Optimal Estimation when Researcher and Social Preferences are Misaligned

Jann Spiess

February 2024

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 as a requirement on the estimation of the average treatment effect can align 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 treatment-effect estimators with fixed bias as sample-splitting procedures. Third, I discuss the implementation of second-best estimators that leave room for beneficial specification searches.

Keywords: program evaluation, randomized experiments, regression adjustments, statistical decision theory, delegation, machine learning, pre-analysis plans.

Stanford University, jspiess@stanford.edu. For their guidance, I am indebted to Sendhil Mullainathan, Alberto Abadie, Elie Tamer, and Gary Chamberlain. For valuable comments and discussions, I also thank Laura Blattner, Avi Feller, Ed Glaeser, Nathan Hendren, Simon Jäger, Hiro Kaido, Maximilian Kasy, Larry Katz, Scott Kominers, Eben Lazarus, Shengwu Li, Mikkel Plagborg-Møller, Ashesh Rambachan, Jon Roth, Liz Santorella, Klaus M. Schmidt, 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, Toronto, the AEA meetings, and Oxford, as well as the editor and three anonymous referees.

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 choice of estimator may also reflect a preference for specific estimates or an incentive to get published 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 illustrates the properties of flexible estimators that leverage the data and researcher expertise, without reflecting 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, in Section 1, I set up a stylized game between a designer and an investigator who are engaged in point estimation of an average treatment effect. The designer aims 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 an estimate close to a preferred value other than the truth. The investigator picks an estimator based on her private information about the specific experiment. The designer chooses optimal constraints on the estimation.[1]

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 precisely, in Section 2, I show that fixing the bias aligns the incentives of the investigator and the designer and is a minimax optimal solution to the designer's problem in my stylized setting with a specific class of preferences. I also argue that it approximately aligns choices in practically relevant settings such as wanting to obtain a significant test or a particularly large estimate. Allowing the investigator to choose the bias can, in principle, improve overall precision through a reduction in the variance. However, an investigator could use her control over the bias to reflect her preferences rather

[1]The designer could represent professional norms, 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.

than her private information. Having argued that the designer should therefore fix the bias to minimize mean-squared error, I show that optimal biases are generally not too large and that unbiased estimation is justified by a minimax argument. Among such unbiased estimators, even an investigator who wants to obtain an estimate close to the same large value will still choose an estimator that minimizes the variance.

Second, having motivated a bias restriction, in Section 3, I establish 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 splitting 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. Specifically, unbiased estimators can differ from a difference in averages only by leave-one-out or leave-two-out regression adjustments of individual outcomes.[2]

Third, in Section 4, I show how incentive-aligned estimators can be implemented in a sampling model. This sampling model explicitly considers the role of covariates in achieving precise estimates of the average treatment effect. Focusing on the practically relevant case of unbiased estimation, I relate the investigator's solution to a prediction problem. By the sample-splitting representation, I can write every unbiased estimator of the average treatment effect in terms of a set of regression adjustments. Focussing on a computationally tractable implementation using a limited number of sample splits, I argue that regression adjustments that minimize prediction risk for a specific loss function represent a feasible strategy by the investigator that ensures a minimax optimal estimation outcome. My results motivate and describe flexible yet robust pre-analysis plans for the analysis of experimental data that allow for specification searches either by a pre-specified algorithm or by giving the investigator successive access to the data. Through sample splitting, this flexibility comes without adding

[2]In particular, for known treatment probability, all unbiased estimators of the sample-average treatment effect take the form of the "leave-one-out potential outcomes" (LOOP, by Wu and Gagnon-Bartsch, 2018), which is a special case of Aronow and Middleton's (2013) extension of the Horvitz and Thompson (1952) estimator.

any bias to the resulting estimator.

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) shows its relevance in achieving valid and efficient inference in high-dimensional observational data. Schorfheide and Wolpin (2012, 2016) provides a justification for 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 on regression adjustments in 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 article, Wu and Gagnon-Bartsch (2018) proposes the "leave-one-out potential outcomes" (LOOP) estimator that yields regression adjustments without bias, which takes the same form the estimator chosen by the researcher in my setting for specific choices of the bias. Wager et al. (2016) proposes 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. Bloniarz et al. (2016) uses the LASSO to select among control variables in experiments. Balzer et al. (2016) proposes a data-adaptive procedure that selects among specifications to minimize the variance of treatment-effect estimators. Relative to this literature, I motivate a bias restriction and fully characterize estimators with a given bias.

A 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. (Brodeur et al., 2016) documents anomalies in published p-values, and (Andrews and Kasy, 2019) provide empirical evidence for publication biases. Christensen and Miguel (2018) surveys evidence and discusses practices that aim to improve transparency and reproducibility. Young (2019) documents the sensitivity of treatment-effect estimates in experiments to the choice of specification.

Finally, I contribute to a literature that explicitly models preferences over estimators (e.g. Glaeser, 2006). Like Di Tillio et al. (2017), I model a researcher analyzing an

experiment within a game with misaligned preferences. Like Schorfheide and Wolpin (2012, 2016), I obtain a formal justification for sample splitting from a principal-agent model. Banerjee et al. (2020, 2017) also approach experimental design and analysis as a decision-theoretic problem. Andrews and Shapiro (2021) embeds estimation in a decision-theoretic model to motivate simple, otherwise inadmissible estimators as the optimal solution to a communication problem. Recently, work across computer science and statistics has been adapting similar delegation models to address conflicts of interest in data-driven decision-making (such as Bates et al., 2022, 2023).

1 Setup

This article approaches 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.

1.1 Target Parameter

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

$$\tau_\theta = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(y_i(1) - y_i(0))}_{=\tau_i} \qquad\qquad \theta = (y_i(1), y_i(0))_{i=1}^n$$

in a given sample of n units. In the Rubin (1974) causal model interpretation, $y_i(d_i) \in \mathcal{Y}$ is the potential outcome of unit i had they received treatment status $d_i \in \{0, 1\}$, and τ_i the respective causal effect. Here, I write $\mathcal{Y} \subseteq \mathbb{R}$ for the support of the outcomes. We may sometimes be interested in the treatment effect averaged over a population distribution of potential outcomes. I consider such extensions to a sampling-based framework in Section 4.

1.2 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 $y_i = y_i(d_i) \in \mathcal{Y}$. But because I assume that the distribution of treatment assignment $(d_i)_{i=1}^n \in \{0,1\}^n$ does not vary with the potential outcomes $\theta = (y_i(1), y_i(0))_{i=1}^n$, 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 $z = (y_i, d_i)_{i=1}^n$ is distributed according to P_θ as follows. $(d_i)_{i=1}^n$ is generated from a known distribution over $\{0,1\}^n$ that does not depend on θ and is one of:*

1. *Each unit i is independently assigned to treatment with known probability $p = \mathrm{P}(d_i = 1)$ (where $0 < p < 1$).*
2. *$(d_i)_{i=1}^n$ 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 the d_i, $y_i = y_i(d_i)$ for all $i \in \{1, \dots, n\}$.

In this notation, I do not explicitly include the covariates $x_1, \dots, x_n$ in the data z, since I condition on the covariates and therefore treat them as constant and not as a random variable. In Section 4, I explicitly consider the case where units themselves are sampled and we observe such pre-treatment covariates. While the distribution of treatment is assumed not to depend on the identity or covariates of units in the experiment, my results extend to distributions with known (non-degenerate) propensity scores, including stratified or conditional random sampling.

1.3 Covariate Adjustments

How can we estimate the sample-average treatment effect τ_θ from the data? 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 τ_θ conditional on the number n_1 of treated units (provided $0 < n_1 < n$). However, $\hat{\tau}^*$ leaves information about individual units on the table, such as information encoded in covariates.

In econometric practice, τ_θ is therefore often estimated from a linear regression of the outcome on treatment and controls. However, 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; Lin, 2013). Second, if the investigator does not document that he picked among multiple covariate adjustments, an unsuspecting observer's inference may be biased towards stronger treatment effects and unjustified confidence (Lenz and Sahn, 2017).

1.4 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 $\hat{\tau} : \mathcal{Z} \to \mathbb{R}$ that maps experimental data $z = (y_i, d_i)_{i=1}^n \in (\mathcal{Y} \times \{0,1\})^n = \mathcal{Z}$ into an estimate $\hat{\tau}(z)$ of the sample-average treatment effect τ_θ. Since my analysis is conditional on the control covariates, this estimator encodes in particular how the estimate of the treatment effect is adjusted for realizations of control variables.

Designer and investigator preferences are expressed by risk functions $r^D, r^I : \Theta \times \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}$ that encode the expected loss $r_\theta^D(\hat{\tau}), r_\theta^I(\hat{\tau})$ of an estimator $\hat{\tau} \in \mathbb{R}^{\mathcal{Z}}$ given the full potential outcomes $\theta = (y_i(1), y_i(0))_{i=1}^n \in \mathcal{Y}^{2n} = \Theta$. Designer and investigator aim to minimize their respective risk given the potential outcomes θ. Throughout, I assume that the designer's risk function expresses a desire to obtain precise estimates of the true treatment effect τ_θ.

Assumption 2 (Designer's risk function)**.** *The designer's risk for an estimator $\hat{\tau} : \mathcal{Z} \to \mathbb{R}$ is the estimator's mean-squared error $r_\theta^D(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tau_\theta)^2]$, which averages over treatment assignment given potential outcomes $\theta \in \Theta$.*

Notably, I do not assume that the designer has an inherent preference for unbiased estimators.[3] 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 $r_\theta^I(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tilde{\tau})^2]$, which

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

expresses a desire to obtain a certain estimate $\tilde{\tau}$ irrespective of the true treatment effect τ_θ. The designer knows only that $r^I \in \mathcal{R}$ for some set of risk functions.

1.5 Prior Information

Since generally no single estimator $\hat{\tau}$ minimizes risk for all potential outcomes $\theta \in \Theta$ and θ 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 π over potential outcomes governs this tradeoff.[4]

The investigator receives the prior distribution π over potential outcomes as a private signal before the data is realized. This private information models researcher expertise. For example, the investigator may have run previous studies or a pilot and synthesized relevant results in the literature. The investigator therefore has a sense of which variables are important and which regression specifications likely work well.

The uninformed designer does not observe the prior π, but instead only has some vague ex-ante information about it. 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 the first-best estimator.

1.6 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:

[4]One alternative approach would involve putting restrictions on the distribution of potential outcomes and discussing efficient estimation in a large-sample approximation. However, since researchers may disagree about these choices, they would add degrees of freedom to the estimation. In the main part of this article, I therefore consider an exact finite-sample decision-theoretic framework.

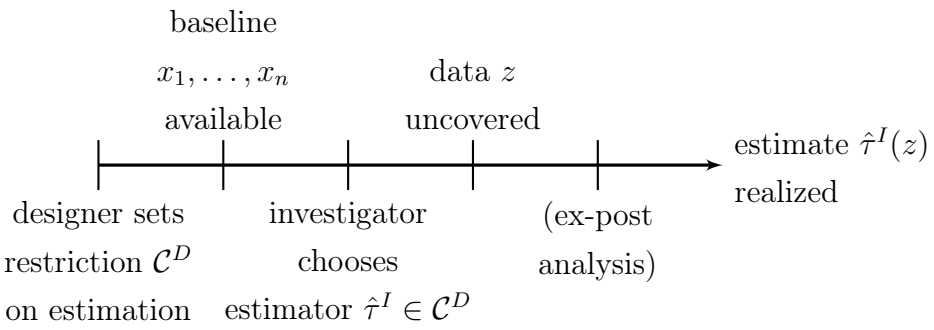

Figure 1: Estimation timeline

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 π and sends a message $m(r^I, \pi)$.
3. The potential outcomes θ 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 modeling 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 is verifiable, the investigator can always choose that message to obtain said estimator. Hence, the designer directly restricts estimators to some set $\mathcal{C}^D$. Subject to the constraint, the investigator specifies an estimator $\hat{\tau}^I \in \mathcal{C}^D$ before data becomes available. Once the data $z \in \mathcal{Z}$ is realized, the investigator reports the estimate $\hat{\tau}^I(z)$ (Figure 1). Since my econometric analysis is conditional on any control variables, 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. I return to discussions of pre-specification when discussing feasible implementations in Section 4.2.

1.7 Investigator and Designer Choices

Having set up the actions available to the investigator and designer, I now describe their preferences. The investigator minimizes average risk over her prior.

Assumption 3 (Investigator's choice)**.** *Given the prior distribution π over potential outcomes Θ, the investigator chooses an estimator that minimizes average risk subject to the constraint $\mathcal{C}^D \subseteq \mathbb{R}^{\mathcal{Z}}$ set by the designer, $\hat{\tau}^I = \hat{\tau}^I(\mathcal{C}^D, \pi) \in \arg\min_{\hat{\tau} \in \mathcal{C}^D} \mathrm{E}_\pi[r^I_\theta(\hat{\tau})]$.*

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. For the private information π, I consider two different solutions. First, I assume that the designer may have a (hyper-) prior η over π, which encodes her belief about the investigator's prior.

Definition 1 (Designer's minimax delegation solution)**.** *Given some set $\mathcal{R}$ of investigator risk functions, I say that a constraint $\mathcal{C}^D \subseteq \mathbb{R}^{\mathcal{Z}}$ solves the designer's delegation problem for a hyperprior η if $\mathcal{C}^D \in \min_{\mathcal{C} \subseteq \mathbb{R}^{\mathcal{Z}}} \sup_{r^I \in \mathcal{R}} \mathrm{E}_\eta[\mathrm{E}_\pi[r^D_\theta(\hat{\tau}^I)]]$.*

In some cases, assuming that the designer has an informative ex-ante belief in the form of such a hyperprior may be unrealistic or impractical to formalize, in which case I also consider the case where the designer instead only assumes that the prior is included in some set Π, and extend the minimax criterion to $\mathcal{R}$ and Π jointly.

Definition 2 (Designer's minimax delegation solution for a worst-case prior)**.** *Given investigator risk functions $\mathcal{R}$ and priors Π, a constraint $\mathcal{C}^D \subseteq \mathbb{R}^{\mathcal{Z}}$ solves the designer's delegation problem for a worst-case prior from Π if if $\mathcal{C}^D \in \min_{\mathcal{C} \subseteq \mathbb{R}^{\mathcal{Z}}} \sup_{r^I \in \mathcal{R}, \pi \in \Pi} \mathrm{E}_\pi[r^D_\theta(\hat{\tau}^I)]$.*

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 $\mathcal{C}^D$ that resolve this tradeoff between flexibility and robustness optimally.

2 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. I then discuss optimal and minimax optimal choices of said biases.

2.1 The Role of Bias

Without misalignment, the first-best estimator that minimizes average mean-squared error generally has bias that changes with the prior. To understand how being flexible on bias can improve estimation, note that bias and variance contribute to the risk

$$r_\theta^D(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau} - \tau_\theta)^2] = \underbrace{(\mathrm{E}_\theta[\hat{\tau}] - \tau_\theta)^2}_{\text{bias}} + \underbrace{\mathrm{Var}_\theta(\hat{\tau})}_{\text{variance}}$$

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

But if the designer leaves the decision over bias to the investigator, then an investigator with misaligned preferences will be inclined to bias the estimator in the direction of her preferences, not of her prior. Consider an investigator with risk $r_\theta^I(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - (\tau_\theta + \delta))^2]$ who would like to show that the treatment effect is higher than it is ($\delta > 0$). The investigator's unconstrained solution is now shifted upward by δ, which is added to the bias term. While reducing the variance relative to an unbiased estimator, the bias component of the designer's risk is increased.

For choices among estimators with given 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, $r_\theta^D(\hat{\tau}) = \mathrm{Var}_\theta(\hat{\tau})$. The biased investigator's risk is $r_\theta^I(\hat{\tau}) = \delta^2 + \mathrm{Var}_\theta(\hat{\tau})$. While risks are not the same, they are shifted by a constant

that is not affected by the choice of estimator, and there is no distortion in choices.

2.2 Fixed-Bias Estimation as Second-Best Solution

Having motivated in an example that the designer choosing 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, he 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. Specifically, I will argue below that fixing the bias everywhere,

$$\mathcal{C}^{\beta} = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_{\theta}[\hat{\tau}] - \tau_{\theta} = \beta_{\theta} \forall \theta \in \Theta\},$$

for some function $\beta \in \mathbb{R}^{\Theta}$ of biases (such as $\beta_{\theta} \equiv 0$, yielding unbiased estimators), is not too permissive provided that investigators all choose as if they minimized mean-squared error relative to *some* target, which may differ from the true treatment effect.

Assumption 4 (Investigator risk restriction)**.** *The investigator has a risk function from the set* $\mathcal{R}^{*} = \{r^{I}; r^{I}_{\theta}(\hat{\tau}) = \mathrm{E}_{\theta}[(\hat{\tau}(z) - \tilde{\tau}_{\theta})^2] \text{ for some } \tilde{\tau} : \Theta \to \mathbb{R}\}$.

The target $\tilde{\tau}_{\theta}$ can vary arbitrarily with the potential outcomes. These risk functions include wanting to systematically overestimate treatment effects ($\tilde{\tau}_{\theta} = \tau_{\theta} + \delta$) or fixed estimation targets ($\tilde{\tau}_{\theta} \equiv$ const.). They can equivalently be written in terms of maximizing expected utility for concave utility functions

$$U^{D}_{\theta}(\hat{\tau}(z)) = a_{\theta} + b_{\theta}\hat{\tau}(z) + c\hat{\tau}^2(z)$$

with parameters $a_{\theta}, b_{\theta}, c$ (where $c < 0$), for which $-\mathrm{E}_{\theta}[U^{D}_{\theta}(\hat{\tau}(z))]$ is equivalent to risk in $\mathcal{R}^{*}$ (up to scaling and the addition of a constant that does not depend on the choice of estimator). Such preferences can also serve as an approximation to wanting to obtain a particularly large estimate (large $\tilde{\tau}_{\theta}$ or large b_{θ}). Nevertheless, considering this specific class of risk functions is restrictive and closely tied to the mean-squared error goal of the designer. I discuss alignment over additional preferences below in Section 2.3, and focus on the preferences in $\mathcal{R}^{*}$ for now.

Lemma 1 (Fixing the bias aligns estimation)**.** *Under Assumptions 1–4 the investigator chooses from the fixed-bias estimators* $\mathcal{C}^{\beta}$ *as if he has the same risk function* $r^{D}_{\theta}(\hat{\tau}) = \mathrm{E}_{\theta}[(\hat{\tau}(z) - \tau_{\theta})^2]$ *as the designer, for any biases* $\beta \in \mathbb{R}^{\Theta}$.

Choices from estimators with fixed bias will be the same for any $r^I \in \mathcal{R}^*$, but there could be a larger set of estimators that provide alignment, or full alignment of preferences could be too costly even if we only consider this specific set of risks. I next show that fixing the bias is not too restrictive in a minimax sense.

Theorem 1 (Fixed bias is minimax optimal)**.** *Assume that $\mathcal{Y}$ is finite and that Assumptions 1–4 hold. Write $\Delta^*(\Theta)$ for all distributions over Θ with full support. Then for every hyperprior η with support within $\Delta^*(\Theta)$ there is a vector of biases $\beta^\eta : \Theta \to \mathbb{R}$ such that the fixed-bias restriction $\mathcal{C}^{\beta^\eta} = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_\theta[\hat{\tau}] = \tau_\theta + \beta^\eta_\theta\}$ solves the designer's delegation problem for the hyperprior η in the sense of Definition 1.*

This theorem states that it is optimal for the designer to fix the bias of the estimator ex-ante, in the sense that otherwise there would be some investigator with worst-case preferences who would exploit any freedom to choose the bias of the estimator in a way that makes the designer worse off. The result is therefore driven by the specific class of preferences from Assumption 4, which is an important limitation. The biases themselves reflect the designer's ex-ante belief, as captured by her hyperprior η. Once the designer has fixed the bias, the instigator's choices are fully aligned, and he chooses a variance-minimizing estimator.

Here and in some results below, I assume that the support of (potential) outcomes is finite to adapt results from the mechanism-design literature and to derive intuitive combinatorial proofs for my results. Since the number of support points is otherwise unrestricted, I think of the finite-support assumption as allowing for flexible approximations to arbitrary distributions, and not as a fundamental restriction.

This econometric finding that fixed-biased estimation is minimax optimal builds upon a mechanism-design result by Frankel (2014), which I adapt to prove of Theorem 1. There, a principal delegates decisions to an agent who observes states. Frankel (2014) characterizes optimal delegation mechanisms without transfers. Similar to my setting here, these mechanisms take the form of budget constants. These budget constraints, just like the bias restrictions, are chosen so that choices are fully aligned.

2.3 Alignment Beyond Simple Loss Functions

Fixing the biases aligns the specific set of preferences in $\mathcal{R}^*$, and Theorem 1 shows that it is a minimax optimal solution for this class of risk functions. In practice, this class of preferences is restrictive, and we may want to also consider choices beyond

these specific risks, including preferences that include testing and inference. In many such cases, fixing the bias still approximately aligns choices. As a first example of alignment beyond these specific preferences, assume that a researcher obtains utility $U(\hat{\tau}(z))$ from an estimate $\hat{\tau}(z)$, for some smooth utility function U. Based on the second-order Taylor approximation

$$U(t) = U(t_0) + U'(t_0)\,(t - t_0) + \frac{1}{2}U''(t_0)\,(t - t_0)^2 + o((t - t_0)^2),$$

for $t{=}\hat{\tau}(z), t_0{=}\,\mathrm{E}_\theta[\hat{\tau}(z)]$, an estimator with bias β_θ yields approximate expected utility

$$\mathrm{E}_\theta[U(\hat{\tau}(z))] \approx U(\tau_\theta + \beta_\theta) + \frac{1}{2}U''(\tau_\theta + \beta_\theta)\ \mathrm{Var}_\theta(\hat{\tau}(z)).$$

Consider the case where the investigator wants to obtain a large estimate (U increasing), but has decreasing returns (U'' negative). In this case, letting the investigator make a choice that affects the bias may lead to estimators with large, positive biases β_θ. But restricting the investigator to estimators with a given bias implies that he wants to choose an estimator with low variance $\mathrm{Var}_\theta(\hat{\tau}(z))$ among such estimators, which is aligned with the minimization of mean-squared error.

As a second example, assume that the investigator employs the estimator $\hat{\tau}$ to test a null hypothesis, such as $\tau_\theta = 0$, and aims to maximize the probability of a rejection. A common test based on approximate Normality would compare the test statistic $\hat{\tau}(z)/\hat{\sigma}(z)$ for a consistent estimator $\hat{\sigma}^2(z)$ of the variance $\mathrm{Var}_\theta(\hat{\tau})$ to the quantiles of a Normal distribution, only rejecting if the absolute value of the test statistic is above some threshold c. Assuming that the estimator $\hat{\tau}$ is approximately Normal and treatment effect and potential bias local to zero (in the sense that they are of a similar order of magnitude as $\sqrt{\mathrm{Var}_\theta(\hat{\tau})}$), the power of the resulting test is approximately $\Phi\left((\tau_\theta + \beta_\theta)/\sqrt{\mathrm{Var}_\theta(\hat{\tau})} - c\right) + \Phi\left(-(\tau_\theta + \beta_\theta)/\sqrt{\mathrm{Var}_\theta(\hat{\tau})} - c\right)$, where Φ is the cdf of the standard Normal distribution. Among potentially biased estimators, the investigator thus may prefer an estimator with a large bias. However, among estimators with fixed bias, an investigator who aims to maximize the probability of a rejection will choose one with low variance, which is aligned with minimizing mean-squared error. Similarly, an investigator aiming to minimize the size of a standard error or length of a confidence interval would want to minimize the variance of $\hat{\tau}$, provided that his inference is based on consistent variance estimates.

Of course, not all plausible loss functions are approximately aligned by unbiased estimation, especially when we take inference into account. For example, an investigator who aims to produce a confidence interval that *includes* a certain, preferred

value may choose an estimator with a large variance. Similarly, if the investigator wants to reject a one-sided null hypothesis, say $\tau_\theta \leq 0$, but believes that the null hypothesis is likely to be true, or does not want to reject a null hypothesis, or obtains convex (rather than concave) utility from estimates, then he may aim to use a more noisy estimator. Here, unbiasedness would not be sufficient to align preferences.

While precise statements about approximate alignment will require additional structure, these illustrations suggest that fixing the bias can approximately align choices among a larger class of preferences. In Appendix A, I provide a more precise analysis of such cases in a large-sample framework based on randomly sampled units. Here, I stay within the conditional finite-sample framework of Section 1, and focus on the tractable preferences from Assumption 2.

2.4 Optimal Choices of Biases

I now discuss the optimal choice of biases by the designer. Theorem 1 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, and that choosing biases that depend on her hyperprior achieves a minimax optimal solution. In general, if the designer has a hyperprior that is quite informative about treatment effects, he 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.

In general, the biases and associated second-best estimators can be complex under the very general setup of Theorem 1. Before narrowing the analysis to specific worst-case priors, I discuss their general properties relative to the estimator $\hat{\tau}^\eta = \arg\min_{\hat{\tau}} \mathrm{E}_\eta[(\hat{\tau}(z) - \tau_\theta)^2]$ (yielding $\hat{\tau}^\eta(z) = \mathrm{E}_\eta[\tau_\theta|z]$) that the designer with hyperprior η would if she could not delegate the estimation to the investigator.[5] The following results on optimal biases, which hold without any additional restrictions on the hyperprior, show that optimal delegation implies *lower* biases than those without delegation (or, equivalently, lower than the average biases of the first-best estimator.

[5]Here and below, I omit writing inner expectations explicitly, and implicitly consider distributions of the implied prior and implied potential outcomes. For example, here I omit E_π in $\mathrm{E}_\eta[\mathrm{E}_\pi[r^D_\theta(\hat{\tau}^I)]]$, and in other places I omit E_θ in writing $\mathrm{E}_\pi[(\hat{\tau}(z) - \tau_\theta)^2]$ instead of $\mathrm{E}_\pi[\mathrm{E}_\theta[(\hat{\tau}(z) - \tau_\theta)^2]]$, unless this omission leads to ambiguity.

Proposition 1 (Bias reduction through delegation)**.** *Under the assumptions of Theorem 1, the optimal biases β_θ^η are the unique solution to a (strictly) convex quadratic minimization problem, and the resulting delegated estimator $\hat{\tau}^I = \hat{\tau}^I(\mathcal{C}^{\beta^\eta}, \pi)$ is η-almost surely unique. The average resulting (squared) optimal bias of the delegated estimator is at most that of the estimator without delegation, $\mathrm{E}_\eta[(\beta_\theta^\eta)^2] \leq \mathrm{E}_\eta[(\mathrm{E}_\theta[\hat{\tau}^\eta(z)] - \tau_\theta)^2]$ with strict inequality unless there is no gain from delegation and $\hat{\tau}^I = \hat{\tau}^\eta$.*[6]

This result characterizes some properties of the optimal biases that emerge from Theorem 1 as a minimax solution for the designer. These biases are the solution to a well-behaved optimization problem, and they are generally lower than the biases of first-best optimal estimators. Specifically, the first-best Bayesian estimators would shrink estimates more towards the prior. However, since the information of the designer is limited, the ex-ante biases set by the designer are less aggressive.

As a practical implication, this result bounds biases in cases where the hyperprior contains little directional information about treatment effects and $\hat{\tau}^\eta(z)$ is approximately unbiased. While the strength of the result lies in its generality, it falls short of a meaningful characterization of biases beyond that case.

2.5 Minimax Optimal Biases

Having discussed some general properties of optimal biases from Theorem 1, I now consider the case where the designer has little information about the distribution. A natural approach to this case is to assume a worst-case prior from some set Π, as in Definition 2. Without any restrictions on the support $\mathcal{Y}$ or set of priors Π, worst-case risks are generally unbounded. I therefore now assume that each prior implies limited variation of potential outcomes. This modeling choice represents the assumption that priors are informative about the distribution of potential outcomes, even in the worst case, while there is no ex-ante knowledge about the location of treatment effects. In this case, we directly obtain that unbiased estimation is a minimax optimal restriction.

[6]The result is a consequence of the specific structure of the nested optimization problem (see Appendix C). The average variance can increase or decrease by delegation, and the average (squared) bias of the first-best estimator can be higher or lower than either. Further, the average (non-squared) biases are all zero across these cases.

Proposition 2 (Minimax optimality of unbiased estimation)**.** *Assume outcomes are from $\mathcal{Y} = \mathbb{R}$, that Assumption 1 holds with iid randomization, and that Assumptions 2–4 hold. Write $\Delta_{\varsigma^2} = \{\pi \in \Delta(\Theta)$ with finite support; $\|\operatorname{Var}_\pi(\theta)\|_2 \leq \varsigma^2\}$, where we treat $\theta \in \mathbb{R}^{2n}$ as a $2n$-dimensional vector and $\|\cdot\|_2$ is the spectral norm. Then unbiased estimation solves the designer's problem for worst-case priors from Δ_{ς^2} in the sense of Definition 2.*

This proposition shows that a restriction to unbiased estimators guarantees maximal utility for the designer when faced with worst-case investigator preferences and priors. The result is driven by a class of priors that does not restrict treatment effects. As a consequence, a bias in any direction would be suboptimal since a worst-case true treatment effect may be just towards the opposite. At the same time, limiting the variation of outcomes for each prior ensures that worst-case average risks are bounded, and restricting these priors to finite support allows us to use the result from Theorem 1 and apply later representation results.

This minimax result demonstrates a gap between the worst-case performance of the (infeasible) first-best estimator, the delegated estimator chosen by the investigator subject to a bias restriction, and the best non-delegation estimator chosen by the designer directly. In this sense, there is a non-trivial cost of misalignment as well as a non-trivial gain from delegation, even in the worst case.

3 Investigator's Solution

Above, I have argued that the designer restricting the investigator to estimators with a given bias is a minimax optimal restriction on estimation. In this section, I establish that this restriction is equivalent to splitting the sample in a particular way. In solving the investigator's constrained optimization problem, optimal estimation corresponds to a choice of out-of-sample regression adjustments.

3.1 Characterization of Fixed-Bias Estimators

When does an estimator have a given bias, conditional on potential outcomes? The designer requires that the investigator use 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 $\hat{\tau}^{\text{HT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} y_i$ is unbiased for any pair of potential outcome vectors because $\mathrm{E}_\theta\left[\frac{d_i-p}{p(1-p)} y_i\right] = y_i(1) - y_i(0)$. If we replace outcomes y_i by adjusted outcomes $y_i - \phi_i(z_{-i})$ with regression adjustments that do not vary with (y_i, d_i), where z_{-i} denotes the data $(y_j, d_j)_{j\neq i}$ from all units other than i, then the resulting estimator is still unbiased. Wu and Gagnon-Bartsch (2018) call the resulting estimator for known p the "leave-one-out potential outcomes" (LOOP) estimator. Since the adjustment $\phi_i(z_{-i})$ is the same whether unit i is treated or not and $\mathrm{E}_\theta\left[\frac{d_i-p}{p(1-p)}\middle| z_{-i}\right] = 0$, the addition of such regression adjustments averages out to zero, no matter the potential outcomes or realized treatment of the other units.

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 $d_i = n_1 - \sum_{j\neq i} d_j$. If we start with the difference in averages $\hat{\tau}^* = \frac{1}{n_1}\sum_{d_i=1} y_i - \frac{1}{n_0}\sum_{d_i=0} y_i = \frac{1}{n_1 n_0}\sum_{d_i=1, d_j=0}(y_i - y_j)$, then we can similarly obtain unbiased estimators that differ from $\hat{\tau}^*$ only by leave-two-out regression adjustments $\phi_{ij}(z_{-ij})$. In every split, they leave out one treated and one untreated unit.

While the leave-one-out and leave-two-out construction appears specific, I show that these sample-splitting estimators are also *all* estimators that are unbiased conditional on potential outcomes. More generally, I prove that any estimator with fixed bias can be written in terms of similar adjustments, for a finite support of outcomes.

Theorem 2 (Representation of fixed-bias estimators)**.** *Let $\hat{\tau}^D$ be a fixed estimator of τ_θ with biases $\beta_\theta = \mathrm{E}_\theta[\hat{\tau}^D(z)] - \tau_\theta$. Then under Assumption 1 and for finite support $\mathcal{Y}$ of the potential outcomes an estimator $\hat{\tau}$ has biases β_θ 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 $\hat{\tau}(z) = \hat{\tau}^D(z) - \frac{1}{n}\sum_{i=1}^n \frac{d_i-p}{p(1-p)}\phi_i(z_{-i})$.*
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 $\hat{\tau}(z) = \hat{\tau}^D(z) - \frac{1}{n_1 n_0}\sum_{i<j}(d_i - d_j)\phi_{ij}(z_{-ij})$, where $\phi_{ij}(z_{-ij})$ may be undefined outside $\mathbf{1}'d_{-ij} = n_1 - 1$.*

This theorem says that any deviation from the leave-one-out or leave-two-out form necessarily leads to bias. This is because we are not imposing any assumptions on the distribution of the data beyond the randomization in the experiment. As a consequence, fixing the bias is equivalent to the designer choosing an estimator $\hat{\tau}^D$

with the desired biases $\mathrm{E}_\theta[\hat{\tau}^D(z)] - \tau_\theta = \beta_\theta$ for all $\theta \in \Theta$, and letting the investigator choose a zero-expectation adjustment of the above form.

The specific representation of this zero-expectation adjustment is restrictive, but not unique. Applied to the Horvitz and Thompson (1952) leave-one-out and sample-average estimators, this result yields a tractable representation of unbiased estimators.

Corollary 1 (Representation of unbiased estimators)**.** *Under Assumption 1 and for finite $\mathcal{Y}$ the estimator $\hat{\tau}$ is unbiased, $\mathrm{E}_\theta[\hat{\tau}(z)] = \tau_\theta$ for all $\theta \in \Theta$, if and only if*

$$\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)}(y_i - \phi_i(z_{-i})) \quad \text{or} \quad \hat{\tau}(z) = \frac{1}{n_1 n_0}\sum_{i<j}(d_i - d_j)(y_i - y_j - \phi_{ij}(z_{-ij})),$$

respectively, with adjustments as in Theorem 2.

Notably, linear regression can not generally be represented this way, as it is not generally unbiased in my setting (Freedman, 2008).

3.2 Optimal Regression Adjustments

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 his subjective prior to minimize average variance over the regression adjustments from Theorem 2. The resulting estimator is a Bayes estimator in the sense of Wald (1950).

Focusing on unbiased estimation for the case of a known treatment probability, which is minimax optimal in the sense of Proposition 2, I consider estimators

$$\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{d_i}{p}(y_i - \phi_i(z_{-i})) - \frac{1-d_i}{1-p}(y_i - \phi_i(z_{-i}))\right) \tag{1}$$

for d_i iid with $\mathrm{P}(d_i = 1) = p$. This estimator is guaranteed to obey the unbiasedness restrictions. We also know from Theorem 2 that *any* estimator can be represented this way for potential outcomes drawn from a prior with finite support. Among this class of estimators, the investigator chooses the one that minimizes the average variance

$$\mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}(z)) = \mathrm{E}_\pi[(\hat{\tau}(z) - \tau_\theta)^2] = \mathrm{E}_\pi\left[\left(\frac{1}{n}\sum_{i=1}^{n}\frac{d_i - p}{p(1-p)}(\phi_i^* - \phi_i(z_{-i}))\right)^2\right]. \tag{2}$$

Here, $\phi_i^* = (1-p)y_i(1) + py_i(0)$ are the oracle adjustments that the investigator would want to choose for $\phi_i(z_{-i})$, but are generally infeasible. These oracle adjustments are averages over potential outcomes that put larger weight on the treated or control

outcomes from the relatively smaller group. For example, if there are more treated than control units on average ($p > \frac{1}{2}$), then the optimal adjustment puts more weight on control outcomes, since it is relatively more prudent to reduce variation among the scarcer outcomes that contribute more variation.

The investigator now minimizes the objective in (2) to choose optimal adjustments. While this minimization problem is convex and quadratic, its solution generally takes complex forms that require full knowledge of the joint distribution of the oracle adjustments ϕ_i^* under the prior π. In the next section, I discuss the structure of optimal and approximately optimal adjustments in a sampling framework, which allows me to derive intuitive expressions in terms of a prediction problem.

4 Sampling-Based Implementation and Extensions

The main results from the previous sections characterize optimal restrictions that a designer can impose to align estimation choices by an investigator with misaligned preferences. In this section, I apply these results to a sampling model in order to derive practical estimators that can serve as the basis for effective pre-analysis plans. I focus on the case of unbiased estimation, for which I show how alignment in terms of sample splitting extends to sampling-based analysis. I then provide practically implementable unbiased estimators in terms of simple prediction solutions that are minimax optimal. In Appendix A, I complement these results with a large-sample approximation that derives the asymptotic distribution of the resulting estimators and argues that choosing among such estimators approximately aligns choices across a larger set of preferences.

4.1 Aligned Estimation in a Sampling Model

The setup of Section 1 analyzes the estimation of sample-average treatment effects conditional on the units in the sample, including their covariates. In many applications, we may model the units as themselves sampled from some population distribution and distinguished ex-ante only by covariates $x_i \in \mathcal{X}$ that are unaffected by the treatment. Furthermore, we may care about average treatment effects defined at the population level, rather than conditional on the drawn sample. In order to capture such cases, I now assume that there is some distribution μ over potential outcomes and covariates $(y_i(1), y_i(0), x_i) \in \mathcal{Y}^2 \times \mathcal{X}$. I then consider potential outcomes and co-

variates $(y_i(1), y_i(0), x_i)_{i=1}^n$ that come from n iid draws from this distribution, where for every observation i we observe the treatment assignment d_i, the realized outcome $y_i = y_i(d_i)$, and the covariate x_i.

Assumption 1' (Random sampling and treatment)**.** *Given a population distribution μ over $(y_i(1), y_i(0), x_i)$, the data $\bar{z} = (y_i, d_i, x_i)_{i=1}^n \in \bar{\mathcal{Z}} = (\mathcal{Y} \times \{0,1\} \times \mathcal{X})^n$ is distributed according to P_μ as follows. Independently for each $i \in \{1, \dots, n\}$, $(y_i(1), y_i(0), x_i)$ is sampled from μ, assigned to treatment with known probability $p = \mathrm{P}(d_i{=}1)$ (where $0{<}p{<}1$) independently of $(y_i(1), y_i(0), x_i)$, and $y_i = y_i(d_i)$ is realized.*

As a natural analog to the setup in Section 1, the investigator's private information is now given by a prior $\bar{\pi}$ over the distribution μ. We can then connect the sampling framework to the more general conditional approach in two ways. As a first option, we can still consider the estimation of the sample-average treatment effect $\tau_\theta = \frac{1}{n}\sum_{i=1}^n y_i(1) - y_i(0)$ conditional on realized potential outcomes and covariates $(y_i(1), y_i(0), x_i)_{i=1}^n$ with preferences as in Section 1 (which may also depend on x). In this case, the prior $\bar{\pi}$ implies a prior $\pi(x)$ over $\theta = (y_i(1), y_i(0))_{i=1}^n$ given $x = (x_i)_{i=1}^n$. The sampling framework then merely implies a specific structure of the prior $\pi(x)$, and the main results go through. In particular, restrictions of the bias (conditional on θ and x) are still second-best solutions per Theorem 1. Furthermore, an inspection of the proof of Proposition 2 reveals that zero conditional bias is still a minimax optimal restriction, even for priors $\pi(x)$ that come from iid sampling.

Within the sampling framework, it may be more natural to consider estimands that are defined at the population level. As a second option for connecting the sampling framework back to the results above, we can explicitly consider the estimation of the average treatment effect $\bar{\tau}_\mu = \mathrm{E}_\mu[y_i(1) - y_i(0)]$ defined on the level of the distribution μ rather than conditional on realized potential outcomes, with designer preferences and researcher choices defined conformably. In this case, restricting the estimator to be unbiased under the assumptions of Assumption 1' aligns preferences similar to those in Assumption 4, with mean-squared error and bliss points defined in terms of the population distribution.

I now discuss applying unbiasedness restrictions at the sampling level. If we impose unbiasedness with respect to sampling, $\mathrm{E}_\mu[\hat{\tau}(\bar{z})] = \bar{\tau}_\mu$, rather than conditional on potential outcomes (and covariates), then we run into a challenge in applying our results about the investigator solution from Section 3. Specifically, the characteriza-

tion of unbiased estimators from Corollary 1 does not apply directly anymore because there are now additional unbiased estimators. For example, an unbiased estimator that only uses data from the first few units is still unbiased for the population-average treatment effect. Similarly, for $n \geq 2$ the estimator $\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} y_i + y_1 - y_2$ is unbiased for $\bar{\tau}_\mu$ under the sampling assumption, since now $\mathrm{E}_\mu[y_1] = \mathrm{E}_\mu[y_2]$ and the last two terms cancel out in expectation. At the same time, these estimators generally have bias conditional on potential outcomes, and do not fall within the class of unbiased estimators characterized by Corollary 1. However, neither estimator is optimal, since they add unnecessary noise. The following result shows that we can generally ignore such conditionally biased estimators.

Lemma 2 (Relationship of unconditionally and conditionally unbiased estimators)**.** *Assume that Assumption 1' holds, and let $\mathcal{M}$ be the class of distributions μ with $\mathrm{E}[y_i^2(1)], \mathrm{E}[y_i^2(0)] < \infty$. Then for every estimator $\hat{\tau}$ that is unconditionally unbiased for all $\mu \in \mathcal{M}$ there is a conditionally (on $(y_i(1), y_i(0), x_i)_{i=1}^n$) unbiased estimator $\hat{\tau}^\dagger$ with $\mathrm{Var}_\mu(\hat{\tau}^\dagger(\bar{z})) \leq \mathrm{Var}_\mu(\hat{\tau}(\bar{z}))$ for all distributions $\mu \in \mathcal{M}$.*

The idea behind this lemma is that any variation in the conditional bias of an unconditionally unbiased estimator is noise, and we can do better by removing it by a type of Rao–Blackwellization. As a consequence, we can still focus on *conditionally* unbiased estimators even when we care about the estimation of population quantities, and even if we only need to impose unconditional unbiasedness to align preferences. Our main results on the characterization of unbiased estimators and their properties then apply directly since the unconditional expected loss is equivalent (up to some irreducible noise) to the average of the conditional variance,

$$\mathrm{E}_\mu[(\hat{\tau}(\bar{z}) - \bar{\tau}_\mu)^2] = \mathrm{E}_\mu[\mathrm{Var}_\theta(\hat{\tau}(\bar{z})|x)] + \mathrm{Var}_\mu(\tau_\theta) = \mathrm{E}_\mu[(\hat{\tau}(\bar{z}) - \tau_\theta)^2] + \frac{\mathrm{Var}_\mu(y_i(1) - y_i(0))}{n},$$

provided that the estimator is conditionally unbiased. Specifically, under the sampling assumptions of Assumption 1' and for finite support, we can restrict ourselves to the unbiased leave-one-out regression-adjustment estimators from (1) of the particularly intuitive form

$$\hat{\tau}(\bar{z}) = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} (y_i - f(x_i|\bar{z}_{-i})), \tag{3}$$

with regression adjustments $f(x_i|\bar{z}_{-i})$ that are allowed to depend on the covariate x_i in addition to data $\bar{z}_{-i} = (y_j, d_j, x_j)_{j \neq i}$ from all other units.

4.2 Feasible Implementations and Pre-Analysis Plans

The main results from Section 2 and Section 3 applied to the sampling setup above show that minimax optimal estimation when researcher and social preferences are misaligned can be achieved by a restriction to sample-splitting estimators of the form in (3) and delegating the estimation of the regression adjustments $f(x_i|\bar{z}_{-i})$ to the investigator. I now discuss how we can ensure that the investigator adheres to such sample-splitting plans, consider feasible implementations that allow for additional flexibility, and demonstrate the form of optimal adjustments in specific cases.

The sample-splitting representation of estimators with fixed bias implies that the investigator is not allowed to use the outcome and treatment assignment of a unit to construct its adjustment. One way to ensure this prohibition is to require that he pre-specifies the construction of regression adjustments before having access to any of the data. This pre-specification is implicit in my setup, since the investigator commits to a mapping from data to estimate. This rules out cases where the investigator introduces additional bias when using data to choose among multiple estimators that are ex-ante unbiased, but may end up being chosen selectively. For the estimator from (3), the investigator would thus be required to file a pre-analysis plan that pre-specifies the full mapping $f(\cdot|\cdot)$, obtained from minimizing $\mathrm{E}_{\bar{\pi}}[(\hat{\tau}(\bar{z})-\tau_\theta)^2]$ based on the prior $\bar{\pi}$ before having access to any data (except possibly for x).

Pre-specifying a procedure by which adjustments are estimated optimally for each unit separately may lead to prohibitive ex-ante optimization and ex-post computation. In practice, we may instead consider a simpler K-fold version

$$\hat{\tau} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\frac{d_i-p}{p(1-p)}(y_i-\hat{f}_k(x_i)), \qquad \hat{f}_k(x_i) = f(x_i|\bar{z}_{-I_k}) \tag{4}$$

for K folds I_k that form a partition $\bigcup_{k=1}^{K} I_k = \{1,\dots,n\}$ and adjustment functions $\hat{f}_k : \mathcal{X} \to \mathbb{R}$ that only depend on data $(y_i, d_i, x_i)_{i\notin I_k}$ from other folds. In this case, the average expected loss in estimating the sample-average treatment effect is

$$\mathrm{E}_{\bar{\pi}}[(\hat{\tau}(z)-\tau_\theta)^2] = \mathrm{E}_{\bar{\pi}}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\frac{d_i-p}{p(1-p)}(\hat{f}_k(x_i)-f^*_\mu(x_i))\right)^2\right] + \text{const.} \tag{5}$$

for oracle adjustments $f^*_\mu(x_i) = \mathrm{E}_\mu[\phi^*_i|x_i] = \mathrm{E}_\mu[(1-p)y_i(1)+py_i(0)|x_i]$ and a constant that does not depend on the choice of the $\hat{f}_k$. This loss suggests the feasible

approximation

$$\hat{f}_k^*(x_i) = \mathrm{E}_{\bar{\pi}}[f_\mu^*(x_i)|x_i, \bar{z}_{-I_k}] = \mathrm{E}_{\bar{\pi}}[(1-p)y_i(1) + py_i(0)|x_i, \bar{z}_{-I_k}]$$

of the infeasible optimal adjustments $f_\mu^*(x_i)$, which can be obtained as the solution to the out-of-sample prediction problem

$$\hat{f}_k^* = \underset{f:\mathcal{X}\to\mathbb{R}}{\arg\min}\, \mathrm{E}_{\bar{\pi}}\left[\frac{(d_i-p)^2}{p(1-p)}(y_i - f(x_i))^2 \middle| \bar{z}_{-I_k}\right]. \tag{6}$$

Here, the loss function for choosing adjustments is a weighted mean-squared error loss, with weights $\frac{(d_i-p)^2}{p(1-p)} = \begin{cases} \frac{1-p}{p}, & d_i = 1 \\ \frac{p}{1-p}, & d_i = 0 \end{cases}$ that put different weights on treatment and control observations, effectively upweighting the larger of the two groups.[7]

In general, the adjustment obtained as the prediction solution in (6) is not exactly optimal. However, Proposition 3 below shows that it still yields a minimax optimal estimator in the case of $K = 2$ folds, and Appendix A shows that for balanced folds the average risk for the estimation of τ_θ is asymptotically equivalent to prediction risk for this loss function. There, I also connect this estimator to cross-fitted Augmented Inverse Propensity Weighted (AIPW) and Double Machine Learning (DML) estimators. This connection offers a practical way for the investigator to commit to an estimator: he pre-specifies the procedure by which the prediction problems are solved for each of the K steps of the cross-fitting procedure. This procedure may include beneficial specification searches to solve the (Bayesian) prediction problem in (6). Thanks to sample splitting, such model selection does not introduce any bias, but may instead be aligned with the minimization of the variance.

An alternative approach that ensures that the investigator cannot introduce any bias, while also avoiding extensive pre-specification, is to give the investigator sequential access to the data.[8] Specifically, I now assume that at each successive step $k \in \{1, \ldots, K\}$, the investigator chooses adjustments $g(x_i|z_{J_k})$ for the units $i \in I_k$

[7] This mirrors Lin's (2013) "tyranny of the minority" estimator, which puts similar weights into a least-squares regression.

[8] Following a similar logic, Anderson and Magruder (2017) propose a hybrid pre-analysis plan for multiple testing. The investigator pre-specifies some hypotheses he will test, and then selects additional hypotheses from a training sample. These are only evaluated on a hold-out sample. I adopt this idea to my estimation setting.

using data only from *earlier* observations $J_k = \bigcup_{\ell<k} I_\ell$, yielding an estimator

$$\hat{\tau}(z) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\frac{d_i - p}{p(1-p)}(y_i - \hat{g}_k(x_i)), \qquad \hat{g}_k(x_i) = g(x_i|\bar{z}_{-J_k}). \tag{7}$$

This means that the designer only has to ensure that the investigator accesses the data in a specific order before committing to the respective regression adjustments, and this leaves more flexibility for interactive specification searches.

For the case of sequential-access estimators, the optimal adjustments are now *exactly* the predictions

$$\hat{g}^*_k(x_i) = \mathrm{E}_{\bar{\pi}}[(1-p)y_i(1) + py_i(0)|x_i, \bar{z}_{-J_k}] = \arg\min_{\hat{y}} \mathrm{E}_{\bar{\pi}}\left[\frac{(d_i - p)^2}{p(1-p)}(y_i - \hat{y})^2 \middle| x_i \bar{z}_{-J_k}\right]$$

of conditional expectations of the oracle adjustments given the respective training data z_{J_ℓ} and the prior. While the additional constraint on each adjustment only using data from *previously available* units rather than those from all others introduces inefficiencies, moving to this sequential-access scheme does not matter in the worst case, and is still a minimax optimal solution from the perspective of the designer.

Proposition 3 (Minimax optimality of sample-splitting estimators)**.** *Assume that Assumption 1' holds along with Assumptions 2–4 conditional on x. Then the following still represent minimax optimal solution in the sense of Definition 2 for worst-case priors $\pi(x) \in \Delta^*_{\varsigma^2} = \{\pi \in \Delta(\Theta); \mathrm{E}_\pi[\|\theta\|^2] < \infty, \|\operatorname{Var}_\pi(\theta)\|_2 \leq \varsigma^2\}$, for some ς^2:*

1. *A restriction to sample-splitting estimators of the form (3);*
2. *A restriction to K-fold estimators of the form (4);*
3. *A restriction to 2-fold estimators of the form (4) for $K = 2$ and with the (potentially suboptimal) prediction adjustments $\hat{f}^*_k$ from (6);*
4. *A restriction to successive-access estimators of the form (7).*

Furthermore, the unique optimal adjustments for the latter are given by $\hat{g}^*_k$.

This result states that the sample-splitting estimators considered in this section are mostly minimax optimal solutions, meaning that they leave enough flexibility to the investigator to find adjustments that ensure that the worst-case average loss does not exceed the bound of Proposition 2. This includes the case when the adjustments for the K-fold estimator are fitted using the prediction adjustments from (6), even when they are not exactly optimal given the prior. However, the result only applies to the $K = 2$ case of having two folds. This is because, for $K \geq 3$, the overlap between training samples leads to additional terms in the variance bound in the proof.

The results of Proposition 3 do not rely on finite support of the outcomes (neither for the support of priors nor for sample-splitting form of estimators), and thus shows that a restriction to sample-splitting estimators is minimax optimal, notwithstanding the finite-support assumption in Theorem 2. It shows that the estimators in this section represent flexible, practically implementable pre-analysis plans with limited pre-commitment that are minimax optimal solutions for a designer delegating the estimation of treatment effects to an investigator with misaligned preferences.

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 algorithmic and exploratory data analysis without leaving room for excess biases. In particular, I characterize unbiased estimators of an average treatment effect as sample-splitting procedures that permit flexible regression adjustments.

My results shed light on the role of bias and variance in treatment-effect estimation from experimental data. Allowing for bias can improve precision. But when incentives are misaligned, giving a researcher the freedom to choose the bias may instead reduce accuracy. 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 control variables can improve unbiased estimation.

These results suggest extensions and generalizations. First, this article focuses on point estimation, but many concerns around researcher bias include inference and significance testing. While the main results extend to aligning preferences over significance in an asymptotic large-sample approximation, a careful finite-sample treatment of preferences over testing and other estimands and loss functions is beyond the scope of this article. Second, detailed commitment to an estimator in the form of pre-specification may not always be practical, in which case a delegation to multiple researchers may provide flexibility while preserving control over bias. Third, the investigator may also have control over the design of the experiment, including the way units are sampled and treatment is assigned. Finally, issues of preference misalignment appear not just in randomized experiments, but also in observational data. I hope that my approach can foster research that extends to these questions.

References

Anderson, Michael L and Jeremy Magruder (2017). Split-sample strategies for avoiding false discoveries. *NBER working paper*. [24 and 54]

Andrews, Isaiah and Maximilian Kasy (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–2794. [4]

Andrews, Isaiah and Jesse M Shapiro (2021). A model of scientific communication. *Econometrica*, 89(5):2117–2142. [5]

Angrist, J D, G W Imbens, and A B Krueger (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67. [4]

Aronow, Peter M and Joel A Middleton (2013). A Class of Unbiased Estimators of the Average Treatment Effect in Randomized Experiments. *Journal of Causal Inference*, 1(1). [3]

Balzer, Laura B, Mark J van der Laan, Maya L Petersen, and the SEARCH Collaboration (2016). Adaptive pre-specification in randomized trials with and without pair-matching. *Statistics in Medicine*, 35(25):4528–4545. [4]

Banerjee, Abhijit V, Sylvain Chassang, Sergio Montero, and Erik Snowberg (2020). A theory of experimenters: Robustness, randomization, and balance. Technical Report 4. [5]

Banerjee, Abhijit V, Sylvain Chassang, and Erik Snowberg (2017). Decision theoretic approaches to experiment design and external validity. In *Handbook of Economic Field Experiments*, volume 1, pages 141–174. Elsevier. [5]

Bates, Stephen, Michael I Jordan, Michael Sklar, and Jake A Soloff (2022). Principal-agent hypothesis testing. *arXiv preprint arXiv:2205.06812*. [5]

Bates, Stephen, Michael I Jordan, Michael Sklar, and Jake A Soloff (2023). Incentive-theoretic bayesian inference for collaborative science. *arXiv preprint arXiv:2307.03748*. [5]

Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S Sekhon, and Bin Yu (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390. [4]

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg (2016). Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics*, 8(1):1–32. [4]

Carroll, Gabriel (2015). Robustness and linear contracts. *The American Economic Review*, 105(2):536–563. [10]

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey (2017a). Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review*, 107(5):261–65. [47]

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2017b). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*. [4]

Christensen, Garret and Edward Miguel (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–80. [4]

Di Tillio, Alfredo, Marco Ottaviani, and Peter Norman Sørensen (2017). Persuasion Bias in Science: Can Economics Help? *The Economic Journal*, 127(605):F266–F304. [4]

Diebold, Francis X (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics*, 33(1):1–1. [4]

Frankel, Alexander (2014). Aligned Delegation. *American Economic Review*, 104(1):66–83. [8, 9, 10, 13, and 31]

Freedman, David A (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193. [4, 7, and 19]

Glaeser, Edward L (2006). Researcher incentives and empirical methods. *NBER working paper*. [4]

Glynn, Adam N and Kevin M Quinn (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56. [47]

Hahn, Jinyong (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2):315. [47]

Hájek, Jaroslav (1962). Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics*, pages 1124–1147. [4]

Hansen, Bruce E and Jeffrey S Racine (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46. [4]

Hirano, Keisuke and Jonathan H Wright (2017). Forecasting With Model Uncertainty:

Representations and Risk Reduction. *Econometrica*, 85(2):617–643. [4]

Holland, Paul W (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945. [6]

Holmström, Bengt Robert (1984). On the theory of delegation. *Bayesian models in economic theory*. [9]

Horvitz, Daniel G and Donovan J Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685. [3, 18, and 19]

Hurwicz, Leonid and Leonard Shapiro (1978). Incentive structures maximizing residual gain under incomplete information. *The Bell Journal of Economics*, pages 180–191. [10]

Leamer, Edward E (1974). False models and post-data model construction. *Journal of the American Statistical Association*, 69(345):122–131. [4]

Leamer, Edward E (1978). *Specification searches: Ad hoc inference with nonexperimental data*. Wiley. [4]

Lehmann, Erich L and Joseph P Romano (2006). *Testing Statistical Hypotheses*. Springer Science & Business Media. [7]

Lenz, Gabriel and Alexander Sahn (2017). Achieving statistical significance with covariates. [7]

Lin, Winston (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318. [4, 7, and 24]

Neyman, Jerzy (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science, 1990*. [5]

Rothe, Christoph (2018). Flexible covariate adjustments in randomized experiments. [4]

Rubin, Donald B (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688. [5]

Rubin, Donald B (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34–58. [6]

Schorfheide, Frank and Kenneth I Wolpin (2012). On the use of holdout samples for model selection. *The American Economic Review*, 102(3):477–481. [4 and 5]

Schorfheide, Frank and Kenneth I Wolpin (2016). To hold out or not to hold out. *Research in Economics*, 70(2):332–345. [4 and 5]

Sterling, Theodore D (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance–Or Vice Versa. *Journal of the American Statistical Association*, 54(285):30. [4]

Tullock, Gordon (1959). Publication decisions and tests of significance—a comment. *Journal of the American Statistical Association*, 54(287):593–593. [4]

Wager, Stefan and Susan Athey (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*. [4]

Wager, Stefan, Wenfei Du, Jonathan Taylor, and Robert J Tibshirani (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678. [4 and 48]

Wald, Abraham (1950). *Statistical decision functions.* Wiley. [8 and 19]

Wu, Edward and Johann A Gagnon-Bartsch (2018). The loop estimator: Adjusting for covariates in randomized experiments. *Evaluation Review*, 42(4):458–488. [3, 4, 18, and 48]

Young, Alwyn (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. [4]

Appendix

Proof of Lemma 1. Take any investigator risk function $r^I \in \mathcal{R}^*$, unbiased estimator $\hat{\tau} \in \mathcal{C}^\beta$ with $\beta \in \mathbb{R}^\Theta$, 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

$$\begin{aligned}
&\mathrm{E}_\pi[r_\theta^D(\hat{\tau})] \overset{r^I \in \mathbb{R}^*}{=} \mathrm{E}_\pi[(\hat{\tau}(z) - \tilde{\tau}_\theta)^2] = \mathrm{E}_\pi[((\hat{\tau}(z) - \mathrm{E}_\theta[\hat{\tau}(z)]) - (\mathrm{E}_\theta[\hat{\tau}(z)] - \tilde{\tau}_\theta))^2] \\
&= \mathrm{E}_\pi[(\hat{\tau}(z) - \mathrm{E}_\theta[\hat{\tau}(z)])^2] + \mathrm{E}_\pi[(\mathrm{E}_\theta[\hat{\tau}(z)] - \tilde{\tau}_\theta)^2] \\
&\overset{\hat{\tau} \in \mathcal{C}^\beta}{=} \mathrm{E}_\pi[\mathrm{Var}_\theta(\hat{\tau}(z))] + \mathrm{E}_\pi[(\tau_\theta + \beta_\theta - \tilde{\tau}_\theta)^2]
\end{aligned}$$

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

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 $N = (2|\mathcal{Y}|)^n$ sample points $z \in \mathcal{Z}$. (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 $\pi \in \Delta(\Theta)$. π is drawn from the (hyper-)prior η.

In the parlance of Frankel (2014), I consider the Φ-moment mechanisms where the agent chooses from estimators $\mathcal{C}_\beta = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_\theta[\hat{\tau}] = \tau_\theta + \beta_\theta \forall \theta \in \Theta\}$ 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 β, I establish that:

1. Any feasible such Φ-moment mechanism (i.e. any bias vector β with $\mathcal{C}_\beta \neq \emptyset$) induces aligned delegation over $\mathcal{R}^*$, that is, subject to the restriction $\hat{\tau} \in \mathcal{C}_\beta$ agents of all risk types $r^I \in \mathcal{R}^*$ choose as if they were of risk type r^D.
2. $\mathcal{R}^*$ is Φ-rich, that is, for any mechanism there exists some $\overline{\beta} \in \mathbb{R}^\Theta$ and a sequence of risk types $(r^{I_k})_{k=1}^\infty \in (\mathcal{R}^*)^{\mathbb{N}}$ such that for all realized $\pi \in \Delta^*(\Theta)$ and all corresponding sequences $(\hat{\tau}_k)_{k=1}^\infty$ of chosen estimators, $\lim_{k\to\infty} \mathrm{E}_\theta[\hat{\tau}_k(z)] = \tau_\theta + \overline{\beta}_\theta$ for all θ in the support of π. (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 $\mathcal{C}_\beta$ is then minimax optimal provided that β minimizes the designer's average risk, for some distribution (hyperprior) η over π. I will develop this deduction below for my specific case (where type and action spaces are not compact) once aligned delegation and richness are established.

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

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

as in the proof of Lemma 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 $r^{I_1}, r^{I_2}, \ldots$ 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 $\mathcal{C} \subseteq \mathbb{R}^{\mathcal{Z}}$ of estimators $\hat{\tau}$ that are outcomes for some investigator risk function $r^I \in \mathcal{R}^*$ and prior π in the support of η. Note that the outcomes of the mechanism are the investigator choices

$$\hat{\tau}_\pi(r^I) \in \arg\min_{\hat{\tau} \in \mathcal{C}} \mathrm{E}_\pi r^I_\theta(\hat{\tau}) \tag{8}$$

where by assumption ties are broken in favor of the designer. I first show that $\mathcal{C}$ in (8) is wlog closed. Since the minimizers are already included in $\mathcal{C}$, taking the closure of $\mathcal{C}$ does not change investigator risk at their optimal choices. Replacing $\mathcal{C}$ by its closure thus does not affect investigator risk at choices (8), 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 $\mathcal{C}$ is closed.

I first assume that $\mathcal{C}$ is also bounded. Define the set $\mathcal{D} = \{\theta \mapsto \mathrm{E}_\theta[\hat{\tau}(z)]; \hat{\tau} \in \mathcal{C}\} \subseteq \mathbb{R}^\Theta$ of vectors of expectations achieved by estimators in $\mathcal{C}$. By linearity of expectation, $\mathcal{D}$ is wlog compact by the above reasoning. Fix some ordering $\theta_1, \ldots, \theta_J$ of Θ (where $J = |\Theta|$). Let δ^0 be the maximal element in $\mathcal{D}$ with respect to the corresponding lexicographic ordering (so that, in particular, $\delta^0_{\theta_1} \geq \delta_{\theta_1}$ for all $\delta \in \mathcal{D}$). For every $h \in \{2, \ldots, J\}$, there exists a function $f_h : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ such that for all $\varepsilon > 0$

$$\delta \in \mathcal{D}, \sum_{j=1}^{h-1} |\delta_{\theta_j} - \delta^0_{\theta_j}| < f_h(\varepsilon) \qquad \Rightarrow \qquad \delta_{\theta_h} < \delta^0_{\theta_h} + \varepsilon. \tag{9}$$

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 $\sum_{j=1}^{h-1} |\delta^\nu_{\theta_j}| < 1/k$ and $\delta^k_{\theta_h} \geq \delta^0_{\theta_j} + \varepsilon$. Since $\mathcal{D}$ is compact, δ^k must have a convergent subsequence with limit $\delta^\varepsilon \in \mathcal{D}$. But $\delta^\varepsilon_{\theta_j} = \delta^0_{\theta_j}$ for $j < h$ and $\delta^\varepsilon_{\theta_h} \geq \delta^0_{\theta_h} + \varepsilon > \delta^0_{\theta_h}$, contradicting that δ^0 is maximal in $\mathcal{D}$ with respect to the lexicographic order. Hence there exists such f_h, and we can assume wlog $\frac{f_h(\varepsilon)}{\varepsilon}$ is monotonically increasing in $\varepsilon > 0$ (otherwise we can choose an f_h that is smaller for small values of ε).

Given the target $\delta^0 \in \mathcal{D}$ and the functions $f_h, h \geq 2$, I construct a sequence of risk functions r^{I_k} such that the expectation of the corresponding investigator choices converges to δ^0 for all $\pi \in \Delta^*(\Theta)$. For $k \in \mathbb{N}$ define $\alpha^k \in \mathbb{R}^\Theta$ recursively by

$$\alpha^k_{\theta_J} = k \qquad\qquad \alpha^k_{\theta_j} = k / \min_{h>j} f_h(1/\alpha^k_{\theta_h}), j < J$$

and consider the sequence of investigator risk functions $r^{I_k}_\theta(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tilde{\tau}^k_\theta)^2], \tilde{\tau}^k_{\theta_j} = \delta^0_{\theta_j} + \alpha^k_{\theta_j}$ 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 $\hat{\tau}_\pi(r^{I_k})$ converges to δ^0. Write $\delta^k_\theta = \mathrm{E}_\theta \hat{\tau}_\pi(r^{I_k})$. Assume for contradiction that δ^k_θ does not converge to δ^0_θ. Since also $\delta^k_\theta \in \mathcal{D}$ for all k and $\mathcal{D}$ compact, $(\delta^k_\theta)_{k=1}^\infty$ must have a converging subsequence $(\delta^{k_\ell}_\theta)_{h=1}^\infty$ with $\delta^{k_\ell} \to \delta^1 \in \mathcal{D} \setminus \{\delta^0\}$ as $h \to \infty$. The average investigator loss along the sequence is

$$\mathrm{E}_\pi r^{I_{k_\ell}}_\theta(\hat{\tau}_\pi(r^{I_{k_\ell}})) = \mathrm{E}_\pi \underbrace{\mathrm{Var}_\theta(\hat{\tau}_\pi(r^{I_{k_\ell}}))}_{\leq \text{const. } (\mathcal{C} \text{ bounded})} + \mathrm{E}_\pi(\delta^{k_\ell}_\theta - (\delta^0_\theta + \alpha^{k_\ell}_\theta))^2. \tag{10}$$

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

$$\begin{aligned}\Delta_\ell &= \mathrm{E}_\pi r^{I_{k_\ell}}_\theta(\hat{\tau}_\pi(r^{I_{k_\ell}})) - \mathrm{E}_\pi r^{I_{k_\ell}}_\theta(\hat{\tau}^0) \overset{(10)}{=} \mathrm{E}_\pi(\delta^{k_\ell}_\theta - (\delta^0_\theta + \alpha^{k_\ell}_\theta))^2 - \mathrm{E}_\pi(\alpha^{k_\ell}_\theta)^2 + \mathcal{O}(1) \\ &= \mathrm{E}_\pi \underbrace{(\delta^{k_\ell}_\theta - \delta^0_\theta)^2}_{\to(\delta^1_\theta - \delta^0_\theta)^2} - 2\,\mathrm{E}_\pi(\delta^{k_\ell}_\theta - \delta^0_\theta)\alpha^{k_\ell}_\theta + \mathcal{O}(1) = -2\sum_{j=1}^J \pi(\theta_j)\alpha^{k_\ell}_{\theta_j}(\delta^{k_\ell}_{\theta_j} - \delta^0_{\theta_j}) + \mathcal{O}(1).\end{aligned}$$

Denote by h the smallest index of for which $\delta^0_{\theta_h} \neq \delta^1_{\theta_h}$. Since δ^0 is maximal with respect to the lexicographic ordering of $\mathcal{D}$ and δ^1 also in $\mathcal{D}$, we must have $\delta^0_{\theta_h} - \delta^1_{\theta_h} > 0$. By revealed preference and since $\alpha^k_{\theta_{j+1}} = o(\alpha^k_{\theta_j})$ for all j, it follows that

$$0 \geq \Delta_\ell / \alpha^{k_\ell}_{\theta_h} = -2\sum_{j=1}^{h-1} \pi(\theta_j)\frac{\alpha^{k_\ell}_{\theta_j}}{\alpha^{k_\ell}_{\theta_h}}(\delta^{k_\ell}_{\theta_j} - \delta^0_{\theta_j}) - 2\pi(\theta_h)(\delta^1_{\theta_h} - \delta^0_{\theta_h}) + o(1).$$

In particular, for $\varepsilon = \pi(\theta_h)(\delta^0_{\theta_h} - \delta^1_{\theta_h})$,

$$\liminf_{\ell \to \infty} \sum_{j=1}^{h-1} \underbrace{\pi(\theta_j)\frac{\alpha^{k_\ell}_{\theta_j}}{\alpha^{k_\ell}_{\theta_h}}(\delta^{k_\ell}_{\theta_j} - \delta^0_{\theta_j})}_{=a^\ell_j} \geq \varepsilon > 0. \tag{11}$$

Hence there must exists some h^* and a subsequence ℓ_s such that

$$a^{\ell_s}_{h^*} \to \nu \in (0, \infty], \qquad\qquad \limsup_{s \to \infty} \frac{a^{\ell_s}_j}{a^{\ell_s}_{h^*}} \leq 1 \ \forall j < h. \tag{12}$$

(That is, $a^{\ell_s}_{h^*}$ 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 $\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_j} > 0$ for all s. By (10), $\sum_{j=1}^{h^*-1} |\delta^{k_s}_{\theta_j} - \delta^0_{\theta_j}| \geq f_{h^*}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})$, so there must exist some $j^* < h^*$ and a refinement of the subsequence along which $|\delta^{k_s}_{\theta_{j^*}} - \delta^0_{\theta_{j^*}}| \geq f_{h^*}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})/(h^*-1)$. Note that

$$\frac{\pi(\theta_{j^*})\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_h}}|\delta^{k_s}_{\theta_{j^*}} - \delta^0_{\theta_{j^*}}|}{\pi(\theta_{h^*})\frac{\alpha^{k_s}_{\theta_{h^*}}}{\alpha^{k_s}_{\theta_h}}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})} \geq \frac{\pi(\theta_{j^*})}{\pi(\theta_{h^*})(h^*-1)}\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_{h^*}}}\frac{f_{h^*}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})}{\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}}}.$$

By (12) there exists some $\nu_0 \in (0, \infty)$ such that $a^{\ell_s}_{h^*} \geq \nu_0$ for all large s. By the definition of $a^{\ell}_{h^*}$ we find, again for large s, that $\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}} = \frac{a^{\ell_s}_{h^*}}{\pi(\theta_{h^*})}\frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}} \geq \frac{\nu_0}{\pi(\theta_{h^*})}\frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}}$. By monotonicty of $\frac{f_{h^*}(\varepsilon)}{\varepsilon}$ therefore for large s

$$\frac{\pi(\theta_{j^*})\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_h}}|\delta^{k_s}_{\theta_{j^*}} - \delta^0_{\theta_{j^*}}|}{\pi(\theta_{h^*})\frac{\alpha^{k_s}_{\theta_{h^*}}}{\alpha^{k_s}_{\theta_h}}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})} \geq \frac{\pi(\theta_{j^*})}{\pi(\theta_{h^*})(h^*-1)}\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_{h^*}}}\frac{f_{h^*}\left(\frac{\nu_0}{\pi(\theta_{h^*})}\frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}}\right)}{\frac{\nu_0}{\pi(\theta_{h^*})}\frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}}}.$$

By construction of the rates α^k_θ, we have that for every triple $j^* < h^* < h$ and every constant $c > 0$ and all large k

$$\frac{\alpha^k_{\theta_{j^*}}}{\alpha^k_{\theta_h}} f_{h^*}\left(c\frac{\alpha^k_{\theta_h}}{\alpha^k_{\theta_{h^*}}}\right) \geq \frac{\alpha^k_{\theta_{j^*}}}{\alpha^k_{\theta_J}} f_{h^*}\left(c\frac{\alpha^k_{\theta_J}}{\alpha^k_{\theta_{h^*}}}\right) = \frac{\alpha^k_{\theta_{j^*}}}{k} f_{h^*}\left(\frac{ck}{\alpha^k_{\theta_{h^*}}}\right) \geq c\alpha^k_{\theta_{j^*}} f_{h^*}\left(\frac{1}{\alpha^k_{\theta_{h^*}}}\right) \geq ck \to \infty.$$

It follows that

$$\frac{\pi(\theta_{j^*})\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_h}}|\delta^{k_s}_{\theta_{j^*}} - \delta^0_{\theta_{j^*}}|}{\pi(\theta_{h^*})\frac{\alpha^{k_s}_{\theta_{h^*}}}{\alpha^{k_s}_{\theta_h}}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})} \to \infty.$$

By (12), $\delta^{k_s}_{\theta_{j^*}} - \delta^0_{\theta_{j^*}} < 0$ for all but at most finitely many s. Hence $a^{\ell_s}_{j^*}/a^{\ell_s}_{h^*} \to -\infty$, and thus $\sum_{j=1}^{h-1} a^{\ell_s}_j \to -\infty$, contradicting (11). 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 δ^0 available (and the investigator minimizes variance given expectation), so unbounded variance along the investigator path can only make the choice with expectation δ^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 δ^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, it is not possible that both average variance and average expectation remain bounded along the ray. If the expectation vector $\mathrm{E}_\theta[\hat{\tau}(z)]$ 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 $\mathrm{E}_\theta[\hat{\tau}_\pi(r^{I_k})] - \tau_\theta \to \bar{\beta}_\theta$ for all $\pi \in \Delta^*(\Theta)$ and all $\theta \in \Theta$. The expected average designer's risk along this sequence is

$$\mathrm{E}_\eta[(\hat{\tau}_\pi(r^{I_k}) - \tau_\theta)^2] = \mathrm{E}_\eta \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k})) + \mathrm{E}_\eta \underbrace{(\mathrm{E}_\theta[\hat{\tau}_\pi(r^{I_k})] - \tau_\theta)^2}_{\to \bar{\beta}^2_\theta \forall \theta \in \Theta, \pi \in \Delta^*(\Theta)},$$

where I omit the argument z of the estimators. Since biases are bounded (since $\mathcal{D}$ is) and the support of η is in $\Delta^*(\Theta)$, by dominated convergence

$$\begin{aligned}\liminf_{k\to\infty} \mathrm{E}_\eta[(\hat{\tau}_\pi(r^{I_k}) - \tau_\theta)^2] &= \liminf_{k\to\infty} \mathrm{E}_\eta \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k})) + \mathrm{E}_\eta \bar{\beta}^2_\theta \\ &\geq \mathrm{E}_\eta \liminf_{k\to\infty} \mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k})) + \mathrm{E}_\eta \bar{\beta}^2_\theta.\end{aligned}$$

For fixed $\pi \in \Delta^*(\Theta)$, $\liminf_{k\to\infty} \mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k}))$ 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 $\mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}^k_\pi)$ 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 $\mathbb{R}^{\mathcal{Z}}$ for which by continuity also $\mathrm{E}_\theta[\hat{\tau}^0_\pi] - \tau_\theta = \bar{\beta}_\theta$.

But then the variance of $\hat{\tau}_\pi^0$ must be at least the variance of a variance-minimizing estimator subject to the bias constraint. Taken together,

$$\begin{aligned}\inf_{r^I \in \mathcal{R}^*} \mathrm{E}_\eta[r_\theta^D(\hat{\tau}_\pi(r^I))] &\geq \liminf_{k\to\infty} \mathrm{E}_\eta[(\hat{\tau}_\pi(r^{I_k}) - \tau_\theta)^2] \\ &\geq \mathrm{E}_\eta \min_{\hat{\tau}\in\mathcal{C}_{\bar{\beta}}} \mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}) + \mathrm{E}_\eta \bar{\beta}_\theta^2.\end{aligned}$$

Now, by aligned delegation, $\min_{\hat{\tau}\in\mathcal{C}_{\bar{\beta}}} \mathrm{E}_\pi(\operatorname{Var}_\theta(\hat{\tau}) + \bar{\beta}_\theta^2) = \min_{\hat{\tau}\in\mathcal{C}_{\bar{\beta}}} \mathrm{E}_\pi r_\theta^D(\hat{\tau}) = \mathrm{E}_\pi r_\theta^D(\hat{\tau}_\pi(r^I))$ 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 β^η, 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 $\mathrm{E}_\eta \beta_\theta^2 \leq \mathrm{E}_\eta r_\theta^D(z \mapsto 0) < \infty$) and the expected average risk is continuous in the choice of bias. □

Proof of Proposition 1. The explicit construction in Appendix C shows that β^η as well as $\hat{\tau}^I = \hat{\tau}^{\eta\to\pi}$ solve (strictly) convex quadratic optimization problems, with uniqueness explicitly established by Remark C.1. Here, we use this construction to establish the bias bound. Specifically, I compare the properties of $\hat{\tau}^I = \hat{\tau}^{\eta\to\pi}$ to the estimator $\hat{\tau}^\eta$ that the designer would be able to achieve without delegation.

The estimator $\hat{\tau}^\eta$ is equal to the estimator $\hat{\tau}^{\eta^*\to\pi^*}$ that is delegated to an investigator who has a prior $\pi^*(\theta) = \mathrm{E}_\eta[\pi(\theta)]$ (where the expectation is over draws of π) by a designer with hyperprior $\eta^* = \mathrm{P}(\pi = \pi^*) = 1$. Since U_η in (22) only depends on η through the distribution π^*, we can choose $U_\eta = U_{\eta^*}$. Since U_η is invertible and the average risk in (23) always non-negative, no matter the choice of reference parameter $\hat{\alpha}^*$, we must have that $\left(\mathbb{I} - \sqrt{U_\eta}(U_\eta + W)^{-1}\sqrt{U_\eta}\right)$ is positive-semidefinite for both $W = W_\eta$ and $W = W_{\eta^*}$. Furthermore,

$$\left(\mathbb{I} - \sqrt{U_\eta}(U_\eta + W_\eta)^{-1}\sqrt{U_\eta}\right) \preceq \left(\mathbb{I} - \sqrt{U_\eta}(U_\eta + W_\eta^*)^{-1}\sqrt{U_\eta}\right)$$

by revealed preference (since delegation decreases average risk), which also implies

$$\begin{aligned}&\operatorname{Bias}^2(\hat{\tau}^{\eta\to\pi}) = \hat{\alpha}^{*\prime}\sqrt{U_\eta}\left(\mathbb{I} - \sqrt{U_\eta}(U_\eta + W_\eta)^{-1}\sqrt{U_\eta}\right)^2\sqrt{U_\eta}\hat{\alpha}^* \\ &\leq \hat{\alpha}^{*\prime}\sqrt{U_\eta}\left(\mathbb{I} - \sqrt{U_\eta}(U_\eta + W_{\eta^*})^{-1}\sqrt{U_\eta}\right)^2\sqrt{U_\eta}\hat{\alpha}^* = \operatorname{Bias}^2(\hat{\tau}^{\eta^*\to\pi^*}).\end{aligned} \tag{13}$$

Finally, if $\hat{\tau}^{\eta^*\to\pi^*} = \hat{\tau}^\eta \neq \hat{\tau}^I = \hat{\tau}^{\eta\to\pi}$, we must have that

$$\begin{aligned}&\operatorname{Risk}^2(\hat{\tau}^{\eta\to\pi}) = \hat{\alpha}^{*\prime}\sqrt{U_\eta}\left(\mathbb{I} - \sqrt{U_\eta}(U_\eta + W_\eta)^{-1}\sqrt{U_\eta}\right)\sqrt{U_\eta}\hat{\alpha}^* \\ &< \hat{\alpha}^{*\prime}\sqrt{U_\eta}\left(\mathbb{I} - \sqrt{U_\eta}(U_\eta + W_{\eta^*})^{-1}\sqrt{U_\eta}\right)\sqrt{U_\eta}\hat{\alpha}^* = \operatorname{Risk}^2(\hat{\tau}^{\eta^*\to\pi^*})\end{aligned}$$

by uniqueness, which implies that (13) is also strict. We obtain the claimed inequality between biases, which is strict when the two estimators are different. □

Proof of Proposition 2. To show that the given biases are minimax optimal, I first establish that they lead to upper bounds on the average risk that hold uniformly across priors within the respective set of priors. I then show that there are specific groups of priors for which we also cannot do any better. I finally put both pieces together to establish minimaxity.

Uniform upper bound. For any prior π, I establish an upper bound on average risk. The following estimator fulfills the bias restrictions $\beta_\theta = -\lambda_1\bar{y}(1) + \lambda_0\bar{y}(0)$:

$$\hat{\tau}^\dagger(z) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{d_i}{p}((1-\lambda_1)(y_i - \mathrm{E}_\pi[\phi_i^*]) - \frac{1-d_i}{1-p}((1-\lambda_0)y_i - \mathrm{E}_\pi[\phi_i^*])\right) \tag{14}$$

for $\phi_i^* = (1-p)(1-\lambda_1)y_i(1)+p(1-\lambda_0)y_i(0)$ (where the expectations are all well-defined because the potential outcomes have bounded second moments). This estimator follows the structure of estimators I discuss in Section 3, but the choices of adjustments may generally be inefficient (which does not affect the worst-case upper bound). For $\pi \in \Delta_{\varsigma^2}$, the average risk is

$$\begin{aligned}
&\mathrm{E}_\pi[r_\theta^D(\hat{\tau}^\dagger)] = \mathrm{E}_\pi[\mathrm{E}_\theta[\hat{\tau}^\dagger(z) - \tau_\theta]] = \mathrm{E}_\pi[\beta_\theta^2] + \mathrm{Var}_\theta(\hat{\tau}^\dagger(z)) \\
&= \mathrm{E}_\pi\left[\left(\frac{1}{n}\sum_{i=1}^{n}(-\lambda_1 y_i(1) + \lambda_0 y_i(0))\right)^2\right] + \mathrm{E}_\pi\Bigg[\underbrace{\mathrm{E}_\theta\left[\left(\frac{1}{n}\sum_{i=1}^{n}\frac{d_i - p}{p(1-p)}(\phi_i^* - \mathrm{E}_\pi[\phi_i^*])\right)^2\right]}_{=\frac{1}{n^2 p(1-p)}\sum_{i=1}^{n}(\phi_i^* - \mathrm{E}_\pi[\phi_i^*])}\Bigg] \\
&= \frac{\mathrm{E}_\pi^2\left[\sum_{i=1}^{n}(-\lambda_1 y_i(1) + \lambda_0 y_i(0))\right]}{n^2} + \frac{\mathrm{Var}(\sum_{i=1}^{n}(-\lambda_1 y_i(1) + \lambda_0 y_i(0)))}{n^2} + \frac{\sum_{i=1}^{n}\mathrm{Var}_\pi(\phi_i^*)}{n^2 p(1-p)} \\
&\leq \frac{\mathrm{E}_\pi^2\left[\sum_{i=1}^{n}(-\lambda_1 y_i(1) + \lambda_0 y_i(0))\right]}{n^2} + \frac{(\lambda_1^2 + \lambda_0^2)\varsigma^2}{n} + \frac{((1-p)^2(1-\lambda_1)^2 + p^2(1-\lambda_0)^2)\varsigma^2}{np(1-p)}.
\end{aligned}$$

For the specific case of unbiasedness ($\lambda_1 = 0 = \lambda_0$), we obtain the bound

$$\sup_{\pi \in \Delta_{\varsigma^2}} \mathrm{E}_\pi[r_\theta^D(\hat{\tau}^I)] \leq \frac{(1-p)^2 + p^2}{np(1-p)}\varsigma^2 = \frac{\frac{1-p}{p} + \frac{p}{1-p}}{n}\varsigma^2.$$

Since the above estimator always obeys the bias restrictions and is always feasible, this bound represents an upper bound of the average risk of all priors in the set.

A hyperprior that yields high best-case average risk. Next, I show that there are hyperpriors that yield average risk close to the above bound, even when

biases are chosen optimally. For these hyperpriors, I consider priors with support (for the potential outcomes) $y_i(1) \in \mathcal{Y}_1 = \{-c_1, +c_1\}, y_i(1) \in \mathcal{Y}_1 = \{-c_0, +c_0\}$ with sufficiently large $c_1, c_0 > 0$. (I note here that none of the finite-support results in this article depend on the support being the same across treatment and control potential outcomes, and that, in particular, Theorem 1 extends to this setting.) I assume that priors are parameterized by $m = (m(1), m(0)) \in \{-1, +1\}^{2n}$. For $q_1 = \sqrt{1 - \frac{\varsigma^2}{c_1^2}}, q_0 = \sqrt{1 - \frac{\varsigma^2}{c_0^2}}$, each prior π_m is iid across $(y_i(1), y_i(0))$ with $\mathrm{P}_{\pi_m}(y_i(1) = +c_1) = \frac{1+mq_1}{2}, \mathrm{P}_{\pi_m}(y_i(0) = +c_0) = \frac{1-mq_0}{2}$, which is chosen such that $\mathrm{E}_{\pi_m}[y_i(1)] = m\, c_1 q_1, \mathrm{E}_{\pi_m}[y_i(1)] = -m\, c_0 q_0, \mathrm{Var}_{\pi_m}(y_i(1)) = \varsigma^2 = \mathrm{Var}_{\pi_m}(y_i(0))$.

Each hyperprior η now puts a distribution over m, which we assume is independent across the $m_i = (m_i(1), m_i(0))$, and we assume is symmetrical around zero. Since the optimal biases β^η corresponding to each hyperprior are unique, they must retain the invariances of the hyperprior. To see that this implies that the biases and the implied optimal estimators must be additive separable across i, note that for any set of biases β_θ with estimators $\hat{\tau}_\pi$ that fulfill $\mathrm{E}_\theta[\hat{\tau}_\pi(z)] = \tau_\theta + \beta_\theta$ (and can depend on π) and any partition of units into two sets I and J, setting the biases to $\beta^*_{\theta_I,\theta_J} = \mathrm{E}_\eta[\beta_{\theta_I,\theta_J}|\theta_I] + \mathrm{E}_\eta[\beta_{\theta_I,\theta_J}|\theta_J] - \mathrm{E}_\eta[\beta_{\theta_I,\theta_J}]$ and corresponding estimators to $\hat{\tau}^*_{\pi_I,\pi_J}(z_I, z_J) = \mathrm{E}_\eta[\hat{\tau}_{\pi_I,\pi_J}(z_I, z_J)|z_I, \pi_I] + \mathrm{E}_\eta[\hat{\tau}_{\pi_I,\pi_J}(z_I, z_J)|z_J, \pi_J] - \mathrm{E}_\eta[\hat{\tau}_{\pi_I,\pi_J}(z_I, z_J)]$ (that fulfill the bias restrictions $\mathrm{E}_\theta[\hat{\tau}^*_\pi(z)] = \tau_\theta + \beta^*_\theta$) and has lower overall average risk because

$$\begin{aligned}
&\mathrm{E}_\eta[(\tau_\pi(z) - \tau_\theta)^2] = \mathrm{E}_\eta[(\tau_\pi(z) - \tau^*_\pi(z) + \tau^*_\pi(z) - \mathrm{E}_\pi[\tau_\theta|z])^2] \\
&= \mathrm{E}_\eta[(\tau_\pi(z) - \tau^*_\pi(z))^2] + \mathrm{E}_\eta[(\tau^*_\pi(z) - \tau_\theta)^2] \geq \mathrm{E}_\eta[(\tau^*_\pi(z) - \tau_\theta)^2],
\end{aligned}$$

where we have used that τ_θ is additive separable, all distributions are independent between I and J, and $\tau_\pi(z) - \tau^*_\pi(z)$ orthogonal to any function that does not include interactions between the two parts. By iteration, there is a set of biases that is additively separable across i and improves weakly over any given bias schedule. Hence, the optimal bias is additively separable, and we can consider biases of the form $\beta_\theta = \frac{1}{n}\sum_{i=1}^n \beta^*_{\theta_i}$ with $\beta^*(+c, +c) = -\beta^*(-c, -c), \beta^*(+c, -c) = -\beta^*(-c, +c)$ by symmetry of the hyperprior. For fixed c, any such bias function can be expressed by the parametrization $\beta^*(y_i(1), y_i(0)) = -\lambda_1 y_i(1) + \lambda_0 y_i(0)$ for suitable λ_1, λ_0. We can therefore limit ourselves to biases $\beta_\theta = \frac{1}{n}\sum_{i=1}^n -\lambda_1 y_i(1) + \lambda_0 y_i(0)$. By Theorem 2, any unbiased estimator with these biases can be written as

$$\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^n \left(\frac{d_i}{p}((1-\lambda_1)(y_i - \phi_i(z_{-i})) - \frac{1-d_i}{1-p}((1-\lambda_0)y_i - \phi_i(z_{-i})) \right).$$

We now consider the optimal adjustments for a given prior π_m. By additive separability (or solving the first-order conditions explicitly), the optimal adjustments do not depend on the data, $\phi_i(z_{-i}) = \phi_i$. For $\phi_i^* = (1-p)(1-\lambda_1)y_i(1) + p(1-\lambda_0)y_i(0)$ as above, the average variance (which the investigator with prior $\pi = \pi_m$ minimizes) is then

$$\mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}(z)) = \mathrm{E}_\pi\left(\frac{1}{n}\sum_{i=1}^n \frac{d_i - p}{p(1-p)}(\phi_i^* - \phi_i)\right)^2 = \frac{1}{p(1-p)n^2}\sum_{i=1}^n \mathrm{E}_\pi(\phi_i^* - \phi_i)^2,$$

which is minimized by $\phi_i = \mathrm{E}_\pi[\phi_i^*]$. Averaging over the hyperprior η, we obtain

$$\begin{aligned}
&\mathrm{E}_\eta[r^D(\hat{\tau})] = \mathrm{E}_\eta\, \mathrm{E}_{\pi_m}[\beta_\theta^2 + \operatorname{Var}_\theta(\hat{\tau})] \\
&= \mathrm{E}_\eta\left[\mathrm{E}_{\pi_m}^2\left[\frac{1}{n}\sum_{i=1}^n -\lambda_1 y_i(1) + \lambda_0 y_i(0)\right] + \operatorname{Var}_{\pi_m}\left(\frac{1}{n}\sum_{i=1}^n -\lambda_1 y_i(1) + \lambda_0 y_i(0)\right)\right. \\
&\qquad \left. + \frac{\frac{1-p}{p}(1-\lambda_1)^2 + \frac{p}{1-p}(1-\lambda_0)^2}{n}\varsigma^2\right] \\
&= \frac{1}{n}\mathrm{E}_\eta\left[(-\lambda_1 m_i(1)c_1 q_1 + \lambda_0 m_i(0)c_0 q_0)^2\right] + \frac{\lambda_1^2 + \lambda_0^2 + \frac{1-p}{p}(1-\lambda_1)^2 + \frac{p}{1-p}(1-\lambda_0)^2}{n}\varsigma^2.
\end{aligned}$$

For $m_i(1), m_i(0)$ independent and $c_1 = c = c_0$ (which yields $q_1 = q = q_0$) chosen such that $cq = \bar{y}$, we obtain

$$\mathrm{E}_\eta[r^D(\hat{\tau})] = \frac{\lambda_1^2 + \lambda_0^2}{n}\bar{y}^2 + \frac{\lambda_1^2 + \lambda_0^2 + \frac{1-p}{p}(1-\lambda_1)^2 + \frac{p}{1-p}(1-\lambda_0)^2}{n}\varsigma^2.$$

For this hyperprior, the optimal choices for the shrinkage parameters are

$$\lambda_1 = \frac{\frac{1-p}{p}\varsigma^2}{(\frac{1-p}{p}+1)\varsigma^2 + \bar{y}^2} = \frac{(1-p)\varsigma^2}{\varsigma^2 + p\bar{y}^2}, \qquad \lambda_0 = \frac{p\varsigma^2}{\varsigma^2 + (1-p)\bar{y}^2},$$

which yields $\mathrm{E}_\eta[r^D(\hat{\tau})] = \frac{(1-p)\varsigma^2(\varsigma^2+\bar{y}^2)}{\varsigma^2+p\bar{y}^2} + \frac{p\varsigma^2(\varsigma^2+\bar{y}^2)}{\varsigma^2+(1-p)\bar{y}^2}$ We can choose $\bar{y}$ (and thus c) large to get arbitrarily close to the bound.

Minimax optimality. Since there is a sequence of hyperpriors over admissible priors for which an optimal solution gets arbitrarily close to the bound from the first part, no minimax optimal solution can achieve a better worst-case outcome. Since the bias restrictions in the proposition attain these bounds, they are minimax optimal. Finally, Theorem 1 shows that bias restrictions are minimax optimal (over preferences) for any hyperprior over priors with full support. Since this result applies to the construction in the second part, we cannot obtain a better worst-case outcome over preferences *and* priors even relative to alternative mechanisms. □

Proof of Theorem 2. 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 $y = d \circ y(1) + (\mathbf{1} - d) \circ y(0) \in \mathcal{Y}^n$. ($\circ$ denotes the Hadamard (entry-wise) product.)

Known treatment probability, binary outcomes. I start with known treatment probability $p = \mathrm{E}_\theta[d_i]$ with d_i iid and binary support.

Theorem 3. *For $\mathcal{Y} = \{0,1\}$, assume that the estimator $\hat{\delta} : (y,d) \mapsto \hat{\delta}(y,d)$ has expectation zero (conditional on $\theta = (y(1), y(0))$). Then, $\hat{\delta}(y,d) = \frac{1}{n}\sum_{i=1}^n \frac{d_i - p}{p(1-p)} \phi_i(y_{-i}, d_{-i})$ for a set of functions $\phi_i : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R}$.*

Proof. Take $\phi_i(y_{-i}, d_{-i})$ such that

$$\hat{\delta}(y,d) = \frac{1}{n}\sum_{i=1}^n \frac{d_i - p}{p(1-p)} \phi_i(y_{-i}, d_{-i}) \tag{15}$$

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 $\phi_i(\mathbf{1}_{n-1}, \mathbf{1}_{n-1})$ in any way that has (15) hold for $\hat{\delta}(\mathbf{1}_n, \mathbf{1}_n)$.
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 (15) 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 (15), 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 $y'_i d_i = 1$, $\hat{\delta}(y,d)$ has the term $\phi_i(y_{-i}, d_{-i})$ appear on the right in (15), 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 $y'_i d_i = 1$. 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 (15) 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 (15) 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 (15) extends to (y,d) with $y'd = 0$.

Write $\hat{\delta}^\phi$ for the function defined by (15) 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, $\mathrm{E}_\theta[\hat{\delta}(y,d)] = 0$ for all $\theta = (y(1), y(0))$, so $0 = \mathrm{E}_\theta[\hat{\delta}(y,d)] = \sum_{d\in\{0,1\}^n} \mathrm{P}(d)\ \hat{\delta}(d \circ y(1) + (1-d) \circ y(0), d)$. Fixing (y^*, d^*), it follows for any $\tilde{y}$ that

$$\hat{\delta}(y^*, d^*) = -\sum_{d\in\{0,1\}^n\setminus\{d^*\}} \mathrm{P}(d)/\,\mathrm{P}(d^*)\ \hat{\delta}((\mathbb{1}_{d_i=d_i^*})_{i=1}^n \circ y^* + (\mathbb{1}_{d_i\neq d_i^*})_{i=1}^n \circ \tilde{y}, d) \tag{16}$$

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 $\tilde{y}^*(d) = (\mathbb{1}_{d_i=d_i^*})_{i=1}^n \circ y^* + (\mathbb{1}_{d_i\neq d_i^*})_{i=1}^n \circ \tilde{y}, d)$ for some $\tilde{y}$ and all $d \neq d^*$, then $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$.

We are ready to show (15) for all (y^*, d^*), by induction over $\mathbf{1}'d^*$. We let $\tilde{y} = \mathbf{1}$ throughout. At $k = 0$, $d^* = \mathbf{0}$. 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 (16), $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$. Assume now that the claim holds for all (y^*, d^*) with $\mathbf{1}'d^* \leq k$, and consider some (y^*, d^*) with $\mathbf{1}'d^* = k+1$. Then, for any $d \neq d^*$ with $\mathbf{1}'d \leq k$, $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$ by the induction assumption. For any $d \neq d^*$ with $\mathbf{1}'d \geq k+1$ 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^*). ■

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.

Theorem 4. *Let $\mathcal{Y} = \{0,1\}$. Assume that $\hat{\delta} : (y,d) \mapsto \hat{\delta}(y,d)$ has expectation zero (for all θ). Then $\hat{\delta}(y,d) = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} \phi_{ij}(y_{-ij}, d_{-ij}), \phi_{ij} = -\phi_{ji}$ for functions $\phi_{ij} : (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R}$.*

Note that we can also write $\hat{\delta}(y,d) = \frac{1}{n_1 n_0} \sum_{i=1}^n \sum_{j=i+1}^n (d_i - d_j)\ \phi_{ij}(y_{-ij}, d_{-ij})$, where we sum over each pair once and ϕ_{ij} is only defined for $j > i$.

We first establish a lemma that adopts the proof strategy from Theorem 3 to the setting at hand. To this end, for $(y(1), y(0)) \in (\mathcal{Y}^2)^n$ write

$$N(y(1), y(0)) = \{(d \circ y(1) + (1-d) \circ y(0), d); d \in \mathcal{D}_{n_1}\}$$

(the 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 $c : \mathcal{C} \to \mathcal{C}^-$ be the surjective correspondence $(y,d) \mapsto \{(ij, (y_{-ij}, d_{-ij})); i<j, d_i \neq d_j\}$.

Lemma 3. *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 $\mathcal{D}_t = \mathcal{C}_t^- \setminus \bigcup_{s<t} \mathcal{C}_s^-$, there exists injections $b_t : \mathcal{C}_t \to \mathcal{D}_t$ for $t \leq T^*$ and*
2. *for all $t > T^*$ and $(y,d) \in \mathcal{C}_t$, there exists some $(y(1), y(0)) \in (\mathcal{Y}^2)^n$ both $(y,d) \in N(y(1), y(0))$ and $(N(y(1), y(0)) \setminus \{(y,d)\}) \cap \bigcup_{s\geq t} \mathcal{C}_s = \emptyset$*

then for any $\hat{\delta}$ that is mean-zero there exist a function $\phi : \mathcal{C}^- \to \mathbb{R}$ 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 ϕ with $\hat{\delta}^\phi(y,d) = \hat{\delta}$ for all $(y,d) \in \bigcup_{t\leq T^*} \mathcal{C}_t$, and then establish that this implies $\hat{\delta}^\phi(y,d) = \hat{\delta}$ also for $(y,d) \in \bigcup_{t>T^*} \mathcal{C}_t$.

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

For the second part, by assumption $\mathrm{E}_\theta[\hat{\delta}(y,d)] = 0$ for all $\theta = (y(1), y(0))$, so $0 = \mathrm{E}_\theta[\hat{\delta}(y,d)] = \sum_{(y,d)\in N(y(1),y(0))} \hat{\delta}(y,d)$. Fixing (y^*, d^*) it follows for any $(y(1), y(0))$ with $(y^*, d^*) \in N(y(1), y(0))$ that

$$\hat{\delta}(y^*, d^*) = -\sum_{(y,d)\in N(y(1),y(0))\setminus\{(y^*,d^*)\}} \hat{\delta}(y,d) \tag{17}$$

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 $(y,d) \in \mathcal{C}_t$, by induction over t. For some $t > T^*$, assuming $\hat{\delta}^\phi = \hat{\delta}$ holds for all $(y,d) \in \mathcal{C}_s$ with $s < t$ (as is the case for all $s \leq T^*$), take any $(y^*, d^*) \in \mathcal{C}_t$. By the second part of the lemma, (17) and the induction assumption we must have $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$. This completes the proof. ∎

Proof of Theorem 4. Define $a, b : \mathcal{C} \to \mathbb{N}_0$ by $a(y,d) = y'd, b(y,d) = (\mathbf{1} - y)'(\mathbf{1} - d)$. Note that $a(y,d) + b(y,d) \leq n$. First, set $T^* = n - 1$ and for every $t \leq T$

$$\mathcal{C}_t = \{(y,d) \in \mathcal{C}; \min(a(y,d), b(y,d)) \geq 1, a(y,d) + b(y,d) = n + 1 - t\}.$$

Then the first assumption of Lemma 3 is fulfilled, as for every $(y,d) \in \mathcal{C}_t$ there exists some $(ij, (y_{-ij}, d_{-ij}) \in \mathcal{C}_t$ with $y'_{-ij} d_{-ij} + (\mathbf{1} - y_{-ij})'(\mathbf{1} - d_{-ij}) = n - 1 - t = a(y,d) + b(y,d) - 2$, but (y,d) is also the unique element in $\mathcal{C}_t$ covering that element of $\mathcal{D}_t$ under the correspondence c (as indeed necessarily $y_i = d_i, y_j = d_j$, which pins

down (y,d) from $(ij,(y_{-ij},d_{-ij}))$. Second, with $T=n+1$ and

$$\mathcal{C}_n=\{(y,d)\in\mathcal{C};a(y,d)=0,b(y,d)\geq 1\},\qquad \mathcal{C}_{n+1}=\{(y,d)\in\mathcal{C};b(y,d)=0\},$$

note that for each $(y^*,d^*)\in\mathcal{C}_n\cup\mathcal{C}_{n+1}$ we have that $(y(1),y(0))=(y^*\circ d^*+\mathbf{1}\circ(1-d^*),y^*\circ(1-d^*))$ produces $N(y(1),y(0))\cap\{(y,d)\in\mathcal{C};\min(a(y,d),b(y,d))=0\}=\{(y^*,d^*)\}$ for $(y^*,d^*)\in\mathcal{C}_n$ and $N(y(1),y(0))\cap\{(y,d)\in\mathcal{C};b(y,d)=0\}=\{(y^*,d^*)\}$ for $(y^*,d^*)\in\mathcal{C}_{n+1}$. This verifies the second assumption of Lemma 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 $y=d\circ y(1)+(\mathbf{1}-d)\circ y(0)$. Our goal now is to extend a representation for binary outcomes to one for finite (but arbitrarily large) support $\mathcal{Y}$.

Lemma 4. *Assume that for $\mathcal{Y}=\{0,1\}$ any $\hat{\delta}$ with $\mathrm{E}_\theta[\delta(y,d)]=0$ for all $\theta=(y(1),y(0))$ permits a representation $\hat{\delta}=\hat{\delta}^\phi$ with $\hat{\delta}^\phi(y,d)=\sum_{\iota\in\mathcal{I}}w_\iota(d_{S_\iota})\phi_\iota(y_{-S_\iota},d_{-S_\iota})$ for fixed $\mathcal{I},(w_\iota)_{\iota\in\mathcal{I}},(S_\iota)_{\iota\in\mathcal{I}}$ (where $\mathcal{I}$ finite) and variable $(\phi_\iota)_{\iota\in\mathcal{I}}$ where $\phi_\iota:(\mathcal{Y}\times\{0,1\})^{\{1,\dots,n\}\setminus S_\iota}\to\mathbb{R}$. Then the representation result extends to any finite $\mathcal{Y}\subseteq\mathbb{R}$ (with the same $\mathcal{I},(w_\iota)_{\iota\in\mathcal{I}},(S_\iota)_{\iota\in\mathcal{I}}$).*

Proof. Write $\mathcal{Y}_\ell=\{0,1,\dots,\ell\}$ 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 $(s,m)\in(\mathbb{N}_0\times\{1,\dots,n\})\cup\{(0,0)\}$ for $\ell=2^s+1$ any $\hat{\delta}$ with $\mathrm{E}_\theta[\hat{\delta}(y,d)]=0$ for all $\theta=(y(1),y(0))\in\mathcal{Y}_{\ell,m}^2$ permits a representation $\hat{\delta}=\hat{\delta}^\phi$ as above with $\phi_\iota:\bigtimes_{i\in\{1,\dots,n\}\setminus S_\iota}(\mathcal{Y}_{\ell,m})_i\to\mathbb{R}$

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 $(s^+,m^+)\in\mathbb{N}_0\times\{1,\dots,n\}$ with $ns^++m^+=t+1$, and write $\ell^+=2^{s^+}+1$ Fix an estimator $\hat{\delta}$ with $\mathrm{E}_\theta[\hat{\delta}(y,d)]=0$ 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

$$\hat{\delta}_1(y,d)=\hat{\delta}(y^+,d)-\hat{\delta}(y^-,d)\qquad\qquad\hat{\delta}_2(y,d)=\hat{\delta}(y,d)$$

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 $\tilde{y}_{m^+}=\max(y_{m^+}-\ell^++1,0),\tilde{y}_{-m^+}=y_{-m^+}$ to obtain $\bar{y},\tilde{y}\in\mathcal{Y}_{\ell,m}^2$ for which

$$\hat{\delta}(y,d)=\hat{\delta}(y,d)-\hat{\delta}(\bar{y},d)+\hat{\delta}(\bar{y},d)=\hat{\delta}_1(\tilde{y},d)+\hat{\delta}_2(\bar{y},d).^{9}$$

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

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

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

$$\hat{\delta}(y,d) = \sum_{\iota \in \mathcal{I}} w_\iota(d_{S_\iota})(\phi^1_\iota(\tilde{y}_{-S_\iota}, d_{-S_\iota}) + \phi^2_\iota(\bar{y}_{-S_\iota}, d_{-S_\iota})$$

for any $(y,d) \in \mathcal{Y}_{\ell^+,m^+} \times \{0,1\}^n$. For $\phi_\iota(y_{-S_\iota}, d_{-S_\iota}) = \phi^1_\iota(\tilde{y}_{-S_\iota}, d_{-S_\iota}) + \phi^2_\iota(\bar{y}_{-S_\iota}, d_{-S_\iota})$ 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 $\mathcal{Y} = \mathcal{Y}_{2^s+1}$. Since it will always hold for subsets, it holds for all $\mathcal{Y} = \mathcal{Y}_\ell$. Now take arbitrary $\mathcal{Y} = \{z_1, \ldots, z_\ell\}$, and define for $(y,d) \in (\mathcal{Y}_\ell \times \{0,1\})^n$ $\tilde{\delta}(y,d) = \hat{\delta}(z_y, d)$ where $(z_y)_i = z_{y_i} \in \mathcal{Y}$. By the intermediate result there is some $\tilde{\phi}$ such that $\tilde{\delta} = \hat{\delta}^{\tilde{\phi}}$. Setting $\phi_\iota(y_{-S_\iota}, d_{-S_\iota}) = \tilde{\phi}(\tilde{y}_{-S_\iota}, d_{-S_\iota})$ with $\tilde{y}$ such that $z_{\tilde{y}} = y$ yields $\hat{\delta}(y,d) = \hat{\delta}^\phi(y,d)$. ■

The representation for general finite support follows from Lemma 4 applied to the binary representation results in Theorem 3 and Theorem 4, respectively, where the results extend to any fixed bias by taking differences. □

Proof of Corollary 1. The corollary follows from Theorem 2 applied to the estimators $\frac{1}{n}\sum_{i=1}^n \frac{d_i - p}{p(1-p)} y_i$, $\frac{1}{n_1 n_0}\sum_{i<j}(d_i - d_j)(y_i - y_j)$, which are unbiased. □

Proof of Lemma 2. Write Σ_n for the set of permutations of $\{1, \ldots, n\}$, $\bar{z}_i = (y_i, d_i, x_i)$, $x = (x_i)_{i=1}^n$, and for $\bar{z} = (\bar{z}_i)_{i=1}^n$ and $\varsigma \in \Sigma_n$ let $\bar{z}_\varsigma = (\bar{z}_{\varsigma(i)})_{i=1}^n$. For an estimator $\hat{\tau} : \bar{z} \mapsto \hat{\tau}(\bar{z}) \in \mathbb{R}$ define $\hat{\tau}^\dagger : \bar{z} \mapsto \hat{\tau}^\dagger(\bar{z}) \in \mathbb{R}$ by $\hat{\tau}^\dagger(\bar{z}) = \frac{1}{|\Sigma_n|}\sum_{\varsigma \in \Sigma_n} \hat{\tau}(\bar{z}_\varsigma)$. Then:

Lemma 5. *$\hat{\tau}$ is unbiased for $\bar{\tau}_\mu$ for all $\mu \in \mathcal{M}$ if and only if $\hat{\tau}^\dagger$ is unbiased for τ_θ conditional on all potential outcome and covariate vectors $(\theta, x) \in \mathcal{Y}^{2n} \times \mathcal{X}$.*

Proof. By sampling, $\mathrm{E}_\mu[\hat{\tau}(\bar{z})] = \mathrm{E}_\mu[\hat{\tau}^\dagger(\bar{z})]$. Thus from $\mathrm{E}_\theta[\hat{\tau}^\dagger(\bar{z})|x] = \tau_\theta$ for all (θ, x) it follows that $\mathrm{E}_\mu[\hat{\tau}(\bar{z})] = \mathrm{E}_\mu[\hat{\tau}^\dagger(\bar{z})] = \mathrm{E}_\mu[\mathrm{E}_\theta[\hat{\tau}^\dagger(\bar{z})|x]] = \mathrm{E}_\mu[\tau_\theta] = \bar{\tau}_\mu$ for all $\mu \in \mathcal{M}$.

Assume now that $\mathrm{E}_\mu[\hat{\tau}(\bar{z})] = \bar{\tau}_\mu = \mathrm{E}_\mu[\tau_\theta]$ for all $\mu \in \mathcal{M}$. Given some $(\theta^\dagger, x^\dagger) = (y_i^\dagger(1), y_i^\dagger(0), x_i^\dagger)_{i=1}^n$, we now show that $\mathrm{E}_{\theta^\dagger}[\hat{\tau}^\dagger(\bar{z})] = \tau_{\theta^\dagger}$.

To this end, let $\mu(\nu) \in \mathcal{M}$ denote the distribution over $(y_I^\dagger(1), y_I^\dagger(0), x_I^\dagger)$ for a random variable I that takes values in $\{1, \dots, n\}$ according to probabilities $\nu_i = \mathrm{P}_\nu(I = i)$. Then the distribution $\mathrm{P}_{\mu(\nu)}$ over θ, x, and data $\bar{z}$ can be seen as coming from an iid distribution P_ν over indices $I_1, \dots, I_n$ all drawn according to ν, leading to realizations $(y_i(1), y_i(0), x_i)_{i=1}^n = (y_{I_i}^\dagger(1), y_{I_i}^\dagger(0), x_{I_i}^\dagger)_{i=1}^n$ underlying the data $(y_i, d_i, x_i)_{i=1}^n$.

By the assumption of (unconditional) unbiasedness, $\mathrm{E}_{\mu(\nu)}[\mathrm{E}_\theta[\hat{\tau}(\bar{z})|x] - \tau_\theta] = 0$ for all ν. Writing $\beta_\theta(x) = \mathrm{E}_\theta[\hat{\tau}(\bar{z})|x] - \tau_\theta$ for the conditional bias and $J = \{I_1, \dots, I_n\}$ for the set of realized indices (which is a random variable with respect to P_ν), we now show, by induction over $|J^\dagger|$, that it follows that $\mathrm{E}_\nu[\beta_\theta(x)|J{=}J^\dagger] = 0$ for all sets $J^\dagger \subseteq \{1, \dots, n\}$ and all ν with $\mathrm{P}_\nu(J = J^\dagger) > 0$. For such $J^\dagger$ and ν, define $\nu(J^\dagger)$ by $\nu_i(J^\dagger){=}\frac{\nu_i}{\sum_{j \in J^\dagger} \nu_j}$, and note that $\mathrm{E}_\nu[\beta_\theta(x)|J{\subseteq}J^\dagger] = \mathrm{E}_{\mu(\nu(J^\dagger))}[\beta_\theta(x)] = 0$. For $|J^\dagger|$, the statement is then immediate, since $\mathrm{E}_\nu[\beta_\theta(x)|J = J^\dagger] = \mathrm{E}_\nu[\beta_\theta(x)|J{\subseteq}J^\dagger] = 0$. Now assuming the claim for all sets smaller than $|J^\dagger| \geq 2$, note that

$$0 = \mathrm{E}_\nu[\beta_\theta(x)|J{\subseteq}J^\dagger] = \mathrm{P}_\nu(J{=}J^\dagger)\,\mathrm{E}_\nu[\beta_\theta(x)|J{=}J^\dagger] + \sum_{J' \subsetneq J^\dagger} \underbrace{\mathrm{P}(J{=}J')\,\mathrm{E}_\nu[\beta_\theta(x)|J{=}J']}_{=0}$$

and thus $\mathrm{E}_\nu[\beta_\theta(x)|J{=}J^\dagger] = 0$ whenever $\mathrm{P}_\nu(J{=}J^\dagger) > 0$, which concludes the induction.

It follows from the previous step that $\mathrm{E}_\nu[\mathrm{E}_\theta[\hat{\tau}(\bar{z})|x] - \tau_\theta|J = \{1, \dots, n\}] = 0$ for ν with full support. For such a choice,

$$0 = \mathrm{E}_\nu[\mathrm{E}_\theta[\hat{\tau}(\bar{z})|x] - \tau_\theta|J = \{1, \dots, n\}] = \frac{1}{|\Sigma_n|} \sum_{\varsigma \in \Sigma_n} \mathrm{E}_{\theta_\varsigma^\dagger}[\hat{\tau}(\bar{z})|x_\varsigma^\dagger] - \tau_{\theta_\varsigma^\dagger} = 0$$

with $\theta_\varsigma^\dagger = (y_{\varsigma(i)}^\dagger(1), y_{\varsigma(i)}^\dagger(0))_{i=1}^n$, $x_\varsigma^\dagger = (x_{\varsigma(i)}^\dagger)_{i=1}^n$. Since $\frac{1}{|\Sigma_n|} \sum_{\varsigma \in \Sigma_n} \mathrm{E}_{\theta_\varsigma^\dagger}[\hat{\tau}(\bar{z})|x_\varsigma^\dagger] = \mathrm{E}_{\theta^\dagger}[\hat{\tau}^\dagger(\bar{z})|x^\dagger]$ and $\tau_{\theta_\varsigma^\dagger} = \tau_{\theta^\dagger}$ for all $\varsigma \in \Sigma_n$, we obtain that $\mathrm{E}_{\theta^\dagger}[\hat{\tau}^\dagger(\bar{z})] = \tau_{\theta^\dagger}$. ■

Finally, let $S \sim \mathcal{U}(\Sigma_n)$ denote a random permutation. Under μ, $\hat{\tau}(\bar{z}) \overset{d}{=} \hat{\tau}(\bar{z}_S)$ and

$$\mathrm{Var}_\mu(\hat{\tau}(\bar{z})) = \mathrm{Var}_\mu(\hat{\tau}(\bar{z}_S)) = \mathrm{Var}_\mu(\underbrace{\mathrm{E}[\hat{\tau}(\bar{z}_S)|\bar{z}]}_{=\hat{\tau}^\dagger(\bar{z})}) + \underbrace{\mathrm{E}_\mu[\mathrm{Var}(\hat{\tau}(\bar{z}_S)|\bar{z})]}_{\geq 0} \geq \mathrm{Var}_\mu(\hat{\tau}^\dagger(\bar{z}))$$

by the law of total variance. □

Proof of Proposition 3. The result follows from the analogous results in Proposition B.1 in the conditional setup of Section 1, which is developed in Appendix B. To establish that the results are still minimax in the context of Section 4.1, note that the priors in the construction in the proof of Proposition 2 are independent across units i and thus still feasible in the setting of Assumption 1'. □

Supplementary Appendix

A Asymptotic Approximations and Alignment

In Section 4, I consider unbiased estimators in a sampling framework. Here, I use large-sample approximations to derive the large-sample distribution of the resulting estimators and formally discuss alignment beyond the specific loss functions from Assumption 4.

A.1 Asymptotically Optimal Unbiased Estimators

Having recast the main results within a sampling framework, argued that conditionally unbiased estimation remains a solution to aligning designer and investigator preferences, and documented the form of simple minimax optimal unbiased estimators, we can employ large-sample approximations to further clarify the structure of feasible and approximately optimal regression adjustments and analyze the distribution of resulting estimators. Specifically, I now describe the asymptotic distribution of the K-fold estimator (4) for K approximately equally-sized folds for n large. Assuming that the adjustment functions $\hat{f}_k$ are consistent for some limit f, for a given true parameter μ and all k, implies that we obtain asymptotic Normality of the estimator from (4).[10]

Proposition A.1 (Asymptotic Normality)**.** *Assume that Assumption 1' holds, that* $\mathrm{E}_\mu[y_i^2(1)], \mathrm{E}_\mu[y_i^2(0)] < \infty$*, that there is a function* $\bar{f} : \mathcal{X} \to \mathbb{R}$ *with* $\mathrm{E}_\mu[\bar{f}^2(x_i)] < \infty$ *for which* $\mathrm{E}_\mu[(\hat{f}_k(x_i) - \bar{f}(x_i))^2] \to 0$ *for all* k *and* $x_i \in I_k$*, and write* $\bar{\tau}_\mu(x_i) = \mathrm{E}_\mu[y_i(1) - y_i(0)|x_i]$ *for the conditional average treatment effect. Then we have that*

$$\sqrt{n}(\hat{\tau}(\bar{z}) - \bar{\tau}_\mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2_\mu(\bar{f})),$$

where $\sigma^2_\mu(\bar{f}) = \frac{\mathrm{E}_\mu[\mathrm{Var}_\mu(y_i(1)|x_i)]}{p} + \frac{\mathrm{E}_\mu[\mathrm{Var}_\mu(y_i(0)|x_i)]}{1-p} + \mathrm{Var}_\mu(\bar{\tau}_\mu(x_i)) + \frac{\mathrm{E}_\mu[(\bar{f}(x_i) - f^*_\mu(x_i))^2]}{p(1-p)}$*, with oracle adjustments* $f^*_\mu(x_i) = \mathrm{E}_\mu[\phi^*_i|x_i] = \mathrm{E}_\mu[(1-p)y_i(1) + py_i(0)|x_i]$*. Furthermore, a consistent estimator of* $\sigma^2_\mu(\bar{f})$ *is* $\hat{\sigma}^2(\bar{z}) = \frac{1}{n}\sum_{i \in I_k}\left(\frac{d_i - p}{p(1-p)}(y_i - \hat{f}_k(x_i)) - \hat{\tau}(\bar{z})\right)^2 \xrightarrow{p} \sigma^2_\mu(\bar{f})$.

[10] Here, we consider a fixed distribution μ, but we could also assume that the distribution μ and target function f both depend on the sample size n. Similarly, we could assume different limits of f across folds k, such as for the case of adjustments $\hat{g}_k$ for sequential-access estimators in Equation 7, but we focus here on the simple case where the estimated functions are the same across folds.

From a frequentist perspective, if there are regression adjustment functions $\hat{f}_k$ that are consistent for the optimal (generally infeasible) oracle adjustments f^*_μ in the sense that $\mathrm{E}_\mu[(\hat{f}_k(x_i) - f^*_\mu(x_i))^2] \to 0$, then the asymptotic variance is minimal and the resulting estimator achieves the Hahn (1998) semi-parametric efficiency bound for estimating average treatment effects. We can also connect this estimator to Augmented Inverse Propensity Weighted (AIPW; see e.g. Glynn and Quinn, 2010) and Double Machine Learning (DML; Chernozhukov et al., 2017a) estimators. Specifically, if we assume that we separately estimate $f^1_\mu(x_i) = \mathrm{E}_\mu[y_i(1)|x_i], f^0_\mu(x_i) = \mathrm{E}_\mu[y_i(1)|x_i]$ by $\hat{f}^1_k(x_i), \hat{f}^0_k(x_i)$ to obtain the estimator $\hat{f}_k(x_i) = (1-p)\hat{f}^1_k(x_i) + p\hat{f}^0_k(x_i)$, then the resulting estimator is the familiar cross-fitted doubly-robust AIPW/DML estimator

$$\hat{\tau}(\bar{z}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{i\in I_k}(\hat{f}^1_k(x_i) - \hat{f}^0_k(x_i)) + \frac{d_i - p}{p(1-p)}(y_i - \hat{f}^{d_i}_k(x_i))$$

The construction here suggests that essentially all unbiased estimators take a similar form, that sample splitting is not only sufficient, but also necessary to ensure unbiasedness, that this type of estimator only requires the estimation of a single nuisance function $\hat{f}$ rather than separate $\hat{f}^1, \hat{f}^0$, and that simple consistency of $\hat{f}$ is sufficient for $\sqrt{n}$ consistency and asymptotic Normality in the experimental case.

From the Bayes perspective in this article, how should an investigator with prior $\bar{\pi}$ choose among estimators of the form (4), or equivalently, which adjustments should he choose? Assuming that the investigator cares about average risk in large samples and is only choosing among adjustments $\hat{f}_k$ that are consistent for some function f, the investigator would choose adjustments that are consistent for functions f_μ that minimize $\mathrm{E}_{\bar{\pi}}[\sigma^2_\mu(f_\mu)]$, or equivalently, that minimize $\mathrm{E}_{\bar{\pi}}[(f_\mu(x_i) - f^*_\mu(x_i))^2]$. The following result also shows that this minimization problem corresponds to choosing adjustments that solve a weighted out-of-sample prediction problem.

Proposition A.2 (Asymptotic alignment between variance minimization and prediction)**.** *Under the assumptions of Proposition A.1 for all k and $i \in I_k$*

$$n\,\mathrm{E}_\mu[(\hat{\tau}(\bar{z}) - \bar{\tau}_\mu)^2] = \sigma^2_\mu(\bar{f}) + o(1) = \mathrm{E}_\mu\left[\frac{(d_i - p)^2}{p(1-p)}(y_i - \hat{f}_k(x_i))^2 \middle| \hat{f}_k\right] + o_p(1).$$

If the assumptions of Proposition A.1 hold for $\bar{\pi}$-almost all μ with limit f_μ and also $\mathrm{E}_{\bar{\pi}}[y^2_i(1)], \mathrm{E}_{\bar{\pi}}[y^2_i(0)] < \infty$, $\mathrm{E}_{\bar{\pi}}[\bar{f}^2(x_i)] < \infty$, and $\mathrm{E}_\mu[(\hat{f}_k(x_i) - f_\mu(x_i))^2] \leq C$ for some constant C that does not depend on μ, then the same holds on average over $\bar{\pi}$,

$$n\,\mathrm{E}_{\bar{\pi}}[(\hat{\tau}(\bar{z}) - \bar{\tau}_\mu)^2] = \mathrm{E}_{\bar{\pi}}[\sigma^2_\mu(f_\mu)] + o(1) = \mathrm{E}_{\bar{\pi}}\left[\frac{(d_i - p)^2}{p(1-p)}(y_i - \hat{f}_k(x_i))^2\right] + o(1).$$

This result shows the asymptotic equivalence of minimizing mean-squared error in the estimation of the treatment effect among unbiased estimators and the solution to an out-of-sample prediction problem. Specifically, the investigator may choose among some (possibly restricted) class of functions $\mathcal{F}$ those functions that solve the weighted mean-squared error minimization problem

$$\hat{f}_k = \arg\min_{f \in \mathcal{F}} \mathrm{E}_{\bar{\pi}} \left[\frac{(d_i - p)^2}{p(1-p)} (y_i - f(x_i))^2 \middle| \bar{z}_{-I_k} \right],$$

as in (6). Hence, the result creates a direct connection between better prediction and better estimation, showing that resolving bias–variance trade-offs more effectively in estimating nuisance components (here, the estimation of f^*_μ) can improve the unbiased estimation of an estimation target.[11]

These results on asymptotically optimal regression adjustments suggest an explicit, approximately optimal solution for the investigator in some specific cases. If the functions $\mathcal{F}$ are unrestricted (that is, taking the form $\mathbb{R}^{\mathcal{X}}$), the result shows that the posterior expectations $\hat{f}_k(x_i) = \mathrm{E}_{\bar{\pi}}[\phi^*_i | x_i, \bar{z}_{-I_k}]$ of the oracle adjustments $\phi^*_i = (1-p)y_i(1) + py_i(0)$ discussed in Section 3.2 provide an asymptotically optimal solution, provided that the distribution $\bar{\pi}$ is such that they are consistent in the sense of $\mathrm{E}_{\bar{\pi}}[(\mathrm{E}_{\bar{\pi}}[\phi^*_i | \bar{z}_i] - \mathrm{E}_\mu[\phi^*_i])^2] \to 0$. Since this choice achieves the asymptotic efficiency bound for $\bar{\pi}$-almost all μ, it is asymptotically optimal among all unbiased estimators, notwithstanding the restriction to K-folds estimation. In this case, there is value from delegation to an investigator with a prior $\bar{\pi}$, provided that this prior is concentrated enough to ensure consistency, while the designer's ex-ante knowledge would not be enough to guarantee the same. For example, the investigator may know which of a large number of covariates to include in the model, while the designer does not.

A.2 Asymptotic Alignment

The asymptotic approximation from the previous section is helpful in discussing alignment across a larger class of loss functions. In Section 2, I argue that fixing the bias

[11] Wager et al. (2016) using a similar sample-splitting construction note that "the precision of the treatment effect estimates obtained by such regression adjustments depends only on the prediction risk of the fitted regression adjustment." Similarly, Wu and Gagnon-Bartsch (2018) show that the variance of the LOOP estimator is approximately the average mean-squared error in predicting the oracle adjustments.

aligns choices across preferences that come from mean-squared error-type loss functions. While we can see such preferences as approximations of wanting to obtain specific or particularly large estimates, in practice we may consider utility functions that take more complicated forms, especially when we care about testing or other downstream decisions based on point estimates and standard errors. Here, I show that the set of utility functions over which unbiasedness asymptotically aligns choices is considerably larger.

Proposition A.3 (Asymptotic risk equivalence)**.** *Assume that the investigator chooses among unbiased estimators of the form in* (4)*, and that the assumptions from Proposition A.1 hold.*

1. *Assume that the investigator aims to maximize utility $U(\hat{\tau}(\bar{z}))$, for a three times continuously differentiable utility function U with $U'' < 0$. Also assume that U''' is bounded and that $\mathrm{E}_\mu[|\sqrt{n}(\hat{\tau}(\bar{z}) - \bar{\tau}_\mu)|^3] \leq B$ for some constant B. Then maximizing expected utility, in the sense of maximizing $n\,\mathrm{E}_\mu[U(\hat{\tau}(\bar{z})) - U(\bar{\tau}_\mu)]$, is asymptotically equivalent to minimizing risk $r^U_\mu(\hat{\tau}) = -\frac{1}{2}\,U''(\bar{\tau}_\mu)\,\sigma^2_\mu(\bar{f})$.*
2. *Assume that the investigator aims to maximize utility $V(\hat{\tau}(\bar{z}), \hat{\sigma}^2(\bar{z}))$ for a continuously differentiable function $V(\hat{\tau}(\bar{z}), \hat{\sigma}^2(\bar{z}))$ with bounded first derivatives and $V(\bar{\tau}_\mu, \cdot)$ monotonically decreasing. Then maximizing expected utility, in the sense of maximizing $\mathrm{E}_\mu[V(\hat{\tau}(\bar{z}), \hat{\sigma}^2(\bar{z}))]$, is asymptotically equivalent to minimizing risk $r^V_\mu(\hat{\tau}) = -V(\bar{\tau}_\mu, \sigma^2_\mu(\bar{f}))$.*

This result presents cases in which maximizing investigator utility is asymptotically aligned with minimizing the asymptotic variance $\sigma^2_\mu(\bar{f})$ (and thus with minimizing the mean-squared error $\mathrm{E}_\mu[(\hat{\tau}(\bar{z}) - \bar{\tau}_\mu)^2]$). While the sufficient conditions are overly restrictive as stated, they suggest connections to cases like those discussed in Section 2.3. The first case could be seen as a formalization of a researcher who, for example, wants to obtain a large estimate $\hat{\tau}(\bar{z})$, where the returns to larger and larger estimates are decreasing. The second case also captures cases where the variance estimate also matters, such as the calculation of confidence intervals and the construction of tests. Specifically, it could capture minimizing the length of a 95% confidence interval for $\bar{\tau}_\mu$, which, scaled up by a factor of $\sqrt{n}$, yields $V(\hat{\tau}(\bar{z}), \hat{\sigma}^2(\bar{z})) = 2 \cdot 1.96 \cdot \hat{\sigma}(\bar{z})$. Or it could capture maximizing the appropriately re-scaled (absolute) t-statistic $V(\hat{\tau}(\bar{z}), \hat{\sigma}^2(\bar{z})) = \frac{|\hat{\tau}(\bar{z})|}{\hat{\sigma}(\bar{z})}$. We could obtain similar results for the power of a test itself, although a complete treatment would require an ap-

propriate limit experiment based on localized sequences. As a simple illustration, we could consider a two-sided hypothesis test of the null hypothesis $\hat{\tau}(\bar{z}) = \bar{\tau}_n$ with size 5% that rejects whenever $\sqrt{n}\frac{|\hat{\tau}(\bar{z})-\bar{\tau}_n|}{\hat{\sigma}^2(\bar{z})} > 1.96$, and consider its power for local alternatives $\bar{\tau}_n = \bar{\tau}_\mu + \frac{\Delta}{\sqrt{n}}$. In this case, maximizing power is asymptotically equivalent to minimizing $\Phi(\Delta/\sigma_\mu(\bar{f}) - 1.96) + \Phi(-\Delta/\sigma_\mu(\bar{f}) - 1.96)$, which is strictly increasing in $\sigma^2_\mu(\bar{f})$ for all $\Delta \neq 0$.

The above result establishes sufficient conditions for the investigator wanting to minimize the variance of an unbiased estimator, but it does not generally imply that choices are fully aligned with the minimization of average mean-squared error. This is because the risk functions in Proposition A.3 are not generally the same as the squared variance $\sigma^2_\mu(\bar{f})$, but instead represent a monotonic transformation. Once we take averages over the prior $\bar{\pi}$, minimizing average risk is therefore not generally equivalent to minimizing average risk $\mathrm{E}_{\bar{\pi}}[\sigma^2_\mu(\bar{f})]$ (except for specific cases like the minimization of the *squared* length of confidence intervals). We can then see the choices of investigator and designer as partially aligned. Or we can impose additional assumptions that yield full alignment, such as in the following corollary.

Corollary A.1 (Asymptotic alignment when efficient estimation is feasible)**.** *Assume that the investigator chooses among unbiased estimators of the form in (4) for which the assumptions from the second part of Proposition A.2 hold $\bar{\pi}$-almost surely across μ, and that there are efficient adjustment functions $\hat{f}^*_k$ with $\mathrm{E}_{\bar{\pi}}[(\hat{f}^*_k(x_i) - f^*_\mu(x_i))^2] \to 0$. Then investigator preferences of the form in Proposition A.3 under the respective additional assumptions (where B does not vary with μ) are asymptotically aligned with the minimization of the average mean-squared error $\mathrm{E}_{\bar{\pi}}[(\hat{\tau}(\bar{z}) - \bar{\tau}_\mu)^2]$.*

The idea behind this corollary is that if variance can be minimized across essentially all μ, then an investigator with the above preferences would indeed do so. This choice is also the choice that asymptotically minimizes the average variance of an unbiased estimator, thus aligning preference between investigator and designer.

A.3 Proofs

Proof of Proposition A.1. The following lemma will help establish the result:

Lemma A.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, $\mathrm{E}[a_i a_j] = 0$ for all $i, j \in I_k$. Then $\mathrm{Var}\left(\sum_{i=1}^n a_i\right) \leq K \sum_{i=1}^n \mathrm{Var}(a_i)$.*

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

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

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

For the K-fold estimator from (4) we have that

$$\sqrt{n}(\hat{\tau}(\bar{z}) - \bar{\tau}_\mu) = \frac{1}{\sqrt{n}} \sum_{k=1}^K \sum_{i\in I_k} \frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i)) + \frac{d_i - p}{p(1-p)}(\bar{f}(x_i) - \hat{f}_k(x_i)).$$

By Lemma A.1 we have that

$$\begin{aligned}&\mathrm{E}_\mu\left[\left(\frac{1}{\sqrt{n}} \sum_{k=1}^K \sum_{i\in I_k} \frac{d_i - p}{p(1-p)}(\bar{f}(x_i) - \hat{f}_k(x_i))\right)^2\right] \\ &\leq \frac{K}{n} \sum_{k=1}^K \sum_{i\in I_k} \underbrace{\mathrm{E}\left[\left(\frac{d_i - p}{p(1-p)}\right)^2 (\bar{f}(x_i) - \hat{f}_k(x_i))^2\right]}_{= \frac{1}{p(1-p)}\mathrm{E}\left[(\bar{f}(x_i) - \hat{f}_k(x_i))^2\right]} \to 0\end{aligned}$$

and by the Central Limit Theorem we find

$$\frac{1}{\sqrt{n}} \sum_{k=1}^K \sum_{i\in I_k} \frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i)) \xrightarrow{p} \mathcal{N}\left(0, \operatorname{Var}_\mu\left(\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i))\right)\right)$$

with

$$\begin{aligned}&\operatorname{Var}_\mu\left(\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i))\right) \\ &= \operatorname{Var}_\mu\left(\mathrm{E}_\mu\left[\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i))\middle| x_i\right]\right) + \mathrm{E}_\mu\left[\operatorname{Var}_\mu\left(\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i))\middle| x_i\right)\right] \\ &= \operatorname{Var}_\mu(\bar{\tau}_\mu(x_i)) \\ &\quad + \mathrm{E}_\mu\left[\operatorname{Var}_\mu\left(\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i))\middle| x_i, d_i\right) + \operatorname{Var}_\mu\left(\frac{d_i - p}{p(1-p)}\mathrm{E}[y_i - \bar{f}(x_i)|x_i, d_i]\middle| x_i\right)\right] \\ &= \operatorname{Var}_\mu(\bar{\tau}_\mu(x_i)) + \mathrm{E}_\mu\left[\frac{\operatorname{Var}_\mu(y_i(1))}{p} + \frac{\operatorname{Var}_\mu(y_i(1))}{1-p} + \frac{(1-p)\,\mathrm{E}_\mu[y_i(1)|x_i] + p\,\mathrm{E}_\mu[y_i(0)|x_i]}{p(1-p)}\right],\end{aligned}$$

as claimed. Consistency of the variance estimator follows with $\mathcal{L}^2$ consistency of $\hat{\tau}(\bar{z})$ and $\hat{f}_k$, and the Law of Large Numbers. □

Proof of Proposition A.2. Following the proof of Proposition A.1, we have that

$$\mathrm{E}_\mu[(\sqrt{n}(\hat{\tau}(\bar{z}) - \bar{\tau}_\mu))^2] = \frac{1}{n}\,\mathrm{E}_\mu\left[\left(\sum_{k=1}^{K}\sum_{i\in I_k}\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i)) + \frac{d_i - p}{p(1-p)}(\bar{f}(x_i) - \hat{f}_k(x_i))\right)^2\right]$$

$$= \frac{1}{n}\,\mathrm{E}_\mu\left[\left(\sum_{k=1}^{K}\sum_{i\in I_k}\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i))\right)^2\right] + o(1)$$

$$= \frac{1}{n}\,\mathrm{Var}_\mu\left(\sum_{i=1}^{n}\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i))\right) + o(1) = \underbrace{\mathrm{Var}_\mu\left(\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i))\right)}_{=\sigma^2_\mu(\bar{f})} + o(1).$$

At the same time, for all j and a representative $i \in I_k$,

$$\mathrm{E}_\mu\left[\left(\frac{d_i - p}{p(1-p)}(y_i - \hat{f}(x_i))\right)^2\middle|\hat{f}_k\right] = \mathrm{E}_\mu\left[\left(\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i) + \bar{f}(x_i) - \hat{f}(x_i))\right)^2\middle|\hat{f}_k\right]$$

$$= \mathrm{E}_\mu\left[\left(\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i))\right)^2\right] + \mathrm{E}_\mu\left[\left(\frac{d_i - p}{p(1-p)}(\bar{f}(x_i) - \hat{f}(x_i))\right)^2\middle|\hat{f}_k\right]$$

$$+ 2\,\mathrm{E}_\mu\left[\left(\frac{d_i - p}{p(1-p)}\right)^2(y_i - \bar{f}(x_i))(\bar{f}(x_i) - \hat{f}(x_i))\middle|\hat{f}_k\right]$$

with

$$\mathrm{E}_\mu\left[\left(\frac{d_i - p}{p(1-p)}(\bar{f}(x_i) - \hat{f}(x_i))\right)^2\middle|\hat{f}_k\right] = \frac{1}{p(1-p)}\,\mathrm{E}_\mu\left[(\bar{f}(x_i) - \hat{f}(x_i))^2\middle|\hat{f}_k\right] \xrightarrow{\mathcal{L}^1} 0$$

$$\mathrm{E}_\mu\left[\left(\frac{d_i - p}{p(1-p)}\right)^2(y_i - \bar{f}(x_i))(\bar{f}(x_i) - \hat{f}(x_i))\middle|\hat{f}_k\right]$$

$$\leq \underbrace{\sqrt{\mathrm{E}_\mu\left[\left(\frac{d_i - p}{p(1-p)}(\bar{f}(x_i) - \hat{f}(x_i))\right)^2\middle|\hat{f}_k\right]}}_{\xrightarrow{p}0}\underbrace{\sqrt{\mathrm{E}_\mu\left[\left(\frac{d_i - p}{p(1-p)}(y_i - \bar{f}(x_i))\right)^2\right]}}_{\to\sigma_\mu(\bar{f})} \xrightarrow{p} 0,$$

which establishes the first result. For establishing the same result on average over $\bar{pi}$, note that under the assumptions we can apply the same arguments under the expectations with respect to $\bar{\pi}$, where we note that the assumptions imply $\mathrm{E}_{\bar{\pi}}[(\hat{f}_k(x_i) - f_\mu(x_i))^2] \to 0$ by dominated convergence. □

Proof of Proposition A.3. In the first case, for $t = \hat{\tau}(\bar{z}), t_0 = \bar{\tau}_\mu, |\tilde{t} - t_0| \leq |t - t_0|$,

$$\begin{aligned} n\,\mathrm{E}_\mu[U(t) - U(t_0)] &= n\,\mathrm{E}_\mu[U'(t_0)(t-t_0) + \frac{1}{2}U''(t_0)(t-t_0)^2 + \frac{1}{6}U'''(\tilde{t})(t-t_0)^3] \\ &= nU'(t_0)\underbrace{\mathrm{E}_\mu[t - t_0]}_{=0} + \frac{1}{2}U''(t_0)\underbrace{\mathrm{E}_\mu[n(t-t_0)^2]}_{\to\sigma^2_\mu(\bar{f})} + \frac{1}{6}\frac{1}{\sqrt{n}}\underbrace{\mathrm{E}_\mu[U'''(\tilde{t})(\sqrt{n}(t-t_0))^3]}_{\text{bounded}} \\ &= \frac{1}{2}U''(t_0)\sigma^2_\mu(\bar{f}) + o(1) \end{aligned}$$

In the second case, also writing $s = \hat{\sigma}^2(\bar{z}), s_0 = \sigma^2_\mu(\bar{f}), (\tilde{s} - s_0)^2 + (\tilde{t} - t_0)^2 \leq (s - s_0)^2 + (t - t_0)^2$,

$$\mathrm{E}_\mu[V(t,s)] = \mathrm{E}_\mu\left[V(t_0, s_0) + \frac{\partial V(\tilde{t}, \tilde{s})}{\partial(t,s)}\begin{pmatrix} t - t_0 \\ s - s_0 \end{pmatrix}\right] = V(t_0, s_0) + o(1),$$

where $\hat{\sigma}^2(\bar{z}) \xrightarrow{\mathcal{L}^1} \sigma^2_\mu(\bar{f})$ and $\hat{\tau}(\bar{z}) \xrightarrow{\mathcal{L}^1} \bar{\tau}_\mu$ under the conditions of Proposition A.1. □

Proof of Corollary A.1. Under the assumptions, the equivalence from Proposition A.3 translates into asymptotic equivalence to $\mathrm{E}_{\bar{\pi}}[r^U_\mu(\hat{\tau})]$ and $\mathrm{E}_{\bar{\pi}}[r^V_\mu(\hat{\tau})]$ respectively, by the same arguments as in the proof of Proposition A.3 extended to expectations with respect to $\bar{\mu}$ under the assumptions of the second part of Proposition A.2. These average risks, as well as the limit $n\,\mathrm{E}_{\bar{\pi}}[(\hat{\tau}(\bar{z}) - \bar{\tau}_\mu)^2] \to \mathrm{E}_{\bar{\pi}}[\sigma^2_\mu(\bar{f})]$, are all minimized by choosing $\hat{f}^*_k$ as in the statement of the corollary that are consistent for f^*_μ, since this choice minimizes $\sigma^2_\mu(\bar{f})$ and thus also $r^U_\mu(\hat{\tau})$ and $r^V_\mu(\hat{\tau})$ μ-almost surely. □

B Minimax Optimality of Practical Sample-Splitting Plans

The sample-splitting representation of estimators with fixed bias implies that the investigator is not allowed to use the outcome and treatment assignment of a unit to construct its adjustment, which may require extensive pre-specification, ex-ante optimization, and ex-post computation of regression adjustments. In Section 4.2, I discuss alternatives based on K-fold sample splitting and consider the use of simple prediction solutions as adjustments, focussing on the sampling model from Section 4.1 for concreteness. Here, I argue that the simplifications to simple, unbiased K-fold estimators and sequential access protocols are still minimax optimal within the main setup and notation of Section 1. Throughout, I assume that the sample is partitioned into folds $\{1, \ldots, n\} = \bigcup_{k=1}^K I_k$. Since I allow for the case of $K = n$ folds, this includes the case of leave-one-out estimation of regression adjustments as a special case.

Specifically, I consider the following restrictions of the unbiased estimators (1):

– A first practical restriction is to assume that adjustments are constructed using K-folds cross-fitting (rather than leave-one-out estimation), implying a restriction of unbiased estimators to

$$\hat{\tau}(z) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\frac{d_i - p}{p(1-p)}(y_i - \phi_i(z_{-I_k})) \tag{18}$$

with adjustments of units $i \in I_k$ that only depend on data z_j from units $j \notin I_k$ from other folds. Among such estimators, the investigator can then minimize the average loss in (2).

– A further restriction is to assume that the adjustments in (18) are obtained from simple posterior expectations

$$\phi_i(z_{-I_k}) = \mathrm{E}_\pi[\phi_i^*|z_{-I_k}] = \mathrm{E}_\pi[(1-p)y_i(1) + py_i(0)|z_{-I_k}] \tag{19}$$

of the oracle adjustments $\phi_i^* = (1-p)y_i(1) + py_i(0)$, which can be obtained by solving the prediction problem $\phi_i(z_{-I_k}) = \arg\min_{\hat{y}\in\mathbb{R}} \mathrm{E}_\pi[w(d_i)(\hat{y}-y_i)^2|z_{-I_k}$ as in Section 4.2. While these adjustments are not generally optimal, they represent a natural analog of the infeasible oracle adjustments.

– A third restriction that ensures that the investigator cannot introduce any bias, while also avoiding extensive pre-specifiction, is to follow the approach of Anderson and Magruder (2017) and give the investigator sequential access to the data. Specifically, at each successive step $\ell \in \{1, \ldots, K\}$, I assume that the investigator chooses adjustments $\phi_i(z_{J_\ell})$ for the units $i \in I_\ell$ using data only from observations $J_\ell = \bigcup_{k<\ell} I_k$, yielding an estimator

$$\hat{\tau}(z) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\frac{d_i - p}{p(1-p)}(y_i - \phi_i(z_{J_\ell})). \tag{20}$$

In this case, optimal adjustments take the form

$$\phi_i(z_{J_\ell}) = \mathrm{E}_\pi[\phi_i^*|z_{J_\ell}] = \mathrm{E}_\pi[(1-p)\, y_i(1) + p\, y_i(0)|z_{J_\ell}] \tag{21}$$

similar to (19).

While none of these restrictions lead to variance-minimizing unbiased estimators (with the exception of the first restriction in the leave-one-out case $K = n$), they all represent minimax optimal restrictions for worst-case priors similar to the general unbiased restriction in Proposition 2.

Proposition B.1 (Minimax optimality of sequential-access protocols and optimal adjustments)**.** *Assume outcomes are from* $\mathcal{Y} = \mathbb{R}$*, that Assumption 1 holds with*

iid randomization, and that Assumptions 2–4 hold. Sor some ς^2*, let* $\Delta_{\varsigma^2} = \{\pi \in \Delta(\Theta)$ *with finite support;* $\|\operatorname{Var}_\pi(\theta)\|_2 \leq \varsigma^2\}$*, as in Proposition 2, and also write* $\Delta^*_{\varsigma^2} = \{\pi \in \Delta(\Theta); \mathrm{E}_\pi[\|\theta\|^2] < \infty, \|\operatorname{Var}_\pi(\theta)\|_2 \leq \varsigma^2\}$*, which drops the requirement of finite support. Then for each* $K \geq 1$ *the following restrictions each solve the designer's problem for worst-case priors from* Δ_{ς^2} *or from* $\Delta^*_{\varsigma^2}$ *in the sense of Definition 2:*

1. *A restriction to estimators of the leave-one-out form in (1);*
2. *A restriction to* K*-fold estimators of the form in (18);*
3. *A restriction to 2-fold estimators of the form in (18) with* $K = 2$ *and with the (potentially suboptimal) adjustments from (19);*
4. *A restriction to sequential-access estimators of the form in (20).*

Furthermore, the optimal adjustments that minimize (19) among sequential-estimators estimators as in (20) take the form from (21) of conditional expectations of the oracle adjustments given the respective training data z_{J_ℓ} *and the prior.*

This proposition shows that each of these restrictions is still optimal in the worst case. It also bridges a gap between finite and infinite support in the minimax results from Proposition 2 and in the representation of unbiased estimators from Corollary 1: while the minimax results and representation there require finite support, the result here shows that minimaxity extends to unbounded support and that leave-one-out estimators are still minimax optimal with infinite support (even though the characterization result in Proposition 2 leaves open whether there are additional unbiased estimators in this case).

Proof of Proposition B.1. For minimax optimality in each case, it is sufficient to bound the worst-case average loss by the minimax bound from the proof of Proposition 2, which is $\frac{\frac{1-p}{p}+\frac{p}{1-p}}{n}\varsigma^2$. If this bound holds for the larger class of priors $\Delta^*_{\varsigma^2}$ (which does not have a restriction to finite support), then it must be minimax optimal for both classes of priors.

For restrictions to estimators to the leave-one-out form in (1), K-fold estimators of the form in (18), and sequential-access estimators of the form in (20), the investigator can choose adjustments $\phi_i = \mathrm{E}_\pi[\phi_i^*]$, since they do not depend on other units and thus fulfill the restrictions in all three cases. Average loss is then bounded by the performance using these adjustments, since the adjustments chosen by the investigator can only improve average loss (even in the case of a worst-case prior). Plugging into

(2) leads to the upper bound

$$\mathrm{E}_\pi\left(\frac{1}{n}\sum_{i=1}^n \frac{d_i-p}{p(1-p)}(\phi_i^*-\mathrm{E}_\pi[\phi_i^*])\right)^2 = \frac{1}{n^2}\sum_{i=1}^n \frac{1}{p(1-p)}\mathrm{E}_\pi[(\phi_i^*-\mathrm{E}_\pi[\phi_i^*])^2]$$
$$= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}_\pi((1-p)y_i(1)+py_i(0)) \le \frac{1}{n^2}\sum_{i=1}^n \frac{(1-p)^2+p^2}{p(1-p)}\varsigma^2 = \frac{\frac{1-p}{p}+\frac{p}{1-p}}{n}\varsigma^2,$$

as desired.

For the K-fold estimator with the (suboptimal) adjustments from (19), write

$$A_k = \frac{1}{n}\sum_{i\in I_k}\frac{d_i-p}{p(1-p)}(\phi_i^*-\mathrm{E}_\pi[\phi_i^*]),$$
$$B_k = \frac{1}{n}\sum_{i\in I_k}\frac{d_i-p}{p(1-p)}(\phi_i^*-\mathrm{E}_\pi[\phi_i^*|z_{-k}]),$$
$$C_k = \frac{1}{n}\sum_{i\in I_k}\frac{d_i-p}{p(1-p)}(\mathrm{E}_\pi[\phi_i^*|z_{-k}]-\mathrm{E}_\pi[\phi_i^*]),$$

for which average loss of the K-fold estimator is $\mathrm{E}_\mu\left[\left(\sum_{k=1}^K B_k\right)^2\right] = \mathrm{Var}_\mu\left(\sum_{k=1}^K B_k\right)$, and for $d_{I_k} = (d_i)_{i\in I_k}$ we have that

$$B_k = A_k - \mathrm{E}_\pi[A_k|z_{-I_k}, d_{I_k}], \qquad C_k = \mathrm{E}_\pi[A_k|z_{-I_k}, d_{I_k}] - \mathrm{E}_\pi[A_k|d_{I_k}],$$
$$A_k = B_k + C_k = A_k - \mathrm{E}_\pi[A_k|d_{I_k}].$$

By construction, for all k and $k' \neq k$,

$$\mathrm{Cov}_\pi(B_k, C_k) = \mathrm{Cov}_\pi(\mathrm{E}_\pi[B_k|z_{-I_k}, d_{I_k}], C_k) = 0,$$
$$\mathrm{Cov}_\pi(A_k, A_{k'}) = \mathrm{E}_\pi\underbrace{\mathrm{Cov}_\theta(A_k, A_{k'})}_{=0} + \mathrm{Cov}_\pi(\underbrace{\mathrm{E}_\theta[A_k]}_{=0}, \mathrm{E}_\theta[A_{k'}]) = 0,$$
$$\mathrm{Cov}_\pi(A_k, C_{k'}) = \mathrm{E}_\pi\underbrace{\mathrm{Cov}_\theta(A_k, C_{k'})}_{=\mathrm{Cov}_\theta(\mathrm{E}_\theta[A_k|d_{I_{k'}}], C_{k'}) = \mathrm{Cov}_\theta(0, C_{k'}) = 0} + \mathrm{Cov}_\pi(\underbrace{\mathrm{E}_\theta[A_k]}_{=0}, \mathrm{E}_\theta[C_{k'}]) = 0,$$
$$\mathrm{Cov}_\pi(B_k, B_{k'}) = \mathrm{Cov}_\pi(A_k - C_k, A_{k'} - C_{k'}) = \mathrm{Cov}_\pi(C_k, C_{k'}).$$

Hence, the average loss of the K-fold estimator is bounded by

$$\begin{aligned}
&\mathrm{Var}_\mu\left(\sum_{k=1}^K B_k\right) = \sum_{k=1}^K \mathrm{Var}_\mu(B_k) + \sum_{k\neq k'} \mathrm{Cov}_\mu(B_k, B_{k'}) = \sum_{k=1}^K \mathrm{Var}_\mu(B_k) + \sum_{k\neq k'} \mathrm{Cov}_\mu(C_k, C_{k'}) \\
&= \sum_{k=1}^K (\mathrm{Var}_\mu(B_k) - \mathrm{Var}_\mu(C_k)) + \mathrm{Var}_\mu\left(\sum_{k=1}^K C_k\right) \\
&\leq \sum_{k=1}^K (\mathrm{Var}_\mu(B_k) - \mathrm{Var}_\mu(C_k)) + K\sum_{k=1}^K \mathrm{Var}_\mu\left(C_k\right) \\
&= \sum_{k=1}^K (\mathrm{Var}_\mu(B_k) + \mathrm{Var}_\mu(C_k)) + (K-2)\sum_{k=1}^K \mathrm{Var}_\mu\left(C_k\right) \\
&= \sum_{k=1}^K \mathrm{Var}_\mu(A_k) + (K-2)\sum_{k=1}^K \mathrm{Var}_\mu\left(C_k\right) = \underbrace{\mathrm{Var}_\mu\left(\sum_{k=1}^K A_k\right)}_{=\mathrm{E}_\pi\left(\frac{1}{n}\sum_{i=1}^n \frac{d_i-p}{p(1-p)}(\phi_i^* - \mathrm{E}_\pi[\phi_i^*])\right)^2} + (K-2)\sum_{k=1}^K \mathrm{Var}_\mu\left(C_k\right) \\
&\leq \frac{\frac{1-p}{p} + \frac{p}{1-p}}{n}\varsigma^2 + \frac{K-2}{n^2 p(1-p)} \sum_{k=1}^K \sum_{i\in I_k} \mathrm{E}_\pi[(\mathrm{E}_\pi[\phi_i^*|z_{-k}] - \mathrm{E}_\pi[\phi_i^*])^2].
\end{aligned}$$

In particular, for $K = 2$, the average loss of the estimator with simple adjustments, $\phi_i = \mathrm{E}_\pi[\phi_i^*]$, is an upper bound and the restrictions to 2-fold estimators with such adjustments ensures the minimax bound, even though the adjustments may be sub-optimal.

To show that for estimators of the form (20) the optimal regression adjustments take the claimed form (21), note that the average loss from (2) becomes

$$\begin{aligned}
&\mathrm{E}_\pi\left[\left(\frac{1}{n}\sum_{k=1}^K \sum_{i\in I_k} \frac{d_i-p}{p(1-p)}(\phi_i^* - \phi_i(z_{J_k}))\right)^2\right] = \mathrm{E}_\pi \,\mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^K \sum_{i\in I_k} \frac{d_i-p}{p(1-p)}(\phi_i^* - \phi_i(z_{J_k}))\right) \\
&= \frac{1}{n^2}\sum_{k=1}^K \mathrm{E}_\pi \,\mathrm{Var}_\theta\Bigg(\underbrace{\sum_{i\in I_k} \frac{d_i-p}{p(1-p)}(\phi_i^* - \phi_i(z_{J_k}))}_{\text{independent mean-zero conditional on } z_{J_k}}\Bigg) \\
&\quad + \frac{2}{n^2}\sum_{k>k'} \mathrm{E}_\pi \,\mathrm{Cov}_\theta\Bigg(\underbrace{\sum_{i\in I_k} \frac{d_i-p}{p(1-p)}(\phi_i^* - \phi_i(z_{J_k}))}_{\text{mean-zero conditional on } z_{J_k}}, \underbrace{\sum_{i\in I_{k'}} \frac{d_i-p}{p(1-p)}(\phi_i^* - \phi_i(z_{J_{k'}}))}_{z_{J_k}\text{-measurable (since } I_{k'}\subseteq J_k \text{ for } k>k')}\Bigg) \\
&= \frac{1}{n^2}\sum_{k=1}^K \sum_{i\in I_k} \mathrm{E}_\pi \,\mathrm{Var}_\theta\left(\frac{d_i-p}{p(1-p)}(\phi_i^* - \phi_i(z_{J_k}))\right) = \frac{1}{n^2 p(1-p)}\sum_{i=1}^n \mathrm{E}_\pi[(\phi_i^* - \phi_i(z_{J_k}))^2],
\end{aligned}$$

with minimizer $\phi_i(z_{J_k}) = \mathrm{E}_\pi[\phi_i^*|z_{J_k}] = \arg\min_{\hat\phi} \mathrm{E}_\pi[(\phi_i^* - \hat\phi)^2|z_{J_k}]$, as claimed. □

C General Solution of Optimal Biases and Second-Best Estimators

Theorem 1 shows that fixing the biases β_θ of the estimator is a minimax optimal solution of the designer's delegation problem, but it does not answer the question what these biases should be. In this section, under the assumptions of Theorem 1, I provide a general high-level solution to the designer's problem of setting biases β_η (given the hyperprior η) as well as the investigator's problem of choosing an optimal estimator $\hat\tau$ given these biases (and prior π). These results apply to any estimand $\tau_\theta \in \mathbb{R}$ and (known) randomization $z|\theta$ with $\theta \in \Theta, z \in \mathcal{Z}$ with $|\Theta|, |\mathcal{Z}| < \infty$. They are helpful to derive general properties of the biases, such as those presented in Proposition 1 in the main article.

C.1 A Transformation to Zero-Bias Adjustments

I first transform the designer's problem of choosing biases β_θ and the investigator's problem of choosing among estimators with $\mathrm{E}_\theta[\hat\tau(z)] = \tau_\theta + \beta_\theta$. The problems of designer and investigator are equivalent to the designer with hyperprior η choosing an estimator $\hat\gamma^\eta : \mathcal{Z} \to \mathbb{R}$, and the investigator with prior π then picking and adjustment $\hat\delta^\pi : \mathcal{Z} \to \mathbb{R}$ subject to the zero-bias condition $\mathrm{E}_\theta[\hat\delta^\pi(z)] = 0\ \forall\theta \in \Theta$. The final estimator is $\hat\tau^{\eta\to\pi} : \mathcal{Z} \to \mathbb{R}$ given by $\hat\tau^{\eta\to\pi}(z) = \hat\gamma^\eta(z) + \hat\delta^\pi(z)$. This construction is equivalent to the designer choosing the biases $\beta_\theta = \mathrm{E}_\theta[\hat\gamma^\eta(z)] - \tau_\theta$ and the investigator choosing among estimators with these biases.

I next consider optimal choices of $\hat\gamma^\eta$ and $\hat\delta^\pi$. Since fixing the biases (or equivalently, restricting $\hat\delta^\pi$ to be zero bias) aligns preferences, we can assume that both designer and investigator aim to minimize the average of the risk

$$r_\theta^D(\hat\tau) = \mathrm{E}_\theta[(\hat\tau^{\eta\to\pi}(z) - \tau_\theta)^2] = (\mathrm{E}_\theta[\hat\tau^{\eta\to\pi}(z)] - \tau_\theta)^2 + \mathrm{Var}_\theta(\hat\tau^{\eta\to\pi}(z)).$$

Investigator and designer then (simultaneously) solve

$$\hat\delta^\pi \in \arg\min_{\hat\delta\in\mathcal{C}^0} \mathrm{E}_\pi r_\theta^D(\hat\tau) = \mathrm{E}_\pi[(\hat\gamma^\eta(z) + \hat\delta(z) - \tau_\theta)^2] = \mathrm{E}_\pi\,\mathrm{Var}_\theta(\hat\gamma^\eta(z) + \hat\delta(z)),$$

$$\hat\gamma^\eta \in \arg\min_{\hat\gamma:\mathcal{Z}\to\mathbb{R}} \mathrm{E}_\eta r_\theta^D(\hat\tau) = \mathrm{E}_\eta[(\hat\gamma(z)+\hat\delta^\pi(z)-\tau_\theta)^2] = \mathrm{E}_\eta[(\mathrm{E}_\theta[\hat\gamma(z)]-\tau_\theta)^2] + \mathrm{E}_\eta\,\mathrm{Var}_\theta(\hat\gamma(z)+\hat\delta^\pi(z)),$$

where $\mathcal{C}^0 = \{\hat\delta : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_\theta[\hat\delta(z)] = 0\ \forall\theta \in \Theta\}$ denotes mean-zero estimators.

Before constructing a solution, I note some general properties of optimal biases and estimators.

Remark C.1 (Uniqueness of second-best estimators)**.** *If π has full support η-almost surely, then any biases $\beta_\theta = \mathrm{E}_\theta[\hat{\gamma}^\eta(z)]$, adjustments $\hat{\delta}^\pi$ given $\hat{\delta}^\pi$, and compound estimator $\hat{\tau}^{\eta\to\pi}$ that solve the above optimization problem are a.s. unique.*

Note that $\hat{\gamma}^\eta$ is not generally unique, as choices that do not differ in their biases $\beta_\theta = \mathrm{E}_\theta[\hat{\gamma}^\eta(z)]$ yield the same compound estimator $\hat{\tau}^{\eta\to\pi}$.

Proof of Remark C.1. For uniqueness of $\hat{\tau}^{\eta\to\pi}$, assume that there are two optimal compound estimators with $\mathrm{E}_\eta[(\hat{\tau}_1^{\eta\to\pi}(z) - \hat{\tau}_2^{\eta\to\pi}(z))^2] > 0$. Consider the compound estimator $\hat{\tau}_*^{\eta\to\pi}(z) = \hat{\gamma}_*^\eta(z) + \hat{\delta}_*^\pi(z)$ with $\hat{\gamma}_*^\eta(z) = \frac{1}{2}\hat{\gamma}_1^\eta(z) + \frac{1}{2}\hat{\gamma}_2^\eta(z)$ and $\hat{\delta}_*^\pi(z) = \frac{1}{2}\hat{\delta}_1^\pi(z) + \frac{1}{2}\hat{\delta}_2^\pi(z)$ (which is feasible since $\hat{\delta}_*^\pi(z)$ is still mean-zero). Then

$$\mathrm{E}_\eta[(\hat{\tau}_*^{\eta\to\pi}(z) - \tau_\theta)^2] + \frac{\mathrm{E}_\eta[(\hat{\tau}_1^{\eta\to\pi}(z) - \hat{\tau}_2^{\eta\to\pi}(z))^2]}{4} = \frac{\mathrm{E}_\eta[(\hat{\tau}_1^{\eta\to\pi}(z) - \tau_\theta)^2 + (\hat{\tau}_2^{\eta\to\pi}(z) - \tau_\theta)^2]}{2}$$

and thus $\mathrm{E}_\eta[(\hat{\tau}_*^{\eta\to\pi}(z) - \tau_\theta)^2] < \max_j \mathrm{E}_\eta[(\hat{\tau}_j^{\eta\to\pi}(z) - \tau_\theta)^2]$. Hence, for one of the j the choice of $\hat{\gamma}_j^\eta$ or of $\hat{\tau}_j^\pi$ must be suboptimal, which is a contradiction. $\hat{\tau}^{\eta\to\pi}$ is therefore a.s. unique, and therefore also the adjustments $\delta^\pi(z) = \hat{\tau}^{\eta\to\pi}(z) - \hat{\gamma}^\eta(z)$. The biases are unique at $\beta_\theta = \mathrm{E}_\theta[\hat{\tau}^{\eta\to\pi}(z)]$. □

C.2 An Explicit Solution Based on Quadratic Forms

Since $|\mathcal{Z}| < \infty$, we can express $\hat{\tau}^{\eta\to\pi}, \hat{\gamma}^\eta, \hat{\delta}^\pi$ as vectors in the finite-dimensional real vector space $\mathbb{R}^{\mathcal{Z}}$. Writing $P \in [0,1]^{\Theta\times\mathcal{Z}}$ for the matrix that has as its rows the probability distributions over $\mathcal{Z}$ for each $\theta \in \Theta$ induced by randomization, we can write $P\hat{\tau} \in \mathbb{R}^\Theta$ for the vector of expectations $\mathrm{E}_\theta[\hat{\tau}(z)]$ for an estimator $\hat{\tau} \in \mathbb{R}^{\mathcal{Z}}$.

The restriction $\delta \in \mathcal{C}^0$ can be expressed as the linear restriction $P\hat{\delta} = \mathbf{0}$. We now choose an orthonormal matrix $Q = (A, B) \in \mathbb{R}^{\mathcal{Z}\times|\mathcal{Z}|}$ such that $\hat{\delta} = \mathbf{0}$ if and only if $\hat{\delta} = B\hat{\beta}$ (and we can write $\hat{\beta} = B'\hat{\delta}$).[12] We have that $B \in \mathbb{R}^{\mathcal{Z}\times b}$ with $b = |\mathcal{Z}| - \mathrm{rank}(P)$, and each of the b columns of B correspond to a mean-zero estimator $B_j : \mathcal{Z} \to \mathbb{R}$,

[12]The specific structure of B is not relevant to this section, and the argument applies more generally than for the specific randomization scheme in the main article. The results in Section 3 provide one specific characterization for the span of B (or equivalently, kernel of P) in the case of the specific randomization considered there.

allowing us to write $\hat{\delta}(z) = \sum_{j=1}^{r} B_j(z)\,\hat{\beta}_j$. When choosing $\hat{\delta}^\pi$ subject to the zero-expectation constraint, we can equivalently assume that the investigator chooses $\hat{\beta}^\pi$ without constraint to obtain $\hat{\delta}^\pi = B\hat{\beta}^\pi$.

We can similarly express the designer choice $\hat{\gamma}^\eta$ in terms of the basis Q as $\hat{\gamma}^\eta = A\hat{\alpha}^\eta + B\hat{\beta}_0^\eta$. Since the investigator can choose and mean-zero estimator $\hat{\delta}^\pi = B\hat{\beta}^\pi$ to achieve $\hat{\tau}^{\eta\to\pi} = \hat{\gamma}^\eta + \hat{\delta}^\pi = A\hat{\alpha}^\eta + B(\hat{\beta}_0^\eta + \hat{\beta}^\pi)$, the designer's choice of $B'\hat{\gamma}^\eta = \hat{\beta}_0^\eta$ is inconsequential, since it will be overridden by the investigator's choice of $\hat{\beta}^\pi$. We can therefore assume wlog that the designer chooses $\hat{\alpha}^\eta$ to obtain $\hat{\gamma}^\eta = A\hat{\alpha}^\eta$.

We first consider the variance term and the investigator solution. Since $\mathrm{E}_\pi \mathrm{Var}_\theta(\hat{\tau}(z))$ for given π and $\hat{\tau}$ is a quadratic non-negative polynomial that is zero for $\hat{\tau}(z) \equiv 0$ (equivalently, $\hat{\tau} = \mathbf{0}$), we can write it as the quadratic form $\mathrm{E}_\pi \mathrm{Var}_\theta(\hat{\tau}(z)) = \hat{\tau}' V_\pi \hat{\tau}$ with $V_\pi \in \mathbb{R}^{\mathcal{Z}\times\mathcal{Z}}$ symmetric positive-semidefinite. Given $\hat{\gamma}^\eta = B\hat{\beta}$, the investigator chooses an mean-zero estimator $\hat{\delta} = B\hat{\beta}$ to minimize $\mathrm{E}_\pi \mathrm{Var}_\theta(\hat{\tau}(z))$ over $\hat{\tau} = \hat{\gamma}^\eta + B\hat{\beta}$, that is, $\hat{\beta}^\pi \in \arg\min_{\hat{\beta}} (A\hat{\alpha}^\eta + B\hat{\beta})' V_\pi (A\hat{\alpha}^\eta + B\hat{\beta})$. The unique solution of this convex minimization problem is located by the first-order condition $(B'V_\pi B)\hat{\beta} = -B'V_\pi A\hat{\alpha}^\eta$ where $B'V_\pi B$ is invertible by uniqueness and we can therefore express

$$\begin{aligned}\hat{\delta}^\pi = B\hat{\beta}^\pi &= -B(B'V_\pi B)^{-1}B'V_\pi A\hat{\alpha}^\eta \\ \hat{\gamma}^\eta + \hat{\delta}^\pi &= \left(\mathbb{I} - B(B'V_\pi B)^{-1}B'V_\pi\right) A\hat{\alpha}^\eta,\end{aligned}$$

where the latter can be seen as a specific projection onto a π-dependent subspace of estimators with the same expectations as $\hat{\gamma}^\eta$. Given the designer's choice $\hat{\alpha}^\eta$, this investigator choice yields average (over π) variance

$$\begin{aligned}\mathrm{E}_\pi \mathrm{Var}_\theta(\hat{\gamma}^\eta(z) + \hat{\delta}^\pi(z)) &= \hat{\alpha}^{\eta\prime} A' \left(\mathbb{I} - V_\pi B(B'V_\pi B)^{-1}B'\right) V_\pi \left(\mathbb{I} - B(B'V_\pi B)^{-1}B'V_\pi\right) A\hat{\alpha}^\eta \\ &= \hat{\alpha}^{\eta\prime} \underbrace{A' \left(V_\pi - V_\pi B(B'V_\pi B)^{-1}B'V_\pi\right) A}_{=W_\pi} \hat{\alpha}^\eta = \hat{\alpha}^{\eta\prime} W_\pi \hat{\alpha}^\eta\end{aligned}$$

where we know that W_π is symmetric positive-semidefinite (since the average variance is non-negative, no matter the designer choice).

We next consider the bias term and the designer solution. For a given η, $\mathrm{E}_\eta \mathrm{E}_\pi[(\mathrm{E}_\theta[\hat{\gamma}(z)] - \tau_\theta)^2]$ for $\hat{\gamma}(z) = \sum_{j=1}^{\mathrm{rank}(P)} A_j \hat{\alpha}_j$ is a quadratic polynomial in the $\hat{\alpha}_j$, non-negative for all $\hat{\alpha}$, and zero for $\hat{\alpha}^* = A'\hat{\tau}^*$, where $\hat{\tau}^*$ is the difference-in-averages (or generally any unbiased estimator, for which $\mathrm{E}_\theta[\hat{\tau}^*(z)] = \tau_\theta \; \forall \theta \in \Theta$). We can therefore write

$$\mathrm{E}_\eta \mathrm{E}_\pi[(\mathrm{E}_\theta[\hat{\gamma}(z)] - \tau_\theta)^2] = (\hat{\alpha} - \hat{\alpha}^*)' U_\eta (\hat{\alpha} - \hat{\alpha}^*) \tag{22}$$

with U_η symmetric positive definite and $\hat{\gamma} = A\hat{\alpha}$.[13] The designer then solves

$$\begin{aligned}\hat{\alpha}^\eta &\in \arg\min_{\hat{\alpha}} \mathrm{E}_\eta \mathrm{E}_\pi[(\mathrm{E}_\theta[\hat{\gamma}(z)] - \tau_\theta)^2] + \mathrm{E}_\eta \mathrm{E}_\pi \mathrm{Var}_\theta(\hat{\gamma}(z) + \hat{\delta}^\pi(z)) \\ &= \arg\min_{\hat{\alpha}} (\hat{\alpha} - \hat{\alpha}^*)' U_\eta (\hat{\alpha} - \hat{\alpha}^*) + \hat{\alpha}' \underbrace{\mathrm{E}_\eta[W_\pi]}_{=W_\eta} \hat{\alpha}\end{aligned}$$

with U_η, W_η symmetric positive semi-definite. Since $\hat{\tau}^{\eta\to\pi}$ is unique, so is $\hat{\alpha}^\eta = A'\hat{\tau}^{\eta\to\pi}$, and the first-order condition $(U_\eta + W_\eta)\hat{\alpha} = U_\eta \hat{\alpha}^*$ (where $U_\eta + W_\eta$ invertible by uniqueness or by positive definiteness) locates this optimal choice $\hat{\alpha}^\eta = (U_\eta + W_\eta)^{-1} U_\eta \hat{\alpha}^*$ which yields the second-best biases and estimator

$$\begin{aligned}\beta^{\eta\to\pi} &= P\hat{\gamma}^\eta - P\hat{\tau}^* = P\left(A - A(U_\eta + W_\eta)^{-1}U_\eta\right)\hat{\alpha}^* = PA(U_\eta + W_\eta)^{-1} W_\eta \hat{\alpha}^* \\ \hat{\tau}^{\eta\to\pi} &= \hat{\gamma}^\eta + \hat{\delta}^\pi = A\hat{\alpha}^\eta + B\hat{\beta}^\pi = (\mathbb{I} - B(B'V_\pi B)^{-1}B'V_\pi)A(U_\eta + W_\eta)^{-1}U_\eta\hat{\alpha}^*.\end{aligned}$$

For the (unique) resulting estimator $\hat{\tau}^{\eta\to\pi}(z) = \hat{\gamma}^\eta(z) + \hat{\delta}^\pi(z)$, which is equivalent to the designer fixing biases β_θ and the investigator with knowledge of the prior π choosing an estimator subject to that restriction, average expectation, bias, and variance contributions to overall average risk are

$$\begin{aligned}&\mathrm{Bias}^2(\hat{\tau}^{\eta\to\pi}) = \mathrm{E}_\eta \mathrm{E}_\pi[(\mathrm{E}_\theta[\hat{\tau}^{\eta\to\pi}(z)] - \tau_\theta)^2] = (\hat{\alpha}^\eta - \hat{\alpha}^*)' U_\eta (\hat{\alpha}^\eta - \hat{\alpha}^*) \\ &= \hat{\alpha}^{*\prime}(\mathbb{I} - U_\eta(U_\eta + W_\eta)^{-1}) U_\eta (\mathbb{I} - (U_\eta + W_\eta)^{-1}U_\eta)\hat{\alpha}^* \\ &= \hat{\alpha}^{*\prime}\sqrt{U_\eta}\left(\mathbb{I} - \sqrt{U_\eta}(U_\eta + W_\eta)^{-1}\sqrt{U_\eta}\right)^2 \sqrt{U_\eta}\hat{\alpha}^*, \\ &\mathrm{Variance}(\hat{\tau}^{\eta\to\pi}) = \mathrm{E}_\eta \mathrm{E}_\pi[\mathrm{Var}_\theta(\hat{\tau}^{\eta\to\pi}(z))] = \hat{\alpha}^{\eta\prime} W_\eta \hat{\alpha}^\eta \\ &= \hat{\alpha}^{*\prime} U_\eta (U_\eta + W_\eta)^{-1} W_\eta (U_\eta + W_\eta)^{-1} U_\eta \hat{\alpha}^* \\ &= \hat{\alpha}^{*\prime}\sqrt{U_\eta}\left(\sqrt{U_\eta}(U_\eta + W_\eta)^{-1}\sqrt{U_\eta}\right)\left(\mathbb{I} - \sqrt{U_\eta}(U_\eta + W_\eta)^{-1}\sqrt{U_\eta}\right)\sqrt{U_\eta}\hat{\alpha}^*,\end{aligned}$$

where total average risk is

$$\begin{aligned}&\mathrm{Risk}(\hat{\tau}^{\eta\to\pi}) = \mathrm{Bias}^2(\hat{\tau}^{\eta\to\pi}) + \mathrm{Variance}(\hat{\tau}^{\eta\to\pi}) \\ &= \hat{\alpha}^{*\prime}\sqrt{U_\eta}\left(\mathbb{I} - \sqrt{U_\eta}(U_\eta + W_\eta)^{-1}\sqrt{U_\eta}\right)\sqrt{U_\eta}\hat{\alpha}^*.\end{aligned} \tag{23}$$

Here, I note that these expressions are functions of the unique parametrization $\hat{\alpha}^*$ of the expectation-contributing target estimator $\hat{\tau}^*$, where the estimand is implicitly defined as $\tau = P\hat{\tau}^* = PA\hat{\tau}^*$, and that all results extend to any other estimand that can be expressed as the expectation of an estimator. In this way, the above

[13] U_η is invertible since $0 = \hat{\alpha}' U_\eta \hat{\alpha} = \mathrm{E}_\eta \mathrm{E}_\pi[\mathrm{E}_\theta^2[\hat{\tau}(z)]]$ with $\hat{\alpha} = A'\hat{\tau}$ implies $\mathrm{E}_\theta[\hat{\tau}(z)] = 0$ for all $\theta \in \Theta$ and therefore $\hat{\tau} = B\hat{\beta}$ for some $\hat{\beta}$ from which we conclude that $\hat{\alpha} = A'B\hat{\beta} = \mathbf{0}$, where we maintain the assumption that π has full support η-a.s.

construction provides a second-best estimator that (weakly) improves the average performance of any estimator with respect to the estimand for which it is unbiased. This characterization allows us to derive general properties of the second-best estimator $\hat{\tau}^{\eta\to\pi}(z) = \hat{\gamma}^{\eta}(z) + \hat{\delta}^{\pi}(z)$ that the designer (who knows the hyperprior η) can achieve by delegating estimation to the investigator (who has access to the prior π), which is done in the proof of Proposition 1.