_0}{n} + \frac{\hat{T}_0\hat{\pi}_1 - \hat{T}_1}{\hat{\pi}_0\hat{\pi}_1 - 1}\left(1 - \frac{\hat{n}_0}{n}\right) \end{bmatrix}.$$

Finally, solving these equations to recover $\hat{\mu}_{1,\mathrm{AN}}$, $\hat{\mu}_{0,\mathrm{AN}}$, and $\hat{\tau}_{\mathrm{AN}^2} \coloneqq \hat{\mu}_{1,\mathrm{AN}} - \hat{\mu}_{0,\mathrm{AN}}$ can be done using any standard matrix inversion technique. We refer to the resulting estimator as $\hat{\tau}_{\mathrm{AN}^2}$ to distinguish it from $\hat{\tau}_{\mathrm{AN}}$.

Somewhat surprisingly, $\hat{\tau}_{\mathrm{AN}^2}$ does not always improve on the MSE of $\hat{\tau}_{\mathrm{AN}}$. The results of a simulation are shown in Figure A3. Both the normal model and power law model simulations suggest that $\hat{\tau}_{\mathrm{AN}^2}$ and $\hat{\tau}_{\mathrm{AN}}$ are preferable in different regimes. This suggests that, although $\hat{\tau}_{\mathrm{AN}^2}$ is estimating the optimal values of $\lambda_1$ and $\lambda_0$, the complicated functional form of the data-dependent estimates of the optimal values sometimes inflates the variance. Although $\hat{\tau}_{\mathrm{AN}}$ is estimating a sub-optimal pair of values $\lambda_1$ and $\lambda_0$, its estimates of those values are lower variance and thus sometimes lead to a lower variance estimator. Examining the results in Figure A3, it seems that $\hat{\tau}_{\mathrm{AN}^2}$ is preferable to $\hat{\tau}_{\mathrm{AN}}$ in the presence of strong correlation between $Y_k$ and $p_k$, but not otherwise.

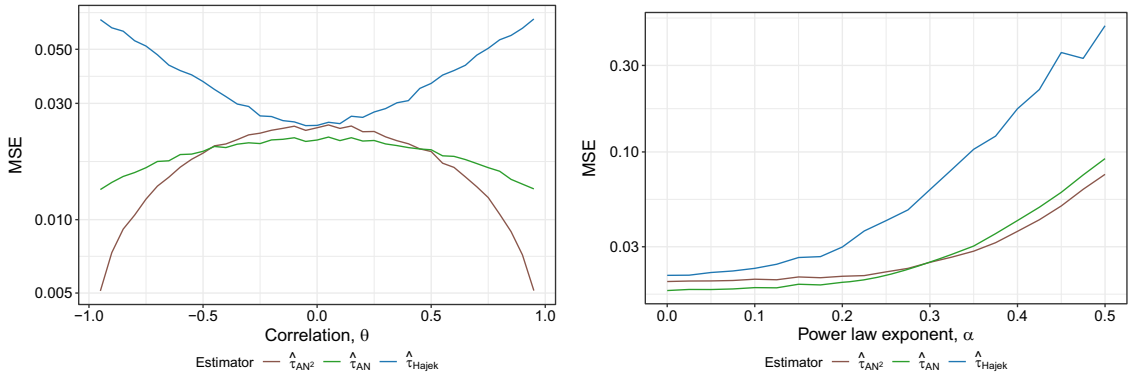


**Figure A3:** Comparison of $\hat{\tau}_{\mathrm{AN}^2}$ and $\hat{\tau}_{\mathrm{AN}}$ in the normal model with $n = 500$, $\mu = 1$ (left) and power law model with $n = 500$ (right). In both models, we see that $\hat{\tau}_{\mathrm{AN}^2}$ is preferable to $\hat{\tau}_{\mathrm{AN}}$ when there is a strong correlation between responses and treatment probabilities.