Optimal Estimation when Researcher and Social Preferences are Misaligned

Jann Spiess

First version: August 2017
This version: December 8, 2018

Abstract

Econometric analysis typically focuses on the statistical properties of fixed estimators and ignores researcher choices. In this article, I approach the analysis of experimental data as a mechanism-design problem that acknowledges that researchers choose between estimators, sometimes based on the data and often according to their own preferences. Specifically, I focus on covariate adjustments, which can increase the precision of a treatment-effect estimate, but open the door to bias when researchers engage in specification searches. First, I establish that unbiasedness is a requirement on the estimation of the average treatment effect that aligns researchers’ preferences with the minimization of the mean-squared error relative to the truth, and that fixing the bias can yield an optimal restriction in a minimax sense. Second, I provide a constructive characterization of all treatment-effect estimators with fixed bias as sample-splitting procedures. Third, I show how these results imply flexible pre-analysis plans for randomized experiments that include beneficial specification searches and offer an opportunity to leverage machine learning.

Jann Spiess, Microsoft Research New England. [jspiess@stanford.edu] For their guidance I am indebted to Sendhil Mullainathan, Alberto Abadie, Elie Tamer, and Gary Chamberlain. For valuable comments or discussions I also thank Laura Blattner, Avi Feller, Edward Glaeser, Nathaniel Hendren, Simon Jäger, Maximilian Kasy, Lawrence Katz, Scott Kominers, Eben Lazarus, Shengwu Li, Mikkel Plagborg-Møller, Ashesh Rambachan, Jonathan Roth, Elizabeth Santorella, Klaus M. Schmidt, and seminar audiences at Harvard/MIT, MSR, Northwestern, Princeton, Yale, NYU, Columbia, Chicago, Penn, Penn State, UCL, LSE, Stanford, Queen Mary, CERGE-EI, Copenhagen, UT Austin, and Toronto.
INTRODUCTION

There is a tension between flexibility and robustness in empirical work. Consider an investigator who estimates a treatment effect from experimental data. If the investigator has the freedom to choose a specification that adjusts for control variables, her choice can improve the precision of the estimate. However, the investigator’s specification search may also produce an estimate that reflects a preference for publication or ideology instead of a more precise guess of the truth. To solve this problem, we sometimes tie the investigator’s hands and restrict her to a simple specification, like a difference in averages. In contrast, this article characterizes flexible estimators that leverage the data and researchers’ expertise, and do not also reflect researchers’ preferences.

To characterize optimal estimators when researcher and social preferences are misaligned, I approach the analysis of experimental data as a mechanism-design problem. Concretely, I consider a designer and an investigator who are engaged in the estimation of an average treatment effect. As the designer, we aim to obtain a precise estimate of the truth (which I capture in terms of mean-squared error). I assume however that the investigator may care about the value of the estimate and not only its precision. For example, the investigator may have a preference for large estimates in order to get published. The investigator picks an estimator based on her private information about the specific experiment. The designer chooses optimal constraints on the estimation by the investigator.

First, I argue that we should not leave the decision over the bias of an estimator to the investigator, and motivate a restriction to estimators with fixed bias. More

1A literature in statistics dating back to at least Sterling (1959) and Tullock (1959), and most strongly associated with the work of Leamer (e.g. 1974, 1978), acknowledges that empirical estimates reflect not just data, but also researcher motives. Fears of biases have been fueled more recently by replication failures (Open Science Collaboration, 2015), anomalies in published p-values (Brodeur et al. 2016), and empirical evidence for publication biases (Andrews and Kasy 2017). Christensen and Miguel (2018) survey evidence and discuss practices that aim to improve transparency and reproducibility. Young (2017) documents the sensitivity of treatment-effect estimates in experiments to the choice of specification.


Abstractly, the designer could represent professional norms. Concretely, it could represent a journal setting standards for the analysis of randomized controlled trials, or the U.S. Food and Drug Administration (FDA) imposing rules for the evaluation of new drugs.
precisely, I prove that setting the bias aligns the incentives of the investigator and
the designer and is a minimax optimal solution to the designer’s problem under
suitable assumptions on preferences. Allowing the investigator to choose the bias can, in principle, improve overall precision through a reduction in the variance. But
an investigator could use her control over the bias to reflect her preferences rather
than her private information. Among unbiased estimators, for example, even an
investigator who wants to obtain an estimate close to some large, fixed value will
still choose an estimator that minimizes the variance.

Second, having motivated a bias restriction, I prove that every estimator of the
average treatment effect with fixed bias has a sample-splitting representation. As the
starting point for this representation, consider a familiar estimator that is unbiased,
namely the difference in averages between treatment and control groups. We can
adjust this estimator for control variables by a procedure that splits the sample
into two groups. From the first group, we calculate regression adjustments that we
subtract from the outcomes in the second group. The updated difference in averages
is still unbiased by construction. Though this procedure appears specific, I prove that
any estimator with fixed bias can be represented by multiple such sample-splitting
steps. Unbiased estimators, for example, can differ from a difference in averages only
by leave-one-out or leave-two-out regression adjustments of individual outcomes.

Third, focusing specifically on estimation with a fixed shrinkage factor (including
unbiased estimation), I relate the investigator’s solution to a prediction problem.
By the sample-splitting representation, I can write every fixed-bias estimator of the
average treatment effect in terms of a set of regression adjustments. When choosing
from this restricted set of estimators, under a balance assumption on her prior,
for a fixed shrinkage factor, and in the case of a known treatment probability, the
investigator picks regression adjustments that minimize prediction risk for a specific
loss function.

As a practical implication, my results motivate and describe flexible yet robust

---

4. This result echoes Frankel’s (2014) characterization of simple delegation mechanisms that align
an agent’s choices with a principal’s preferences by fixing budgets. In Section 4, I explore the
similarities of my solution to results in the mechanism-design literature on delegation that goes
back to Holmström (1978, 1984), and I exploit these parallels in the proof of my minimax result.

5. In particular, for known treatment probability, I show that all unbiased estimators of the
sample-average treatment effect take the form of the “leave-one-out potential outcomes” (LOOP)
estimator from Wu and Gagnon-Bartsch (2017), which is a special case of Aronow and Middleton’s
pre-analysis plans for the analysis of experimental data. There are two types of flexible pre-analysis plan that achieve precise estimation of treatment effects without leaving room for bias from specification searches. In the first type, the investigator commits to an algorithm that predicts outcomes from covariates. This algorithm can engage in automated specification searches to learn a good model from the data. Adjusting outcomes by its fitted out-of-sample predictions will yield an unbiased estimator.

There is a second, more flexible type of pre-analysis plan that achieves unbiased and precise estimation without the investigator committing to her specification searches in advance. In this second type of pre-analysis plan, the investigator only commits to splitting the data and distributing subsamples to her research team. Each researcher then engages in specification searches on a part of the data and reports back a prediction function. I characterize estimators of the treatment effect with fixed shrinkage (including unbiased estimators) that delegate the estimation of some or all regression adjustments in this way. Delegation to one researcher improves over simple pre-analysis plans. Delegation to at least two researchers asymptotically attains the semi-parametric efficiency bound of Hahn (1998) under assumptions that apply to most parametric and many semi- and non-parametric estimators of the regression adjustments.

The results in this article relate to the practice of sample splitting in econometrics, statistics, and machine learning. From Hájek (1962) to Jackknife IV (Angrist et al., 1999), model selection (e.g. Hansen and Racine, 2012), and time-series forecasting (see e.g. Diebold, 2015, Hirano and Wright, 2017), sample splitting is used as a tool to avoid bias by construction. Wager and Athey (2017) highlight the role of sample splitting in the estimation of heterogeneous treatment effects. Chernozhukov et al. (2017b) show its relevance in achieving valid and efficient inference in high-dimensional observational data. Schorfheide and Wolpin (2012, 2016) provide a justification of sample splitting in a principal-agent framework. In a similar spirit,
my results show that sample splitting is a feature of optimal estimators.

Moreover, I build upon an active literature in statistics on regression adjustments to experimental data. Freedman (2008) and Lin (2013) discuss the bias of linear-least squares regression adjustments. Most closely related to the investigator’s solution in my paper, Wu and Gagnon-Bartsch (2017) propose the “leave-one-out potential outcomes” (LOOP) estimator that yields regression adjustments without bias, which coincides with the estimator chosen by the researcher in my setting for the case of zero bias and known treatment probability. Wager et al. (2016) propose a related sample-splitting estimator based on separate prediction problems in the treatment and control groups. Rothe (2018) obtains a similar family of estimators from the efficient influence function. Relative to this literature, I motivate a bias restriction and fully characterize estimators with given bias.

Relatedly, this article contributes to a growing literature that employs machine learning in program evaluation (see e.g. Mullainathan and Spiess, 2017). As in Wager et al. (2016) and Wu and Gagnon-Bartsch (2017), the sample-splitting construction allows researchers to leverage machine learning in estimating average treatment effects in experimental data. Bloniarz et al. (2016) specifically use the LASSO to select among control variables in experiments. Athey and Imbens (2016) use regression trees to estimate heterogeneous treatment effects. Chernozhukov et al. (2017) estimate treatment effects from high-dimensional observational data. I discuss machine learning in a principal-agent framework, which is mostly agnostic about specific algorithms.

My analysis is limited in three ways. First, I assume randomization, and thus that identification is resolved by design. My findings extend to known propensity scores, stratified and conditional randomization, and corresponding identification from quasi-experiments. Second, I focus on the analysis of a single experiment, and neither on repeated interactions between designer and investigator, nor on the publication policies that may shape investigators’ preferences. Third, I characterize optimal estimators in terms of sample splitting and regression adjustments, but I do not discuss in depth the construction of these adjustments. A large and active literature that straddles econometrics, statistics, and machine learning provides guidance.

---

When treatment is not random, endogeneity creates auxiliary prediction tasks in the propensity score that interact with fitting regression adjustments (Robins and Rotnitzky, 1995; Chernozhukov et al., 2017b). Finite-sample unbiased estimation may then be infeasible absent strong parametric assumptions, and inference may be invalid when these additional prediction tasks are ignored (Belloni et al., 2014).
and tools to provide efficient prediction functions.

The remaining article is structured as follows. Section 1 introduces the main ideas behind my theoretical results in a stylized example. In Section 2 I formally lay out the specific estimation setting and my mechanism-design approach. I preview my main theoretical results in Section 3. In Section 4 I solve for optimal restrictions on the investigator’s estimation. Section 5 characterizes estimators with fixed bias and discusses the investigator’s second-best choice. For the case that full ex-ante commitment is infeasible or impractical, Section 6 considers unbiased estimators that permit ex-post researcher input.

1 A Simple Example

I consider the estimation of a sample-average treatment effect. But the main features of my analysis are already apparent when we focus on a single unit within that sample. As an example, I discuss the estimation of the effect of random assignment to a job-training program on the earnings of one specific worker.\footnote{Throughout, I focus on intent-to-treat effects, so I do not consider take-up or the use of random assignment as an instrument.}

1.1 Estimating the Unit-Level Causal Effect

The causal effect on unit $i$ is $\tau_i = y_i(1) - y_i(0)$, where $y_i(1), y_i(0)$ are the potential outcomes when assigned to treatment or control, respectively. For assignment to a job-training program, $y_i(1) = $1,190 could be the earnings of worker $i$ when he is offered the training program, and $y_i(0) = $1,080 the earnings of the same worker without access to this training, so $\tau_i = $110. We do not observe both potential outcomes for one unit simultaneously, but observe only the treatment status $d_i$ and the realized outcome

$$y_i = \begin{cases} y_i(1), & d_i = 1, \\ y_i(0), & d_i = 0. \end{cases}$$

But since treatment is assigned randomly (with probability $p = P(d_i = 1)$), we can still obtain an unbiased estimate of the unit treatment effect.\footnote{Here, I assume that we know that treatment has been assigned with known probability $p = P(d_i = 1)$. Throughout the remaining article, I also consider random assignment with a fixed number of treated units rather than a known ex-ante probability of treatment. The case of known number $n_1$ of treated units has structurally similar features, but is not the same as the case} Indeed, I will
note below that \( \frac{d_i - p}{p(1 - p)} y \) is an unbiased estimator for \( \tau_i \). (Throughout, by “unbiased” I mean that, for fixed potential outcomes \( y_i(1) \) and \( y_i(0) \), the treatment-effect estimator averages out to \( y_i(1) - y_i(0) \) over random draws of treatment \( d_i \).

In addition to the realized outcome \( y_i \) and treatment status \( d_i \), I assume that we also have access to some pre-treatment characteristics \( x_i \) of unit \( i \). Estimating the treatment effect \( \tau_i = y_i(1) - y_i(0) \) for, say, a treated unit \( (d_i = 1) \) amounts to imputing the missing, counterfactual control outcome \( y_i(0) \). When we have additional information about that unit, we can hope to use it together with the outcome, treatment, and characteristic data \( z_{-i} = (y_j, d_j, x_j)_{j \neq i} \) of all other units to estimate \( y_i(0) \), and thus \( \tau_i \). The investigator could, for example, run a linear regression of earnings on treatment, pre-assignment earnings, and some basic demographic characteristics to impute the counterfactual outcome \( y_i(0) \). She could then estimate that worker’s treatment effect by the difference between realized and imputed earnings.

If we do not put any restriction on estimation and investigator and social preferences agree, then the investigator’s estimator will represent her expertise as well as the data. I model the investigator’s expertise as a prior distribution \( \pi \) over potential outcomes \( y_i(1), y_i(0) \) given characteristics \( x_i \). (To be more precise, this prior will be over the joint distribution of the potential outcomes of all units given all their controls.) If the investigator aims to minimize the average mean-squared error \( E_\pi (\hat{\tau}_i - \tau_i)^2 \), then for \( d_i = 1 \) she will estimate \( \tau_i \) by

\[
\hat{\tau}_i = E_\pi [\tau_i | y_i, d_i, x_i, z_{-i}] = y_i(1) - E_\pi [y_i(0) | y_i(1), x_i, z_{-i}].
\]

This estimator represents the investigator’s best guess of the treatment effect given her prior and all information in the data. In the training-program example, one specific prior could imply the use of Mincer polynomials in imputing the missing counterfactual outcome by its posterior expectation \( E_\pi [y_i(0)| y_i(1), x_i, z_{-i}] \).

with known probability \( p = \frac{n_1}{n} \). The reason for the difference is that knowledge of all other units’ treatment status is not informative about a given unit’s treatment status for known \( p \), but perfectly determines the left-out unit’s treatment status for known \( n_1 \). Instead of leave-one-out regression adjustments, for fixed \( n_1 \) I therefore show in Section 5 that leave-two-out regression adjustments fully characterize treatment-effect estimators with given bias.
1.2 Specification Searches and Optimal Restrictions on Estimation

If investigator and social preferences are misaligned, then the investigator’s estimator may represent her incentives more than her expertise and the data. Even if the investigator commits to an estimator ex-ante, she could still choose one that is biased towards her preference rather than her prior. As the designer, we therefore should not only require that the investigator commits to an estimator before she has seen all of the data, but also restrict the estimators the investigator can choose from.

We face a tradeoff between flexibility and robustness. Constraints that are too permissive may lead to publication bias. One extreme solution would restrict the investigator to simple specifications that do not use control covariates, or use them only in simple linear regressions. Conventional pre-analysis plans often take this form. But restricting the investigator to a few estimators may forfeit experiment-specific knowledge about the relationship of control variables to outcomes in the prior, which I assume encodes the private information of the investigator.

I show that fixing the bias is a restriction on estimation that resolves this tradeoff. The bias of the first-best optimal estimator usually varies with the prior. Indeed, the posterior expectation of the treatment effect $\tau_i$ is usually biased towards the investigator’s prior expectation $E_\pi \tau_i$. But when we leave the decision over bias to the investigator, then the investigator may shrink her estimator to her preferred estimate instead of her prior.

Once we restrict the investigator to, say, unbiased estimators of $\tau_i$, even an investigator who wants to minimize mean-squared error relative to some fixed target $\tilde{\tau}_i$ (rather than the true treatment effect) will minimize average mean-squared error relative to the true treatment effect among unbiased estimators, since the investigator’s average risk (or cost in the nomenclature of mechanism design) is then

$$E_\pi (\hat{\tau}_i - \hat{\tau}_i)^2 = E_\pi (\hat{\tau}_i - \tau_i)^2 + E_\pi (\tau_i - \tilde{\tau}_i)^2.$$ 

My first main result is that fixing the bias represents an optimal restriction in a minimax sense (Theorem 1) over a set of investigator preferences that generalize this risk function (Assumption 5). That is, the designer’s average mean-squared error is minimal for an investigator that minimizes mean-squared error relative to some worst-case target, given some (hyper-)prior over the investigator’s private informa-
tion. Specifically, if an uninformed designer has little systematic information about the location of the treatment effect, they may want to set the bias close to zero.

1.3 Optimal Unbiased Estimation

Now that investigator and social preferences are aligned, how can the investigator choose an estimator with given bias and low variance? Focusing on the case of zero bias, a simple unbiased estimator of the unit-level treatment effect \( \tau_i \) is available. Indeed, as e.g. noted by Athey and Imbens (2016) (where \( \hat{\tau}_i \) is called the “transformed outcome”), the estimator

\[
\hat{\tau}_i = \frac{d_i - p}{p(1 - p)} y_i = \begin{cases} 
\frac{1}{p} y_i & d_i = 1, \\
-\frac{1}{1 - p} y_i & d_i = 0,
\end{cases}
\]

is unbiased because \( E[\hat{\tau}_i] = \frac{1}{p} y_i(1) - \frac{1}{1 - p} y_i(0) = \tau_i \). But this estimator can have very high variance. Assume that job training is assigned with probability \( p = .5 \), and that the potential earnings are \( y_i(1) = $1,190 \) and \( y_i(0) = $1,080 \). Then

\[
\hat{\tau}_i = \begin{cases} 
+$2,380 & d_i = 1, \\
-$2,160 & d_i = 0,
\end{cases}
\]

is an unbiased, but extremely variable estimator of the treatment effect \( \tau_i = $110 \). Indeed, the variance of \( \hat{\tau}_i \) under treatment assignment is

\[
\text{Var}(\hat{\tau}_i) = p(1 - p)(\hat{\tau}_i(d_i = 1) - \hat{\tau}_i(d_i = 0))^2,
\]

so in the example the standard error amounts to \( \sqrt{\text{Var}(\hat{\tau}_i)} = $2,270 \).

We can modify this estimator by regression adjustments \( \hat{y}_i \) to obtain

\[
\hat{\tau}_i = \frac{d_i - p}{p(1 - p)} (y_i - \hat{y}_i).
\]

As long as \( \hat{y}_i \) only uses information from \( x_i \) and \( z_{-i} = (y_j, d_j, x_j)_{j \neq i} \), and not the outcome \( y_i \) or treatment effect \( d_i \), \( \hat{\tau}_i \) will still be unbiased. Averaging over all \( \hat{\tau}_i \) and for an appropriate choice of the adjustments, Wu and Gagnon-Bartsch (2017) introduce this estimator as the “leave-one-out potential outcomes” (LOOP) estimator. My second main result shows that all estimators of the treatment effect with a given
bias can be written in this way (Lemma 1). Concretely, any unbiased estimator of
the sample-average treatment effect is the average over estimators $\hat{\tau}_i$ for all $i$ that
each include an adjustment that uses data only from all other units. All unbiased
estimators are thus equivalent to a repeated sample-splitting procedure. Conversely,
if $\hat{y}_i$ is fitted, for example, by a regression of $y$ on $x$ that violates the sample-splitting
construction by also including $y_i$, then overfitting of $\hat{y}_i$ to $y_i$ would bias the treatment-
effect estimate towards zero.

Among unbiased estimators (or, more generally, estimators with fixed shrinkage),
which regression adjustment minimizes variance? As Wu and Gagnon-Bartsch (2017)
also note, the investigator would optimally set $\hat{y}_i$ to $(1 - p)y_i(1) + py_i(0)$, since this
leads to $\hat{\tau}_i = \tau_i$. But without using $y_i(1)$ or $y_i(0)$, the investigator’s best choice (in
the sense that it minimizes the variance of the treatment effect estimate for the given
unit) is the posterior expectation

$$\hat{y}_i = E_x[(1 - p)y_i(1) + py_i(0)|x_i, z_{-i}].$$

In the example, if the investigator’s best guess of the expected potential earnings,
$\frac{y_i(1) + y_i(0)}{2}$, based on her prior and data on all other units is $\hat{y}_i = $1,100, then

$$\hat{\tau}_i = \begin{cases} +2($1,190 - $1,100) = $180 & d_i = 1, \\ -2($1,080 - $1,100) = $40 & d_i = 0 \end{cases}$$

is still unbiased for $\tau_i = $110, but has much lower variance (the standard error is
now $\sqrt{\text{Var}(\hat{\tau}_i)} = $70). My third main result shows that, for a known treatment
probability $p$, among estimators with fixed shrinkage (including the special case of
unbiased estimators) the investigator’s solution for the regression adjustments takes
this form (Theorem 2) provided that an additional balance assumption on the prior
is met.

1.4 Machine Learning

By construction, the estimator in (1) of the unit-level treatment effect $\tau_i$ is unbiased
whatever the regression adjustment is. In particular, the sample-splitting construction
ensures that prior information only affects variance. Even a misspecified or

\[12\] The addition of this balance assumption corrects an error in an earlier version of this manuscript
and ensures that the exact duality between unbiased (or fixed-shrinkage) estimation holds in finite
samples. The duality also holds in a standard large-sample approximation.
dogmatic prior does not systematically bias what we learn about $\tau_i$. As also used in Wager et al. (2016) and Wu and Gagnon-Bartsch (2017), this robust construction offers an opportunity to leverage tools that produce good predictions of potential outcomes even when they come with little guarantees that would otherwise ensure unbiasedness.

The specific regression adjustments $\hat{y}_i = E_\pi[(1 - p)y_i(1) + py_i(0)|x_i, z_{-i}]$ solve an out-of-sample prediction problem. Take the special case $p = .5$. Then $\hat{f}_i(x_i) = E_\pi[.5y_i(1) + .5y_i(0)|x_i, z_{-i}]$ minimizes average prediction risk for the loss $(\hat{f}_i(x_i) - y_i)^2$ where $\hat{f}_i$ uses outcome and treatment data from all other units only. This is a regression problem where the quality of fit is measured at a new sample point, and not inside the training sample. Supervised machine-learning algorithms are built to solve exactly such out-of-sample prediction problems. For example, shrinkage methods like ridge regression of the LASSO can have better out-of-sample prediction performance than a linear least-squares regression that optimizes the in-sample fit.

I also obtain an intuitive formula for calculating standard errors. The variance of $\hat{\tau}_i$ is the expected loss in predicting the weighted potential outcome sum $(1 - p)y_i(1) + py_i(0)$ by the adjustment $\hat{y}_i$, which can be estimated from the realized outcome $y_i$ that has been excluded from the construction of $\hat{y}_i$. When units are sampled randomly, I show that, under mild conditions on the construction of regression adjustments, standard errors can be calculated from estimated prediction loss.

1.5 Unbiased Estimation without Pre-Specification

Regression adjustments incorporate flexibly the investigator’s expertise as well as the data, but to ensure that they do not add bias, the investigator must commit to their construction in advance. Indeed, once the investigator has seen the full sample data, she cannot credibly claim that some adjustment uses data only from other units. Practically, the investigator could pre-specify a machine-learning algorithm that learns regression adjustments from the data. But that may be impractical when the construction of adjustments requires input by the researcher.

However, complete pre-specification is not necessary to ensure unbiasedness (or, more generally, a given bias). Instead the investigator could commit to splitting and distributing the sample. Assume there is a researcher in the investigator’s research team that has not yet seen the data. To obtain a regression adjustment for unit $i$,  

\footnote{When treatment is not balanced, $p \neq .5$, additional weights in the prediction loss express that adjustments for the smaller group effectively get weighted up in \textcircled{1}. For details, see (3) in Section 5.}

11
the investigator could give that researcher access to data only from all other units. That researcher then obtains a good adjustment \( \hat{y}_i \) from that subsample and returns that regression adjustment to the investigator, who estimates the treatment effect according to (1). In that case, that researcher’s choice will not introduce bias even if the researcher does not commit to the construction of the regression adjustments in advance.

Of course, estimating the average treatment effect on all \( n \) sample units in this way would require a team of \( n \) researchers. But my fourth main result characterizes all estimators with given bias that remain feasible without detailed pre-specification and when only \( K \) researchers are available (Corollary 1). Even ex-post analysis by a single researcher improves over simple pre-analysis plans without the need for detailed pre-specification. I also show that delegating estimation to two researchers approximates optimal estimation in that it ensures asymptotic efficiency under mild conditions.

2 Setup

Having given a simple example, I now lay out formally how I approach causal inference as a mechanism-design problem. A designer delegates the estimation of an average treatment effect in a randomized experiment to an investigator. The investigator receives a private signal about the distribution of potential outcomes, but has unknown preferences that can be biased. The designer does not analyze the dataset herself, but instead sets constraints on the investigator’s estimator.

In this section, I first define the data-generating process and target parameter before introducing the investigator’s and designer’s problems. To simplify the further analysis, I then argue that we can restrict the analysis to direct restrictions by the designer on the space of estimators the investigator commits to.

2.1 Target Parameter

Following Neyman (1923), I am interested in the average treatment effect

\[
\tau_\theta = \frac{1}{n} \sum_{i=1}^{n} \left( y_i(1) - y_i(0) \right)
\]

where \( \theta = (y_i(1), y_i(0))_{i=1}^{n} \).
in a given sample of $n$ units. In the [Rubin (1974, 1975, 1978)] causal model interpretation, $y_i(d_i)$ is the potential outcome of unit $i$ had they received treatment status $d_i \in \{0, 1\}$, and $\tau_i$ the respective causal effect.

The $n$ units may be randomly sampled from a population distribution,

$$(y_i(1), y_i(0), x_i) \overset{iid}{\sim} P,$$

with pre-treatment characteristics $x_i \in X$. In this case, my analysis will extend to the estimation of the population-average treatment effect $\tau = E[y_i(1) - y_i(0)]$ and the conditional average treatment effect (given characteristics $x \in X^n$) $\frac{1}{n} \sum_{i=1}^{n} E[y_i(1) - y_i(0) | x_i]$. My main analysis is conditional on $(y_i(1), y_i(0), x_i)_{i=1}^{n}$ and therefore focuses on the sample-average treatment effect $\tau_0$, but I will return to $\tau$ when I discuss inference.

### 2.2 Experimental Setup

I assume that treatment is assigned randomly to overcome the missing-data problem central to causal inference [Holland (1986)]. For a unit with treatment status $d_i$, we only observe the realized outcome $y_i = y_i(d_i)$. But because I assume that the distribution of treatment assignment $d \in \{0, 1\}^n$ does not vary with the potential outcome vectors $y(1), y(0) \in \mathbb{R}^n$ [Cochran (1972)], we can estimate the treatment effect without bias. The stable-unit treatment effect assumption [Rubin (1978)] of no interference between units is implicit.

**Assumption 1 (Random Treatment).** Given potential outcomes $\theta = (y_i(1), y_i(0))_{i=1}^{n}$, the data $z = (y, d)_{i=1}^{n}$ is distributed according to $P_\theta$ as follows. $d$ is generated from a known distribution over $\{0, 1\}^n$ that does not depend on $(y(1), y(0))$ and is one of:

1. Each unit is independently assigned to treatment with known probability $p = P(d_i = 1)$ (where $0 < p < 1$).
2. $d$ is drawn uniformly at random from all assignments with known number $n_1 = \sum_{i=1}^{n} d_i$ of treated units (where $0 < n_1 < n$).

Given $d$, $y_i = y_i(d_i)$ for all $i \in \{1, \ldots, n\}$.

In this notation, I do not explicitly include the covariates $x_1, \ldots, x_n$ in the data $z$, since I condition on the controls and therefore treat $(x_i)_{i=1}^{n}$ as a constant and not as a random variable. While neither of the distributions of $d$ depends on the controls,
my results will extend to distributions that are known functions of $x_i$ if they ensure identification of $\tau_\theta$. These include stratified or conditional random sampling, and sampling according to known propensity scores.

### 2.3 Covariate Adjustments

How can we estimate the sample-average treatment effect $\tau_\theta$ from data $(y_i, d_i, x_i)_{i=1}^n$? Since treatment is exogenous, the average difference

$$\hat{\tau}^*(z) = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} (y_i - y_j) = \frac{1}{n_1} \sum_{d_i=1} y_i - \frac{1}{n_0} \sum_{d_i=0} y_i$$

between treatment and control outcomes is an unbiased estimator of $\tau_\theta$ conditional on the number $n_1$ of treated units (provided $0 < n_1 < n$).

Of course this difference in averages $\hat{\tau}^*$ leaves information in the covariates $x_1, \ldots, x_n$ on the table and is likely inefficient. In econometric practice, $\tau_\theta$ is therefore often estimated from a linear regression of the outcome on treatment and controls. But the researcher’s choice of control strategy can bias published results. First, implicit model assumptions may bias estimates. Even simple linear regressions can be biased (Freedman, 2008), although this bias vanishes asymptotically if interactions are included (Lin, 2013). Second, if the investigator does not document that she picked among multiple covariate adjustments, an unsuspecting observer’s inference may be biased towards stronger treatment effects and unjustified confidence (Lenz and Sahn, 2017).

### 2.4 Estimation Preferences

I explicitly consider the choice of the control specification in a mechanism-design framework. A designer and an investigator face a choice of an estimator

$$\hat{\tau} : Z \rightarrow \mathbb{R}$$

that maps experimental data $z = (y, d) \in (Y \times \{0, 1\})^n = \mathcal{Z}$ into an estimate $\hat{\tau}(z)$ of the sample-average treatment effect $\tau_\theta$. Since my analysis is conditional on the control covariates, this estimator encodes in particular how the estimate of the treatment effect is adjusted for the realizations $x_1, \ldots, x_n$ of the control variables.

Designer and investigator preferences are expressed by risk functions $r^D, r^I$:
\(\Theta \times \mathbb{R}^Z \rightarrow \mathbb{R}\) that encode the expected loss \(r^D_\theta(\hat{\tau}), r^I_\theta(\hat{\tau})\) of an estimator \(\hat{\tau} \in \mathbb{R}^Z\) given the full matrix \(\theta = (y(1), y(0)) \in \mathbb{Y}^{2n} = \Theta\) of 2n potential outcomes in the sample at hand. Both designer and investigator aim to minimize their respective risk given the potential outcomes \(\theta\). Throughout this article, I specifically assume that the designer’s risk function expresses a social desire to obtain precise estimates of the true treatment effect \(\tau_\theta\).

**Assumption 2 (Social risk function).** The designer’s risk for an estimator \(\hat{\tau} : \mathbb{Z} \rightarrow \mathbb{R}\) is the estimator’s mean-squared error

\[
r^D_\theta(\hat{\tau}) = \mathbb{E}_\theta[(\hat{\tau}(z) - \tau_\theta)^2],
\]

where the expectation averages over random treatment assignment given potential outcomes \(\theta \in \Theta\).

Notably, I do not assume that the designer has an inherent preference for unbiased estimators\(^{14}\). While my characterization results will depend on this specific form of the social risk function, the general mechanism-design approach extends to alternative risk (or equivalently utility) functions.

The investigator’s risk function can differ from the designer’s risk function. For example, I will later consider risk functions that include \(r^I_\theta(\hat{\tau}) = \mathbb{E}_\theta[(\hat{\tilde{\tau}}(z) - \tilde{\tau})^2]\), which expresses a desire to obtain a certain estimate \(\tilde{\tau}\) irrespective of the true treatment effect \(\tau_\theta\). The designer knows only that \(r^I \in \mathbb{R}\) for some set of risk functions.

### 2.5 Prior Information

Since generally no single estimator \(\hat{\tau}\) minimizes risk for all potential outcomes \(\theta \in \Theta\) and \(\theta\) is not known, a good estimator has to trade off risk performance across different draws of potential outcomes. Following [Wald (1950)], I assume that a prior distribution \(\pi\) over potential outcomes governs this tradeoff\(^{15}\).

The investigator receives the prior distribution \(\pi\) over potential outcomes \(\theta\) as a private signal before the data \(z\) is realized. This private information models re-

\(^{14}\)Still, the minimization of squared-error loss is associated with unbiasedness, as e.g. in [Lehmann and Romano (2006), Example 1.5.6].

\(^{15}\)One alternative approach to finding a good estimator would involve putting restrictions on the distribution of potential outcomes and discussing efficient estimators under some large-sample approximation. But since researchers may reasonably disagree about these choices, this would itself add an additional degree of freedom to estimation. I instead consider estimation in an exact finite-sample decision-theoretic framework that does not restrict the distribution of potential outcomes.
searcher expertise. For example, the investigator may have run previous studies or a pilot and synthesize relevant results in the literature. The investigator therefore has a sense which variables are important and which regression specifications are more likely to work well.

The uninformed designer does not observe the prior $\pi$, but only has a diffuse (hyper-)prior $\eta$ for $\pi$. The designer therefore designs a mechanism that elicits the investigator’s prior information. Optimally, the designer would want to obtain an estimator that minimizes average mean-squared error given the investigator’s private prior, but since the investigator’s preferences may differ from the designer’s, the latter cannot generally achieve a first-best estimator.

2.6 Mechanism Structure and Timeline

I assume that the designer has the authority to set rules in the form of a mechanism without transfers. The designer cannot verify the investigator’s risk type or private prior information. The investigator follows whatever mapping from investigator decisions to final estimator the designer sets, and the designer follows through on the mapping she commits to. Similar to Frankel’s (2014) delegation setup, the game between designer and investigator plays out in the following steps:

1. The designer chooses a mechanism that consists of a message space $M$ and a mapping from messages $m$ into estimators $\hat{\tau}_m : \mathcal{Z} \to \mathbb{R}$.

2. The investigator observes the prior distribution $\pi$ and sends a message $m(r^I, \pi)$.

3. The potential outcomes $\theta$ are realized, the data $z$ drawn according to the experiment, and the estimate $\hat{\tau}_{m(r^I, \pi)}(z)$ formed.

In econometric terms, I think of the investigator’s message as a modelling decision. The designer then restricts the space of models the investigator can choose from.

For simplicity, I assume that the investigator’s message given her risk type and private information and the mapping of her message to the final estimator are deterministic, but the setup extends to stochastic actions as in Frankel (2014). By the revelation principle, the specific form of the mechanism is not a substantial restriction, since it includes direct mechanisms in which the investigator reveals her risk type and her private information (as e.g. in Holmström 1984).
Since the investigator controls the estimator with her choice of message, we can assume without loss of generality that the message space is a set of estimators (and the mapping from message to estimator the identity). Indeed, take any estimator that is an outcome for some message. Since neither risk type nor prior are verifiable, the investigator can always choose that message to obtain said estimator.

Hence, the designer directly restricts estimators to some set $C^D$. Subject to the constraint, the investigator specifies an estimator $\hat{\tau}^I \in C^D$ before data becomes available. Once the data $z \in Z$ is realized, the investigator reports the estimate $\hat{\tau}^I(z)$ (Figure 1). Since my econometric analysis is conditional on the control variables $x_1, \ldots, x_n$, this baseline information can be available to the investigator and inform her choice of estimator.

Optimal estimation in this framework will require some degree of commitment by the investigator before the data is available. Otherwise, any restriction on estimation would be cheap talk, since the investigator could choose an estimator ex post that justifies their preferred estimate at the realized data. But I will show that optimal commitment is less constraining than restricting the investigator to pre-analysis plans with simple specifications that are chosen ex ante. First, the investigator’s estimator can still contain (automated) specification searches. Second, in Section 6, I show that it is not generally necessary to specify the full estimator ex ante, and that additional exploratory analysis after the data has become available can improve estimation.

2.7 Investigator and Designer Choices

Having set up the actions available to the investigator and designer, I now describe their preferences. The investigator chooses an estimator to minimize average risk subject to her prior.
Assumption 3 (Investigator’s choice). Given the prior distribution $\pi$ over potential outcomes $\theta \in \Theta$, the investigator minimizes average risk subject to the constraint $C^D \subseteq \mathbb{R}^Z$ set by the designer,

$$\hat{\tau}' = \hat{\tau}'(C^D, \pi) \in \arg\min_{\hat{\tau} \in C^D} \mathbb{E}_\pi[\tau'_\phi(\hat{\tau})].$$

The designer does not know the risk function of the investigator, but only assumes that it falls within some set $\mathcal{R}$ of risk functions. Adapting the maxmin criterion from the mechanism-design literature (e.g., Hurwicz and Shapiro 1978; Frankel 2014; Carroll 2015), I assume that the designer chooses a constraint that minimizes average risk at a worst-case investigator type within that set.

Definition 1 (Designer’s minimax delegation problem). Given some set $\mathcal{R}$ of investigator risk functions, the designer picks a constraint $C^D \subseteq \mathbb{R}^Z$ to minimize average mean-squared error,

$$C^D = C^D(\mathcal{R}, \eta) \in \min_{C \subseteq \mathbb{R}^Z} \sup_{r^I \in \mathcal{R}} \mathbb{E}_\eta[r^D_\theta(\hat{\tau}^I)],$$

where I assume that the investigator breaks ties in the designer’s favor.

The minimax criterion can be seen as a game between designer and nature. For every choice of restriction that the designer picks, nature responds with an investigator who produces maximal average mean-squared error. In this game, the designer picks a constraint that ensures that the average risk at a worst-case outcome is minimal.

Without constraints, the investigator’s estimator may be a poor fit from the designer’s perspective. But if the constraints are too restrictive, for example if we reduce the allowed set of estimators to the difference in averages $\hat{\tau}^*$, we will use the investigator’s expertise inefficiently. I therefore solve for constraints $C^D$ that resolve this tradeoff between flexibility and robustness optimally.

2.8 Support Restriction

Throughout this article, I assume that the support of (potential) outcomes is finite, for three reasons. First, I adapt results from the mechanism-design literature that involve finite sums. Second, I use and provide complete-class theorems that fully
characterize admissible (non-dominated) estimators provided their support is finite. Third, I derive intuitive combinatorial proofs for my characterization results.

**Assumption 4** (Finite support). *The support \( \mathcal{Y} \) of potential outcomes \( y_i(1), y_i(0) \) is finite.*

Since the number of support points is otherwise unrestricted, the finite-support assumption allows for flexible approximations to arbitrary distributions.

### 3 Overview of Main Results

In this section, I preview my main theoretical results. Under specific restrictions on investigator preferences, I show that fixing the bias is a minimax optimal constraint on estimation. I then present a representation of treatment-effect estimators with given bias, characterize the investigator’s optimal choice from this restricted class for the case of estimators with fixed shrinkage in a special case, and extend the analysis to estimators with limited pre-specification.

I assume that investigator risk functions express mean-squared error relative to some target which may not be the true treatment effect.

**Assumption 5** (Investigator risk restriction). *The investigator has a risk function from the set*

\[
\mathcal{R}^* = \{ r_I^*; r_I^*(\hat{\tau}) = E_\theta[(\hat{\tau}(z) - \tilde{\tau}_\theta)^2] \text{ for some } \tilde{\tau} : \Theta \rightarrow \mathbb{R} \}.
\]

The target function \( \tilde{\tau}_\theta \) is unrestricted in this definition. For example, the investigator may want to achieve a constant target no matter what the true potential outcomes are (\( \tilde{\tau}_\theta = \text{const.} \)). Or the investigator may prefer to obtain estimates above the true treatment effect (\( \tilde{\tau}_\theta = \tau_\theta + \varepsilon \)).

In any of these cases, restricting investigators to unbiased estimators (or more generally, estimators with given bias, \( E_\theta[\hat{\tau}(z)] = \tau_\theta + \beta_\theta \)) ensures that they choose among these estimators as if they had the designer’s preference, i.e. they minimize average variance. Once I have established tools for asymptotically valid inference, it will also follow that unbiasedness aligns the choices of investigators who want to obtain a small standard error or a low \( p \)-value.

While fixing the bias aligns preferences, this restriction may be too strong. However, I establish that it is minimax optimal for an appropriate choice of biases.
Theorem 1 (Fixed bias is minimax optimal). Write $\Delta^*(\Theta)$ for all distributions over $\Theta$ with full support. For every hyperprior $\eta$ with support within $\Delta^*(\Theta)$ there is a set of biases $\beta^n : \Theta \to \mathbb{R}$ such that the fixed-bias restriction

$$C^n = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; E_\theta[\hat{\tau}] = \tau_\theta + \beta^n_\theta\}$$

is a minimax optimal mechanism in the sense of Definition 1, i.e.

$$C^n \in \arg \min_{C} \sup_{r \in \mathbb{R}^*} E_\eta \left[ r \mathbb{D} \left( \arg \min_{\hat{\tau} \in C} E_\pi \left[ r \mathbb{D} \left( \hat{\tau} \right) \right] \right) \right].$$

This result implies that the designer should not leave the choice of bias to the investigator. If the designer has an informative hyperprior, she may set biases to reflect that information. But with little information on the designer’s side, the designer may want to set them close to zero. I discuss setting the bias in Appendix F.

With a restriction to given bias (in the sense of an ex-ante fixed vector of biases), the investigator chooses the estimators to minimize variance. The next result specifically characterizes estimators with a given shrinkage $\lambda$ and overall bias $\alpha$ (including unbiased estimators, for which $\alpha = 0 = \lambda$), and therefore the choice set of the investigator if the designer sets $\lambda$ and $\alpha$.

**Lemma 1** (Representation of fixed-shrinkage estimators). The estimator $\hat{\tau}$ has fixed shrinkage,

$$E_\theta[\hat{\tau}(\cdot)] = \alpha + (1 - \lambda)\tau_\theta$$

for all potential outcomes $\theta \in \Theta$ (where $\lambda \in [0, 1]$), if and only if:

1. For a known treatment probability $p$, there exist leave-one-out regression adjustments $(\phi_i : (\mathcal{Y} \times \{0, 1\})^{n-1} \to \mathbb{R})_{i=1}^n$ such that

$$\hat{\tau}(\cdot) = \alpha + (1 - \lambda)\frac{1}{n} \sum_{i=1}^n \frac{d_i - p}{p(1 - p)} (y_i - \phi_i(z_{-i})).$$

2. For a fixed number $n_1$ of treated units, there exist leave-two-out regression

---

\[I\] conjecture that such fixed-shrinkage estimators emerge as optimal solutions of the designer for minimax preferences over researcher preferences and priors.

20
adjustments \((\phi_{ij} : (\mathcal{Y} \times \{0, 1\})^{n-2} \to \mathbb{R})_{i<j}\) such that

\[
\hat{\tau}(z) = \alpha + (1 - \lambda) \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}(z_{-ij})),
\]

where \(\phi_{ij}(z_{-ij})\) may be undefined outside \(1'd_{-ij} = n_1 - 1\).

For the case of known probability \(p\), any unbiased estimator \((\alpha = 0 = \lambda)\) can therefore be written in the leave-one-out form that Wu and Gagnon-Bartsch (2017) obtain as a special case of the unbiased estimators introduced by Aronow and Middleton (2013).

The result directly extends to a characterization of estimators with fixed bias. Indeed, fixing the bias is equivalent to the designer choosing an estimator \(\hat{\tau}^D\) with the desired biases \(E_\theta[\hat{\tau}^D(z)] - \tau_\theta = \beta_\theta\) for all \(\theta \in \Theta\), and letting the investigator choose a zero-expectation adjustments \(\hat{\delta}^I\) \((E_\theta[\hat{\delta}^I(z)] = 0\) for all \(\theta \in \Theta\)\) to form the estimator \(\hat{\tau} = \hat{\tau}^D + \hat{\delta}^I\). Given \(\hat{\tau}^D\), any estimator with the associated bias profile can thus be written as

\[
\hat{\tau}^D(z) - \frac{1}{n} \frac{d_i - p}{p(1-p)} \sum_{i=1}^n \phi_i(z_{-i}), \quad \hat{\tau}^D(z) - \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j) \phi_{ij}(z_{-ij}),
\]

respectively, with adjustments as in the lemma. The class of unbiased estimators underlying the lemma corresponds to the unbiased choices

\[
\hat{\tau}^D(z) = \frac{1}{n} \sum_{i=1}^n \frac{d_i - p}{p(1-p)} y_i, \quad \hat{\tau}^D(z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j).
\]

All estimators with given bias are hence sample-splitting estimators that leave one or two units out, respectively, when calculating their regression adjustments. But when is an estimator not just of this form, but also precise? As a general solution, the investigator would now pick one set of adjustments that minimize variance averaged over their prior, yielding constrained optimal solution from the perspective of the designer.

For the specific case of fixed-shrinkage estimators and under an additional balance assumption, the adjustments take a particularly simple form in the case of known treatment probability. The investigator would optimally want to set regres-
sion adjustments to the oracle solutions

$$\bar{y}_i = (1 - p)y_i(1) + py_i(0),$$

$$\Delta \bar{y}_{ij} = \left( \frac{n_0}{n} y_i(1) + \frac{n_1}{n} y_i(0) \right) - \left( \frac{n_0}{n} y_j(1) + \frac{n_1}{n} y_j(0) \right),$$

respectively, but since the potential outcomes are unknown, these adjustments are infeasible. Minimizing average loss relative to these oracle adjustments generally leads to a linear system of first-order conditions that locate an optimal estimator with minimal average variance. In a specific case, I show that the investigator chooses leave-one-out or leave-two-out expectations of these adjustments.

**Theorem 2** (Choice of the investigator from fixed-shrinkage estimators). For a known treatment probability $p$, an investigator with risk $r \in \mathbb{R}^*$ and prior $\pi$ over $\Theta$ with

$$E_\pi \left[ E_\pi \left[ \bar{y}_j | y_i(1), z_{-ij} \right] | z_{-i} \right] = E_\pi \left[ E_\pi \left[ \bar{y}_j | y_i(0), z_{-ij} \right] | z_{-i} \right]$$

(2)

for all $i \neq j$ for given shrinkage $\lambda \in [0, 1)$ and overall bias $\alpha$ chooses the Bayes estimator

$$\hat{\tau}(z) = \alpha + (1 - \lambda) \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1 - p)} (y_i - E_\pi \left[ \bar{y}_i | z_{-i} \right]).$$

Hence, in this case all optimal fixed-shrinkage estimators take as regression adjustments conditional expectations of potential outcomes. These conditional expectations can be obtained as solutions to a prediction problem. Independently of the mechanism-design setup, and for the general Bayes solutions (even when the balance assumption does not hold), the set of investigator solutions across different priors completely characterize the class of admissible unbiased estimators of the sample-average treatment effect.

---

17 Due to a mistake in the proof, a previous version of this theorem lacked the balance assumption in Equation 2 for known $p$, and falsely claimed a direct extension to fixed $n_1$. To see that the theorem does not generally apply without this assumption, consider the following example: For $\alpha = 0 = \lambda, n = 2, p = .5, \pi : y_1(1) = y_2(1), y_i(0) = y_2(0) \sim U(\{+1, -1\})$ (where $y_1(1), y_1(0)$ independent), which violates Equation 2 the adjustments in the theorem are $y_{-i}/2$ (yielding average loss 3), while optimal adjustments are given by $y_{-i}/4$ (yielding average loss 2.75). (The average loss without adjustment is 4.)

18 For known $p$ and zero bias, this exact solution mirrors Wu and Gagnon-Bartsch’s (2017) LOOP estimator, for which the authors discuss estimating the adjustments using different prediction methods.
Theorem 3 (Complete-class theorem for unbiased estimators). For any unbiased estimator $\hat{\tau}$ of the sample-average treatment effect that is not dominated with respect to variance, there is a converging sequence of priors $(\pi_i)_{i=1}^\infty$ with full support such that $\hat{\tau}$ equals the limit of the respective optimal Bayes estimators. Conversely, for any converging sequence of priors $(\pi_i)_{i=1}^\infty$ that put positive weight on every state $\theta \in \Theta$, every converging subsequence of corresponding Bayes estimators is admissible among unbiased estimators.

Now that I have characterized the optimal solution of the designer and the investigator (with an explicit expression for the case of fixed-shrinkage estimators under an additional assumption), I return to the question of commitment. The representation of fixed-bias estimators in Lemma 1 requires that the construction of regression adjustments does not involve the adjusted unit. In Theorem 2, the investigator would therefore have to commit to their construction before she has access to the full sample. This pre-specification leaves room for automated specification searches in constructing the adjustments. But fully pre-specifying all specification searches may be impractical.

I also characterize estimators that ensure fixed bias not by the investigator fully pre-specifying adjustments, but by a commitment to a sample-splitting scheme. I consider estimation contracts that have the investigator delegate estimation tasks on subsamples to $K$ researchers who do not share information about the data they receive.

Definition 2 ($K$-distribution contract). A $K$-distribution contract $\hat{\tau}^\Phi$ distributes data $z = (y, d) \in (Y \times \{0, 1\})^n = Z$ to $K$ researchers. Researcher $k$ receives data $g_k(z) \in A_k$ and returns the intermediate output $\hat{\phi}_k(g_k(z)) \in B_k$. The estimate is

$$\hat{\tau}^\Phi(\hat{\phi}_k^{K}_{k=1}; z) = \Phi((\hat{\phi}_k(g_k(z)))^{K}_{k=1}; z).$$

The investigator chooses the functions $g_k$ (from data in $Z$ to researcher input in $A_k$) and $\Phi$ (from the researcher outputs in $\times_{k=1}^{K} B_k$ and data in $Z$ to estimates in $\mathbb{R}$) before accessing the data.

As one special case of my general representation result of fixed-bias $K$-distribution contracts, I characterize estimators with given bias that divide the sample into $K$ folds and then give each researcher access to all but one of these folds. In that case, I deduce from the representation of fixed-shrinkage estimators in Lemma 1 that the
estimator always has the given bias if and only if each researcher only controls the regression adjustments for the respective left-out fold.

**Corollary 1** (Characterization of fixed-bias $K$-fold distribution contracts). For $K$ disjoint folds $\mathcal{I}_k \subseteq \{1, \ldots, n\}$ with projections $g_k : (y, d) = z \mapsto z - I_k = (y_i, d_i)_{i \not\in I_k}$, a $K$-distribution contract $\hat{\tau}^\Phi$ has given bias if and only if:

1. For a known treatment probability $p$, there exist a fixed estimator $\hat{\tau}_0(z)$ with the given bias and regression adjustment mappings $(\Phi_k)_{k=1}^K$ such that

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \frac{d_i - p}{p(1 - p)} \phi_i^k(z_i)$$

where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z - I_k))$.

2. For a fixed number $n_1$ of treated units, there exist a fixed estimator $\hat{\tau}_0(z)$ with the given bias and regression adjustment mappings $(\Phi_k)_{k=1}^K$ such that

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n_1 n_0} \sum_{k=1}^K \sum_{\{i < j\} \subseteq \mathcal{I}_k} (d_i - d_j) \phi_{ij}^k(z_{ij})$$

where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z - I_k))$.

These sample-distribution contracts achieve the given bias without detailed commitments by the researchers. For $K = 1$, I show that giving one researcher (with risk function in the set $\mathcal{R}^*$) access to part of the sample for exploratory ex-post analysis can improve over simple pre-analysis plans. For $K = 2$, I show that a flexible, unbiased pre-analysis plan that specifies distribution to two researchers asymptotically achieves semi-parametric efficiency when the units are sampled iid under conditions on the population distribution.

### 4 Designer’s Solution

Having set up the estimation of a sample-average treatment effect as a mechanism-design problem, I justify a restriction to estimators with fixed bias by solving the designer’s delegation problem. Subject to fixed bias, the investigator pre-specifies an estimator according to the designer’s preferences. I prove minimax optimality of fixed-bias restrictions, echoing a result from mechanism design on optimal delegation.
4.1 The Role of Bias

When there is no misalignment of preferences, then the resulting first-best estimator that minimizes average mean-squared error will generally have bias that changes with prior. To understand how being flexible on bias can improve estimation, note that both bias and variance contribute to the risk

\[ r^D_D(\hat{\tau}) = \mathbb{E}_\theta[(\hat{\tau} - \tau)^2] = \underbrace{(\mathbb{E}_\theta[\hat{\tau}] - \tau)^2}_{\text{bias}} + \underbrace{\mathbb{V}_\theta(\hat{\tau})}_{\text{variance}} \]

the designer aims to minimize. We can often improve an estimator with fixed bias by moving along this bias-variance tradeoff. Indeed, consider the first-best solution \( \hat{\tau_\pi} = \arg \min_{\hat{\tau}} \mathbb{E}_\pi[r^D_D(\hat{\tau})] \) of the designer. The estimate \( \hat{\tau_\pi}(z) = \mathbb{E}_\pi[\tau_\theta|z] \) comprises the posterior expectations \( \mathbb{E}_\pi[y_i(1) - y_i(0)|z] \), which are usually biased towards the prior expectation of unit treatment effects when the prior is informative along this dimension.

But if the designer leaves the decision over bias to the investigator, then an investigator who has biased preferences will be inclined to bias the estimator in the direction of her preferences, not of her prior. Consider an investigator with risk

\[ r^I_I(\hat{\tau}) = \mathbb{E}_\theta[(\hat{\tau}(z) - (\tau_\theta + \varepsilon))^2] \quad (\varepsilon > 0) \]

who would like to show that the treatment effect is higher than it is. The investigator’s unconstrained solution is now shifted upward by \( \varepsilon \), which is added to the bias term. While reducing the variance relative an unbiased estimator, the designer’s risk may also be increased through additional bias.

For choices among estimators with fixed bias, however, the investigator’s and designer’s preferences in this example are perfectly aligned. With bias fixed at zero, say, mean-squared error is variance, \( r^D_D(\hat{\tau}) = \mathbb{V}_\theta(\hat{\tau}) \). The \( \varepsilon \)-biased investigator’s risk is \( r^I_I(\hat{\tau}) = \varepsilon^2 + \mathbb{V}_\theta(\hat{\tau}) \). While risks are not the same, they are shifted by a constant. There is no distortion in choices between estimators with fixed bias for this investigator loss function.

4.2 Fixed-Bias Estimation as Second-Best

Having motivated in an example that fixing the bias can align investigator choices, I extend alignment to a minimax result. If the investigator has constant bias, I have
argued that among estimators with fixed bias she will still commit to a variance-minimizing estimator. To show that this example extends to an optimal solution, I have to establish that the bias restriction is neither too permissive nor too restrictive.

A restriction that fixes the bias, for example to zero,

\[ C^* = \{ \hat{\tau} : Z \rightarrow \mathbb{R}; E_\theta[\hat{\tau}] = \tau_\theta \forall \theta \in \Theta \}, \]

is not too permissive provided that investigators all choose as if they minimized mean-squared error relative to some target, albeit not necessarily relative to the true treatment effect.

**Assumption 5 (Investigator risk restriction).** The investigator has a risk function from the set

\[ \mathcal{R}^* = \{ r^I, r^I_\theta(\hat{\tau}) = E_\theta[(\hat{\tau}(z) - \tilde{\tau}_\theta)^2] \text{ for some } \tilde{\tau} : \Theta \rightarrow \mathbb{R} \}. \]

The target \( \tilde{\tau}_\theta \) can vary arbitrarily with the potential outcomes. In particular, permissible risk functions include constant biases relative to the truth (\( \tilde{\tau} = \tau + \varepsilon \)) or fixed estimation targets (\( \tilde{\tau} = \text{const.} \)). \( \mathcal{R}^* \) also includes the designer’s risk function \( r^D \) at \( \tilde{\tau} = \tau \).

**Lemma 4.1 (Unbiasedness aligns estimation).** If the investigator has risk from \( \mathcal{R}^* \) then the investigator will choose from the unbiased estimators \( C^* \) according to the designer’s preferences.

Note that the result extends to restrictions to fixed bias (that can vary with \( \theta \)).

Once I have established asymptotically valid inference for unbiased estimators in [Appendix E](#), I will also show in [Remark E.3](#) that the unbiasedness restriction aligns the choices of investigators who want to obtain small standard errors or tight confidence intervals. For a local-to-null alternative, by [Remark E.4](#) unbiasedness also insures asymptotic alignment in large samples when the investigator wants to obtain a low \( p \)-value (that is, wants to maximize the power of a test against some null hypothesis \( \tau_\theta = \tau_0 \)).

Note, however, that there are many risk (or equivalently utility) functions for which fixing the bias does not provide alignment. In particular, it may be a poor alignment device for non-convex loss functions. Take an investigator who wants to produce an estimate that does not reject some null hypothesis, for example when...
running a balance or robustness check. In that case, if some valid way of calculating standard errors is available, the investigator would want to obtain high variance even among unbiased estimators in order to weaken the evidence against her preferred null hypothesis.

For the class $\mathcal{R}^*$ of investigator risk functions, fixing the bias is not too restrictive because it is minimax optimal over investigator preferences. While Lemma 4.1 establishes that choices from unbiased estimators will be the same for any $r^I \in \mathcal{R}^*$, there could be a larger set of estimators that provide alignment, or full alignment of preferences could be too costly.

**Theorem 1** (Fixed bias is minimax optimal). Write $\Delta^*(\Theta)$ for all distributions over $\Theta$ with full support. For every hyperprior $\eta$ with support within $\Delta^*(\Theta)$ there is a set of biases $\beta^\eta : \Theta \to \mathbb{R}$ such that the fixed-bias restriction

$$C^\eta = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathbb{E}_{\theta}[\hat{\tau}] = \tau_{\theta} + \beta^\eta_{\theta}\}$$

is a minimax optimal mechanism in the sense of [Definition 7](#) i.e.

$$C^\eta \in \arg \min_C \sup_{r^I \in \mathcal{R}^*} \mathbb{E}_{\eta} \left[ r^D_{\theta} \left( \arg \min_{\hat{\tau} \in \mathcal{C}} \mathbb{E}_{\pi}[r^I_{\theta}(\hat{\tau})] \right) \right].$$

This minimax result shows that the gains from variance reduction of being flexible on bias are fully undone by the cost of misalignment for a worst-case risk function, for any relaxation of the fixed-bias restriction. Once we allow the bias to track the prior, it could as well reflect the preference of a worst-case investigator. The designer therefore chooses fixed biases that reflect her hyperprior.

If the designer has a hyperprior $\eta$ that is quite informative about treatment effects, she could introduce biases towards expected treatment effects under that hyperprior. Crucially, however, these biases would be fixed ex-ante and not chosen by the investigator. But when the hyperprior contains little systematic information about the treatment effect at $\theta$, then $\beta_{\theta}$ close to zero is a natural choice. In [Appendix F](#) I highlight one construction that shows how an (approximately) uninformative hyperprior delivers (approximately) unbiased estimation as the support grows. There, I also lay out how being minimax over a specific, uninformative class of hyperpriors yields zero bias. I further conjecture that a specific class of biases,

---
namely those representing fixed shrinkage,

$$\beta_\theta = \alpha + \lambda \theta,$$

emerge as minimax solutions against a worst-case prior over bounded support.

4.3 Connection to Aligned Delegation

My econometric finding that fixed-biased estimation is minimax optimal (Theorem 1) builds upon a mechanism-design result by Frankel (2014). There, a principal delegates decisions to an agent who observes states. Frankel (2014) characterizes optimal delegation mechanisms without transfers. In a class of maxmin optimal, simple mechanisms, the agent behaves according to the principal’s preferences.

In a leading example from Frankel (2014), a school principal delegates the grading of a group of students to a teacher. The teacher may prefer to give more skewed or better grades than the principal, who does not observe the students’ performance. However, the principal can exploit that the teacher’s biased preferences are consistent across students. If the teacher and the principal agree on the ranking of students, fixing the distribution of grades obtains a second-best grade assignment. If the teacher has a constant bias, fixing the average grade already achieves agreement between principal and teacher. In both cases, the teacher chooses from the restricted grade assignments according to the principal’s preferences.

What a fixed average is to grading in Frankel (2014), constant bias is to estimation in my setting. More precisely, I identify Frankel’s (2014) school principal with my designer, the teacher with the investigator, and individual students with different draws of the data. In the school example, the performance of students is the private information of the teacher. For estimation, the prior distribution over potential outcomes is the private information of the investigator. Where the teacher chooses a grade for each student, the investigator commits to an estimator, that is, the investigator chooses an estimate for each (potential) draw of the data.

Frankel (2014) shows that fixing the average over grades is a maxmin (in utility terms) optimal mechanism for a class of biased squared-error preferences. Analogously, my fixed-bias restriction fixes weighted sums over estimates. But since fixing the bias requires setting many sums at once, and the designer’s and investigator’s preferences involve weights determined by the prior, additional work is required to establish the minimax optimality in Theorem 1. In Appendix A, I show how
Frankel’s (2014) result carries over to the designer’s problem across all \( \theta \in \Theta \), where the investigator sets all \((2|Y|)^n\) values of \( \hat{\tau}(y,d) \) simultaneously.

4.4 Design of Experiment vs Design of Estimator

In Theorem 1, I have assumed that treatment is assigned randomly according to some fixed rule, but my results extend to the design of treatment assignment itself. The investigator may leverage prior knowledge about potential outcomes to adjust propensity scores (Kasy, 2016). For example, if the prior distribution of treated outcomes has larger variance than that of controls, the investigator may want to assign more units to treatment. Under the fixed-bias restriction, the investigator’s preference over this additional decision remains aligned with the goal of the designer.

5 Investigator’s Solution

The designer restricts the investigator to estimators with given bias. I establish that this restriction is equivalent to splitting the sample in a particular way. In solving the investigator’s constrained optimization problem in the specific case of fixed shrinkage, known treatment probability, I show that optimal estimation is equivalent to a set of out-of-sample prediction tasks under a balance assumption on the prior. In that case, I obtain a complete-class theorem that characterizes admissible unbiased estimators of the sample-average treatment effect.

Throughout this section, I assume that the investigator fully specifies her estimator before it is applied to outcome and treatment data \( z = (y,d) \). Although the estimator is pre-specified, it can still include (automated) specification searches. The pre-specified estimator thus plays the role of a flexible pre-analysis plan. Since my results hold conditional on potential outcomes, the covariates \( x_1, \ldots, x_n \) can be common knowledge before this pre-analysis plan is filed. In Section 6, I show how the results in this section extend when full pre-specification is impractical. There, I provide a constructive characterization of pre-analysis plans that only commit to the way the sample is split and distributed.

5.1 Characterization of Fixed-Bias Estimators

When does an estimator have a given bias, conditional on potential outcomes? The designer requires that the investigator provides a fixed-bias estimator. In this section,
I provide an intuitive representation of estimators of a given bias that the investigator can achieve transparently by construction.

For the case of zero bias, a class of estimators that ensures unbiasedness is obtained by sample splitting. For known treatment probability $p$, the Horvitz and Thompson (1952) estimator $\hat{\tau}^{HT} = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1 - p)} y_i$ is unbiased for any pair of potential outcome vectors because

$$E_{\theta} \left[ \frac{d_i - p}{p(1 - p)} y_i \right] = y_i(1) - y_i(0).$$

If we replace outcomes $y_i$ by adjusted outcomes $y_i - \phi_i(z_{-i})$ with regression adjustments that do not vary with $(y_j, d_j)_{j \neq i}$ from all units other than $i$, then the resulting estimator is still unbiased. (Recall that I condition on controls $x_1, \ldots, x_n$ throughout.) Wu and Gagnon-Bartsch (2017) call the resulting estimator for known $p$ the “leave-one-out potential outcomes” (LOOP) estimator. This estimator is a special case of Aronow and Middleton’s (2013) modification of the Horvitz and Thompson (1952) estimator. Since the adjustment $\phi_i(z_{-i})$ is the same whether unit $i$ is treated or not and $E_{\theta} \left[ \frac{d_i - p}{p(1 - p)} z_{-i} \right] = 0$, their addition averages out to zero, no matter the potential outcomes or realized treatment of the other units.\footnote{It would not be enough to exclude the treatment status $d_i$ from the construction of unit $i$’s regression adjustment, and thus use $y_i$, since $y_i$ can be correlated with $d_i$.}

I show that these sample-splitting estimators are also all estimators that are unbiased conditional on potential outcomes. If an estimator cannot be written as a Horvitz and Thompson (1952) estimator with leave-one-out regression adjustments (i.e. in the form of Wu and Gagnon-Bartsch’s (2017) LOOP estimator), it must have bias for some matrix of potential outcomes.\footnote{If instead we considered estimators that are unbiased given some distribution of potential outcomes (for example, we may want to model noise terms in potential outcomes that we do not want to condition on), then the result would trivially extend as long as we do not restrict this distribution.}

A leave-one-out estimator can have bias conditional on the number of treated units. If the number $n_1$ of treated units is known, the leave-one-out adjustment $\phi_i(z_{-i})$ implicitly depends on $d_i = n_1 - \sum_{j \neq i} d_j$. For permutation randomization, I therefore start with the difference in averages $\hat{\tau}^* = \frac{1}{n_1 n_0} \sum_{d_i = 1, d_j = 0} (y_i - y_j)$ and establish that all unbiased estimators differ from $\hat{\tau}^*$ only by leave-two-out regression adjustments $\phi_{ij}(z_{-ij})$. In every sample split, these unbiased estimators leave out
one treated and one untreated unit.

Lemma 1 (Representation of fixed-shrinkage estimators). The estimator $\hat{\tau}$ has fixed shrinkage,

$$E_\theta[\hat{\tau}(z)] = \alpha + (1 - \lambda)\tau_\theta$$

for all potential outcomes $\theta \in \Theta$ (where $\lambda \in [0, 1)$), if and only if:

1. For a known treatment probability $p$, there exist leave-one-out regression adjustments $(\phi_i : (Y \times \{0, 1\})^{n-1} \to \mathbb{R})_{i=1}^n$ such that

$$\hat{\tau}(z) = \alpha + (1 - \lambda)\frac{1}{n} \sum_{i=1}^n \frac{d_i - p}{p(1 - p)}(y_i - \phi_i(z - i)).$$

2. For a fixed number $n_1$ of treated units, there exist leave-two-out regression adjustments $(\phi_{ij} : (Y \times \{0, 1\})^{n-2} \to \mathbb{R})_{i<j}$ such that

$$\hat{\tau}(z) = \alpha + (1 - \lambda)\frac{1}{n_1n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}(z - ij)),$$

where $\phi_{ij}(z - ij)$ may be undefined outside $1'd_{-ij} = n_1 - 1$.

While I have derived this result for fixed-shrinkage estimators, the characterization directly carries over to estimators with fixed bias. Indeed, fixing the bias is equivalent to the designer choosing an estimator $\hat{\tau}^D$ with the desired biases $E_\theta[\hat{\tau}^D(z)] - \tau_\theta = \beta_\theta$ for all $\theta \in \Theta$, and letting the investigator choose a zero-expectation adjustments $\hat{\delta}^I$ ($E_\theta[\hat{\delta}^I(z)] = 0$ for all $\theta \in \Theta$) to form the estimator $\hat{\tau} = \hat{\tau}^D + \hat{\delta}^I$. Given $\hat{\tau}^D$, any estimator with the associated bias profile can thus be written as

$$\hat{\tau}^D(z) - \frac{1}{n} \sum_{i=1}^n \frac{d_i - p}{p(1 - p)}\phi_i(z - i), \quad \hat{\tau}^D(z) - \frac{1}{n_1n_0} \sum_{i<j} (d_i - d_j)\phi_{ij}(z - ij),$$

respectively, with adjustments as in the lemma. The statement of the lemma for

\footnote{Wager et al. (2016) consider leave-one-out estimators separately in the treatment and control groups, and use a leave-two-out construction to derive asymptotic unbiasedness.}
α = 0 = λ corresponds to the unbiased choices

\[ \hat{\tau}^D(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} y_i, \quad \hat{\tau}^D(z) = \frac{1}{n_1n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j). \]

The representations are restrictive, but not unique. In the minimal non-trivial case \( n = 2 \) and \(|Y| = 2\) for known treatment probability, the leave-one-out representation reduces the dimension of estimators \( \hat{\tau} \in \mathbb{R}^{(|Y|\times\{0,1\})^n} \) from 16 to 8. Unbiased estimators form a 7-dimensional affine linear subspace, and equivalent representations lie on lines in Euclidean space.

Notably, linear regression can not generally be represented in this way, as it is not generally unbiased in my setting (Freedman 2008). In Appendix D I provide a simple example of a biased OLS regression. Also, I make a connection between overfitting and bias, and show that bias can persist even under sampling from a population distribution and in large samples with high-dimensional controls.

We usually associate sample splitting with losses in efficiency in return for robustness. Since all unbiased estimators must split the sample, this logic applies here only through the robustness of the unbiasedness assumption to any distribution of potential outcomes. As long as we do not impose additional structure, all admissible (with respect to variance or equivalently mean-squared error) unbiased estimators must be among the sample-splitting estimators.

This result implies that the set of fixed-bias estimators the investigator chooses from is characterized by prohibitions. When we represent an estimator by a sum over adjusted outcomes, then there must be one such representation for which the investigator is not allowed to use the outcome and treatment assignment of a unit to construct its adjustment. For this prohibition to apply, in practice the investigator has to commit how the adjustment is constructed before she has access to the respective outcome and treatment status. I show below that this commitment leaves room for automated specification searches, and discuss in Section 6 that human specification searches also remain feasible.

### 5.2 Solution to the Investigator’s Problem

Given the restriction to a given bias, what is the optimal solution of the investigator? The sample-splitting representation provides an objective criterion for fixed bias. Since preferences are aligned, the investigator applies their subjective prior to
minimize average variance over the regression adjustments from Lemma 1. The resulting estimator is a Bayes estimator in the sense of Wald (1950), and can generally be obtained from a system of linear equations representing the associated first-order conditions.

In the specific case of estimators with fixed shrinkage, the adjustments take a particularly simple form as solutions to prediction problems provided an additional assumption is met and treatment is assigned with known probability. If the investigator knew the potential outcomes, a set of variance-minimizing regression adjustments would be given by the infeasible oracle solutions

\[ \bar{y}_i = (1 - p)y_i(1) + py_i(0), \]

\[ \Delta \bar{y}_{ij} = \left( \frac{n_0}{n} y_i(1) + \frac{n_1}{n} y_i(0) \right) - \left( \frac{n_0}{n} y_j(1) + \frac{n_1}{n} y_j(0) \right) = \bar{y}_i - \bar{y}_j. \]

For known \( p \), I establish a condition under which the respective Bayesian leave-one-out posterior expectations minimize average risk.\(^{22}\) The resulting estimator is a constrained Bayes estimator in the sense of Wald (1950).

**Theorem 2** (Choice of the investigator from fixed-shrinkage estimators). For a known treatment probability \( p \), an investigator with risk \( r \in \mathbb{R}^* \) and prior \( \pi \) over \( \Theta \) with

\[ E_\pi [E_\pi [\bar{y}_j | y_i(1), z_{-ij}] | z_{-i}] = E_\pi [E_\pi [\bar{y}_j | y_i(0), z_{-ij}] | z_{-i}] \] (2)

for all \( i \neq j \) for given shrinkage \( \lambda \in [0, 1) \) and overall bias \( \alpha \) chooses the Bayes estimator

\[ \hat{\tau}(z) = \alpha + (1 - \lambda) \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1 - p)} (y_i - E_\pi [\bar{y}_i | z_{-i}]). \]

The balance assumption in Equation 2 expresses that, on average, what we learn about a given adjustment does not depend on whether we see the treatment or control outcome of another unit, conditional on data from other units. The theorem is non-trivial because one adjustment appears in the estimate for multiple draws of the data.

\(^{22}\)This is similar to Wu and Gagnon-Bartsch (2017) LOOP estimator, which estimates \( y_i(1) \) and \( y_i(0) \) separately from all other units and then averages these estimates with weights \( 1 - p \) and \( p \) to obtain an adjustment that estimates \( \bar{y}_i \).
In particular, if two sample draws only differ in one unit, then the adjustments to that unit are the same. Key to the proof (which I develop in Appendix C) is solving a system of first-order conditions jointly for all potential draws of the data.

While the objective bias restriction dictates sample splitting and guarantees preference alignment, the prior picks one suitable estimator that trades off risk optimally between different unobserved states. If the prior assigns low probability to the realized set of potential outcomes, then the estimator still has the given bias, but may have high variance. In any case, the investigator wants to reveal her best guess given prior knowledge.

Sample splitting guards not just against misaligned preferences, but also against priors that are dogmatic in the treatment effect. From a Bayesian point of view, we only use the prior information orthogonal to the treatment effect. Hence, even if the investigator’s prior is very informative about the treatment effect, the estimator will not reflect this ex-ante bias. The definition of investigator risk functions $R^*$ as mean-squared error with respect to some pseudo-target therefore plays a second role. Alignment with respect to these preferences also implies robustness against misspecification of priors in the direction of the treatment effect. Hence, wrong preconceptions about treatment effects will not lead to systematic distortions in estimates if we restrict researchers to, say, unbiased estimators.

5.3 Complete Class and Estimation-Prediction Duality

Since there is generally no single best estimator for all values of the truth, we have minimized average loss for some prior. If instead we consider admissible estimators that are not dominated by any other estimator in a purely frequentist sense, the same conclusions apply. Indeed, a duality result connects admissible unbiased estimation and admissible prediction.

For finite support any admissible estimator is the limit of a Bayes estimator that minimizes posterior loss given the data for some prior with full support (e.g. Ferguson, 1967). I apply this complete-class argument to unbiased estimators by applying it to the representation in Lemma 1.

**Theorem 3** (Complete-class theorem for unbiased estimators). For any unbiased estimator $\hat{\tau}$ of the sample-average treatment effect that is not dominated with respect to variance, there is a converging sequence of priors $(\pi_t)_{t=1}^\infty$ with full support such that $\hat{\tau}$ equals the limit of the respective optimal Bayes estimators. Conversely, for any
converging sequence of priors \( (\pi_t)_{t=1}^{\infty} \) that put positive weight on every state \( \theta \in \Theta \), every converging subsequence of corresponding Bayes estimators is admissible among unbiased estimators.

Under the assumption of Theorem 2, the individual increments

\[
\phi_i(z_{-i}) = E_{\pi} \left[ \bar{y}_i | z_{-i} \right],
\]

solve a leave-one-out out-of-sample prediction problem. Indeed, \( \phi_i \) minimizes the average of the forecast risk

\[
r_i^\theta(\hat{y}_i) = E_{\theta} \left[ w(d_i)(\hat{y}_i - y_i)^2 \right]
\]

given the respective data and the prior \( \pi \). The weights

\[
w(d_i) = \left( \frac{(d_i - p)}{p(1 - p)} \right)^2,
\]

put higher emphasis on the smaller of the treatment and control groups.

5.4 Constrained Cross-Fold Solutions

It may be infeasible to estimate all regression adjustments optimally. Mimicking machine-learning practice, one could instead partition the sample into \( K \) folds and estimate adjustments in one fold jointly from the units in all other folds. The resulting estimator resembles Wager et al.’s (2016) “cross-estimation” and Chernozhukov et al.’s (2017a) “cross-fitting” estimator.

Remark 5.1 (Exact K-fold cross-fitting). For a partition of the sample

\[
\{1, \ldots, n\} = \bigcup_{k=1}^{K} \mathcal{X}^{(k)}
\]

into \( K \) folds with \( n^{(k)} \geq 2 \) units each of which \( n_1^{(k)} > 0 \) treated and \( n_0^{(k)} > 0 \) untreated, \(^{23}\) This mirrors Lin’s (2013) “tyranny of the minority” estimator, which puts similar weights into a least-squares regression.
the estimator

$$
\hat{\tau}(z) = \frac{1}{n} \sum_{k=1}^{K} n^{(k)} \sum_{i \in I^{(k)}} \frac{d_i n^{(k)} - n^{(k)}_1}{n^{(k)}_1 n^{(k)}_0} \left( y_i - \phi_i^{(k)}(z - I^{(k)}) \right)
$$

is unbiased for the sample-average treatment effect $\tau$ conditional on $(I^{(k)})_{k=1}^{K}$ and $(n^{(k)}_1)_{k=1}^{K}$ under either randomization.

Randomization could be within folds or folds could be chosen after overall randomization. If $K$ divides $n_1$ and $n_0$, we achieve perfect balance by stratifying folds by treatment (or the other way around), $Kn^{(k)}_1 = n_1$ and $Kn^{(k)}_0 = n_0$.

In asymptotic approximation, the optimal regression adjustments are predictions even when not all adjustments are estimated, and even without the additional balance assumption in Theorem 2. Indeed, Theorem E.1 in Appendix E establishes conditions under which the asymptotic variance in estimating the average treatment effect exactly comes from prediction risk according to the weighted loss function. An unbiased estimator of the risk is the average loss the left-out folds, which thus allows for the valid estimation of variance.

5.5 Machine Learning Algorithms as Agents

When high-dimensional unit characteristics are available, machine learning offers a solution to the prediction problems implicit to unbiased estimation. Effectively, machine learning engages in automated specification searches to find a model that predicts well, which Wager et al. (2016) and Wu and Gagnon-Bartsch (2017) also leverage for variance reduction in the same setting. I take a principal-agent perspective on machine-learning algorithms to provide a formal embedding. The investigator as principal delegates to the machine-learning agent. Through sample splitting, there is no misalignment of preferences between the investigator and the machine-learning agent provided the latter minimizes prediction risk, and the investigator achieves a second-best estimation solution from first-best predictions. This connection holds exactly under the assumptions in Theorem 2, and generally in a standard large-sample approximation (Appendix E).

For randomly sampled units, the implicit prediction solutions forecast outcomes from characteristics. If units are draw according to the population distribution
$(y_i(1), y_i(0), x_i) \stackrel{\text{iid}}{\sim} P$ that includes characteristics $x_i$, then

$$y_i(1), y_i(0) | x_1, \ldots, x_n \sim P(x_i).$$

Increments $\phi_i(y_{T_i}, d_{T_i})$ fitted on $T_i \subseteq \{1, \ldots, n\} \setminus \{i\}$ minimize expected forecast risk

$$E[r^i_\theta(\hat{y}_i) | x_1, \ldots, x_n, y_{T_i}, d_{T_i}] = E[E_\theta[w(d_i)(\hat{y}_i - y_i)^2 | y_i(1), y_i(0)] | x_i]$$

over $\hat{y}_i \in \mathbb{R}$. Writing $\hat{y}_i = \hat{f}_i(x_i)$ with $\hat{f}_i : \mathcal{X} \to \mathbb{R}$ a function of training data $(y_{T_i}, d_{T_i}, x_{T_i})$ evaluated on the test point $x_i$, $\hat{f}_i$ solves the prediction problem

$$L_i(\hat{f}) = E[w(d_i)(\hat{f}_i(x_i) - y_i)^2 | x_i] \to \min_{\hat{f}}.$$  (4)

Here, I conflate the population distribution $P$ with the sampling process to describe the distribution of observable data.

Supervised machine learning offers non-parametric solutions of out-of-sample prediction problems like (4) that are particularly suitable for high-dimensional characteristics $x_i$. Since the test point $(y_i, d_i, x_i)$ follows the same distribution as the training sample $T_i$, sample-splitting techniques within the training sample allow for specification searches (in the form of model regularization and combination) to obtain good average predictions at the test point. Furthermore, the realized loss at $i$ is an unbiased estimate of $L_i(\hat{f}_i)$.

I capture machine learning as an agent who minimizes average forecast risk for weighted loss $w(d_i)(\hat{f}_i(x_i) - y_i)^2$. The machine-learning agent’s choice $\hat{f}_i$ may have complex structure that eludes causal interpretation and its parameters may not even be stable approximations of correlation patterns (Mullainathan and Spiess, 2017). However, the investigator as principal cares only about the forecast properties of the agent’s solution.

Crucially, sample splitting guards against prediction mistakes. Even when the specific prediction method does not minimize forecast risk or makes systematic mistakes, the resulting estimator is still unbiased. Worse predictions can lead to worse estimation performance, but only through variance.
6 Pre-Analysis Plans and Ex-Post Analysis

There are two ways in which we can guarantee that the investigator delivers an unbiased estimator (or, more generally, an estimator with fixed bias). In the previous section, I derived a representation of fixed-bias estimators that require that the investigator’s estimator only uses one part of the sample when constructing regression adjustments for another part. Since the investigator will ultimately work with all of the data, this condition cannot be verified ex-post, but has to be guaranteed by ex-ante commitment. One way to guarantee that the estimator fulfills this condition is to require that the investigator commits to the construction of all regression adjustments before she has seen any of the data.

In this section, I consider instead that the investigator commits to how she will split and distribute the data to one or multiple researchers who have not yet accessed the data. Detailed commitment may be infeasible for methods that require active guidance by the researcher, impractical for very complex algorithms, or inefficient when some prior uncertainty is resolved only after the initial commitment. I therefore consider sample-splitting schemes that leave some or all regression adjustments unspecified, and instead delegate their estimation. Delegating to one researcher can already improve over simple pre-specified estimators. Delegating to two researchers attains semi-parametric efficiency without any commitment beyond sample splitting.

6.1 Automated vs Human Specification Searches

The results in this article imply a constructive characterization of robust yet flexible pre-analysis plans. The two ways of ensuring unbiasedness correspond to two different types of specification searches. The first way in which we can be flexible while also ensuring unbiasedness is that the investigator commits in her pre-analysis plan which algorithm she will use to construct regression adjustments.

The second way in which specification searches remain possible applies when the investigator splits the sample and distributes it to one or multiple researchers. Then each researcher can search through specifications using his full subsample and does not have to commit to an empirical strategy ex ante. As long as the investigator commits to how she will distribute the sample and use the output from the researchers, and follows the procedures I characterize below, the resulting estimator is again guaranteed to be unbiased.

Automated and human specification searches can be combined to ensure precise
and unbiased estimation under logistical constraints. An investigator who analyzes
the data by herself can split the sample into two, apply a pre-specified algorithm
to the first half of the data, and search through specifications by hand only in the
second half.

6.2 Unbiased Estimators without Full Commitment

I show that the class of unbiased estimators includes protocols that do not require full
pre-commitment, but leave additional degrees of freedom open. (I discuss the results
in terms of unbiased estimation, but they carry over to estimation with fixed bias
in the same way as above.) The investigator commits to an estimator that includes
flexible inputs by one or multiple researchers. Each researcher obtains access to a
subset of the data, but does not have to pre-commit to their output.

**Definition 2** (K-distribution contract). A K-distribution contract \( \hat{\tau} \) distributes
data \( z = (y, d) \in (Y \times \{0, 1\})^n = \mathcal{Z} \) to K researchers. Researcher \( k \) receives data
\( g_k(z) \in A_k \) and returns the intermediate output \( \hat{\phi}_k(g_k(z)) \in B_k \). The estimate is

\[
\hat{\tau} \Phi((\hat{\phi}_k(z))_{k=1}^K; z) = \Phi((\hat{\phi}_k(g_k(z)))_{k=1}^K; z).
\]

The investigator chooses the functions \( g_k \) (from data in \( \mathcal{Z} \) to researcher input in \( A_k \))
and \( \Phi \) (from the researcher outputs in \( \times_{k=1}^K B_k \) and data in \( \mathcal{Z} \) to estimates in \( \mathbb{R} \))
before accessing the data.

While the investigator still commits which part of the data individual researchers
receive and how their choices and the data form an overall estimate, the individual
researchers’ actions are not pre-specified. From my results in the previous section, I
obtain a full characterization of K-distribution contracts that are unbiased no matter
the choices of the researchers. Since the resulting estimators are always unbiased, the
preferences of the researchers, the investigator, and the designer over these contracts
are aligned provided that the investigator and the researchers all minimize average
risk for risk functions in \( \mathcal{R}^* \) and have the same prior \( \pi \).

**Lemma 6.1** (Characterization of unbiased K-distribution contracts). A K-distribution
contract \( \hat{\tau} \) is unbiased for the sample-average treatment effect \( \tau_0 \) for any conformable
researcher input \( (\hat{\phi}_k)_{k=1}^K \) if and only if:

1. For known treatment probability \( p \), there exist regression adjustments \( (\hat{\phi}_i :
For a known treatment probability $p$, there exist a fixed estimator $\hat{\tau}_0(z)$ with the given bias and regression adjustment mappings $(\Phi_k)_{k=1}^K$ such that

$$\hat{\tau}(\hat{\phi}_k)_{k=1}^K; z) = \frac{1}{n} \sum_{i=1}^n \frac{d_i - p}{p(1 - p)} (y_i - \phi_i((\hat{\phi}_k(g_k(z))))_{k \in C_i}: z_{-i})$$

where $(\phi_i)_{i \in I_k} = \Phi_k(\hat{\phi}_k(z_{-I_k}))$.

2. For a fixed number $n_1$ of treated units, there exist a fixed estimator $\hat{\tau}_0(z)$ with
the given bias and regression adjustment mappings \((\Phi_k)_{k=1}^K\) such that

\[
\tau^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n_1n_0} \sum_{k=1}^K \sum_{\{i<j\} \subseteq I_k} (d_i - d_j) \phi_{ij}^k(z_{-ij}),
\]

where \((\phi_i^k)_{i \in I_k} = \Phi_k(\hat{\phi}_k(z_{-I_k})).\)

\(K\)-fold distribution contracts are similar to \(K\)-fold cross-fitting from Remark 5.1, but different in terms of motivation and more flexible in terms of application. \(K\)-fold distribution is motivated by ensuring unbiasedness, not by computational limitations. While \(K\)-fold cross-fitting is contained as the special case where a researcher determines the regression adjustments for all units in the target fold directly from their training data (that is, no data from the target fold is used to adjust any of the units in that fold), \(K\)-fold distribution contracts also contain solutions that use additional data without bias. Indeed, for the case of known \(p\), say, if regression adjustments take the form \(\phi_i^k(\lambda_k; z_{-i})\) with a pre-determined function \(\phi_i^k(\cdot; \cdot)\) and some tuning parameter \(\lambda_k\), then the adjustments can be a function of all the data in \(z_{-i}\) as long as the tuning parameter \(\lambda_k\) is fitted only on the other folds.

6.3 Hybrid Pre-Analysis Plans

I apply the previous result to show that a simple pre-analysis plan is dominated by a hybrid pre-analysis plan that allows for additional discretion after part of the data is revealed. The investigator fixes some regression adjustment, but can modify others after access to a subset of the sample. Since sample splitting ensures preference alignment, the hybrid estimator will dominate if the ex-post analysis permits better implementation of prior information.

I now assume that the investigator’s prior \(\pi\) is only realized after the data is available. Before the data is available, the investigator has a prior \(\eta^l\) over \(\pi\). I think of \(\eta^l\) as a crude approximation to \(\pi\). A simple ex-ante prior \(\eta^l\) could come from high costs of fully writing down or automating the way in which the investigator translates prior information and data into predictions of potential outcomes. The ex-post prior \(\pi\) could also represent updated beliefs after the pre-analysis plan has

\(^{24}\) This idea can be applied to the post-LASSO (Belloni et al., 2013) after selection on the training sample. Unlike the cross-fitted LASSO, the post-selection fitting step can include the full sample (provided all regression adjustments are fitted using a leave-one- or leave-two-out construction). Furthermore, the selection step can include researcher intervention that has not been pre-specified.
been filed. In both cases, however, the difference does not represent the information in the collected data itself, which will be incorporated in the posterior distribution instead.

Anderson and Magruder (2017) propose a hybrid pre-analysis plan for multiple testing. The investigator pre-specifies some hypothesis they will test, and then selects additional hypotheses from a training sample. The additional hypotheses are only evaluated on the remaining hold-out sample. I adopt their proposal to my estimation setting.

**Definition 6.1** (Hybrid pre-analysis plan). A hybrid pre-analysis plan is a 1-fold distribution contract, i.e. an estimator

\[ \hat{\tau}^\Phi(\hat{\phi}; z) = \Phi(\hat{\phi}(z_T); z) \]

that pre-specifies a mapping \( \Phi \) from ex-post researcher input \( \hat{\phi}(z_T) \) and realized sample data \( z \) to an estimate of the sample-average treatment effect. The researcher (which here could be the investigator herself) obtains access to training data \( T \subseteq \{1, \ldots, n\} \) before the final estimator is formed.

I assume that the investigator must still pre-commit to an unbiased estimator, so Corollary 1 for \( K = 1 \) fully characterizes the plans available to the investigator. In these sample-splitting plans, the choices of the researcher after gaining access to the training sample are fully aligned with the intentions of the investigator according to their updated prior. The investigator pre-commits all adjustments in the training sample according to \( \eta^I \), while the researcher chooses the remaining regression adjustments according to \( \pi \) and their training data.

**Theorem 6.1** (Hybrid pre-analysis plan dominates rigid pre-analysis plan). Assume that investigator and researcher have risk functions in \( R^* \). The optimal unbiased pre-committed estimator \( \hat{\tau}^{\text{pre}} \) is strictly dominated by an unbiased hybrid pre-analysis plan with respect to average variance, i.e. the hybrid plan is as least as precise on average over any ex-ante prior \( \eta^I \) and strictly better for many non-trivial ex-ante priors \( \eta^I \).

Since the researcher’s and investigator’s preference over unbiased estimators is fully aligned with the designer’s goal, there is no preference misalignment and the variance captures all of their risk functions.
Remark 6.1 (Optimal hybrid pre-analysis plan). The dominating hybrid plan is:

1. For known treatment probability \( p \), the researcher chooses regression adjustments \( (\phi^\text{post}_i : (\mathcal{Y} \times \{0, 1\})^{n-1} \to \mathbb{R})_{i \in \mathcal{T}} = \hat{\phi}(z_T) \) to obtain

\[
\hat{\tau}^{\text{hybrid}}(\hat{\phi}; z) = \hat{\tau}^{\text{pre}}(z) - \frac{1}{n} \sum_{i \in \mathcal{T}} d_i - p \frac{p(1-p)}{p(1-p)} \phi^\text{post}_i(z_{-i})
\]

where \( 1 \leq |\mathcal{T}| \leq n - 1 \).

2. For fixed number \( n_1 \) of treated units, the researcher chooses adjustments \( (\phi^\text{post}_{ij} : (\mathcal{Y} \times \{0, 1\})^{n-2} \to \mathbb{R})_{\{i<j\} \cap \mathcal{T} = \emptyset} = \hat{\phi}(z_T) \) to obtain

\[
\hat{\tau}^{\text{hybrid}}(\hat{\phi}; z) = \hat{\tau}^{\text{pre}}(z) - \frac{1}{n_1 n_0} \sum_{\{i<j\} \cap \mathcal{T} = \emptyset} (d_i - d_j) \phi^\text{post}_{ij}(z_{-ij})
\]

where \( 1 \leq |\mathcal{T}| \leq n - 2 \).

In both cases, the investigator commits to the training sample \( \mathcal{T} \subseteq \{1, \ldots, n\} \) and the unbiased estimator \( \hat{\tau}^{\text{pre}} : \mathcal{Z} \to \mathbb{R} \).

6.4 Many-Researcher Delegation

The hybrid pre-analysis plan is itself dominated by a plan that distributes the data to multiple researchers. If a single researcher has access to the full dataset before committing their estimator, bias can return even if the researcher represents their estimate by regression adjustments. Distribution to multiple researchers reduces inefficiency without introducing misalignment. Even when ex-ante commitment beyond a trivial estimator is infeasible or undesirable, distribution between at least two researchers can produce an ex-post desirable estimator.

Remark 6.2 (More researchers are better). Assume that the investigator and researchers all have risk functions in \( \mathcal{R}^* \), and that the researchers all share the same (ex-post) prior \( \pi \). Then an optimal unbiased \( K \)-distribution contract is dominated by an unbiased \( K + 1 \)-distribution contract in the sense of Theorem 6.1.

I now consider standard large-sample efficiency criteria for the estimation of the population-average treatment effect. There is no unique variance-minimal solution in finite samples, as the class of admissible estimators is large. In the large-sample limit, however, essentially all admissible estimators have approximately equal performance,
and coordination between researchers with different (non-dogmatic) priors is resolved by a common understanding of the truth.

Under random sampling of units, the semi-parametric efficiency bound of Hahn (1998) is achieved at oracle prediction adjustments.\[25\] For \((y_i(1), y_i(0), x_i) \overset{iid}{\sim} P\) with fixed probability \(p\) of treatment, an infeasible estimator of the population average treatment effect \(\tau\) is

\[
\hat{\tau}^P(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1 - p)} (y_i - E[\bar{y}_i|x_i])
\]

where the oracle regression adjustments are optimal given knowledge of \(P\). While we will not generally be able to achieve the variance of \(\hat{\tau}^P\), under assumptions we can achieve a variance that is asymptotically equivalent (i.e. \(\text{Var}(\hat{\tau}) / \text{Var}(\hat{\tau}^P) \to 1\) as \(n \to \infty\)).

**Remark 6.3** (Semi-parametric efficiency). If researchers use prediction algorithms \((A_n : Z \to \mathbb{R}^X, z \mapsto \hat{f}_n)_{n=1}^\infty\) with

\[
E[(\hat{f}_n(x_i) - E[\bar{y}_i|x_i])^2] \to 0
\]

as \(n \to \infty\), then delegation to two researchers with risk functions in \(\mathcal{R}^*\) (who each obtain access to half of the data, say) without further commitment achieves both finite-sample unbiased estimation of \(\tau_\theta\), and large-sample semi-parametric efficient estimation of \(\tau\) for the semi-parametric efficiency bound of Hahn (1998).

In other words, semi-parametric efficiency is achieved from distribution of the data to at least two independent researchers with risk-consistent predictors. Data distribution ensures that there is no misalignment.

**Conclusion**

By taking a mechanism-design approach to econometrics, I account for misaligned researcher incentives in causal inference. I motivate why and how we should pre-commit our empirical strategies, and demonstrate that there exist flexible pre-analysis plans that allow for exploratory data analysis and machine learning without leaving room for biases. In particular, I characterize all unbiased estimators of an

\[25\]See also Imbens (2004) for a discussion of efficient estimation of average treatment effects.
average treatment effect as sample-splitting procedures that permit beneficial specification searches.

My results shed light on the role of bias and variance in treatment-effect estimation from experimental data. Allowing for bias can reduce the variance and thus improve precision. But when incentives are misaligned, giving a researcher the freedom to choose the bias may, in fact, reduce precision. However, once we restrict the researcher to fixed-bias estimators, some bias in return for a substantial variance reduction in the nuisance parameters associated with the control variables can improve unbiased estimation.

References


Appendix

Throughout this appendix, I restate the relevant claims from the main paper with their original numbering. I prepend the letter of the respective section to additional and auxiliary results.

A Minimax Optimality of Fixed Bias

Lemma 4.1 (Unbiasedness aligns estimation). If the investigator has risk from $\mathcal{R}^\ast$ then the investigator will choose from the unbiased estimators $\mathcal{C}^\ast$ according to the designer’s preferences.

Proof of Lemma 4.1. Take any investigator risk function $r^I \in \mathcal{R}^\ast$, unbiased estimator $\hat{\tau} \in \mathcal{C}^\ast$, and prior $\pi \in \Delta(\Theta)$. ($\Delta(\Theta)$ denotes the unit $|\theta| - 1$-simplex in $\mathbb{R}^\Theta$.) Then, the designer’s average risk is

$$E_\pi[r^D_\theta(\hat{\tau})]$$

by a bias-variance decomposition. (I conflate $P_\theta$ into $P_\pi$.) Since $E_\pi[(\tau_\theta - \tilde{\tau}_\theta)^2]$ is constant with respect to $\hat{\tau}$ and $E_\pi[\text{Var}_\theta(\hat{\tau}(z))]$ does not vary with $\tilde{\tau}$, the estimation target $\tilde{\tau}$ does not affect the choice of the estimator from $\mathcal{C}^\ast$. Hence, choices are as if $\tilde{\tau} = \tau$. The investigator chooses from $\mathcal{C}^\ast$ according to the designer’s risk $r^D$.

Theorem 1 (Fixed bias is minimax optimal). Write $\Delta^\ast(\Theta)$ for all distributions over $\Theta$ with full support. For every hyperprior $\eta$ with support within $\Delta^\ast(\Theta)$ there is a set of biases $\beta^\eta : \Theta \rightarrow \mathbb{R}$ such that the fixed-bias restriction

$$\mathcal{C}^\eta = \{\hat{\tau} : Z \rightarrow \mathbb{R}; E_\theta[\hat{\tau}] = \tau_\theta + \beta^\eta_\theta\}$$

is a minimax optimal mechanism in the sense of Definition 1, i.e.

$$\mathcal{C}^\eta \in \arg\min_{\mathcal{C}} \sup_{r^I \in \mathcal{R}^\ast} E_\eta \left[ r^D_\theta \left( \arg\min_{\hat{\tau} \in \mathcal{C}} E_\pi[r^I_\theta(\hat{\tau})] \right) \right].$$

51
Proof of Theorem 1. I apply the strategy from Theorem 1 in Frankel (2014) to establish that the unbiasedness restriction yields a minimax (maxmin in utility terms) optimal mechanism. Relative to the quadratic-loss constant-bias setup in Frankel (2014), average risk yields weighted sums where the prior changes weights and the bias changes across decisions (sample draws) and states (posterior expectations). Rather than using Lemma 3 on quadratic-loss constant-bias utilities in Frankel (2014) as stated there, I therefore appeal directly to the logic of his more general Theorem 1, which I extend to deal with the non-compact type and action spaces in my application.

The agent’s (investigator’s) actions are the estimates \(\hat{\tau}(z)\) at all \(N = (2|Y|)^n\) sample points \(z \in Z\). (I assume that the covariates \(x\) are already known when the investigator commits to their estimator.) The state that only the agent observes is the investigator’s prior \(\pi \in \Delta(\Theta)\). \(\pi\) is drawn from the (hyper-)prior \(\eta\).

In the parlance of Frankel (2014), I consider the \(\Phi\)-moment mechanisms where the agent chooses from estimators

\[
C_\beta = \{ \hat{\tau} : Z \to \mathbb{R}; E_\theta[\hat{\tau}] = \tau_\theta + \beta_\theta \forall \theta \in \Theta \}
\]

for a set of fixed biases \(\beta \in \mathbb{R}^\Theta\). (Each expectation – a weighted sum over actions \(\hat{\tau}(z)\) – is a map from actions to real numbers.) To show that this mechanism is maxmin optimal for some choice of \(\beta\), I establish that:

1. Any feasible such \(\Phi\)-moment mechanism (i.e. any bias vector \(\beta\) with \(C_\beta \neq \emptyset\)) induces aligned delegation over \(R^*\), that is, subject to the restriction \(\hat{\tau} \in C_\beta\) agents of all risk types \(r^I \in R^*\) choose as if they were of risk type \(r^D\).

2. \(R^*\) is \(\Phi\)-rich, that is, for any mechanism there exists some \(\overline{\beta} \in \mathbb{R}^\Theta\) and a sequence of risk types \((r^k_k)_{k=1}^\infty \in (R^*)^N\) such that for all realized \(\pi \in \Delta^*(\Theta)\) and all corresponding sequences \((\hat{\tau}^k_k)_{k=1}^\infty\) of chosen estimators, \(\lim_{k \to \infty} E_\theta[\hat{\tau}^k_k(z)] = \tau_\theta + \overline{\beta}_\theta\) for all \(\theta\) in the support of \(\pi\). (Unlike Frankel (2014) I do not explicitly consider mixed strategies since randomized estimators are dominated in my setting.)

Similar to Frankel’s (2014) Theorem 1, the restriction \(C_\beta\) is then minimax optimal provided that \(\beta\) is chosen to minimize the designer’s average risk, for some distribution (hyperprior) \(\eta\) over \(\pi\). I will develop this deduction below for my specific case (in which type and action spaces are not compact) once I have established aligned
delegation and richness.

1. **Aligned delegation.** For $\beta \in \mathbb{R}^\Theta$ such that $C_\beta \neq \emptyset$, the average over risk $r^I \in \mathcal{R}^*$ for an estimator $\hat{\tau} \in C_\beta$ over the prior $\pi \in \Delta(\Theta)$ is

$$E_{\pi} r^I_\theta(\hat{\tau}) = E_{\pi} [\text{Var}_\theta(\hat{\tau}(z))] + E_{\pi} [((\tau_\theta + \beta_\theta - \bar{\tau}_\theta)^2]$$

as in the proof of Lemma 4.1. Hence, choices do not vary with the risk type of the investigator and are as if the investigator shared the designer’s risk function $r^D$.

2. **Richness.** For some arbitrary, but fixed mechanism, our goal is to find a vector of biases $\bar{\beta}$ and a risk sequence $r^{I1}, r^{I2}, \ldots$ such that biases of mechanism outcomes along this sequence always converge to $\bar{\beta}$. I first justify assumptions on the mechanism, then pick a bias vector $\bar{\beta}$, and finally construct a suitable sequence of risk types that ensures bias convergence.

For some conformal mechanism, consider the set $C \subseteq \mathbb{R}^Z$ of estimators $\hat{\tau}$ that are outcomes for some investigator risk function $r^I \in \mathcal{R}^*$ and prior $\pi$ in the support of $\eta$. Note that the outcomes of the mechanism are the investigator choices

$$\hat{\tau}_\pi(r^I) \in \arg \min_{\hat{\tau} \in C} E_{\pi} r^I_\theta(\hat{\tau})$$

where by assumption ties are broken in favor of the designer. I first show that $C$ in (5) is wlog closed. Since the minimizers are already included in $C$, taking the closure of $C$ does not change investigator risk at their optimal choices. Replacing $C$ by its closure thus does not affect investigator risk at choices (5), and can only improve outcomes for the designer, since additional ties are broken in their favor. For the analysis of minimax optimal mechanisms, we can therefore assume wlog that $C$ is closed.

I first assume that $C$ is also bounded. Define the set

$$D = \{ \theta \mapsto E_{\theta} [\hat{\tau}(z)]; \hat{\tau} \in C \} \subseteq \mathbb{R}^\Theta$$

of vectors of expectations achieved by estimators in $C$. By linearity of expectation, $D$ is wlog compact by the above reasoning. Fix some ordering $\theta_1, \ldots, \theta_J$ of $\Theta$ (where $J = |\Theta|$). Let $\delta^0$ be the maximal element in $D$ with respect to the corresponding lexicographic ordering (so that, in particular, $\delta^0_{\theta_1} \geq \delta_{\theta_1}$ for all $\delta \in D$). For every
\( h \in \{2, \ldots, J\}, \) there exists a function \( f_h : \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0} \) such that for all \( \varepsilon > 0 \)
\[
\delta \in \mathcal{D}, \sum_{j=1}^{h-1} |\delta_{\theta_j} - \delta_{\theta_h}^{0}| < f_h(\varepsilon) \quad \Rightarrow \quad \delta_{\theta_h} < \delta_{\theta_h}^{0} + \varepsilon. \tag{6}
\]

Indeed, assume not, then there must be some \( h \) and some \( \varepsilon > 0 \) such that for every \( k \in \mathbb{N} \) there exists a \( \delta^k \in \mathcal{D} \) with \( \sum_{j=1}^{h-1} |\delta_{\theta_j}^k| < 1/k \) and \( \delta_{\theta_h}^k \geq \delta_{\theta_h}^{0} + \varepsilon. \) Since \( \mathcal{D} \) is compact, \( \delta^k \) must have a convergent subsequence with limit \( \delta^\varepsilon \in \mathcal{D}. \) But \( \delta_{\theta_j}^\varepsilon = \delta_{\theta_j}^{0} \) for \( j < h \) and \( \delta_{\theta_h}^\varepsilon \geq \delta_{\theta_h}^{0} + \varepsilon > \delta_{\theta_h}^{0}, \) contradicting that \( \delta^0 \) is maximal in \( \mathcal{D} \) with respect to the lexicographic order. Hence there exists such \( f_h, \) and we can assume wlog \( f_h(\varepsilon) \) is monotonically increasing in \( \varepsilon > 0 \) (otherwise we can choose an \( f_h \) that is smaller for small values of \( \varepsilon \)).

Given the target \( \delta^0 \in \mathcal{D} \) and the functions \( f_h, h \geq 2, \) I construct a sequence of risk functions \( r^k \) such that the expectation of the corresponding investigator choices converges to \( \delta^0 \) for all \( \pi \in \Delta^*(\Theta). \) Concretely, for \( k \in \mathbb{N} \) define \( \alpha^k \in \mathbb{R}^{\Theta} \) recursively
\[
\alpha^k_{\theta_j} = k \quad \Rightarrow \quad \alpha^k_{\theta_j} = k / \min_{h \geq j} f_h(1/\alpha^k_{\theta_h}), j < J
\]
and consider the sequence of investigator risk functions
\[
r^k_\theta(\hat{\tau}) = E_\theta[(\hat{\tau}(z) - \tilde{\tau}^k_\theta)^2], \quad \tilde{\tau}^k_\theta = \delta_{\theta_j}^{0} + \alpha^k_{\theta_j}
\]
which falls within \( \mathcal{R}^*. \)

For the case of bounded \( \mathcal{C} \) and some arbitrary, but fixed \( \pi \in \Delta^*(\Theta), \) it remains to show that the expectation of \( \hat{\tau}_\pi(r^k) \) converges to \( \delta^0. \) Write \( \delta^k_\theta = E_\theta \hat{\tau}_\pi(r^k). \) Assume for contradiction that \( \delta^k_\theta \) does not converge to \( \delta^0_\theta. \) Since also \( \delta^0_\theta \in \mathcal{D} \) for all \( k \) and \( \mathcal{D} \) compact, \( (\delta^k_\theta)_{k=1}^\infty \) must have a converging subsequence \( (\delta^{k_\ell}_\theta)_{\ell=1}^\infty \) with \( \delta^{k_\ell} \rightarrow \delta^1 \in \mathcal{D} \setminus \{\delta^0\} \) as \( \ell \rightarrow \infty. \) The average investigator loss along the sequence is
\[
E_\pi r^k_\theta(\hat{\tau}_\pi(r^k)) = E_\pi Var_\theta(\hat{\tau}_\pi(r^k)) + E_\pi (\delta^{k_\ell} - (\delta^0_\theta + \alpha^{k_\ell}_\theta))^2. \tag{7}
\]
Note that an estimator \( \hat{\tau}^0 \) with expectation \( \delta^0 \in \mathcal{D} \) would also have been available in \( \mathcal{C} \) by definition of \( \mathcal{D}, \) and the difference in risk between the chosen subsequence
and the alternative is

\[
\Delta_{\ell} = \mathbb{E}_{\pi} r_{\theta} \left( \pi^{(k_{\ell})} \right) - \mathbb{E}_{\pi} r_{\theta} \left( \pi^{0} \right)
\]

\[
= \mathbb{E}_{\pi} \left( \delta_{\theta}^{k_{\ell}} - (\delta_{\theta}^{0} + \alpha_{\theta}^{k_{\ell}}) \right)^2 - 2 \mathbb{E}_{\pi} (\delta_{\theta}^{k_{\ell}} - \delta_{\theta}^{0}) \alpha_{\theta}^{k_{\ell}} + \mathcal{O}(1)
\]

\[
= -2 \sum_{j=1}^{h-1} \pi(\theta) \alpha_{\theta j}^{k_{\ell}} (\delta_{\theta j}^{k_{\ell}} - \delta_{\theta j}^{0}) + \mathcal{O}(1).
\]

Denote by \( h \) the smallest index of for which \( \delta_{\theta h}^{0} \neq \delta_{\theta h}^{1} \). Since \( \delta^{0} \) is maximal with respect to the lexicographic ordering of \( D \) and \( \delta^{1} \) also in \( D \), we must have \( \delta_{\theta h}^{0} - \delta_{\theta h}^{1} > 0 \). By revealed preference and since \( \alpha_{\theta j+1}^{k} = o(\alpha_{\theta j}^{k}) \) for all \( j \), it follows that

\[
0 \geq \Delta_{\ell} / \alpha_{\theta h}^{k_{\ell}} = -2 \sum_{j=1}^{h-1} \pi(\theta) \alpha_{\theta j}^{k_{\ell}} (\delta_{\theta j}^{k_{\ell}} - \delta_{\theta j}^{0}) - 2 \pi(\theta h) (\delta_{\theta h}^{1} - \delta_{\theta h}^{0}) + o(1).
\]

In particular, for \( \varepsilon = \pi(\theta h) (\delta_{\theta h}^{1} - \delta_{\theta h}^{0}) \),

\[
\liminf_{\ell \to \infty} \sum_{j=1}^{h-1} \pi(\theta) \alpha_{\theta j}^{k_{\ell}} (\delta_{\theta j}^{k_{\ell}} - \delta_{\theta j}^{0}) \geq \varepsilon > 0.
\]

Hence there must exists some \( h^* \) and a subsequence \( \ell_s \) such that

\[
a_{\ell s}^{h^*} \to \nu \in (0, \infty], \quad \limsup_{s \to \infty} \frac{a_{\ell s}^{h^*}}{a_{\ell s}^{h^*}} \leq 1 \forall j < h.
\]

(That is, \( a_{h^*}^{\ell_s} \) is a maximal sequence within that subsequence, for a suitable asymptotic notion of maximality; it is not unique, but an instance can be constructed from iterated subsequences.) For simplicity, I write \( k_s = k_{\ell_s} \). I assume wlog that \( \delta_{\theta h^*}^{k_s} - \delta_{\theta j}^{0} > 0 \) for all \( s \). By (7),

\[
\sum_{j=1}^{h^*-1} |\delta_{\theta j}^{k_s} - \delta_{\theta j}^{0}| \geq f_{h^*} (\delta_{\theta h^*}^{k_s} - \delta_{\theta j}^{0}),
\]

55
so there must exist some \( j^* < h^* \) and a refinement of the subsequence along which 
\[ |\delta_{\theta_{j^*}}^{k_s} - \delta_{\theta_{j^*}}^0| \geq f_{h^*} (\delta_{\theta_{h^*}}^{k_s} - \delta_{\theta_{h^*}}^0) / (h^* - 1) \].

Note that

\[
\frac{\pi(\theta_{j^*}) \alpha_{\theta_{j^*}}^{k_s} |\delta_{\theta_{j^*}}^{k_s} - \delta_{\theta_{j^*}}^0|}{\pi(\theta_{h^*}) \alpha_{\theta_{h^*}}^{k_s} (\delta_{\theta_{h^*}}^{k_s} - \delta_{\theta_{h^*}}^0)} \geq \frac{\pi(\theta_{j^*})}{\pi(\theta_{h^*}) (h^* - 1)} \frac{\alpha_{\theta_{j^*}}^{k_s} f_{h^*} (\delta_{\theta_{h^*}}^{k_s} - \delta_{\theta_{h^*}}^0)}{\alpha_{\theta_{h^*}}^{k_s} \delta_{\theta_{h^*}}^{k_s} - \delta_{\theta_{h^*}}^0}.
\]

By \([\rho]\) there exists some \( \nu_0 \in (0, \infty) \) such that \( a_{h^*}^{\ell_s} \geq \nu_0 \) for all large \( s \). By the definition of \( a_{h^*}^{\ell_s} \) we find, again for large \( s \), that

\[
\delta_{\theta_{h^*}}^{k_s} - \delta_{\theta_{h^*}}^0 = \frac{a_{h^*}^{\ell_s} \alpha_{\theta_{h^*}}^{k_s}}{\pi(\theta_{h^*}) \alpha_{\theta_{h^*}}^{k_s}} \geq \frac{\nu_0}{\pi(\theta_{h^*}) \alpha_{\theta_{h^*}}^{k_s}}.
\]

By monotonicity of \( f_{k_s}(\cdot) / \varepsilon \) therefore for large \( s \)

\[
\frac{\pi(\theta_{j^*}) \alpha_{\theta_{j^*}}^{k_s} |\delta_{\theta_{j^*}}^{k_s} - \delta_{\theta_{j^*}}^0|}{\pi(\theta_{h^*}) \alpha_{\theta_{h^*}}^{k_s} (\delta_{\theta_{h^*}}^{k_s} - \delta_{\theta_{h^*}}^0)} \geq \frac{\pi(\theta_{j^*})}{\pi(\theta_{h^*}) (h^* - 1)} \frac{\alpha_{\theta_{j^*}}^{k_s} f_{h^*} \left( \frac{\nu_0}{\pi(\theta_{h^*}) \alpha_{\theta_{h^*}}^{k_s}} \right)}{\alpha_{\theta_{h^*}}^{k_s} \delta_{\theta_{h^*}}^{k_s} - \delta_{\theta_{h^*}}^0}.
\]

By construction of the rates \( \alpha_{\theta}^k \), we have that for every triple \( j^* < h^* < h \) and every constant \( c > 0 \) and all large \( k \)

\[
\frac{\alpha_{\theta_{j^*}}^{k_s} f_{h^*}}{\alpha_{\theta_{h^*}}^{k_s}} \left( \frac{\alpha_{\theta_{j^*}}^{k_s}}{\alpha_{\theta_{h^*}}^{k_s}} \right) \geq \frac{\alpha_{\theta_{j^*}}^{k_s} f_{h^*}}{\alpha_{\theta_{h^*}}^{k_s}} \left( \frac{\alpha_{\theta_{j^*}}^{k_s}}{\alpha_{\theta_{h^*}}^{k_s}} \right) = \frac{\alpha_{\theta_{j^*}}^{k_s} f_{h^*}}{\alpha_{\theta_{h^*}}^{k_s}} \left( \frac{ck}{\alpha_{\theta_{h^*}}^{k_s}} \right) \geq c \frac{\alpha_{\theta_{j^*}}^{k_s} f_{h^*}}{\alpha_{\theta_{h^*}}^{k_s}} \left( \frac{1}{\alpha_{\theta_{h^*}}^{k_s}} \right) \geq ck \rightarrow \infty.
\]

It follows that

\[
\frac{\pi(\theta_{j^*}) \alpha_{\theta_{j^*}}^{k_s} |\delta_{\theta_{j^*}}^{k_s} - \delta_{\theta_{j^*}}^0|}{\pi(\theta_{h^*}) \alpha_{\theta_{h^*}}^{k_s} (\delta_{\theta_{h^*}}^{k_s} - \delta_{\theta_{h^*}}^0)} \rightarrow \infty.
\]

By \([\rho]\), \( \delta_{\theta_{j^*}}^{k_s} - \delta_{\theta_{j^*}}^0 < 0 \) for all but at most finitely many \( s \). Hence \( a_{j^*/h^*}^{\ell_s} \rightarrow -\infty \), and thus \( \sum_{j=1}^{h-1} \delta_{j}^{\ell_s} \rightarrow -\infty \), contradicting \([8]\). Therefore \( \delta^1 = \delta^0 \).
Consider now the case when $\mathcal{C}$ is unbounded. First, if $\mathcal{C}$ is unbounded but $\mathcal{B}$ is still bounded (and thus wlog compact by linearity of the expectation projection), then the same argument as above goes through since there is always an estimator with finite variance and expectation $\delta^0$ available (and the investigator minimizes variance given expectation), so unbounded variance along the investigator path can only make the choice with expectation $\delta^0$ more attractive.

Second, if $\mathcal{B}$ is also unbounded, then $\mathcal{C}$ cannot be minimax optimal. Since $\mathcal{B}$ is unbounded, it must contain a sequence $\delta^k \in \mathcal{B}$ with $\|\delta^k\|$ diverging. The projection of $\delta^k$ on the unit sphere towards the origin must contain a converging subsequence with limit $v$ where $\|v\| = 1$. Consider a sequence of investigators with $\tilde{\tau}^k = v$ along the ray defined by the direction of this cluster point. One, if the average variance along the sequence of investigator choices is unbounded, then so is the average risk of the designer. Two, if the average variance along the sequence of investigator choices is bounded, then the bias diverges and average risk of the designer is again unbounded. Indeed, I show that it is not possible that both average variance and average expectation remain bounded along the ray. If the expectation vector $E_\theta[\hat{\tau}(z)]$ along that sequence of investigators remains bounded, pick a point arbitrarily close to the ray that falls outside that bound. (Such a point exists by construction of $v$.) As investigator preference moves along the ray, the gain in average investigator risk from moving to that point outweigh any cost in terms of variance since the marginal cost of being off the expectation target only increases, while the variance cost remains bounded. Hence, the bias cannot remain bounded and the average risk of the designer diverges.

We therefore have that for any $\pi \in \Delta^*(\Theta)$ the bias of investigator choices along the sequence $r^I_k$ converges to $\bar{\beta}_\theta = \delta^0_\theta - \tau_\theta$ for all $\theta \in \Theta$.

**Proof of minimax optimality.** Given any mechanism, by richness there exists a sequence of investigator risk functions $r^I_k$ in $\mathcal{R}^*$ and a bias vector $\bar{\beta}$ such that $E_\theta[\hat{\tau}_\pi(r^I_k)] - \tau_\theta \to \bar{\beta}_\theta$ for all $\pi \in \Delta^*(\Theta)$ and all $\theta \in \Theta$. The expected average designer’s risk along this sequence is

$$E_\eta[(\hat{\tau}_\pi(r^I_k) - \tau_\theta)^2] = E_\eta \operatorname{Var}_\theta(\hat{\tau}_\pi(r^I_k)) + E_\eta (E_\theta[\hat{\tau}_\pi(r^I_k)] - \tau_\theta)^2,$$

$$\to \bar{\beta}_\theta^2 \forall \theta \in \Theta, \pi \in \Delta^*(\Theta)$$
where I omit the argument $z$ of the estimators. Since biases are bounded (since $D$ is) and the support of $\eta$ is in $\Delta^*(\Theta)$, by dominated convergence

$$\liminf_{k \to \infty} E_\eta[(\hat{\tau}_\pi(r^I_k) - \tau_\theta)^2] = \liminf_{k \to \infty} E_\eta \Var_\theta(\hat{\tau}_\pi(r^I_k)) + E_\eta \bar{\beta}^2_\theta \geq E_\eta \liminf_{k \to \infty} E_\pi \Var_\theta(\hat{\tau}_\pi(r^I_k)) + E_\eta \bar{\beta}^2_\theta.$$  

For fixed $\pi \in \Delta^*(\Theta)$, $\liminf_{k \to \infty} E_\pi \Var_\theta(\hat{\tau}_\pi(r^I_k))$ is at least the minimal asymptotic variance along a sequence $\hat{\tau}^k_\pi$ with bounded bias that converges to $\bar{\beta}$, and is otherwise unrestricted. Take such a sequence for which $E_\pi \Var_\theta(\hat{\tau}^k_\pi(r^I_k))$ converges to its minimal limit. Along this sequence, $\hat{\tau}^k_\pi$ must be bounded, so it must have a convergent subsequence with some limit $\hat{\tau}^0_\pi$ in $\mathbb{R}^Z$ for which by continuity also $E_\theta(\hat{\tau}^0_\pi) - \tau_\theta = \bar{\beta}_\theta$. But then the variance of $\hat{\tau}^0_\pi$ must be at least the variance of a variance-minimizing estimator subject to the bias constraint. Taken together,

$$\inf_{r^I \in \mathcal{R}^*} E_\eta[R^D_\theta(\hat{\tau}_\pi(r^I))] \geq \liminf_{k \to \infty} E_\eta[(\hat{\tau}_\pi(r^I_k) - \tau_\theta)^2] \geq E_\eta \min_{\hat{\tau} \in \mathcal{C}_{\bar{\beta}}} E_\pi \Var_\theta(\hat{\tau}) + E_\eta \bar{\beta}^2_\theta.$$  

Now, by aligned delegation,

$$\min_{\hat{\tau} \in \mathcal{C}_{\bar{\beta}}} E_\pi(\Var_\theta(\hat{\tau}) + \bar{\beta}^2_\theta) = \min_{\hat{\tau} \in \mathcal{C}_{\bar{\beta}}} E_\pi R^D_\theta(\hat{\tau}) = E_\pi R^D_\theta(\hat{\tau}_\pi(r^I))$$

for every $r^I \in \mathcal{R}^*$ for choices from $\mathcal{C}_{\bar{\beta}}$. It follows that for every mechanism there is a set of biases such that the fixed-bias mechanisms has at least weakly better worst-case (over investigator types in $\mathcal{R}^*$) performance. Hence, at an optimal choice of biases $\beta^\eta$ given the hyperprior $\eta$, the fixed-bias restriction $\mathcal{C}^\eta$ is minimax optimal. Such a minimizer exists because the set of biases is wlog compact (indeed, we can assume $E_\eta \beta^2_\theta \leq E_\eta R^D_\theta(z \to 0) < \infty$) and the expected average risk continuous in the choice of bias. \hfill \Box

I conjecture that the restriction of the support to priors with full support is not necessary.
B Representation of Unbiased Estimators

As in the main text, for fixed $n \geq 1$ and finite support $\mathcal{Y}$ I consider potential outcomes $\theta = (y(1), y(0)) \in \Theta = (\mathcal{Y}^2)^n$ from which for treatment $d \in \{0, 1\}^n$ we observe $y = d \circ y(1) + (1 - d) \circ y(0) \in \mathcal{Y}^n$. (Here, $\circ$ denotes the Hadamard (entry-wise) product.) The estimate of interest is $\tau_\theta = 1'(y(1) - y(0))/n$.

**Lemma 1** (Representation of fixed-shrinkage estimators). The estimator $\hat{\tau}$ has fixed shrinkage,

$$E_\theta[\hat{\tau}(z)] = \alpha + (1 - \lambda)\tau_\theta$$

for all potential outcomes $\theta \in \Theta$ (where $\lambda \in [0, 1)$), if and only if:

1. For a known treatment probability $p$, there exist leave-one-out regression adjustments $(\phi_i : (\mathcal{Y} \times \{0, 1\})^{n-1} \to \mathbb{R})_{i=1}^n$ such that

$$\hat{\tau}(z) = \alpha + (1 - \lambda)\frac{1}{n} \sum_{i=1}^n d_i - p \frac{p(1 - p)}{p(1 - p)} (y_i - \phi_i(z_{-i})).$$

2. For a fixed number $n_1$ of treated units, there exist leave-two-out regression adjustments $(\phi_{ij} : (\mathcal{Y} \times \{0, 1\})^{n-2} \to \mathbb{R})_{i<j}$ such that

$$\hat{\tau}(z) = \alpha + (1 - \lambda)\frac{1}{n_1n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}(z_{-ij})).$$

where $\phi_{ij}(z_{-ij})$ may be undefined outside $1'd_{-ij} = n_1 - 1$.

I build up this general representation result in steps from simple estimators with binary outcomes to general estimators with finite support.

**B.1 Known treatment probability, binary outcomes**

I start with known treatment probability $p = E_\theta[d_i]$ with $d_i$ iid and binary support.

A natural class of admissible estimators are Bayes estimators, so a tempting starting point for the analysis of optimal unbiased estimators are (limits of) Bayes estimators that minimize average mean-squared error given the data and are also unbiased. However:
Remark B.1. For $\mathcal{Y} = \{0,1\}$ and $p = .5$, the only unconstrained Bayes estimator (with respect to average mean-squared error) that is unbiased (conditional on $(y(1), y(0))$) is $\hat{\tau}(y, d) = \frac{1}{n}(2d - 1)'(2y - 1)$. For $\mathcal{Y} = \{0,1\}$ and $p \neq .5$, there are no unconstrained Bayes estimators that are also unbiased.

Sketch of proof. For any prior, the unconstrained Bayes estimator with respect to average mean-squared error is the posterior expectation of $\tau_{\theta}$ given the data. Any posterior expectation of $\tau_{\theta}$ is bounded between the maximal treatment effect $+1$ and the minimal treatment effect $-1$. To achieve unbiasedness, any data that is consistent with either of the extremes must therefore yield an estimate of $+1$ or $-1$, respectively. Iterating this argument, the unique unconstrained Bayes estimator is the one achieved from a prior that puts full probability on $(y_i(1), y_i(0)) \in \{(1,0), (0,1)\}$ and zero probability on the configurations $\{(1,1), (0,0)\}$. This yields $E_{\theta}[y_i(1) - y_i(0)|y_i, d_i] = (2d_i - 1)(2y_i - 1)$, which is unbiased for $p = .5$, but not for $p \neq .5$.

The remark implies that searching for unbiased estimators among unconstrained Bayes estimators to characterize the class of admissible unbiased estimators is futile, and I instead first characterize unbiased estimators before returning to optimality by solving for constrained Bayes estimators subject to the resulting representation.

Theorem B.1. For $\mathcal{Y} = \{0,1\}$, assume that the estimators $\hat{\tau}^A, \hat{\tau}^B$ are unbiased $\tau_{\theta}$ (conditional on $\theta = (y(1), y(0))$). Then,

$$\hat{\tau}^B(y, d) - \hat{\tau}^A(y, d) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1 - p)} \phi_i(y_{-i}, d_{-i})$$

for a set of functions $\phi_i : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R}$.

For $n = 2$, the proof of Theorem B.1 can be made on a two-dimensional lattice folded into a torus. The general proof can similarly be understood as summing over hypercubes on the surface of an $n$-torus.

Proof. For $\hat{\delta}(y, d) = \hat{\tau}^B(y, d) - \hat{\tau}^A(y, d)$, take $\phi_i(y_{-i}, d_{-i})$ such that

$$\hat{\delta}(y, d) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1 - p)} \phi_i(y_{-i}, d_{-i})$$

(10)
for all \((y, d)\) with \(y'd > 0\) (that is, all those that include some pair \((y_j, d_j) = (1, 1)\)).
This is always feasible, say by the following inductive construction:

1. Set the \(\phi_i(1_{n-1}, 1_{n-1})\) in any way that has (10) hold for \(\hat{\delta}(1_n, 1_n)\).

2. Assuming that \(\phi_i(y_{-i}, d_{-i})\) has been set for all \(i\) and \((y, d)\) with \(y'd \geq n - k\) such that (10) holds for such \((y, d)\) (as is the case for \(k = 0\) by the previous step), consider \((y, d)\) with \(y'd = n - (k + 1)\). Among the terms \(\phi_i(y_{-i}, d_{-i})\) in (10), those with \(y'_{-i}d_{-i} = n - (k + 1)\) have already been set by the induction assumption, and it remains to show that we can set conformable terms \(\phi_i(y_{-i}, d_{-i})\) for \(y'_{-i}d_{-i} = n - (k + 2)\).

Provided that \(k < n - 1\), note that any \((y, d)\) with \(y'd = n - (k + 1)\) contains at least one \((y_i, d_i)\) with \(y'_id_i = 1\), \(\hat{\delta}(y, d)\) has the term \(\phi_i(y_{-i}, d_{-i})\) appear on the right in (10), where thus \(y'_{-i}d_{-i} = y'd - 1 = n - (k + 2)\) (so it has not yet been set). But note that this specific \(\phi_i(y_{-i}, d_{-i})\) also appears only for that \((y, d)\) among all \((y, d)\) with \(y'd = n - (k + 1)\) as necessarily \(y'_id_i = 1\).

Hence, we can set all previously undetermined \(\phi_i(y_{-i}, d_{-i})\) for all \(i\) and \(y'd\) with \(y'd \geq n - (k + 1)\) in a way that (10) holds for such \((y, d)\).

By induction, we have set all \(\phi_i(y_{-i}, d_{-i})\) for any \(i\) and \(y'd \geq 1\) conformably with (10) for such \((y, d)\). Since this includes all terms of the form \(\phi_i(y_{-i}, d_{-i})\), it remains to show that the unbiasedness assumption implies that (10) extends to \((y, d)\) with \(y'd = 0\).

Write \(\hat{\delta}^\phi\) for the function defined by (10) for all \((y, d)\). We have thus shown that \(\hat{\delta}^\phi(y, d) = \hat{\delta}(y, d)\) for all \((y, d)\) with \(y'd > 0\). By assumption, \(E_\theta[\hat{\delta}(y, d)] = 0\) for all \(\theta = (y(1), y(0))\), so

\[
0 = E_\theta[\hat{\delta}(y, d)] = \sum_{d \in \{0,1\}^n} P(d) \hat{\delta}(d \circ y(1) + (1 - d) \circ y(0), d).
\]

Fixing \((y^*, d^*)\), it follows for any \(\tilde{y}\) that

\[
\hat{\delta}(y^*, d^*) = -\sum_{d \in \{0,1\}^n \setminus \{d^*\}} P(d) / P(d^*) \hat{\delta}((1_{d_i = d_i^*})_{i=1}^n \circ y^* + (1_{d_i \neq d_i^*})_{i=1}^n \circ \tilde{y}, d) \quad (11)
\]

Since \(\hat{\delta}^\phi\) is similarly zero-bias by construction, the same holds for \(\hat{\delta}^\phi\). Thus, if for
some \((y^*, d^*)\) \(\hat{\delta}\) and \(\hat{\delta}^\phi\) agree on

\[
\hat{y}^*(d) = (\mathbb{1}_{d_i = d_i^*})_{i=1}^n \circ y^* + (\mathbb{1}_{d_i \neq d_i^*})_{i=1}^n \circ \tilde{y}, d)
\]

for some \(\tilde{y}\) and all \(d \neq d^*\), then \(\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)\).

We are ready to show \([10]\) for all \((y^*, d^*)\), by induction over \(1'd^*\). We let \(\tilde{y} = 1\) throughout. At \(k = 0, d^* = 0\). For any \(d \neq d^*\), \(\tilde{y}(d)'d \geq 1\), so \(\hat{\delta}(\tilde{y}^*(d), d) = \hat{\delta}^\phi(\tilde{y}^*(d), d)\). By \([11]\), \(\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)\). Assume now that the claim holds for all \((y^*, d^*)\) with \(1'd^* \leq k\), and consider some \((y^*, d^*)\) with \(1'd^* = k+1\). Then, for any \(d \neq d^*\) with \(1'd \leq k\), \(\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)\) by the induction assumption. For any \(d \neq d^*\) with \(1'd \geq k+1\) there must be at least one dimension \(i\) with \(d_i = 1, d_i^* = 0\), thus \(\tilde{y}^*(d)'d \geq 1\) and \(\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)\) follows by construction. We conclude that \(\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)\) for all \((y^*, d^*)\).

Since \(\hat{\tau}(y, d) = \frac{1}{n} \sum_{i=1}^n\left(\frac{d_i - p}{p(1-p)} y_i\right)\) is unbiased for \(\tau_\theta\), the following characterization is immediate:

**Corollary B.1.** For \(\mathcal{Y} = \{0, 1\}\), any unbiased estimator \(\hat{\tau}\) of \(\tau_\theta\) can be expressed as

\[
\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \left(\frac{d_i - p}{p(1-p)} (y_i - \phi_i(y_{-i}, d_{-i})))\right).
\]

The following result for the special case \(n = 2\) shows that the reduction in degrees of freedom in the estimator implied by unbiasedness is substantial:

**Remark B.2.** For \(n = 2\), the \(\phi_i(y_{-i}, d_{-i})\) are unique up to the one-dimensional equivalence class \(\phi_i^*(y_{-i}, d_{-i}) = \phi_i^*(y_{-i}, d_{-i}) + (-1)^i(2d_{3-i} - 1)\Delta\), so unbiasedness reduces the degrees of freedom from \(\hat{\tau} \in \mathbb{R}^{16}\) to \(\hat{\phi} \in \mathbb{R}^7\).

### B.2 Fixed treatment group size, binary

Assume now that instead of the treatment probability, the number of treated is fixed at \(n_1\), so that \(d \sim \mathcal{U}(\mathcal{D}_{n_1})\) with \(\mathcal{D}_{n_1} = \{t \in \{0, 1\}^n; t' n = n_1\}\). Effectively, we assume invariance to permutations in the assignment of treatment, but not more.

The natural, unbiased treatment-control-difference estimator can be written as

\[
\hat{\tau}^\alpha(y, d) = \frac{1}{n_1} \sum_{d_i = 1} y_i - \frac{1}{n_0} \sum_{d_i = 0} y_i = \frac{1}{n_1 n_0} \sum_{d_i = 1, d_j = 0} (y_i - y_j),
\]

62
of which an unbiased extension is

\[ \hat{\tau}^\phi(y, d) = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} (y_i - y_j - \phi_{ij}(y_{-ij}, d_{-ij})) \]

with \( \phi_{ij} = -\phi_{ji} \). I claim that these are also all extensions.

**Theorem B.2.** Let \( \mathcal{Y} = \{0, 1\} \). Assume that \( \hat{\tau}^A, \hat{\tau}^B \) are unbiased for \( \tau_0 \). Then,

\[ \hat{\tau}^B(y, d) - \hat{\tau}^A(y, d) = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} \phi_{ij}(y_{-ij}, d_{-ij}), \quad \phi_{ij} = -\phi_{ji} \]

for functions \( \phi_{ij} : (\mathcal{Y} \times \{0, 1\})^{n-2} \rightarrow \mathbb{R} \).

Note that we can alternatively write

\[ \hat{\tau}^B(y, d) - \hat{\tau}^A(y, d) = \frac{1}{n_1 n_0} \sum_{i=1}^{n} \sum_{j=1}^{n} (d_i - d_j) \phi_{ij}(y_{-ij}, d_{-ij}), \]

where we sum over each pair once and \( \phi_{ij} \) is only defined for \( j > i \).

We first establish a lemma that adopts the proof strategy from **Theorem B.1** to the setting at hand. To this end, for \((y(1), y(0)) \in (\mathcal{Y}^2)^n\) write

\[ N(y(1), y(0)) = \{(d \circ y(1) + (1 - d) \circ y(0), d); d \in D_{n_1}\} \]

(the set of observations consistent with \(y(1), y(0))\) and let

\[ \mathcal{C} = \bigcup_{(y(1), y(0)) \in (\mathcal{Y}^2)^n} N(y(1), y(0)). \]

Let \( c : \mathcal{C} \rightarrow \mathcal{C}^- \) be the surjective correspondence

\[ (y, d) \mapsto \{(ij, (y_{-ij}, d_{-ij})); i < j, d_i \neq d_j\}. \]

**Lemma B.1.** If there exists a partition \( \mathcal{C} = \bigcup_{t=1}^{T*} \mathcal{C}_t \) such that for some \( T^* \)

1. for \( \mathcal{C}_t^- = \bigcup_{(y, d) \in \mathcal{C}_t} c(y, d) \) and

\[ D_t = \mathcal{C}_t^- \setminus \bigcup_{s < t} \mathcal{C}_s^- , \]

there exists injections \( b_t : \mathcal{C}_t \rightarrow D_t \) for \( t \leq T^* \) and

63
2. for all \( t > T^* \) and \((y, d) \in C_t\), there exists some \((y(1), y(0)) \in (Y^2)^n\) both 
\((y, d) \in N(y(1), y(0))\) and 

\[
(N(y(1), y(0)) \setminus \{(y, d)\}) \cap \bigcup_{s \geq t} C_s = \emptyset
\]

then for any \( \hat{\delta} \) that is mean-zero there exist a function \( \phi : C^- \to \mathbb{R} \) such that \( \hat{\delta} = \hat{\delta}^\phi \) with

\[
\hat{\delta}^\phi(y, d) = \frac{1}{n_1 n_0} \sum_{i=1}^n \sum_{j=i+1}^n (d_i - d_j) \phi_{ij}(y_{-ij}, d_{-ij}).
\]

**Proof.** Given some \( \hat{\delta} \), we first construct such a family \( \phi \) with \( \hat{\delta}^\phi(y, d) = \hat{\delta} \) for all \((y, d) \in \bigcup_{s<T^*} C_s\), and then establish that this implies \( \hat{\delta}^\phi(y, d) = \hat{\delta} \) also for \((y, d) \in \bigcup_{t>T^*} C_t\).

For the first part, I argue inductively as follows: Take \( t \leq T^* \) and assume \( \phi \) has been set on \( \bigcup_{s<t} C_s \) such that \( \hat{\delta}^\phi = \hat{\delta} \) on \( \bigcup_{s<t} C_s \) (which is given trivially for \( t = 1 \)) then for every \((y, d) \in C_t\) by the first assumption of the lemma there exists a unique term \( \phi_{ij}(y_{-ij}, d_{-ij}) = \phi(b_t(y, d)) \) with \( b_t(y, d) \in D_t \) that has not yet been set, so we can set the terms \( \phi(D_t) \) in a way that \( \hat{\delta}^\phi = \hat{\delta} \) on \( C_t \) and thus on \( \bigcup_{s\leq t} C_s \). This completes the proof of the first part.

For the second part, note that by assumption \( E_\theta[\hat{\delta}(y, d)] = 0 \) for all \( \theta = (y(1), y(0)) \), so

\[
0 = E_\theta[\hat{\delta}(y, d)] = \sum_{(y, d) \in N(y(1), y(0))} \hat{\delta}(y, d).
\]

Fixing \((y^*, d^*)\) it follows for any \((y(1), y(0))\) with \((y^*, d^*) \in N(y(1), y(0))\) that

\[
\hat{\delta}(y^*, d^*) = - \sum_{(y, d) \in N(y(1), y(0)) \setminus \{(y^*, d^*)\}} \hat{\delta}(y, d) \tag{12}
\]

Since \( \hat{\delta}^\phi \) is similarly zero-bias by construction, the same holds for \( \hat{\delta}^\phi \). We are now ready to show that \( \hat{\delta}^\phi = \hat{\delta} \) for all \((y, d) \in C_t\), by induction over \( t \). For some \( t > T^* \), assuming \( \hat{\delta}^\phi = \hat{\delta} \) holds for all \((y, d) \in C_s\) with \( s < t \) (as is the case for all \( s \leq T^* \)), take any \((y^*, d^*) \in C_t\). By the second part of the lemma, \( \hat{\delta}^\phi_{ij}(y_{-ij}, d_{-ij}) \) and the induction assumption we must have \( \hat{\delta}(y^*, d^*) = \hat{\delta}^\phi_{ij}(y^*, d^*) \). This completes the proof. \( \square \)
We are ready to prove the main result:

**Proof of Theorem B.2.** \( \hat{\delta}(y, d) = \hat{\tau}^B(y, d) - \hat{\tau}^A(y, d) \) is a unbiased estimator of zero.

Define \( a, b : C \to \mathbb{N}_0 \) by

\[
a(y, d) = y' d, \quad b(y, d) = (1 - y)' (1 - d).
\]

Note that \( a(y, d) + b(y, d) \leq n \).

First, set \( T^* = n - 1 \) and for every \( t \leq T \)

\[
C_t = \{(y, d) \in C; \min(a(y, d), b(y, d)) \geq 1, a(y, d) + b(y, d) = n + 1 - t\}.
\]

Then the first assumption of Lemma B.1 is fulfilled, as for every \( (y, d) \in C_t \) there exists some \( (i, j, y_{-ij}, d_{-ij}) \in C_t \) with \( y_{-ij} d_{-ij} + (1 - y_{-ij})'(1 - d_{-ij}) = n - 1 - t = a(y, d) + b(y, d) - 2 \), but \( (y, d) \) is also the unique element in \( C_t \) covering that element of \( D_t \) under the correspondence \( c \) (as indeed necessarily \( y_i = d_i, y_j = d_j \), which pins down \( (y, d) \) from \( (i, j, y_{-ij}, d_{-ij}) \)).

Second, with \( T = n + 1 \) and

\[
C_n = \{(y, d) \in C; a(y, d) = 0, b(y, d) \geq 1\}, \quad C_{n+1} = \{(y, d) \in C; b(y, d) = 0\},
\]

note that for each \( (y^*, d^*) \in C_n \cup C_{n+1} \) we have that \( (y(1), y(0)) = (y^* \circ d^* + 1 \circ (1 - d^*), y^* \circ (1 - d^*)) \) produces

\[
N(y(1), y(0)) \cap \{(y, d) \in C; \min(a(y, d), b(y, d)) = 0\} = \{(y^*, d^*)\}
\]

for \( (y^*, d^*) \in C_n \) and

\[
N(y(1), y(0)) \cap \{(y, d) \in C; b(y, d) = 0\} = \{(y^*, d^*)\}
\]

for \( (y^*, d^*) \in C_{n+1} \). This verifies the second assumption of Lemma B.1.

Unbiased estimators (for binary outcomes) are thus fully characterized by leave-two-out adjustments. Note that leave-one-out adjustments as in the case of known treatment probability \( p \) would not generally be unbiased.
B.3 Extension to finite support

Take some distribution over the treatment assignment vector \( d \in \{0,1\}^n \), data \((y(1),y(0)) \in (\mathcal{Y}^2)^n\) as before where \( \mathcal{Y} \subseteq \mathbb{R} \), and \( y = d \circ y(1) + (1-d) \circ y(0) \).

Our goal now is to extend a representation for binary outcomes to one for finite (but arbitrarily large) support \( \mathcal{Y} \).

**Lemma B.2.** Assume that for \( \mathcal{Y} = \{0,1\} \) any \( \hat{\delta} \) with \( E_\theta[\delta(y,d)] = 0 \) for all \( \theta = (y(1),y(0)) \) permits a representation \( \hat{\delta} = \hat{\delta}_\phi \) with

\[
\hat{\delta}_\phi(y,d) = \sum_{i \in \mathcal{I}} w_i(d_{S_i}) \phi_i(y_{-S_i},d_{-S_i})
\]

for fixed \( \mathcal{I}, (w_i)_{i \in \mathcal{I}}, (S_i)_{i \in \mathcal{I}} \) (where \( \mathcal{I} \) finite) and variable \( (\phi_i)_{i \in \mathcal{I}} \) where

\[
\phi_i : (\mathcal{Y} \times \{0,1\})^{\{1,\ldots,n\} \setminus S_i} \to \mathbb{R}.
\]

Then the representation result extends to any finite \( \mathcal{Y} \subseteq \mathbb{R} \) (with the same \( \mathcal{I}, (w_i)_{i \in \mathcal{I}}, (S_i)_{i \in \mathcal{I}} \)).

**Proof.** Write \( \mathcal{Y}_\ell = \{0,1,\ldots,\ell\} \) and define (for \( \ell \geq 2, m \geq 0 \))

\[
\mathcal{Y}_{\ell,m} = \bigtimes_{i=1}^m \mathcal{Y}_{2\ell-1} \times \bigtimes_{i=m+1}^n \mathcal{Y}_{\ell}
\]

We first establish the following intermediate result by induction over \( t = ns + m \) from \( t = 0 \): For any \((s,m) \in (\mathbb{N}_0 \times \{1,\ldots,n\}) \cup \{(0,0)\}\) for \( \ell = 2^s + 1 \) any \( \hat{\delta} \) with \( E_\theta[\hat{\delta}(y,d)] = 0 \) for all \( \theta = (y(1),y(0)) \in \mathcal{Y}_{\ell,m}^2 \) permits a representation \( \hat{\delta} = \hat{\delta}_\phi \) as above with \( \phi_i : \times_{i \in \{1,\ldots,n\} \setminus S_i} (\mathcal{Y}_{\ell,m})_i \to \mathbb{R} \)

For \( t = 0 \), the statement holds by the assumption of the lemma. Assume now that its holds for \( t \) with such \((s,m)\) such that \( t = ns + m \) and \( \ell = 2^s + 1 \), and consider the \((s^+,m^+) \in \mathbb{N}_0 \times \{1,\ldots,n\} \) with \( ns^+ + m^+ = t + 1 \), and write \( \ell^+ = 2^{s^+} + 1 \). Fix an estimator \( \hat{\delta} \) with \( E_\theta[\hat{\delta}(y,d)] = 0 \) for all \( \theta = (y(1),y(0)) \in \mathcal{Y}_{\ell^+,m^+}^2 \).

For \((y,d) \in \mathcal{Y}_{\ell,m} \times \{0,1\}^n\) define \( y_{m^+}^+ = \ell^+ + y_{m^+} - 1, y_{m^+}^- = y_{-m^+} \) as well as \( y_{m^+}^+ = \ell^+, y_{m^+}^- = y_{-m^+} \) to obtain \( y^+, y^- \in \mathcal{Y}_{\ell^+,m^+} \), and define estimators by

\[
\hat{\delta}_1(y,d) = \hat{\delta}(y^+,d) - \hat{\delta}(y^-,d) \quad \quad \hat{\delta}_2(y,d) = \hat{\delta}(y,d)
\]

where thus \( \hat{\delta}_2 \) is merely a restriction of \( \hat{\delta} \) to \( \mathcal{Y}_{\ell,m} \times \{0,1\}^n \). For \((y,d) \in \mathcal{Y}_{\ell^+,m^+} \times \{0,1\}^n\) define \( \bar{y}_{m^+} = \min(y_{m^+},\ell^+)\), \( \bar{y}_{m^+}^- = y_{-m^+} \) and \( \bar{y}_{m^+}^+ = \max(y_{m^+} - \ell^+ + \)
$1, 0), \tilde{y}_{-m^+} = y_{-m^+}$ to obtain $\tilde{y}, \bar{y} \in Y_{\ell, m}^2$ for which

$$\hat{\delta}(y, d) = \hat{\delta}(y, d) - \hat{\delta}(\bar{y}, d) + \hat{\delta}(\tilde{y}, d)$$

$$= \hat{\delta}_1(\tilde{y}, d) + \hat{\delta}_2(\bar{y}, d)$$

$\hat{\delta}_2$ is unbiased (for $Y_{\ell, m}$) by construction. Note that

$$E_\theta[\hat{\delta}_1(y, d)] = E_\theta[\hat{\delta}(y^+, d)] - E_\theta[\hat{\delta}(y^-, d)] = 0$$

for any $\theta = (y(1), y(0)) \in Y_{\ell, m}$, as they generate $(y^+(1), y^+(0)), (y^-(1), y^-(0)) \in Y_{\ell^+, m^+}$ for which $\hat{\delta}$ is unbiased by assumption, so $\hat{\delta}_1$ is likewise unbiased (for $(y(1), y(0)) \in Y_{\ell, m}$). By the induction assumption, there are thus $\phi^1, \phi^2$ with

$$\hat{\delta}(y, d) = \sum_{i \in I} w_i(d_{S_i}) (\phi_i^1(\tilde{y}^-_{S_i}, d_{-S_i}) + \phi_i^2(y^-_{S_i}, d_{-S_i})$$

for any $(y, d) \in Y_{\ell^+, m^+} \times \{0, 1\}^n$. For

$$\phi_i(y^-_{S_i}, d_{-S_i}) = \phi_i^1(\tilde{y}^-_{S_i}, d_{-S_i}) + \phi_i^2(y^-_{S_i}, d_{-S_i})$$

we therefore have $\hat{\delta} = \hat{\delta} \phi$. This concludes the induction step and thus the proof of the intermediate result.

Setting $m = n$, it is immediate that the statement of the lemma holds for all $Y = Y_{2s+1}$. Since it will always hold for subsets, it holds for all $Y = Y_{\ell}$. Now take arbitrary $Y = \{z_1, \ldots, z_{\ell}\}$, and define for $(y, d) \in (Y_{\ell} \times \{0, 1\})^n$

$$\tilde{\delta}(y, d) = \hat{\delta}(z_y, d)$$

where $(z_y)_i = z_{y_i} \in Y$. By the intermediate result there is some $\tilde{\phi}$ such that $\hat{\delta} = \hat{\phi} \tilde{\phi}$. Setting $\phi_i(y^-_{S_i}, d_{-S_i}) = \tilde{\phi}(\tilde{y}^-_{S_i}, d_{-S_i})$ with $\tilde{y}$ such that $z_{\tilde{y}} = y$ yields $\hat{\delta}(y, d) = \hat{\phi}(y, d)$.

We are now ready to prove the representation result in the main paper.

Proof of Lemma 1. The representation for general finite support follows from Lemma B.2 applied to the binary representation results in Theorem B.1 and Theorem B.2, re-

---

*$26$The explicit definition of $\tilde{y}$ corrects an earlier version of this manuscript, which incorrectly used $y - \tilde{y}_{m^+}$ as an argument in place of $\tilde{y}$. 

67
respectively, where the results extend to any fixed bias by taking differences.

C CHARACTERIZATION OF OPTIMAL UNBIASED ESTIMATORS

When is an estimator not just unbiased, but has also low average mean-squared error? Under an additional balance assumption on the prior, there is a particularly simple representation of optimal regression adjustments.

I start with the representation

\[ \hat{\tau}^\phi(y, d) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} (y_i - \phi_i(y_{-i}, d_{-i})) \]

for known treatment probability \( p \) and consider the error

\[ \Delta_\theta^\phi(y, d) = \hat{\tau}^\phi(y, d) - \tau_\theta \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{d_i - p}{p(1-p)} (y_i - \phi_i(y_{-i}, d_{-i})) - (y(1)_i - y(0)_i) \right) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} (\bar{y}_i - \phi_i(y_{-i}, d_{-i})) \]

for the adjustment oracle \( \bar{y}_i = (1-p)y(1)_i + py(0)_i \), which would be the loss-minimizing choice for \( \phi_i(y_{-i}, d_{-i}) \).

**Proposition C.1.** For some prior \( \pi \) over \( \theta = (y(1), y(0)) \) for which Equation 2 holds, any \( \phi^*_\pi \) with

\[ \phi^*_\pi(y_{-i}, d_{-i}) = E_{\pi}[\bar{y}_i|y_{-i}, d_{-i}] \]

is a (global) minimizer of average loss \( E_{\pi}L_\theta(\phi) \), where \( L_\theta(\phi) = E_{\theta}(\Delta_\theta^\phi(y, d))^2 \).

**Proof.** The restriction that adjustments \( \phi_i(y_{-i}, d_{-i}) \) are functions only of \( y_{-i}, d_{-i} \) (and of \( \pi \)) requires some care, as each such adjustments appears given multiple draws of \( (y, d) \). Write

\[ M_i(y^*_{-i}, d^*_{-i}) = \{(y, d) \in (\mathcal{Y} \times \{0, 1\})^n; (y_{-i}, d_{-i}) = (y^*_{-i}, d^*_{-i})\} \]

for the \( (y, d) \) for which \( \hat{\tau}^\phi(y, d) \) (and thus \( \Delta_\theta^\phi(y, d) \)) includes the term \( \phi_i(y^*_{-i}, d^*_{-i}) \).
Then,
\[
\frac{\partial E_\pi L_\theta(\phi)}{\partial \phi_i(y^*_i, d^*_i)} = \frac{\partial E_\pi \left[ 1_{(y,d) \in M(y^*_i, d^*_i)} (\Delta^\phi(y, d))^2 \right]}{\partial \phi_i(y^*_i, d^*_i)}
= E_\pi \left[ 1_{(y,d) \in M(y^*_i, d^*_i)} \frac{\partial (\Delta^\phi(y, d))^2}{\partial \phi_i(y^*_i, d^*_i)} \right],
\]
where we note that we can exchange differentiation and integration because all sum-
mands are bounded. I omit writing $E_\theta$ explicitly inside $E_\pi$ and consider the joint
distribution of $\theta$ and $z$. Here, for all $(y, d) \in M(y^*_i, d^*_i),
\[
\frac{\partial (\Delta^\phi(y, d))^2}{\partial \phi_i(y^*_i, d^*_i)} = -\frac{2}{n^2} \left( \frac{(d_i - p)^2}{p(1-p)} (\bar{y}_i - \phi_i(y^*_i, d^*_i)) + \sum_{j \neq i} (d_i - p)(d_j^* - p) \frac{(y_j - \phi_j(y_{-j}, d_{-j}))}{(p(1-p))^2} \right).
\]
The first-order condition $\frac{\partial E_\pi L_\theta(\phi)}{\partial \phi_i(y^*_i, d^*_i)} = 0$ is therefore
\[
E_\pi \left[ 1_{(y,d) \in M(y^*_i, d^*_i)} (d_i - p)^2 (\phi_i(y^*_i, d^*_i) - \bar{y}_i) \right]
= -\sum_{j \neq i} (d_j^* - p) E_\pi \left[ 1_{(y,d) \in M(y^*_i, d^*_i)} (d_i - p)(\phi_j(y_{-j}, d_{-j}) - \bar{y}_j) \right].
\]
The condition is trivially fulfilled for $P_\pi((y, d) \in M(y^*_i, d^*_i)) = 0$. Otherwise, equivalently
\[
-p(1-p)(\phi_i(y^*_i, d^*_i) - E_\pi[\bar{y}_i|(y_{-i}, d_{-i}) = (y^*_i, d^*_i)])
= -\sum_{j \neq i} (2d_j^* - 1) E_\pi \left[ (d_i - p)\phi_j(y_{-j}, d_{-j})|(y_{-i}, d_{-i}) = (y^*_i, d^*_i) \right] \quad (13)
\]
Note that this system of first-order conditions will generally have many solutions, as
the $\phi$-representation of $\bar{z}^\theta$ is not unique. I now show that the specific choice
\[
\phi_i(y^*_i, d^*_i) = E_\pi[\bar{y}_i|(y_{-i}, d_{-i}) = (y^*_i, d^*_i)]
\]
(for $E_\pi P_\theta((y, d) \in M(y^*_i, d^*_i)) > 0$, otherwise, say, zero) is a (global) posterior-loss
minimizer under the balance assumption from Equation 2. To that end, for \( i \neq j \)

\[
E_\pi [(d_i - p) E_\pi [\bar{y}_{j} | y_{-j}, d_{-j}] | y_{-i}, d_{-i}]
= E_\pi [(d_i - p) E_\pi [\bar{y}_{j} | y_i, d_i, y_{-ij}, d_{-ij}] | y_j, d_j, y_{-ij}, d_{-ij}]
= E_\pi [(d_j - p) E_\pi [\bar{y}_{j} | y_i(d_i), y_{-ij}, d_{-ij}] | y_j(d_j), y_{-ij}, d_{-ij}]
= p(1 - p) = 0,
\]

which is a sufficient condition for the right-hand side of Equation 13 summing to zero.

The first-order condition follows. Also

\[
\frac{\partial^2 E_\pi L_\phi(\phi)}{\partial \phi_i(y_{-i}^A, d_{-i}^A) \partial \phi_j(y_{-j}^B, d_{-j}^B)}
= \frac{1}{(p(1 - p)n^2)} E_\pi \left[ \mathbb{1}_{(y_i, d_i) \in M(y_{-i}^A, d_{-i}^A) \cap M(y_{-j}^B, d_{-j}^B)} (d_i^B - p)(d_j^B - p) \right] \cdot
\]

\[
= \begin{cases} 
\frac{1}{(p(1 - p)n^2)} P_\pi((y_i, d_i) = (y_i^A, d_i^A), (i, y_i^A, d_i^A) = (j, y_j^B, d_j^B)) & \text{for } i \neq j \text{ and } y_{-i/j}^B, x_{-i/j}^A = (y_{-i/j}^B, d_{-i/j}^B) \\
\frac{1}{(p(1-p)n^2)} P_\pi(y^*, d^*) & \text{if } i \neq j \text{ and } y_{ij}^B \text{ and } x_{ij}^A \text{ are defined} \\
0, & \text{otherwise.}
\end{cases}
\]

Note that \( \frac{\partial^2 E_\pi L_\phi(\phi)}{\partial \phi_i(y_{-i}^A, d_{-i}^A) \partial \phi_j(y_{-j}^B, d_{-j}^B)} \) is two times the variance-covariance matrix of the (mean-zero) random variables \( \mathbb{1}_{(y_i, d_i) \in M(y_{-i}^A, d_{-i}^A) \cap M(y_{-j}^B, d_{-j}^B)} \), and therefore everywhere positive semi-definite. It follows that the first-order conditions locate a (global) minimizer.

The proposition directly yields the general characterization result in the main paper.

**Theorem 2** (Choice of the investigator from fixed-shrinkage estimators). For a known treatment probability \( p \), an investigator with risk \( r \in \mathcal{R}^* \) and prior \( \pi \) over \( \Theta \) with

\[
E_\pi [E_\pi [\bar{y}_{j} | y_i(1), z_{-ij}] | z_{-i}] = E_\pi [E_\pi [\bar{y}_{j} | y_i(0), z_{-ij}] | z_{-i}]
\]

for all \( i \neq j \) for given shrinkage \( \lambda \in [0, 1] \) and overall bias \( \alpha \) chooses the Bayes

---

\[\text{Footnote 17}\]

An earlier version of this manuscript incorrectly claimed that the condition in Equation 2 is unnecessary by arguing that always \( E_\pi [(d_i - p) E_\pi [\bar{y}_{j} | y_{-j}, d_{-j}] | y_{-i}, d_{-i}] = 0 \), which is generally false. For a counterexample see [Footnote 17].
estimator

\[ \hat{\tau}(z) = \alpha + (1 - \lambda) \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} (y_i - E_\pi[\bar{y}_i|z_i]) \].

Proof. By Proposition C.1, this choice of adjustment is a global average-loss minimizer for unbiased estimation. For estimators with fixed shrinkage \( \lambda \) and overall bias \( \alpha \), note that every such estimator can be written as

\[ \hat{\tau}(z) = \alpha + (1 - \lambda) \hat{\tau}_{nb}(z), \]

where \( \hat{\tau}_{nb}(z) \) is unbiased. Since also \( \text{Var}_\theta(\hat{\tau}(z)) = (1 - \lambda)^2 \text{Var}_\theta(\hat{\tau}_{nb}(z)) \), where \( 1 - \lambda \neq 0 \) by assumption, we can wlog consider unbiased estimators. \( \square \)

D OLS is Biased

Consider a sample of \( n \) units \((y_i, d_i, x_i)\), where \( d_i \in \{0, 1\} \) are iid given \( x_1, \ldots, x_n \) with \( P(d_i = 1) = p \in (0, 1) \).

D.1 Conditional on covariates

Conditional on covariates \( x_i = 1 \) for all \( i \) and for \( y_i = x_id_i \), the sample-average treatment effect is \( \tau = 1/n \) (one for the first unit, zero for all other units). The coefficient \( \hat{\tau}_{OLS} \) on \( d \) in a linear regression of \( y \) on \( d \) and \( x \) (with intercept) has expectation \( E[\hat{\tau}_{OLS}|n_1] = 0 \) conditional on any number \( 1 < n_1 < n - 1 \) of treated units. Indeed, \( x \) perfectly explains \( y \), so the coefficient on \( d \) will always be zero (by Frisch-Waugh or otherwise).

D.2 Over the sampling distribution

Assume that \( x_i \in \mathbb{R}^{k_n+1} \) with \( P(x_{i0}) = q \in (0, 1) \) and

\[ x_{i1}, \ldots, x_{ik}|x_{i0} \sim (1 - x_{i0}) \cdot \mathcal{N}(0,1) \]

(that is, \( x_{ij} = 0 \) for all \( j > 0 \) if \( x_{i0} = 1 \), \( x_i \) iid across units. (Alternatively, any non-degenerate distribution will do.) Let \( y_i = x_{i0}d_i \). The average treatment effect
of $d_i$ on $y_i$ is
\[
\tau^{\text{pop}} = \mathbb{E}[y_i|d_i = 1] - \mathbb{E}[y_i|d_i = 0] = q.
\]

Let $\hat{\tau}^{\text{OLS}}$ be the coefficient on $d$ in a linear regression of $y$ on $d$ and $x$ (with intercept). For $k_n/n \to \alpha \in (0, 1 - q)$ as $n \to \infty$ we also find
\[
\hat{\tau}^{\text{OLS}} \overset{p}{\to} \frac{q}{1 - \alpha}.
\]

Indeed, writing $A_x$ for the annihilator matrix with respect to $x$ and the intercept, by Frisch-Waugh $\hat{\tau}^{\text{OLS}} = \frac{d'A_x y}{d'A_x d}$ with
\[
\begin{align*}
\mathbb{E}[d'A_x y|x] &= p(1 - p)(n_{x=1} - 1), \\
\mathbb{E}[d'A_x d|x] &= p(1 - p) \text{trace}(A_x) = p(1 - p)(n - k_n - 1).
\end{align*}
\]

By the law of large numbers (where variances are suitably bounded),
\[
\begin{align*}
\frac{d'A_x y}{n} &\overset{p}{\to} p(1 - p) \mathbb{E}[n_{x=1}/n] = p(1 - p)q, \\
\frac{d'A_x d}{n} &\overset{p}{\to} p(1 - p)(1 - \alpha).
\end{align*}
\]

E ASYMPTOTIC INFERENCE

In this section, I derive asymptotically valid inference of the average treatment effect. These results deviate from the approach in the main paper in two notable, related ways. First, I assume that potential outcomes and controls themselves are sampled iid from a population distribution, and inference will not condition on their realizations. Second, in order to obtain valid inference, I take large-sample approximations. The estimator of interest is still unbiased in finite samples for the sample-average treatment effect. But for efficiency and inference I focus on the estimation of the population-average treatment effect in large samples.

Building up to a characterization of the variance of the treatment-effect estimator in terms of out-of-sample prediction quality, I first state an auxiliary remark that will simplify the proof of the main result.

**Remark E.1** ($K$-fold variance bound). Consider $n$ square-integrable, mean-zero random variables $a_1, \ldots, a_n$ and a partition $\bigcup_{k=1}^{K} I_k = \{1, \ldots, n\}$ such that, for all
$k$, $E[a_i a_j] = 0$ for all $i, j \in I_k$. Then,

$$\text{Var} \left( \sum_{i=1}^{n} a_i \right) \leq K \sum_{i=1}^{n} \text{Var}(a_i).$$

**Proof.** By Cauchy-Schwarz, applied once per row, we find that

$$\text{Var} \left( \sum_{i=1}^{n} a_i \right) = \text{Var} \left( \sum_{k=1}^{K} \sum_{i \in I_k} a_i \right) \leq \left( \sum_{k=1}^{K} \sqrt{\text{Var} \left( \sum_{i \in I_k} a_i \right)} \right)^2 \leq K \sum_{k=1}^{K} \text{Var} \left( \sum_{i \in I_k} a_i \right) = K \sum_{k=1}^{K} \sum_{i \in I_k} \text{Var}(a_i),$$

where the last equality follows because increments are uncorrelated within folds.

I assume that potential outcomes and control variables are drawn iid from a population distribution

$$(y_i(1), y_i(0), x_i) \overset{\text{iid}}{\sim} P,$$

treatment is assigned according to a known treatment probability $P(d_i = 1) = p \in (0, 1)$, and data $(y_i, d_i, x_i)$ obtained from $y_i = y_i(d_i)$.

In this section, I focus on $K$-fold estimators similar to those in [Remark 5.1](#). Specifically, I assume that a sample of size $n$ is divided into $K$ equally-sized folds

$$\bigcup_{k=1}^{K} I_k = \{1, \ldots, n\}$$

(so I implicitly assume that $K$ divides $n$). In this setting, I consider the asymptotic distribution of the estimator

$$\hat{\tau} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \frac{d_i - p}{p(1-p)} (y_i - \hat{f}_k(x_i))$$

of the population-average treatment effect $\tau = E[y(1) - y(0)]$, where each $\hat{f}_k : \mathcal{X} \to \mathbb{R}$ is fitted only on folds other than $I_k$. My first result characterizes the asymptotic distribution of $\hat{\tau}$. Throughout, I use indices $i$ and $k$ outside sums for a representative draw from the respective distribution.

**Theorem E.1** (Asymptotic distribution of $K$-fold estimator). Assume that
1. $E[\text{Var}(\hat{f}_k(x_i)|x_i)] \to 0$ as $n \to \infty$,

2. $E \left[ \left( \frac{1-p}{p} \right)^{2d_i-1} (y_i - \hat{f}_k(x_i))^2 \right] \to L$ (where $i \in I_k$), and

3. $E[(\hat{f}_k(x_i) - y_i)^{2+\delta}] < C < \infty$ for some $\delta, C > 0$.

Then,

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, s^2), \quad s^2 = \frac{L}{p(1-p)} - \tau^2.$$

Note that the distribution of prediction functions $\hat{f}_k$ will depend on the sample size of the training sample, and thus on $n$. Furthermore, the result can be extended to the case where the population distribution itself depends on $n$. While I assume that $K$ is fixed here, the conclusion also holds with $K$ growing provided that $K E[\text{Var}(\hat{f}_k(x_i)|x_i)] \to 0$.

The first condition expresses that the prediction variance vanishes and predictions stabilize in large samples. The second condition defines the asymptotic prediction loss of the algorithm. The third condition is a regularity assumption that will ensure asymptotic convergence. When this condition holds, I do not require the assumption of bounded support of potential outcomes from the main paper. Importantly, I do not assume that the prediction functions approximate the best prediction of $y$ given $x$ or are risk-consistent, only that their variance vanishes.

**Proof of Theorem E.1** Write $t_i = \frac{d_i-p}{p(1-p)}$. I decompose

$$\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} (t_i(y_i - \hat{f}_k(x_i)) - \tau)$$

$$= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} (t_i(y_i - \underbrace{E[\hat{f}_k(x_i)|x_i]}_{=g_n(x_i)}) + t_i(E[\hat{f}_k(x_i)|x_i] - \hat{f}_k(x_i)) - \tau)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_i(y_i - g_n(x_i)) - \tau) + \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} t_i(\hat{f}_k(x_i) - g_n(x_i)).$$

For the first part, note that $E[(t_i(y_i - g_n(x_i)) - \tau)^{2+\delta}]$ is bounded, uniformly in $n$. 

74
Its expectation is zero and its variance is

\[ s_n^2 = \text{Var} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_i(y_i - g_n(x_i))) - \tau \right) = \text{Var} \left( t_i(y_i - g_n(x_i)) \right) \]

\[ = E \left[ t_i^2 (y_i - g_n(x_i))^2 \right] - \left( E[t_i(y_i - g_n(x_i))] \right)^2 \]

\[ = (\frac{d_i-p}{p(1-p)})^2 - \frac{1}{p(1-p)} \]

\[ = E \left[ \frac{(1-p)}{p} (y_i - g_n(x_i))^2 \right] - \tau^2. \]

Hence, by the Lyapunov CLT for triangular arrays,

\[ \frac{1}{\sqrt{n} s_n^2} \sum_{i=1}^{n} (t_i(y_i - g_n(x_i)) - \tau) \overset{d}{\longrightarrow} N(0,1). \]

Combining the first two assumptions,

\[ E \left[ \left( \frac{1-p}{p} \right)^{2d_i-1} (y_i - g_n(x_i))^2 \right] \rightarrow L, \]

so we obtain that \( s_n^2 \rightarrow s^2 = L\frac{p(1-p)}{p(1-p)} - \tau^2 \) and thus

\[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_i(y_i - g_n(x_i)) - \tau) \overset{d}{\longrightarrow} N(0, s^2). \]

For the second part, by Remark E.1

\[ \text{Var} \left( \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} t_i(\hat{f}_k(x_i) - g_n(x_i)) \right) \]

\[ \leq \frac{K}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \text{Var} \left( t_i(\hat{f}_k(x_i) - g_n(x_i)) \right) = K E \left[ t_i^2(\hat{f}_k(x_i) - g_n(x_i))^2 \right] \]

\[ = K E \left[ \left( \frac{d_i-p}{p(1-p)} \right)^2 \right] E \left[ (\hat{f}_k(x_i) - g_n(x_i))^2 \right] \]

\[ = \frac{K}{p(1-p)} E \left[ (\hat{f}_k(x_i) - E[\hat{f}_k(x_i)|x_i])^2 \right] = \frac{K}{p(1-p)} E \left[ \text{Var}(\hat{f}_k(x_i)|x_i) \right] \rightarrow 0 \]

75
as \( n \to \infty \). In particular,

\[
\frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} t_i (\hat{f}_k(x_i) - g_n(x_i)) \xrightarrow{p} 0.
\]

The claim of the theorem follows.

The asymptotic variance is a function of the expected prediction loss and the treatment effect, and can be estimated consistently from the sample analogs.

**Remark E.2** (Asymptotically valid variance estimate). Under the assumptions of Theorem E.1, the asymptotic variance of \( \hat{\tau} \) can be estimated consistently by

\[
\hat{s}^2 = \frac{1}{n-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \left( \frac{d_i - p}{p(1-p)} (y_i - \hat{f}_k(x_i)) - \hat{\tau} \right)^2.
\]

As a consequence, we can construct asymptotically valid standard errors and Normal-theory confidence intervals from \( \hat{s}^2 \). To be more precise, \( \hat{s} \sqrt{\frac{1}{n}} \) is a valid standard error for \( \hat{\tau} \), and

\[
[\hat{\tau} - z_{1-\alpha/2} \hat{s} \sqrt{\frac{1}{n}}, \hat{\tau} + z_{1-\alpha/2} \hat{s} \sqrt{\frac{1}{n}}]
\]

a \( 1 - \alpha \) confidence interval for \( \tau \) (where \( z_{1-\alpha/2} \) is the \( 1 - \alpha/2 \)-quantile of the standard Normal distribution).

The asymptotic results extend to the case of fixed \( n_1 \) (by setting \( p = n_1/n \), provided that \( \mathbb{E} [\hat{f}_k(x_i)] \to \mathbb{E} [\bar{y}_i] \)), exact cross-fitting as in Remark 5.1 with balanced folds, and folds that are only approximately of the same size or only approximately balanced.

Now that we have established asymptotically valid inference, I am ready to return to preference alignment.

**Remark E.3** (Alignment over precision). Assume the investigator chooses among unbiased estimators, that is, by Lemma 1 among regression adjustments. Assume further that she constructs regression adjustments in a \( K \)-fold procedure with (a sequence of) prediction functions that fulfill the regularity assumptions for asymptotically valid inference in Theorem E.1. Then, if the investigator wants to obtain small standard errors or tight confidence intervals, her choices are aligned with the
designer’s preference for low mean-squared error $E[(\hat{\tau} - \tau)^2]$ among these unbiased estimators.

Proof. The asymptotic distribution of $\hat{\tau}$ as well as the probability limit of $s^2$ only depend on the asymptotic loss $L$, the treatment probability $p$, and the treatment effect $\tau$. The investigator through her choice of adjustments can only control $L$, and for these preferences chooses a sequence of prediction functions that minimizes asymptotic prediction loss. This is also the variance-minimizing choice the designer prefers. (Since $L$ is non-random, the specific utility function over the size of standard errors or confidence intervals does not matter here.)

Note that unbiasedness is crucial to reduce the degrees of freedom over the asymptotic distribution to the variance, with respect to which designer and investigator are aligned. Conversely, designer and investigator may have different preferences over the bias-variance trade-off, so allowing for (asymptotic) bias would break alignment even when the estimator is asymptotically Normal.

By the same argument as in the proof of Remark E.3, choices are also aligned over the power of a test against some null hypothesis. Since the investigator cannot move the expectation of the estimator, the best she can do is to pick a sequence of prediction functions for which the asymptotic loss $L$ is minimal.

Remark E.4 (Alignment over power). Consider a sequence of population distributions with $\tau_n = \tau_0 + \frac{\delta}{\sqrt{n}}$. Assume that the investigator constructs a one- or two-sided test against the null hypothesis $\tau = \tau_0$ by comparing the test statistic $\sqrt{n}(\hat{\tau} - \tau_0)$ to the standard Normal distribution, and that the investigator’s (sequence of) prediction functions fulfill the regularity assumptions in Theorem E.1. If the investigator has a preference for rejecting $\tau = \tau_0$, then her choices are aligned with the designer’s goal of minimizing $E[(\hat{\tau} - \tau)^2]$.

Based on the asymptotic approximation from Theorem E.1 I am now ready to prove the result from the main paper that distribution to two researchers attains asymptotic efficiency.

Remark 6.3 (Semi-parametric efficiency). If researchers use prediction algorithms $(A_n : Z \to \mathbb{R}^X, z \mapsto \hat{f}_n)_{n=1}^\infty$ with

$$E[(\hat{f}_n(x_i) - E[\hat{y}_i|x_i])^2] \to 0$$

77
as \( n \to \infty \), then delegation to two researchers with risk functions in \( \mathcal{R}^* \) (who each obtain access to half of the data, say) without further commitment achieves both finite-sample unbiased estimation of \( \tau_\theta \), and large-sample semi-parametric efficient estimation of \( \tau \) for the semi-parametric efficiency bound of Hahn (1998).

**Proof of Remark 6.3** Similar to the proof of Theorem E.1, again setting \( t_i = \frac{d_i - p}{p(1 - p)} \), I decompose, with \( K = 2 \),

\[
\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} (t_i(y_i - \hat{f}_k(x_i)) - \tau)
\]

\[
= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} (t_i(y_i - E[\bar{y}_i|x_i]) + t_i(E[\bar{y}_i|x_i] - \hat{f}_k(x_i)) - \tau)
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_i(y_i - E[\bar{y}_i|x_i])) - \tau) + \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} t_i(\hat{f}_k(x_i) - E[\bar{y}_i|x_i])
\]

The latter part converges to zero in probability by Remark E.1 as in the proof of Theorem E.1. Since the support of potential outcomes is bounded, the first part converges by the standard CLT to a mean-zero Normal distribution with asymptotic variance

\[
\text{Var}(t_i(y_i - E[\bar{y}_i|x_i])) = \frac{E\text{Var}(y_i(1)|x_i)}{p} + \frac{E\text{Var}(y_i(0)|x_i)}{1 - p} + \text{Var}(E[y_i(1) - y_i(0)|x_i])
\]

which is the efficiency bound of Hahn (1998).

**F Hyperpriors and Optimal Biases**

The minimax result in the main paper establishes that fixing the bias is a minimax optimal solution to the designer’s delegation problem that aligns the choices of the investigator with the goal of the designer. An optimal choice of biases, however, depends on the hyperprior of the designer, and zero as a choice is not an optimal solution in general.

In this section, I discuss one justifications for why I put special emphasis on unbiasedness coming from a specific notion of an uninformed designer. I then highlight that some hyperpriors, even on finite support, deliver zero bias as an exact solution. Finally, I discuss how the characterization of unbiased estimators extends to the case of other choices of the bias.
F.1 Uninformativeness and Zero Bias

Intuitively, a designer who has no systematic information about the location of the average treatment effect will set the biases to zero. If we assumed that the support of the outcome variables was continuous and unbounded, then one elegant formalization of this argument would capture uninformativeness about the location of the treatment effect by an invariance to translation actions in that direction, yielding an improper hyperprior that would deliver zero bias under an appropriate criterion. Since my paper is, however, formulated for finite support, and since dealing with improper priors would bring with it additional technical complications, I propose here one way of obtaining (approximately) zero bias under a specific notion of (approximate) uninformativeness in order to highlight the connection between invariances and bias.

In order to illustrate one construction of an approximately uninformative hyperprior, I start with an arbitrary hyperprior \( \eta \) over priors with (full) support in the grid

\[
\Theta_0 = \mathcal{Y}_k^{2n}
\]

and \( \mathcal{Y}_k = \{-k, -k + 1, \ldots, -1, 0, 1, \ldots, k - 1, k\} \) for some \( k \in \mathbb{N} \). (I am choosing an equally-spaced grid for convenience.) From \( \eta \) I construct increasingly uninformative hyperpriors \( \eta^m \) over priors with support \( \Theta_m = \mathcal{Y}_{k+m}^{2n} \) for all \( m \in \mathbb{N}_0 \).

In order to construct the hyperprior \( \eta^m \) for \( m \geq 0 \), consider

\[
g = (r, t) \in \{-1, 1\}^{2n} \times \mathcal{Y}_m^{2n} = G_m
\]

and define the action of \( G_m \) on \( \mathbb{Z}^{2n} \) by \( g \circ \theta = (r_i \cdot \theta_i + t_i) \). (Note that \( g \) maps \( \Theta_0 \) to \( \Theta_m \).) The distribution \( \eta \) over priors \( \pi \) on \( \Theta_0 \) implies a distribution \( g \circ \eta \) over priors \( g \circ \pi \) (defined by \( (g \circ \pi)(g \circ \theta) = \pi(\theta) \)) on \( g \circ \Theta_0 \subseteq \Theta_m \) that extends to a distribution on \( \Theta_m \). The distribution \( \eta^m \) over priors with support in \( \Theta_m \) is then given by the composition of Uniform\((G_m)\) and \( \eta^m \) that first draws a random action \( \tilde{g} \) and then independently draws a prior over \( \Theta_m \) according to \( \tilde{g} \circ \eta^m \).

This construction yields hyperpriors that are increasingly uninformative about the location of outcomes in that they exhibit more and more symmetries with respect to reflection and translation of the data. Writing \( (\beta^m_\theta)_{\theta \in \Theta_m} \) for the biases chosen optimally by the designer according to the hyperprior \( \eta^m \) constructed in this way, any
bias $\beta^m_\theta$ will therefore approach zero as the support grows.

**Remark F.1** (Approximate unbiasedness). For any fixed $\theta \in \mathbb{Z}^{2n}$ (with $m_0$ large enough such that $\theta \in \Theta_m$), $\lim_{m \to \infty, m \geq m_0} \beta^m_\theta = 0$.

Note that the priors drawn from $\eta^m$ do not have full support, which could be rectified by taking appropriate approximations. Note also that in this example I am using invariances in the outcomes, and not just in the treatment effects, so a smaller class of invariances may suffice to obtain a similar result.

A similar approach would start with an (improper) invariant hyperprior over priors with support in $\mathbb{Z}^{2n}$ and then consider restrictions of that distribution to an increasing sequence of finite support sets, showing similarly that as the support grows and the hyperprior approaches the uninformative hyperprior, the optimal biases shrink to zero.

**Proof idea.** By symmetry, the optimal bias at the origin $0$ within any support $\Theta_m$ subject to the hyperprior $\eta^m$ is zero, $\beta^m_0 = 0$. Similarly, at fixed $\theta \in \Theta_{m_0}$ and for $m \geq m_0 + k$, the optimal bias would be zero if we conditioned $\eta^m$ on

$$\max_{i \in \{1, \ldots, 2n\}} |\theta_i - \tilde{t}| \leq (m - m_0)$$

in $\tilde{g} = (\tilde{r}, \tilde{t})$, since this would make $\theta$ the center of symmetry. Since

$$P_{\eta^m} \left( \max_{i \in \{1, \ldots, 2n\}} |\theta_i - \tilde{t}| \leq (m - m_0) \right) \to 1$$

as $m \to \infty$, the argument extends to the unconditional optimization.

**F.2 Zero Bias as a Minimax Solution**

In my main setup, the designer optimizes against a worst-case risk, and averages over a (hyper-)prior over the investigator’s prior information. One approach of fixing the bias would replace the hyperprior with assuming a worst-case prior. However, without restrictions on the priors and for a fixed, finite support, such a minimax solution would be driven by priors that put full weight on extreme outcome values, which is econometrically unappealing.

Rather, I propose a minimax approach to fixing the biases that includes uncertainty about the location of the outcomes. For generality, I formulate this result
on the level of uncertainty about hyperpriors, and then return to implications for uncertainty over priors. I follow the construction and nomenclature from the above discussion of uninformativeness.

Specifically, I start with a set $H$ (which can be a singleton) of hyperpriors, where for each $\eta \in H$ the priors in the support of $\eta$ have as support the grid $Y_2^n$ for

$$Y_2 = \{-k, -k + 1, \ldots, -1, 0, 1, \ldots, k - 1, k\}.$$  

Consider

$$g = (r, t) \in \{-1, 1\}^{2n} \times \mathbb{Z}^{2n} = G$$  

and, similar to the above, define the action of $G$ on $\mathbb{Z}^{2n}$ by $g \circ \theta = (r_i \cdot \theta_i + t_i)$. The distribution $\eta$ over priors $\pi$ on $Y_2^n$ implies a distribution $g \circ \eta$ over priors $g \circ \pi$ (defined by $(g \circ \pi)(g \circ \theta) = \pi(\theta)$) on $g \circ Y_2^n \subseteq \mathbb{Z}^{2n}$. From that, I obtain the set of hyperpriors

$$H^* = G \circ H = \{g \circ \eta; g \in G, \eta \in H\}.$$  

Following the logic of the proof of Remark F.1, the invariances of $H^*$ imply that an investigator who optimizes against a worst-case hyperprior in $H^*$ chooses zero bias as a minimax (in risk and hyperprior) optimal restriction of this form:

**Remark F.2** (Minimax optimality of zero bias). The unbiased estimators are minimax optimal for the invariant set $H^*$ of hyperpriors in the sense that the choice $\beta_\theta = 0$ for all $\theta \in \mathbb{Z}^{2n}$ minimizes (among fixed-bias restrictions)

$$\sup_{\eta \in H^*} \sup_{r^I \in \mathcal{R}^*} \mathbb{E}_\eta \left[ r^D_\theta \left( \arg \min_{\hat{\tau}} \mathbb{E}_\pi \left[ r^I_{\theta}(\hat{\tau}) \right] \right) \right].$$

As a special case, this result includes the case where all hyperpriors are singletons, and the designer thus optimizes against a worst-case prior directly. Hence, for any set of priors $\Pi$ with support $Y_2^n$ and

$$\Pi^* = G \circ \Pi = \{g \circ \pi; g \in G, \pi \in \Pi\},$$
the minimax result yields minimization of

\[
\sup_{\pi \in \Pi^*} \sup_{r^I \in \mathbb{R}^*} \mathbb{E}_\pi \left[ r^D_\theta \left( \arg \min_{\hat{\tau} \in \mathcal{C}} r^I_{\theta} (\hat{\tau}) \right) \right].
\]

Note that the priors in \( \Pi^* \) now have varying support.

\section*{F.3 Hyperpriors with Exactly Zero Bias}

There are hyperpriors that trivially yield zero bias, namely those that by virtue of \( \text{Var}_\pi(\bar{y}_i | z_{-i}) = 0 \) (for known \( p \)) allow the investigator to pick an unbiased estimator with zero loss (such as, in the case of \( p = .5 \), if the investigator knows \( y_i(1) + y_i(0) \) from \( x_i \)). While these constitute extreme examples, they point towards a general intuition: if the investigator has strong private information about the choice of optimal adjustments, then setting a non-zero bias would create a burden by imposing loss that cannot be avoided.

\section*{F.4 Investigator Solution when Some Bias is Optimal}

If the designer has non-trivial information about the distribution of treatment effects, the remaining results in the paper formulated in terms of unbiased estimation extend at least partly.

First, assume that the hyperprior of the designer implies (approximately) that \( \mathbb{E}_\theta[\hat{\tau}(z)] = (1 - \lambda)\tau_\theta \) is an optimal restriction (for \( \lambda \in (0, 1) \)), expressing a fixed shrinkage factor that is set by the designer. Then, the investigator still faces the same unbiased estimation problem as in the main paper since the optimal shrunk estimator is the optimal unbiased estimator multiplied by that factor ex-post.

Second, even if no such structure is available, the results still extend with modifications. Note that the designer’s solution can equivalently be phrased as choosing a reference estimator \( \hat{\tau}^D \) (with the desired biases) and letting the investigator choose mean-zero adjustments \( \hat{\delta}^I \) to obtain an estimator

\[
\hat{\tau}(z) = \hat{\tau}^D(z) + \hat{\delta}^I(z).
\]

My characterization of unbiased regression adjustments directly yields a characterization of mean-zero adjustments that characterize the choice set of the investigator, only that the reference estimator has now changed. However, the optimal adjustments now take a different form. For example, in the case of \( n = 1 \) with known \( p \),
the optimal adjustment now takes the general form

$$\phi_i = p(1-p)E_{\pi}[\hat{\tau}^D(y = y(1), d = 1) - \hat{\tau}^D(y = y(0), d = 0)],$$

which precisely yields the familiar adjustment $E_{\pi}[(1 - p)y(1) + py(0)]$ when applied to the unbiased reference estimator $\hat{\tau}^D(z) = \frac{d-p}{p(1-p)}y$.

**G Additional Proofs**

In this section, I restate and sketch the proofs of the remaining results, which largely follow from the main results proved earlier.

**Theorem 3** (Complete-class theorem for unbiased estimators). For any unbiased estimator $\hat{\tau}$ of the sample-average treatment effect that is not dominated with respect to variance, there is a converging sequence of priors $\pi_t$ with full support such that $\hat{\tau}$ equals the limit of the respective optimal Bayes estimators. Conversely, for any converging sequence of priors $\pi_t$ that put positive weight on every state $\theta \in \Theta$, every converging subsequence of corresponding Bayes estimators is admissible among unbiased estimators.

**Proof.** Note first that, for $\pi$ with full support, the estimator that minimizes average variance among unbiased estimators is unique (even though the representation in Theorem 2 in general is not). Indeed, among unbiased estimators the investigator minimizes (conflating the distribution of $\theta$)

$$E_{\pi}[(\hat{\tau}(z) - \tau_\theta)^2] = E_{\pi}[(\hat{\tau}(z) - E_{\pi}[\tau_\theta|z])^2] + E_{\pi}[(E_{\pi}[\tau_\theta|z] - \tau_\theta)^2].$$

Hence, within the affine linear subspace of $\mathbb{R}^Z$ given by the unbiased estimators, the investigator chooses the point $\hat{\tau}$ closest to $(z \mapsto E_{\pi}[\tau_\theta|z])_{z \in Z}$ according to the weighted (with positive weights) Euclidean distance

$$d(\hat{\tau}_1, \hat{\tau}_2) = E_{\pi}[(\hat{\tau}_2(z) - \hat{\tau}_1(z))^2]$$

(with the distribution over $Z$ implied by $\pi$ through draws of $\theta$). Hence, the investigator’s solution is unique when $\pi$ has full support.

Since every limiting estimator is the limit of Bayes estimators with full support, with finite domain and bounded codomain, any such limiting estimator is admissible. Since the state space is finite, every admissible estimator is Bayes (e.g. Ferguson).
If the estimator is Bayes with respect to a prior with full support, it is unique and therefore has an adjustment representation of the claimed form. If the corresponding prior does not have full support, we can write it as a limit of admissible estimators that are Bayes with respect to priors with full support and thus unique, so the estimator is a limit of estimators of the claimed form.

Corollary 1 (Characterization of fixed-bias $K$-fold distribution contracts). For $K$ disjoint folds $I_k \subseteq \{1, \ldots, n\}$ with projections $g_k : (y,d) = z \mapsto z-I_k = (y_i, d_i)_{i \neq I_k}$, a $K$-distribution contract $\hat{\tau}^\Phi$ has given bias if and only if:

1. For a known treatment probability $p$, there exist a fixed estimator $\hat{\tau}_0(z)$ with the given bias and regression adjustment mappings $(\Phi_k^k)_{k=1}^K$ such that

$$\hat{\tau}^\Phi((\hat{\phi}_i^k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n} \sum_{k=1}^K \sum_{i \in I_k} \frac{d_i - p}{p(1-p)} \phi_i^k(z-i)$$

where $(\phi_i^k)_{i \in I_k} = \Phi_k(\hat{\phi}_i(z-I_k))$.

2. For a fixed number $n_1$ of treated units, there exist a fixed estimator $\hat{\tau}_0(z)$ with the given bias and regression adjustment mappings $(\Phi_k^k)_{k=1}^K$ such that

$$\hat{\tau}^\Phi((\hat{\phi}_i^k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n_1 n_0} \sum_{k=1}^K \sum_{i<j \{i,j\} \subseteq I_k} (d_i - d_j) \phi_{ij}^k(z-ij),$$

where $(\phi_i^k)_{i \in I_k} = \Phi_k(\hat{\phi}_i(z-I_k))$.

Proof. The result is a special case of Lemma 6.1 for this specific choice of the functions $g_k$.

Remark 5.1 (Exact $K$-fold cross-fitting). For a partition of the sample

$$\{1, \ldots, n\} = \bigcup_{k=1}^K I^{(k)}$$

into $K$ folds with $n^{(k)} \geq 2$ units each of which $n_1^{(k)} > 0$ treated and $n_0^{(k)} > 0$ untreated, the estimator

$$\hat{\tau}(z) = \frac{1}{n} \sum_{k=1}^K n^{(k)} \sum_{i \in I^{(k)}} \frac{d_i n^{(k)} - n_1^{(k)}}{n_1^{(k)} n_0^{(k)}} \left( y_i - \phi_i^{(k)}(z-I^{(k)}) \right)$$

84
is unbiased for the sample-average treatment effect \( \tau \) conditional on \((\mathcal{I}^{(k)})_{k=1}^{K}\) and \((n^{(k)})_{k=1}^{K}\) under either randomization.

**Proof.** Unbiasedness is immediate from Lemma 1. Optimality of this choice of adjustments follows as in the proof of Theorem 2. \(\square\)

**Lemma 6.1** (Characterization of unbiased \(K\)-distribution contracts). A \(K\)-distribution contract \(\hat{\tau}^\Phi\) is unbiased for the sample-average treatment effect \(\tau_\theta\) for any conformable researcher input \((\hat{\phi}_k)_{k=1}^{K}\) if and only if:

1. For known treatment probability \(p\), there exist regression adjustments \((\phi_i: (X_{k\in C_i} B_k) \times (\mathcal{Y} \times \{0, 1\})^{n-1} \to \mathbb{R})_{i=1}^{n}\) such that

   \[
   \hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^{K}; z) = \frac{1}{n} \sum_{i=1}^{n} d_i \frac{1 - p}{p(1 - p)} (y_i - \phi_i((\hat{\phi}_k(g_k(z)))_{k\in C_i}; z_{-i}))
   \]

   for \(C_i = \{k; g_k(z) = \tilde{g}(z_{-i})\text{ for some } \tilde{g}\}\).

2. For fixed number \(n_1\) of treated units, there exist regression adjustments \((\phi_{ij}: (X_{k\in C_{ij}} B_k) \times (\mathcal{Y} \times \{0, 1\})^{n-2} \to \mathbb{R})_{i<j}^{n}\) such that

   \[
   \hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^{K}; z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}((\hat{\phi}_k(g_k(z)))_{k\in C_{ij}}; z_{-ij})),
   \]

   for \(C_{ij} = \{k; g_k(z) = \tilde{g}(z_{-ij})\text{ for some } \tilde{g}\}\).

**Proof.** Since the resulting estimator must be unbiased, and researcher choices are themselves functions of the data made available to them, the result follows directly from the general representation result of unbiased estimators (Lemma 1). \(\square\)

**Theorem 6.1** (Hybrid pre-analysis plan dominates rigid pre-analysis plan). Assume that investigator and researcher have risk functions in \(\mathcal{R}^*\). The optimal unbiased pre-committed estimator \(\hat{\tau}^{\text{pre}}\) is strictly dominated by an unbiased hybrid pre-analysis plan with respect to average variance, i.e. the hybrid plan is as least as precise on average over any ex-ante prior \(\eta^I\) and strictly better for many non-trivial ex-ante priors \(\eta^I\).

**Proof.** A researcher with risk function in \(\mathcal{R}^*\) minimizes variance among unbiased estimators. Since the original adjustments corresponding to \(\hat{\tau}^{\text{pre}}\) are available to the researcher, her choice can only reduce variance on average over her prior. Unless the
ex-post changes are ineffectual, this will strictly improve variance averaged over the hyperprior.

Remark 6.1 (Optimal hybrid pre-analysis plan). The dominating hybrid plan is:

1. For known treatment probability $p$, the researcher chooses regression adjustments $(\phi_{i}^{\text{post}} : (\mathcal{Y} \times \{0, 1\})^{n-1} \rightarrow \mathbb{R})_{i \notin T} = \hat{\phi}(z_{T})$ to obtain

$$
\hat{\tau}_{\text{hybrid}}^{\phi}(\phi; z) = \hat{\tau}_{\text{pre}}(z) - \frac{1}{n} \sum_{i \notin T} \frac{d_i - p}{p(1 - p)} \phi_{i}^{\text{post}}(z_{-i})
$$

where $1 \leq |T| \leq n - 1$.

2. For fixed number $n_1$ of treated units, the researcher chooses adjustments $(\phi_{ij}^{\text{post}} : (\mathcal{Y} \times \{0, 1\})^{n-2} \rightarrow \mathbb{R})_{\{i < j\}\cap T = \emptyset} = \hat{\phi}(z_{T})$ to obtain

$$
\hat{\tau}_{\text{hybrid}}^{\phi}(\phi; z) = \hat{\tau}_{\text{pre}}(z) - \frac{1}{n_1 n_0} \sum_{\{i < j\}\cap T = \emptyset} (d_i - d_j) \phi_{ij}^{\text{post}}(z_{-ij})
$$

where $1 \leq |T| \leq n - 2$.

In both cases, the investigator commits to the training sample $T \subseteq \{1, \ldots, n\}$ and the unbiased estimator $\hat{\tau}_{\text{pre}} : Z \rightarrow \mathbb{R}$.

Proof. This is a special case of Corollary 1.

Remark 6.2 (More researchers are better). Assume that the investigator and researchers all have risk functions in $\mathcal{R}^{*}$, and that the researchers all share the same (ex-post) prior $\pi$. Then an optimal unbiased $K$-distribution contract is dominated by an unbiased $K + 1$-distribution contract in the sense of Theorem 6.1.

Proof. The result follows by the revealed-preference argument in the proof of Theorem 6.1.

Note that to obtain this result I assume that all researchers have the same prior, which renders the proof trivial, but represents an unrealistic assumption. A more attractive result would assume that the $K$ researchers each obtain a draw from the same hyperprior $\eta$ (where draws are correlated between each other and with the true distribution, also drawn from $\eta$, of $\theta$), and I conjecture that in this case more researchers still improve average estimation quality.