Why experimenters might not always want to randomize, and what they could do instead

Maximilian Kasy*
Department of Economics, Harvard University

October 22, 2015

Abstract
Suppose that an experimenter has collected a sample as well as baseline information about the units in the sample. How should she allocate treatments to the units in this sample? We argue that the answer does not involve randomization if we think of experimental design as a statistical decision problem. If, for instance, the experimenter is interested in estimating the average treatment effect and evaluates an estimate in terms of the squared error, then she should minimize the expected mean squared error through choice of a treatment assignment. We provide explicit expressions for the expected mean squared error which lead to easily implementable procedures for experimental design.

The objective of controlled trials is to have treatment groups which are similar a priori (balanced), so we can “compare apples with apples.” The expressions for risk derived in this paper provide an operationalization of the notion of balance. The intuition for our non-randomization result is similar to the reasons for not using randomized estimators - adding noise can never decrease risk. There are, however, arguments for randomization outside the decision theoretic framework.

Keywords: experimental design, decision theory, nonparametric Bayesian methods, field experiments

*I thank Alberto Abadie, Ivan Canay, Gary Chamberlain, Raj Chetty, Nathaniel Hendren, Guido Imbens, Larry Katz, Gary King, Michael Kremer, and Don Rubin, as well as seminar participants at the Harvard development retreat, the Harvard Labor economics workshop, the Harvard Quantitative Issues in Cancer Research seminar, the Harvard Applied Statistics Seminar, the UT Austin, Princeton, Columbia, and Northwestern Econometrics seminars, at RAND, and at the 2013 CEME conference at Stanford for helpful discussions.
1 Introduction

Experiments, and in particular randomized experiments, are the conceptual reference point that gives empirical content to the notion of causality. In recent years, actual randomized experiments have become increasingly popular elements of the methodological toolbox in a wide range of social science disciplines. Examples from the recent political science literature abound. Blattman et al. (2014), for instance, provided training in “alternative dispute resolution practices” to residents of a random set of towns in Liberia. Kalla and Broockman (2015) analyze whether the possibility of scheduling a meeting with a congressional office changes depending on whether it is revealed that the person seeking a meeting is a political donor. Findley et al. (2015) study the effect of varying messages to incorporation services in different locations on the possibility of establishing (illegal) anonymous shell corporations. A number of additional examples are discussed below.

Researchers conducting such field experiments in political science (as well as in other fields such as economics, medicine, and other social and bio-medical sciences) are often confronted with variants of the following situation (cf. Morgan and Rubin, 2012). They have selected a random sample from some population and have conducted a baseline survey for the individuals in this sample. Then a discrete treatment is assigned to the individuals in this sample, usually based on some randomization scheme. Finally, outcomes are realized, and the data are used to perform inference on some average treatment effect.

A key question that experimenters have to decide on is how to use covariates from the baseline survey in the assignment of treatments. Intuition and the literature suggest to use stratified randomization conditional on covariates, also known as blocking. Moore (2012), for instance, makes a convincing argument that blocking on continuous as well as discrete covariates is better than full randomization or blocking only on a small number of discrete covariates. We analyze this question – how to use baseline covariates – as a decision problem. The experimenter has to choose a treatment assignment and an estimator, given knowledge of the covariate distribution in the sample. Her objective is to minimize risk based on a loss function such as the mean squared error (MSE) of a point estimator. The decision criteria considered are Bayesian average risk and (conditional) minimax risk.
We show, first, that experimenters might not want to randomize in general. While surprising at first, the basic intuition for this result is simple and holds for any statistical decision problem. The conditional expected loss of an estimator is a function of the covariates and of the treatment assignment. The treatment assignment that minimizes conditional expected loss is in general unique if there are continuous covariates, so that a deterministic assignment strictly dominates all randomized assignments. There are some arguments for randomization which are outside the decision-theoretic framework, in particular the possibility of conducting randomization inference. We will discuss these in section 5.

We next discuss how to implement optimal and near-optimal designs in practice, where near-optimal designs might involve some randomization. The key problem is to derive tractable expressions for the expected mean squared error of estimators for the average treatment effect, given a treatment assignment. Once we have such expressions we can numerically search for the best assignment, or for a set of assignments that are close to optimal. In order to calculate the expected MSE, we need to specify a prior distribution for the conditional expectation of potential outcomes given covariates. We provide simple formulas for the expected MSE for a general class of non-parametric priors.

Our recommendation not to randomize raises the question of identification. We show that conditional independence of treatment and potential outcomes given covariates still holds for the deterministic assignments considered, under the usual assumption of independent sampling. Conditional independence only requires a controlled trial (CT), not a randomized controlled trial (RCT).

To gain some intuition for our non-randomization result, note that in the absence of covariates the purpose of randomization is to pick treatment and control groups which are similar before they are exposed to different treatments. Formally, we would like to pick groups which have the same (sample) distribution of potential outcomes. Even with

\[1\] If experimenters have a preference for randomization for reasons outside the decision problem considered in the present paper, a reasonable variant of the procedure suggested here would be to randomize among a set of assignments which are “near-minimizers” of risk. If we are worried about manipulation of covariates, in particular, a final coin-flip which possibly switches treatment and control groups might be helpful. I thank Michael Kremer for this suggestion.
covariates observed prior to treatment assignment, it is not possible to make these groups identical in terms of potential outcomes. We can, however, make them as similar as possible in terms of covariates. Allowing for randomness in the treatment assignment to generate imbalanced distributions of covariates can only hurt the balance of the distribution of potential outcomes. The analogy to estimation might also be useful to understand our non-randomization result. Adding random (mean 0) noise to an estimator does not introduce any bias. But it is never going to reduce the mean squared error of the estimator.

The purpose of discussing tractable non-parametric priors – and one of the main contributions of this paper – is to operationalize the notion of “balance.” In general, it will not be possible to obtain exactly identical distributions of covariates in the treatment and control group. When picking an assignment, we have to trade off balance across various dimensions of the joint distribution of covariates. Picking a prior distribution for the conditional expectation of potential outcomes, as well as a loss function, allows one to calculate an objective function (Bayesian risk) which performs this trade-off in a coherent and principled way.

The rest of this paper is structured as follows. Section 1.1 reviews some of the relevant literature. Section 2 discusses a motivating example that anticipates the main intuition for our later results. Section 3 provides a brief review of statistical decision theory and makes the general argument that optimal solutions to decision problems do not involve randomization. Section 4 discusses our proposed procedure, which is based on minimizing the expected mean squared error of treatment assignments, where the expectation is taken over a nonparametric prior. Section 6 concludes, and appendix A contains some proofs. An online appendix contains further material, including a discussion of popular priors, a discussion of frequentist inference, a review of the experimental design literature, some simulations, and an empirical application.

1.1 Literature

There is a large and old literature on experimental design in statistics, going back at least to Smith (1918), and receiving broader attention since Kiefer and Wolfowitz (1959)
and related contributions. A good general introduction to the theory of experimental design can be found in Cox and Reid (2000); a formal treatment of the theory of optimal design is given by Shah and Sinha (1989). The fact that deterministic designs might be optimal in the presence of covariates was noted by Atkinson (1982) in the context of a parametric model and sequential experimental design. Some general discussions on the role of randomization in experiments took place a few decades ago, see in particular Stone (1969), Kempthorne (1977), and Rubin (1978). We will revisit some of their arguments, concerning randomization inference, manipulability, and the role of parametric models, in section 5 below. The role of covariates in the analysis of experiments is also discussed by Imai et al. (2008).

Morgan and Rubin (2012) argue that experimenters should re-randomize until satisfactory covariate balance is achieved. Their argument is similar in spirit to ours. Moore (2012) makes an argument for blocking making use of the rich set of baseline covariates often available in field experiments. I very much agree with this argument; one way to think of the present paper is that it provides a formal foundation for this argument and takes it to its logical conclusion. Simultaneous treatment assignment to all units is not always feasible, particularly in settings where participants arrive sequentially. Such settings are discussed by Moore and Moore (2013); sequential design is not considered in the present paper. Bruhn and McKenzie (2009) have studied the relative variance of estimators under various designs using simulations.

In contrast to most of the literature on optimal design, the perspective taken in this paper is nonparametric, while allowing for continuous covariates. Here we draw on the extensive literature on inference on average treatment effects under unconfoundedness, as reviewed in Imbens (2004). Part of this paper takes a nonparametric Bayesian perspective, considering (Gaussian) process priors for conditional expectations of potential outcomes. This follows a long tradition in the literatures on spline estimation (cf. Wahba 1990), on “Kriging” in Geostatistics (cf. Matheron 1973, Yakowitz and Szidarovszky 1985), and in the more recent machine learning literature (cf. Williams and Rasmussen 2006). For a general introduction to Bayesian methods with a focus on their decision theoretic motivation, see Robert (2007). O’Hagan and Kingman (1978) considered Gaussian process priors in the
context of experimental design, taking an approach similar to ours but without allowing for covariates. A forceful argument for a Bayesian perspective on experimental design has been made by Berry (2006).

A few examples of experiments from the recent political science literature were mentioned in the introduction. Additional examples abound. Bolsen et al. (2014) sent messages on water conservation to a random set of individuals in Florida to study differential effects by voting behavior. Michelitch (2015) study whether taxi fare bargaining is affected by differences in ethnicity and/or political affiliation between driver and customer in Ghana. Further examples can be found in Gerber et al. (2014); Hanmer et al. (2014); McClendon (2014); Nyhan and Reifler (2014); Paler (2013).

2 A motivating example and some intuitions

2.1 Setup

Before we present our general results and our proposed procedure, let us discuss a simple motivating example. The example is stylized to allow calculations “by hand,” but the intuitions from this example do generalize. Suppose an experimenter has a sample of 4 experimental units \( i = 1, \ldots, 4 \), and she observes a continuous covariate \( X_i \) for each of them, where it so happens that \( (X_1, \ldots, X_4) = (x_1, \ldots, x_4) = (0, 1, 2, 3) \). She assigns every unit to one of two binary treatments, \( d_i \in \{0, 1\} \).

Our experimenter wants to estimate the (conditional) average treatment effect of treatment \( D \) across these 4 units,

\[
\beta = \frac{1}{n} \sum_{i} E[Y_i^1 - Y_i^0 | X_i].
\]

The experimenter plans to estimate this treatment effect by calculating the difference in means across treatment and comparison group in her experiment, that is

\[
\hat{\beta} = \frac{1}{n_1} \sum_{i} D_i Y_i - \frac{1}{n_0} \sum_{i} (1 - D_i) Y_i.
\]

\(^2\)Consider for instance the setting of Nyhan and Reifler (2014), where legislators \( i \) received letters about fact-checking \( (D_i = 1) \) or didn’t \( (D_i = 0) \). The outcome of interest \( Y \) in this case is the future fact-checking rating of these legislators, an important covariate \( X_i \) might be their past rating.
Since there are 4 experimental units, there are $2^4 = 16$ possible treatment assignments. The 16 rows of table 1 correspond to these assignments. In the first row we have $(d_1, \ldots, d_4) = (0, 1, 1, 0)$, in the second row $(d_1, \ldots, d_4) = (1, 0, 0, 1)$, etc.

Assume now that potential outcomes are determined by the equation

$$Y_i^d = x_i + d + \epsilon_i^d,$$

where the $\epsilon_i^d$ are independent given $X$, have mean 0 and variance 1. This is labeled as “model 1” in table 1. The average treatment effect in this model is equal to 1. For any treatment assignment $(d_1, \ldots, d_n)$, we can calculate the corresponding bias, variance, and mean squared error of $\hat{\beta}$ as an estimator of $\beta$, where

$$\text{Var}(\hat{\beta}) = \frac{1}{n_1} + \frac{1}{n_0}$$

with $n_1$ being the number of units $i$ receiving treatment $d_i = 1$, and $n_0$ being the same for $d_i = 0$. The bias is given by

$$\text{Bias} = E[\hat{\beta}] - \beta = \frac{1}{n_1} \sum_i d_i x_i - \frac{1}{n_0} \sum_i (1 - d_i) x_i$$

for our model.

For the assignment in the first row of table 1, we get $\text{Var}(\hat{\beta}) = \frac{1}{2} + \frac{1}{2} = 1$, for the assignment in row 7, $\text{Var}(\hat{\beta}) = \frac{1}{1} + \frac{1}{3} = 1.33$, and similarly for the other rows. The bias for the first row is given by $\text{Bias} = E[\hat{\beta}] - \beta = \frac{1}{2} \cdot (x_2 + x_3 - x_1 - x_4) = 0$, for the third row by $\text{Bias} = \frac{1}{2} \cdot (x_1 + x_3 - x_2 - x_4) = -1$, etc. The mean squared error (MSE) for each row is given by

$$MSE(d_1, \ldots, d_n) = E[(\hat{\beta} - \beta)^2] = \text{Bias}^2 + \text{Var}.$$

Forgetting the covariates

Suppose now for a moment that the information about covariates got lost – somebody deleted the column of $X_i$ in your spreadsheet. Then every $i$ looks the same, before treatment

\footnote{This is the ex-post bias, given covariates and treatment assignment. This is the relevant notion of bias for us. It is different from the ex ante-bias, which is the expected bias when we don’t know the treatment assignment yet.}
<table>
<thead>
<tr>
<th>assignment</th>
<th>designs</th>
<th>model 1</th>
<th>model 2</th>
<th>EMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>d_1 d_2 d_3 d_4</td>
<td>1 2 3 4 5</td>
<td>bias variance MSE</td>
<td>bias variance MSE</td>
<td></td>
</tr>
<tr>
<td>0 1 1 0</td>
<td>1 1 1 1 1</td>
<td>0.0 1.0 1.0</td>
<td>2.0 1.0 5.0</td>
<td>1.77</td>
</tr>
<tr>
<td>1 0 0 1</td>
<td>1 1 1 1 1</td>
<td>0.0 1.0 1.0</td>
<td>-2.0 1.0 5.0</td>
<td>1.77</td>
</tr>
<tr>
<td>1 0 1 0</td>
<td>1 1 1 1 0</td>
<td>-1.0 1.0 2.0</td>
<td>3.0 1.0 10.0</td>
<td>2.05</td>
</tr>
<tr>
<td>0 1 0 1</td>
<td>1 1 1 1 0</td>
<td>1.0 1.0 2.0</td>
<td>-3.0 1.0 10.0</td>
<td>2.05</td>
</tr>
<tr>
<td>1 1 0 0</td>
<td>1 1 1 0 0</td>
<td>-2.0 1.0 5.0</td>
<td>6.0 1.0 37.0</td>
<td>6.51</td>
</tr>
<tr>
<td>0 0 1 1</td>
<td>1 1 1 0 0</td>
<td>2.0 1.0 5.0</td>
<td>-6.0 1.0 37.0</td>
<td>6.51</td>
</tr>
<tr>
<td>0 1 0 0</td>
<td>1 1 0 0 0</td>
<td>-0.7 1.3 1.8</td>
<td>3.3 1.3 12.4</td>
<td>2.49</td>
</tr>
<tr>
<td>0 0 1 0</td>
<td>1 1 0 0 0</td>
<td>0.7 1.3 1.8</td>
<td>-0.7 1.3 1.8</td>
<td>2.49</td>
</tr>
<tr>
<td>1 1 0 1</td>
<td>1 1 0 0 0</td>
<td>-0.7 1.3 1.8</td>
<td>0.7 1.3 1.8</td>
<td>2.49</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>1 1 0 0 0</td>
<td>0.7 1.3 1.8</td>
<td>-3.3 1.3 12.4</td>
<td>2.49</td>
</tr>
<tr>
<td>1 0 0 0</td>
<td>1 1 0 0 0</td>
<td>-2.0 1.3 5.3</td>
<td>4.7 1.3 23.1</td>
<td>6.77</td>
</tr>
<tr>
<td>1 1 1 0</td>
<td>1 1 0 0 0</td>
<td>-2.0 1.3 5.3</td>
<td>7.3 1.3 55.1</td>
<td>6.77</td>
</tr>
<tr>
<td>0 0 0 1</td>
<td>1 1 0 0 0</td>
<td>2.0 1.3 5.3</td>
<td>-7.3 1.3 55.1</td>
<td>6.77</td>
</tr>
<tr>
<td>0 1 1 1</td>
<td>1 1 0 0 0</td>
<td>2.0 1.3 5.3</td>
<td>-4.7 1.3 23.1</td>
<td>6.77</td>
</tr>
<tr>
<td>0 0 0 0</td>
<td>1 0 0 0 0</td>
<td>NaN Inf NaN</td>
<td>NaN Inf NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>1 1 1 1</td>
<td>1 0 0 0 0</td>
<td>NaN Inf NaN</td>
<td>NaN Inf NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>

MSE model 1: Inf 3.2 2.7 1.5 1.0
MSE model 2: Inf 20.6 17.3 7.5 5.0

Notes: Each row of this table corresponds to one possible treatment assignment (d_1, ..., d_4). The columns for “model 1” correspond to the model $Y_i^d = x_i + d + e_i^d$, the columns for “model 2” to the model $Y_i^d = -x_i^2 + d + e_i^d$. Each row shows the bias, variance, and MSE of $\hat{\beta}$, for the given assignment and model. The designs 1-5 correspond to uniform random draws from the assignments with rows marked by an entry of 1. Design 1 randomizes over all rows, design 2 over rows 1 through 14, etc. The last column of the table shows the Bayesian expected MSE for each assignment for the squared exponential prior discussed below. For details, see section 2.
is assigned. The variance of potential outcomes $Y^d$ now includes both the part due to $\epsilon^d_i$ and the part due to $X_i$, and is equal to $\text{Var}(Y^d_i) = \text{Var}(X_i) + \text{Var}(\epsilon^d_i) = \text{Var}(X_i) + 1$. Since units $i$ are indistinguishable in this case, treatment assignments are effectively only distinguished by the number of units treated. Since we observed no covariates and have random sampling, there is no bias (even ex-post), and the MSE of any assignment with $n_1$ treated units is equal to

$$MSE(d_1, \ldots, d_n) = \text{Var}(\hat{\beta}) = \left( \frac{1}{n_1} + \frac{1}{n_0} \right) \cdot (\text{Var}(X_i) + 1).$$

Any procedure that randomly assigns $n_1 = n/2 = 2$ units to treatment 1 is optimal in this case. This is, of course, the standard recommendation. A similar argument applies when we observe a discrete covariate, with several observations $i$ for each value of the covariate. In this latter case, randomization conditional on $X_i$ is optimal.

### 2.2 Randomized designs

Let us now, and for the rest of our discussion, assume again that we observed the covariates $X_i$. The calculations we have done so far for this case are for a fixed (deterministic) assignment of treatment. What if we use randomization? Any randomization procedure in our setting can be described by the probabilities $p(d_1, d_2, d_3, d_4)$ it assigns to the different rows of table 1. The MSE of such a procedure is given by the corresponding weighted average of MSEs for each row:

$$MSE = \sum_{d_1, \ldots, d_n} p(d_1, \ldots, d_n) \cdot MSE(d_1, \ldots, d_n).$$

We can now compare various randomization procedures (experimental designs) in terms of their expected mean squared error. Let us first consider the completely random procedure where each $D_i$ is determined by an independent coin flip. Such a procedure assigns equal probability to all 16 possible treatment assignments. Since there is a chance of 1 in 8 to get all 0s or all 1s, as in the last two assignments in table 1, the variance of an estimator based on such an assignment is technically infinite. This case corresponds to the first of the designs in table 1.

Any reasonable experimenter would re-randomize if she got all 0s or all 1s. If we do this, we eliminate the last two rows of table 1. This procedure still has a sizable chance of
picking an assignment with a large mean squared error, such as for instance \( d = (0, 0, 0, 1) \). The MSE for this re-randomization procedure, labeled “design 2,” is equal to 3.2, cf. the row labeled “MSE model 1.”

Now, again, most experimenters would not use this procedure either. Standard procedure is to reduce randomness further by only allowing for treatment assignments which have treatment and control groups of equal size. This eliminates all but the first 6 rows and yields the third design considered. The advantage of doing so is to reduce variance relative to designs allowing for unequal group sizes. Correspondingly, the average MSE is reduced to 2.7.

So far we have ignored the information available about the covariate \( X \). Many experimenters use such information by blocking, that is by forming groups based on covariate values, and randomizing within these groups. A special case is pairwise randomization, where groups are of size 2. In our case, we could group observations 1 and 2, as well as observations 3 and 4, and randomly assign one unit within each group to treatment and one to control. This gives the fourth design and leaves us with random assignment over rows 1 through 4, resulting in a further reduction of average mean squared error to 1.5.

This procedure, which already has more than halved the expected mean squared error relative to randomization over all possible assignments, can still be improved upon. The assignment 0, 1, 0, 1 has systematically higher values of \( X \) among the treated relative to the control, resulting in a bias of 1 in our setting, and systematically lower values of \( X \) among the treated for the assignment 1, 0, 1, 0, resulting in a bias of -1. Eliminating these two leaves us with the last design, which only allows for the first two treatment assignments. Since these two have the same MSE, we might as well randomize between them. This is what the present paper ultimately recommends. This yields an average MSE of 1 in the present setting.

**2.3 Other data generating processes**

Let us now take a step back and consider what we just did. We first noted that the MSE of any randomized procedure is given by the weighted average of the MSEs of the deterministic assignments it is averaging over. By eliminating assignments with larger MSE from the
support of the randomization procedure we could systematically reduce the MSE. In the limit, we would only randomize between the two assignments with the lowest MSE.

The careful reader will have noted that all our calculations of the MSE depend on the assumed model determining potential outcomes. So let us consider a different model ("model 2") where the covariate affects outcomes in a different way,

\[ Y^d_i = -x_i^2 + d + \epsilon_i^d. \]

Relative to model 1, we have flipped the sign of the effect of \( X \), and we have made the relationship non-linear. Table 1 shows the corresponding values of bias, variance, and MSE for model 2. Again the ranking over the alternative randomized designs is the same as before (cf. the row labeled "MSE model 2"), even though the values of the average MSE are changed. There appears to be a comforting robustness of the optimal procedure.

The careful reader will now again note that we have demonstrated this for two particular models, but one might be able to construct models where the ranking of deterministic assignments is partially reversed. That is indeed true. Experimental design has to proceed without knowledge of the underlying data generating process. The experimenter therefore must trade off performance across alternative data generating processes, as in any statistical decision problem. This can be done explicitly using a (nonparametric) Bayesian prior. Such a prior allows us to calculate the expected MSE, where the expectation averages over possible data generating processes. That is the approach which we will propose in section 4 below. The result is an expected MSE for each assignment \((d_1, \ldots, d_n)\). The last column of Table 1 shows the expected MSE for each deterministic treatment assignment; randomized procedures can again be evaluated in terms of weighted averages of the deterministic MSEs.\(^4\) We can then pick an optimal such assignment, or randomize over a set of assignments with low MSE. Before we go there, however, we shall take a step back and briefly review general statistical decision theory. This allows us to demonstrate that the arguments against randomization, or for randomization over a selected set of assignments, and the intuition of our little example, hold under very general circumstances for any statistical decision problem.

\(^4\)The prior used to calculate this \( EMSE \) assumes that \( C((x_1, d_1), (x_2, d_2)) = 10 \ast \exp \left( - \left( \|x_1 - x_2\|^2 - (d_1 - d_2)^2 \right) / 10 \right) \), for details see section 4 below.
3 Decision theory tells us we should not randomize

We shall now turn to the general argument against randomization. This argument does not depend on the specifics of the experimental design problem. The conclusion holds whether there are spill-overs or not, whether we assume a parametric model or not, whether treatment is binary or not, whether we are interested in average treatment effects or in other objects, whether we evaluate estimates based on quadratic loss or some other criterion, whether we take a minimax or a Bayesian perspective, etc.

3.1 Brief review of decision theory

Our argument does require that we think of experimental design as a decision problem. Statistical decision theory provides one of the conceptual foundations for modern statistics. Statistical decision theory was pioneered by Abraham Wald in the 1940s. A classic introduction can be found in Berger (1985). We shall provide a brief review of decision theory in this section.

The basic idea of decision theory is to think of statistical problems as involving a choice of some action $a$ (picking an estimate, deciding whether to reject a hypothesis, picking an experimental treatment allocation $d$, ...). The choice is made based on the available evidence (data) $X$, so that $a = \delta(X)$. The action will ultimately be evaluated by a loss function $L$, where loss also depends on the unknown state of the world, $\theta$, so $L = L(a, \theta)$. An estimate might be evaluated, for instance, by how much it deviates from the truth, $L = (\hat{\beta} - \beta)^2$. Loss can be thought of as the negative of a utility function, a familiar object in economic and game theoretic models.

Since we do not know what the true state of the world is, we cannot evaluate a procedure based on its actual loss. We can, however, consider various notions of average loss. If we average loss over the randomness of a sampling scheme, and possibly over the randomness of a treatment allocation scheme, holding the state of the world $\theta$ fixed, we obtain the frequentist risk function

$$R(\delta, \theta) = E[L(\delta(X), \theta)|\theta].$$

We can calculate this frequentist risk function. If loss is the squared error of an estimate $\hat{\beta}$
for some true parameter $\beta$, then the frequentist risk function is equal to the mean squared error we calculated in the example of section 2 for each row of table 1. The models 1 and 2 that we considered correspond to two values of $\theta$.

The risk function by itself does not yet provide an answer to the question of which decision procedure to pick, unfortunately. The reason is that one procedure $\delta$ might be better in terms of $R(\delta, \theta)$ for one state of the world, while another procedure $\delta'$ is better for another state $\theta'$ of the world. Returning again to table 1, we have for instance that the assignment in row 10 is better than row 3 for model 1, but worse for model 2.

We thus have to face a trade-off across states $\theta$. There are different ways to do this. One way is to focus on the worst-case scenario, that is on the state of the world with the worst value of the risk function. This yields the minimax decision criterion. Another way is to assign weights $\pi$ (prior probabilities) to different states of the world, and to minimize the weighted average. This yields Bayesian procedures.

### 3.2 Randomization in general decision problems

Let us now return to the question of randomization. Why does decision theory tell us we should not randomize? Suppose we allow our decision procedure $\delta$ to depend on a randomization device $U$, so that $a = \delta(U, X)$. $U$ does not tell us anything about the state of the world $\theta$, and $U$ does not directly enter the loss function $L$. As a consequence, the frequentist risk function for the randomized procedure is simply a weighted average of the risk functions for the non-random decisions that our random procedure might pick. An example is the average MSE we calculated in section 2, cf. the last two rows of table 1.

Formally, let $\delta^u(x) = \delta(u, x)$ be the decision made when $U = u$ and $X = x$, and let $R(\delta^u, \theta)$ be the risk function of the procedure which picks $\delta^u(x)$ whenever $X = x$. Then we have

$$R(\delta, \theta) = \sum_u R(\delta^u, \theta) \cdot P(U = u),$$

where we have assumed for simplicity of notation that $U$ is discrete.

To get to a global assessment of $\delta$, we need to trade off risk across values of $\theta$. If we do
so using the prior distribution $\pi$, we get the Bayes risk

$$R^\pi(\delta) = \int R(\delta, \theta) d\pi(\theta)$$
$$= \sum_u \int R(\delta^u, \theta) d\pi(\theta) \cdot P(U = u)$$
$$= \sum_u R^\pi(\delta^u) \cdot P(U = u).$$

In terms of table 1, Bayes risk averages the MSE for a given assignment $\delta^u$ and data generating process $\theta$, $R(\delta^u, \theta)$, both across the rows, using the randomization device $U$, and across different models for the data generating process, using the prior distribution $\pi$. If, alternatively, we evaluate each action based on the worst-case scenario for that action, we obtain minimax risk

$$R^{\text{mm}}(\delta) = \sum_u \left( \sup_\theta R(\delta^u, \theta) \right) \cdot P(U = u).$$

In terms of table 1, minimax risk evaluates each row (realized assignment $\delta^u$) in terms of the worst case model for this assignment, $\sup_\theta R(\delta^u, \theta)$, and evaluates randomized designs in terms of the weighted average across the worst case risk for each assignment.

Letting $R^*(\delta)$ denote either Bayes or minimax risk, we thus have

$$R^*(\delta) = \sum_u R^*(\delta^u) \cdot P(U = u). \quad (3)$$

The risk of any randomized procedure is given by the weighted average of risk across the deterministic procedures it averages over. Now consider an alternative decision procedure such that

$$\delta^* \in \arg\min_\delta R^*(\delta), \quad (4)$$

where the argmin is taken over non-randomized decision functions $a = \delta(X)$. It follows that

$$R^*(\delta^*) \leq \min_u R^*(\delta^u) \leq R^*(\delta). \quad (5)$$

The second inequality is strict, unless $\delta$ only puts positive probability on the minimizers, over $u$, of $R^*(\delta^u)$. If the optimal deterministic $\delta^*$ is unique, $R^*(\delta^*) < R^*(\delta)$ unless $\delta$ assigns probability 1 to $\delta^*$, so that $\delta$ does not involve randomization. We have thus shown the following proposition:
Proposition 1 (Optimality of non-randomized decisions)

Consider the general decision problem discussed in section 3.1. Let $R^*$ be either Bayes risk or minimax risk. Then:

1. The optimal risk $R^*(\delta^*)$ when considering only deterministic procedures $\delta(X)$ is no larger than the optimal risk when allowing for randomized procedures $\delta(X,U)$.

2. If the optimal deterministic procedure $\delta^*$ is unique then it has strictly lower risk than any non-trivial randomized procedure.

3.3 Experimental design

So far we have discussed general statistical decision problems in this section. How does our discussion relate to the experimental design problem? We can think of an experiment as a decision problem that proceeds in several stages:

1. Sample observations $i$ from the population of interest, and collect baseline covariates $X_i$.

2. Using the baseline covariates $X$ and possibly a randomization device $U$, allocate units $i$ to treatments $D_i$.

3. After running the experiment, collect data on outcomes $Y_i$ of interest. Use these to calculate estimators $\hat{\beta}$ of the objects $\beta$ that we are interested in. In many settings $\beta$ is the average treatment effect.

4. Finally, loss is realized, $L = L(\hat{\beta}, \beta)$. A common choice is $L = (\hat{\beta} - \beta)^2$.

Since this paper is about experimental design, we are interested in the second step. Taking everything else (sampling and the matrix of baseline covariates $X$, the choice of estimator $\hat{\beta}$ and object of interest $\beta$, the loss function $L$) as given, we can ask what is the best way to assign treatment $D$, using the information in the baseline covariates $X$, and possibly the randomization device $U$?

In order to answer this question, we first calculate the risk function of a deterministic choice of treatment allocation $\delta = (d_1, \ldots, d_n)$ which only depends on $X$ but not on $U$. If
loss is the quadratic estimation error \((\hat{\beta} - \beta)^2\), this gives the mean squared error

\[
R(\delta, \theta) = MSE(d_1, \ldots, d_n) = E[(\hat{\beta} - \beta)^2|X, \theta].
\]

This is the calculation we did in section 2 to get the MSE columns for models 1 and 2. The two models can be thought of as two values of \(\theta\) at which we evaluate the risk function.

As in the general decision problem, we get that the risk function of a randomized procedure is given by the weighted average of the risk functions for deterministic procedures,

\[
R(\delta, \theta) = \sum_u R(\delta^u, \theta) \cdot P(U = u).
\]

Consequently, the same holds for Bayes risk and minimax risk. This implies that optimal deterministic procedures perform at least as well as any randomized procedures.

But do they perform better? Do we lose anything by randomizing? As stated in proposition 1, the answer to this question hinges on whether the optimal deterministic assignment is unique. Any assignment procedure, random or not, is optimal if and only if it assigns positive probability only to assignments \((d_1, \ldots, d_n)\) which minimize risk among deterministic assignments.

If we were to observe no covariates then permutations of treatment assignments would not change risk, since everything “looks the same” after the permutation. In that case randomization is perfectly fine, as long as the size of treatment and control group are fixed. Suppose next that we observe a covariate with small support, so that we get several observations for each value of the covariate. In this case we can switch treatment around within each block defined by a covariate value without changing risk – again, everything “looks the same” after the permutation. So in this case stratified (blocked) randomization is optimal.

Suppose now, however, that we observe continuous covariates or discrete covariates with large support. In that case, in general, no two observations in our sample will look the same. We can’t just switch treatment around without changing risk. In that case, therefore, the optimal assignment will in general be unique (possibly up to a “sign flip,” switching 0s and 1s in treatment assignment). This implies that randomizing comes with a cost of larger mean squared error relative to the optimal deterministic assignment.
4 Our proposed procedure

There is one key question we need to answer in order to implement optimal designs in practice: how to aggregate the risk function $R(\delta, \theta)$ into a global notion of risk $R^*$. We mentioned that this might be done using either a Bayesian or a minimax approach. In this section we take a Bayesian approach, using a nonparametric Bayesian prior over possible data generating processes. This is fairly straightforward in its implementation in the present context, as we shall see.

The key object that we need a prior for is the conditional expectation of potential outcomes given covariates, since it is this function which determines the bias of any given estimate of the average treatment effect. Denote this function by $f$, where

$$f(x, d) = E[Y_i^d|X_i = x, \theta].$$

In model 1 of section 2, we had $f(x, d) = x + d$, in model 2 we had $f(x, d) = -x^2 + d$. We assume that the prior for $f$ is such that expectations and variances exist. We denote

$$E[f(x, d)] = \mu(x, d),$$

$$\text{Cov}(f(x_1, d_1), f(x_2, d_2)) = C((x_1, d_1), (x_2, d_2))$$

for a covariance function $C$. Assume further that we have a prior over the conditional variance of $Y_i$ which satisfies

$$E[\text{Var}(Y_i|X_i, D_i, \theta)|X_i, D_i] = \sigma^2(X_i, D_i) = \sigma^2.$$

Assuming a prior centered on homoskedasticity is not without loss of generally, but is a natural baseline.

How should the prior moments $\mu$ and $C$ be chosen? We provide an extensive discussion in the online appendix; a popular choice in the machine learning literature are so-called squared exponential priors, where $\mu \equiv 0$ and

$$C((x_1, d_1), (x_2, d_2)) = \exp \left( - \left( \|x_1 - x_2\|^2 - (d_1 - d_2)^2 \right) / l \right).$$

The length scale parameter $l$ determines the assumed smoothness of $f$ for such priors. This is the prior we used to calculate the expected MSE in table 1 choosing a length scale of
l = 10. Another popular set of priors builds on linear models. For such priors we show below that expected MSE corresponds to balance as measured by differences in covariate means.

Recall that we are ultimately interested in estimating the conditional average treatment effect,

\[ \beta = \frac{1}{n} \sum_i E[Y^1_i - Y^0_i | X_i]. \]

One more question has to be settled before we can give an expression for the Bayes risk of any given treatment assignment: how is \( \beta \) going to be estimated? We will consider two possibilities. The first possibility is that we use an estimator that is optimal in the Bayesian sense, namely the posterior best linear predictor of \( \beta \). The second possibility is that we estimate \( \beta \) using a simple difference in means, as in the example of section 2.

### 4.1 Bayes optimal estimation

It is useful to introduce the following notation for the relevant prior moments:

\[ \mu_i = \mu(X_i, d_i) \]
\[ C_{i,j} = C((X_i, d_i), (X_j, d_j)) \]
\[ \mu_\beta = \frac{1}{n} \sum_i [\mu(X_i, 1) - \mu(X_i, 0)], \]
\[ C_\beta = \frac{1}{n} \sum_j [C((X_i, d_i), (X_j, 1)) - C((X_i, d_i), (X_j, 0))] \]

Let \( \mu \) and \( C \) be the corresponding vector and matrix with entries \( \mu_i \) and \( C_{i,j} \). Note that both \( \mu \) and \( C \) depend on the treatment assignment \( (d_1, \ldots, d_n) \). Using this notation, the prior mean and variance of \( Y \) are equal to \( \mu \) and \( C + \sigma^2 I \), and the prior mean of \( \beta \) equals \( \mu_\beta \).

Let us now consider the posterior best linear predictor, which is the best estimator (in the Bayesian risk sense) among all estimators linear in \( Y_i \). The posterior best linear

---

5This class of estimators includes all standard estimators of \( \beta \) under unconfoundedness, such as those based on matching, inverse probability weighting, regression with controls, kernel regression, series regression, splines, etc. Linearity of the estimator is unrelated to any assumption of linearity in \( X \); we are considering the posterior BLP of \( \beta \) in \( Y \) rather than the BLP of \( Y_i \) in \( X_i \).
predictor is equal to the posterior expectation if both the prior for $f$ and the distribution of the residuals $Y - f$ are multivariate normal.

**Proposition 2 (Posterior best linear predictor and expected loss)**

The posterior best linear predictor for the conditional average treatment effect is given by

$$\hat{\beta} = \mu_\beta + \bar{C} \cdot (C + \sigma^2 I)^{-1} \cdot (Y - \mu),$$  \hspace{1cm} (6)

and the corresponding mean squared error (Bayes risk) equals

$$MSE(d_1, \ldots, d_n) = \text{Var}(\beta|X) - \bar{C} \cdot (C + \sigma^2 I)^{-1} \cdot \bar{C},$$  \hspace{1cm} (7)

where $\text{Var}(\beta|X)$ is the prior variance of $\beta$.

The proof of this proposition follows from standard characterizations of best linear predictors and can be found in appendix A. The expression for the mean squared error provided by equation (7) is easily evaluated for any choice of $(d_1, \ldots, d_n)$. Since our goal is to minimize the MSE, we can in fact ignore the $\text{Var}(\beta|X)$ term which does not depend on $(d_1, \ldots, d_n)$.

4.2 Difference in means estimation

Let us next consider the alternative case where the experimenter insists on using the simple difference in means estimator. We will need the following notation:

$$\mu^d_i = \mu(X_i, d)$$

$$C^d_{i,j} = C((X_i, d^1), (X_j, d^2)),$$

for $d, d^1, d^2 \in \{0, 1\}$. We collect these terms in the vectors $\mu^d$ and matrices $C^{d^1,d^2}$, which are in turn collected as

$$\tilde{\mu} = (\mu^1, \mu^2)$$

$$\tilde{C} = \begin{pmatrix} C^{00} & C^{01} \\ C^{10} & C^{11} \end{pmatrix}.$$

$\tilde{\mu}$ and $\tilde{C}$ do not depend on $(d_1, \ldots, d_n)$, in contrast to $\mu$ and $C$ defined before. The latter are a sub-vector / sub-matrix of the former, selected through a choice of $(d_1, \ldots, d_n)$. Defining $\tilde{f}$ analogously to $\tilde{\mu}$, we get $E[\tilde{f}] = \tilde{\mu}$ and $\text{Var}(\tilde{f}) = \tilde{C}$. 
Proposition 3 (Expected MSE for designs, using the difference in means estimator)

Assume that $\beta$ is estimated using the difference in means estimator

$$\hat{\beta} = \frac{1}{n_1} \sum_i D_i Y_i - \frac{1}{n_0} \sum_i (1 - D_i) Y_i.$$ 

Then

$$MSE(d_1, \ldots, d_n) = \sigma^2 \cdot \left[ \frac{1}{n_1} + \frac{1}{n_0} \right] + (w' \cdot \tilde{\mu})^2 + w' \cdot \tilde{C} \cdot w. \quad (8)$$

where

$$w = (w^0, w^1),$$

$$w^1_i = \frac{d_i}{n_1} - \frac{1}{n},$$

$$w^0_i = - \frac{1-d_i}{n_0} + \frac{1}{n}.$$

The expression for the MSE in this proposition has 3 terms. The first term is the variance of the estimator. The second and third term together are the expected squared bias. This splits in turn into the square of the prior expected bias, and the prior variance of the bias.

It is interesting to note that we recover the standard notion of balance if, in addition to the assumptions of proposition 3, we impose a linear, separable model for $f$, that is

$$f(x, d) = x' \cdot \gamma + d \cdot \beta,$$

where $\gamma$ has prior mean 0 and variance $\Sigma$. For this model we get that the bias of $\hat{\beta}$ is equal to

$$(\bar{X}^1 - \bar{X}^2)' \cdot \beta,$$

where $\bar{X}^d$ is the sample mean of covariates in either treatment arm, and thus the prior expected squared bias is equal to

$$(\bar{X}^1 - \bar{X}^2)' \cdot \Sigma \cdot (\bar{X}^1 - \bar{X}^2),$$

and the MSE equals

$$MSE(d_1, \ldots, d_n) = \sigma^2 \cdot \left[ \frac{1}{n_1} + \frac{1}{n_0} \right] + (\bar{X}^1 - \bar{X}^2)' \cdot \Sigma \cdot (\bar{X}^1 - \bar{X}^2).$$

Risk is thus minimized by (i) choosing treatment and control arms of equal size, and (ii) optimizing balance as measured by the difference in covariate means $(\bar{X}^1 - \bar{X}^2)$. 

4.3 Discrete optimization

We now have almost all the ingredients for a procedure that can be used by practitioners. One important element is missing, still: How do we find the optimal assignment $d^*$? Or how can we at least find a set of assignments which are close to optimal in terms of expected risk? This question arises since solving the problem by brute force is generally not feasible. We could do so in our stylized example in section 2 since in this example there were only $2^4 = 16$ possible treatment assignments. In general there are $2^n$ assignments for a sample of size $n$, a number that very quickly becomes prohibitive to search by brute force.

So what to do? One option, which is close to a common practice among experimenters, is to re-randomize a feasible number of times, and to pick the best assignment among those randomly picked. A quick calculation shows that this procedure performs reasonably well. Suppose we re-randomize $k$ times, and pick $\hat{d}$ as the best of the $k$ assignments we got. Then the probability that $\hat{d}$ is better then 99% of all assignments, say, is equal to $1 - 0.99^k$. If $k = 500$, this chance is already larger than 99%, and significantly larger $k$ are feasible in practice.

Implementation of our design procedure using re-randomization is simple and can be summarized as follows: First, pick a prior. The online appendix discusses good defaults. For our example in section 2, we used $\mu = 0$ and $C((x_1,d_1),(x_2,d_2)) = 10 \times \exp(-((x_1 - x_2)^2 - (d_1 - d_2)^2)/10)$. Calculate the corresponding matrix $\tilde{C}$, and $\tilde{C}^{-1}$. Then iterate the following:

1. Draw a random treatment assignment.
2. Calculate the objective function of equation (8).
3. Compare it to the best MSE obtained thus far. If the new MSE is better than the old one, store the new treatment assignment and MSE.
4. Iterate for some pre-specified number of times $k$.

Alternative optimization methods

There exist alternative, more sophisticated methods of optimization, of course, and there is an extensive literature discussing algorithms for discrete optimization problems such as
one set of procedures, called greedy algorithms, is based on local search. The idea is to start from some assignment $d^0$, and search over a set of “neighboring” assignments that only differ in a few components to find the best assignment among those. This best assignment is labeled $d^1$ and is used as the starting point for a new local search. The procedure is left to run until a local optimum is found or timeout is reached.

A variation on greedy algorithms is so-called simulated annealing. This is one of the most popular algorithms for discrete optimization and was introduced by Kirkpatrick et al. (1983). The algorithm uses noisy perturbations to a greedy search, to avoid getting stuck in a local minimum. The noise is reduced in later iterations so the algorithm converges.

We have implemented versions of each of these, and Matlab code is available upon request. In practice it appears that re-randomization performs quite well, and might be preferred by practitioners on account of its simplicity.

5 Arguments for randomization

There are a number of arguments that can be made for randomization and against the framework considered in this paper. We shall discuss some of these, and to what extent they appear to be valid.

Randomization inference requires randomization

That is correct. How does this relate to our argument? The arguments of section 3 show that optimal procedures in a decision theoretic sense do not rely on randomization. Randomization inference cannot be rationalized as an optimal procedure under the conceptual framework of decision theory, however. As a consequence, the arguments of section 3 do not apply.

One could take the fact that randomization inference is not justified by decision theory as an argument against randomization inference. But one could also consider a compromise approach which is based on randomization among assignments which have a low expected mean squared error. Such partially random assignments are still close to optimal and allow the construction of randomization based tests.
Is identification assuming conditional independence still valid without randomization?

Yes it is. Selection on observables holds, as the name suggests, for any treatment assignment which is a function of observables. Put differently, conditional independence is guaranteed by any controlled trial, whether randomized or not, as stated in the following proposition.

**Proposition 4 (Conditional independence)**

Suppose that \((X_i, Y^0_i, Y^1_i)\) are i.i.d. draws from the population of interest, which are independent of \(U\). Then any treatment assignment of the form \(D_i = d_i(X_1, \ldots, X_n, U)\) satisfies conditional independence,

\[(Y^0_i, Y^1_i) \perp D_i | X_i.\]

This is true, in particular, for deterministic treatment assignments of the form \(D_i = d_i(X_1, \ldots, X_n)\).

The Bayesian approach to experimental design requires the choice of a prior

That is correct, as far as the ranking of assignments in terms of MSE is concerned. A treatment assignment might be optimal for a particular prior and objective (such as minimizing the MSE of an estimator for the ATE), but not for another prior or objective.

This does not imply, however, that the *interpretation* of the data obtained from the experiment has to rely on the same prior as the one used for *designing* the experiment, or on any prior at all. From the perspective of a researcher analyzing data produced by an experiment where treatment was assigned based on observables as well as a randomization device, it is immaterial how treatment was assigned, as long as the conditional independence condition \((Y^0_i, Y^1_i) \perp D_i | X_i\) holds. This is guaranteed by proposition \[4\]. The design can furthermore be chosen to be optimal for estimators such as the difference in means, which does not use a prior, as shown in proposition \[3\].
Randomized assignments are perceived as fair, and thus politically easier to implement

A possible objection to the practical feasibility of optimal designs might be a political perception that randomized assignments are fair, while assignment based on covariates such as sex or race is not fair, in a similar way that “tagging” in taxation is considered unfair. This might sometimes be a constraint on feasible experiments.

Note however that optimal designs as discussed in this paper seek to balance the covariate distribution across treatments, leading to a more equitable distribution of treatment across demographic or other groups, relative to random assignments.

Does the proposed approach rely on a parametric model?

No, it does not. Nowhere have we imposed restrictions on the distribution of potential outcomes or the distribution of covariates, beyond the assumption of i.i.d. sampling. The family of priors considered in section 4 in particular, is nonparametric, and has support on all sufficiently smooth functions.

6 Conclusion

In this paper we discuss the question how information from baseline covariates should be used when assigning treatment in an experiment. In order to give a well grounded answer to this question, we adopt a decision theoretic and non-parametric framework. The non-parametric perspective and the consideration of continuous covariates distinguish this paper from much of the previous experimental design literature.

A number of conclusions emerge from our analysis. First, randomization is in general not optimal. Rather, we should pick a risk-minimizing treatment assignment, which is generically unique in the presence of continuous covariates. Second, we can consider nonparametric priors which lead to very tractable estimators and expressions for Bayesian risk (mean squared error). The general form of the expected MSE for such priors is $\text{Var}(\beta|X)$ –

---

6I thank Larry Katz for pointing this out.
7To be precise, the support of these priors is the closure of the Reproducing Kernel Hilbert Space corresponding to $C$. For further discussion see the online appendix.
\( \mathbf{C}' \cdot (C + \Sigma)^{-1} \cdot \mathbf{C} \), where \( \mathbf{C} \) and \( C \) are the appropriate covariance vector and matrix from the prior distribution, cf. section 4. We suggest picking the treatment assignment that minimizes this prior risk. Finally, conditional independence between potential outcomes and treatments given covariates does not require random assignment. It is ensured by conducting controlled trials, and does not rely on randomized controlled trials.

A Proofs

**Proof of proposition 2** By the usual formula for best linear predictors,

\[
\hat{\beta} = E[\beta | X, D] + \text{Cov}(\beta, Y | X, D) \cdot \text{Var}(Y | X, D)^{-1} \cdot (Y - E[Y | X, D]).
\]

Using the notation introduced before proposition 2 we can denote these terms by

\[
E[\beta | X, D] = \mu_{\beta},
\]
\[
\text{Cov}(\beta, Y | X, D) = \mathbf{C}',
\]
\[
\text{Var}(Y | X, D) = C + \sigma^2 I,
\]
\[
E[Y | X, D] = \mu,
\]

which yields equation (6). We furthermore have, by the general properties of best linear predictors, that

\[
\text{MSE}(d_1, \ldots, d_n) = E[(\beta - \hat{\beta})^2 | X, D] = \text{Var}(\beta - \hat{\beta} | X, D),
\]

and \( \text{Cov}(\hat{\beta}, \beta - \hat{\beta} | X, D) = 0 \), so that

\[
\text{Var}(\beta | X, D) = \text{Var}(\hat{\beta} | X, D) + \text{Var}(\beta - \hat{\beta} | X, D).
\]

This immediately implies equation (7). \( \square \)

**Proof of proposition 3**

Let \( \epsilon_i^d = Y_i^d - f(X_i, d_i) \). We can write

\[
\Delta := \hat{\beta} - \beta = \sum_i \left[ \left( \frac{d_i}{n_1} - \frac{1-d_i}{n_0} \right) \cdot \left( f(X_i, d_i) + \epsilon_i^d \right) - \frac{1}{n} (f(X_i, 1) - f(X_i, 0)) \right]
\]

25
and

\[ \text{MSE}(d_1, \ldots, d_n) = E[\Delta^2] = \text{Var}(\Delta) + E[\Delta]^2 \]
\[ = E[\text{Var}(\Delta|f)] + \text{Var}(E[\Delta|f]) + E[\Delta]^2. \]

The first term is equal to
\[ \text{Var}(\Delta|f) = \sigma^2 \left[ \frac{1}{n_1} + \frac{1}{n_0} \right]. \]

The second term is equal to the variance of
\[ E[\Delta|f] = w' \cdot \tilde{f}. \]

The third term is equal to the square of
\[ E[\Delta] = E[E[\Delta|f]] \]
\[ = E[w' \cdot \tilde{f}] \]
\[ = w' \cdot E[\tilde{f}]. \]

The claim follows once we recall \( E[\tilde{f}] = \tilde{\mu} \) and \( \text{Var}(\tilde{f}) = \tilde{C}. \) □

**Proof of proposition 4**

The assumption of independent sampling implies that

\[ (X_i, Y_i^0, Y_i^1) \perp (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, U), \]

thus

\[ (Y_i^0, Y_i^1) \perp (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, U)|X_i, \]

and therefore

\[ (Y_i^0, Y_i^1) \perp d_i(X_1, \ldots, X_n, U)|X_i. \]

□
References


