Causal Inference by Minimizing the Dual Norm of Bias: Kernel Matching & Weighting Estimators for Causal Effects

Nathan Kallus
School of Operations Research and Information Engineering
Cornell University and Cornell Tech
New York, NY 10011

Abstract

We consider the problem of estimating causal effects from observational data and propose a novel framework for matching- and weighting-based causal estimators. The framework is based on expressing the bias of a causal estimator as an operator on the unknown conditional expectation function of outcomes and formulating the *dual norm of the bias* as the norm of this operator with respect to a function space that represents the potential structure for outcomes. We give the term *worst-case bias minimizing* (WCBM) to estimators that minimize this quantity for some function space and show that a great variety of existing causal estimators belong to this family, including one-to-one matching (with or without replacement), coarsened exact matching, and mean-matched sampling. We propose a range of new, kernel-based matching and weighting estimators that arise when one minimizes the dual norm of the bias with respect to a reproducing kernel Hilbert space. Depending on the case, these estimators can be solved either in closed form, using quadratic optimization, or using integer optimization. We show that estimators based on *universal* kernels are consistent for the causal effect. In numerical experiments, the new, kernel-based estimators outperform all standard causal estimators in estimation error, providing a successful balance between generality and efficiency.

1 Introduction

Compared to controlled experiments, observational studies are uniquely characterized by a lack of control on membership in the treatment and control groups. While in controlled experimentation, randomization ensures comparability and hence unbiased and consistent estimation of effect; in observational studies, valid inference about a causal effect of treatment requires adjusting the groups so that they become comparable. Comparable for the purpose of causal inference means as similar as possible in some observed covariates. The covariates constitute the relevant information known about each observational subject and, as long as these covariates account for any confounding between the effects of treatment and the effects of self-selection, making the groups comparable with respect to these makes the groups comparable for the purpose of causal inference.

Matching and weighting have been some of the most popular ways to achieve this comparability [6, 32, 44]. In matching, we sample a (multi-)subset from the groups to get samples that are more similar to one another than the original samples. For example, in one-to-one matching [31], one composes a matched sample out of pairs of treated and control subjects so that the total pairwise distance between covariate vectors is small or even minimal, mimicking a randomized matched-pair experiment [14]. If we allow subjects to be paired with replacement, we can have a sample with duplicates. Weighting is a generalization where we can assign weights that are not integer multiples. For example, in coarsened exact matching (CEM) [19], one coarsens the covariates to create strata and re-weights the samples so that they have equal frequency in each stratum, mimicking a randomized block experiment [9].

Matching and weighting is employed for two purposes: (a) to reduce error due to confounding and (b) to reduce error due to imbalance. In a controlled experiment, (a) is achieved by randomization and (b) is achieved by, e.g., blocking (see [21] for more about balance in controlled experiments). For example, in an experiment on the effect of a drug on mortality, a randomized block design that ensures balance in important covariates such as age is generally more powerful than a completely randomized design, while both are unbiased (unconfounded). In observational studies, a very popular matching method to achieve (a) is propensity score matching (PSM) [33]. However, because it does nothing toward purpose (b) and because it depends on strong modelling assumptions to fit a correct propensity model, it is sometimes advocated that one match on the covariates themselves rather than estimated propensity scores [22]. A related weighting method is propensity score weighting (PSW) [15]. However, it too does nothing toward (b) and relies on strong modelling assumptions. Even if the model is correct, the estimated weights can be very unstable leading to a practice of ad hoc trimming, which may re-introduce confounding bias [8, 7]. Here, partly for these reasons, we focus on matching and weighting that address both (a) and (b) by balancing the covariates themselves rather than imputed estimates of propensities.

In this paper, we develop a novel and encompassing framework for estimators that balance the covariates via weighting and matching. There are many different such estimators and each addresses imbalance differently. Our framework teases out how a particular notion of imbalance corresponds to a notion of structure. By decomposing the error of matching and weighting estimators, we formulate the bias of the estimator as an operator on the conditional expectation of outcomes given covariates. This conditional expectation function is unknown (or else there would be no need to experiment) and when one considers what the *worst-case* bias may be over a space of possible such functions one recovers the *dual norm of the bias* if the space is a Banach space. The dual norm of the bias is an observable quantity, expressed only in terms of the given data. We term any estimator that chooses matched subsamples or weights by minimizing the worst-case bias as *worst-case bias minimizing* (WCBM). A surprising result is that a great variety of standard methods used in the practice of causal inference are all WCBM. This observation leads us to consider new methods that are WCBM. Using reproducing kernel Hilbert spaces (RKHS) to express structure we obtain a new class of kernel-based matching and weighting causal estimators. These have desirable properties like consistency and perform exceptionally well in practice.

All proofs are given in Section 9.

2 Set up

We begin by describing the set up. We consider an observational study with n subjects. We index the subjects by $i = 1, \dots, n$. We let this order be arbitrary so that the subjects are exchangeable (later, we consider subjects comprising an iid process). Of these, n_1 received a treatment whose effect is of interest (denoted by $T_i = 1$) and n_0 received a control treatment against which we want to compare (denoted by $T_i = 0$). Let $\mathcal{T}_0 = \{i : T_i = 0\}$ and $\mathcal{T}_1 = \{i : T_i = 1\}$ be the sets of subjects that received treatment and control, respectively. We let $T = (T_1, \dots, T_n)$ denote the collection of treatment assignments, which constitutes part of the observed data.

Using Neyman-Rubin potential outcome notation [39], we let $Y_i(0)$, $Y_i(1)$ be the (real-valued) potential outcomes for subject i. We observe the outcome for the treatment to which subject i was exposed, $Y_i = Y_i(T_i)$. And, $Y(1 - T_i)$ represents the unobserved, counterfactual outcome we would have observed if subject i were exposed to the opposite treatment. $Y(1-T_i)$ is *missing data*. Throughout the paper, for these to be well defined, we assume that the stable unit treatment value assumption (SUTVA) holds [35], which requires that which treatment one of subjects experiences not affect the outcomes of another subject and that potential outcomes are fixed as which treatment is experienced changes (so only *which* one we observe, and hence Y_i, is affected).

Let X_i, taking values in some $\mathcal{X}$, be the side covariates that we observe for subject i. Let $X = (X_1, \dots, X_n)$ denote the collection of all baseline covariates of all n subjects, which constitues part of the observed data. The space $\mathcal{X}$ is general; assumptions about it will be specified as necessary. As an example, it can be composed real-valued vectors $\mathcal{X} \subseteq \mathbb{R}^d$ that include both discrete (dummy) and continuous variables.

We denote by $\mathrm{TE}_i = Y_i(1)-Y_i(0)$ the unobservable causal treatment effect for subject i. The primary quantity of interest for estimation is the *sample average (causal) treatment effect on the treated sample* (SATT):

$$\mathrm{SATT} = \tfrac{1}{n_1}\textstyle\sum_{i\in\mathcal{T}_1} \mathrm{TE}_i = \tfrac{1}{n_1}\sum_{i=1}^n T_i(Y_i(1) - Y_i(0)).$$

We consider estimators for SATT based on *weighting*. We restrict to *honest* weights that only depend on the observed X, T and not on any observed outcome data. (If we used outcome data one might complain that we are mining for an effect that is not there.) In particular, we will consider the choice of a weighting function $W = W(X, T)$ that produces a weight $W_i \in \mathbb{R}$ for each subject i, leading to the estimator

$$\hat{\tau}_W = \textstyle\sum_{i=1}^n (-1)^{T_i+1} W_i Y_i.$$

Because we are estimating SATT and we in fact observe $Y_i(1)$ for each $i \in \mathcal{T}_1$, we always set $W_i = 1/n_1$ for $i \in \mathcal{T}_1$, leading to estimators of the form

$$\hat{\tau}_W = \tfrac{1}{n_1}\textstyle\sum_{i\in\mathcal{T}_1} Y_i - \sum_{i\in\mathcal{T}_0} W_i Y_i.$$

We also always assume $\sum_{i\in\mathcal{T}_0} W_i = 1$.

The bias of the estimator resulting from weights W is the difference between it and SATT, conditioned on all the observable data upon which the weights are based:

$$\mathrm{bias} = \mathbb{E}\left[\hat{\tau}_W - \mathrm{SATT}|\, X, T\right].$$

We let $\mathcal{W} = \mathcal{W}_0 \times \mathcal{W}_1$ denote the space of allowable weights, where $\mathcal{W}_0$ and $\mathcal{W}_1$ are the space of weights for the control and treated sample, respectively. We required that $\mathcal{W}_0 \subseteq \{W_{\mathcal{T}_0} \in \mathbb{R}^{\mathcal{T}_0} : \sum_{i\in\mathcal{T}_0} W_i = 1\}$ and that $\mathcal{W}_1 = \{(1/n_1, \dots, 1/n_1)\}$. If all weights in $\mathcal{W}_0$ are rational with a fixed denominator, we call $\hat{\tau}_W$ a matching estimator because it is equivalent to constructing a (multi-)set from the control subjects to match the treated sample. We note some special cases of $\mathcal{W}_0$ that correspond to a variety of existing classes of estimators for SATT:

- General weights:
$$\mathcal{W}_0^{\text{general}} = \{W_{\mathcal{T}_0} \in \mathbb{R}^{\mathcal{T}_0} : \textstyle\sum_{i\in\mathcal{T}_0} W_i = 1\}.$$
- Nonnegative (probability) weighting:
$$\mathcal{W}_0^{\text{nonnegative}} = \{W_{\mathcal{T}_0} \in \mathbb{R}_+^{\mathcal{T}_0} : \textstyle\sum_{i\in\mathcal{T}_0} W_i = 1\}.$$
- Matching with fixed size n_0' and without replacement:
$$\mathcal{W}_0^{\text{w/o rep.}} = \{W_{\mathcal{T}_0} \in \{0, 1/n_0'\}^{\mathcal{T}_0} : \textstyle\sum_{i\in\mathcal{T}_0} W_i = 1\}.$$
- Matching with fixed size n_0' and with replacement:
$$\mathcal{W}_0^{\text{w/ rep.}} = \{W_{\mathcal{T}_0} \in \{0, 1/n_0', \dots\}^{\mathcal{T}_0} : \textstyle\sum_{i\in\mathcal{T}_0} W_i = 1\}.$$

Note that, as estimators for a *population* effect (if we are to assume random sampling of subjects from a population), both SATT and $\hat{\tau}$ preclude regression adjustments [11]. Nonetheless, without parametric assumptions necessary for such adjustments, assuming random sampling, SATT is the *uniform minimum variance unbiased estimator* for the treatment effect on the treated population [24]. In fact, matching and weighting are often regarded as means of reducing model dependence [16]. Specifically, if one matches very closely, then simple difference estimators and complicated regression estimators are all very close, so the point of which estimator to use after matching is largely moot.

A standing assumption in this paper, essential for causal inference from observational data, is that of *weak ignorability in expectation*.

Assumption 1. For each $t = 0, 1$ and $i = 1, \dots, n$, conditioned on X_i, $Y_i(t)$ is mean-independent of T_i and each value of T_i is possible. That is, for each $t = 0, 1$ and $i = 1, \dots, n$,

$$\mathbb{E}\left[Y_i(t)\middle|\, T_i, X_i\right] = \mathbb{E}\left[Y_i(t)\middle|\, X_i\right], \quad \text{and}$$
$$\mathbb{P}\left(T_i = t\middle| X_i\right) > 0.$$

Ignorability, also known as unconfoundedness, means that we have the right covariates needed to separate the effect of the treatment itself from the effect of self-selection. For example, in an observational study, self-selection for "treatment" might imply affluence, which might imply good outcomes regardless of treatment, so if we control for income in X_i we can isolate this effect. The form of ignorability we use is termed "weak" because it need only apply for each $t = 0, 1$ separately, and it is termed "in expectation" because only mean-independence, rather than full stochastic independence, is assumed.

Aside: alternative frameworks for causal inference

The Neyman-Rubin potential outcome framework is not the only framework used to describe causal relationships. Other frameworks for causality include, most notably, Pearl's framework of causal Bayesian networks and do-calculus [25] as well as structural equation models (SEM) [12]. We do not consider the SEM framework because of its need for a priori models, the common restriction to linear relationships, incompatible notation, and the less clear question of model-free identifiability.

Pearl's framework generalizes both potential outcomes and SEM [28, 27]. Inference in this framework depends on directed acyclic graph (DAG) models to describe a priori causal relationships. The standard practice in applications of the Neyman-Rubin framework is generally to condition on all observed covariates X that are potentially relevant [36], but one can easily come up with DAG constructions where the inclusion of a covariate in such conditioning can (asymptotically) bias causal estimates, the simplest of which is the M-graph [40, 26]. In effect, a causal DAG, correctly specified, can specify the correct subset of the covariates X that should be included in order to achieve Assumption 1. The estimation or validation of a causal DAG from data is an active field of research, e.g. [17, 42, 29].

3 Decomposing the bias

We define the conditional expectation of the control potential outcome given the covariates x as follows:

$$f_0(x) = \mathbb{E}\left[Y_i(0)\middle| X_i = x\right].$$

The non-random function f_0 does not depend on i due to exchangeability. By the law of iterated expectation, the residual $\epsilon_i = Y_i(0) - f_0(X_i)$ has mean 0, is mean-independent of X_i, and is uncorrelated with any function of X_i.

By conditioning on X_i, we can decompose the error of the estimator into two terms: error that can be controlled by matching on X_i and the orthogonal residual error, which cannot be controlled by X_i but which disappears in expectation due to ignorability.

Theorem 1. *Under Assumption 1, the bias of $\hat{\tau}_W$ is*

$$\mathbb{E}\left[\hat{\tau}_W - \text{SATT}\middle|\, X, T\right] = B(W; f_0),$$
where $\quad B(W; f) := \frac{1}{n_1}\sum_{i\in\mathcal{T}_1} f(X_i) - \sum_{i\in\mathcal{T}_0} W_i f(X_i).$

Moreover, letting

$$E(W) = \tfrac{1}{n_1}\textstyle\sum_{i\in\mathcal{T}_1} \epsilon_i - \sum_{i\in\mathcal{T}_0} W_i \epsilon_i,$$

we have that

$$\hat{\tau}_W - \text{SATT} = B(W; f_0) + E(W),$$
$$\mathbb{E}\left[E(W)\middle|\, X, T\right] = 0.$$

4 The dual norm of the bias

The target of weighting or matching for causal inference is to eliminate bias in comparing the treatment and control samples. Theorem 1 provides an explicit form of the bias in terms of the observed covariates X. However, it involves the unknown function $f_0 : \mathcal{X} \to \mathbb{R}$. As alluded to in Sec. 1, we consider weighting schemes that guard against any possible such function by minimizing the worst-case bias over the unit ball of a Banach space. A normed vector space is a Banach space if the corresponding metric space is complete (see [23] and Ch. 10 of [34] for more on Banach spaces).

Let $\mathcal{V}$ denote the vector space of all functions $\mathcal{X} \to \mathbb{R}$ under usual pointwise addition and scaling. Let $\mathcal{F} \subseteq \mathcal{V}$ be a subspace of functions, against which we wish to guard. Endow this space with a semi-norm $\|\cdot\| : \mathcal{F} \to \mathbb{R}$ (a semi-norm can assign zero magnitude to nonzero vectors). For $f \notin \mathcal{F}$, let us write $\|f\| = \infty$. Thus, the assumption that $f_0 \in \mathcal{F}$ is encapsulated by $\|f_0\| < \infty$.

Given only that $\|f_0\| < \infty$, we will consider weighting or matching schemes that choose W to minimize the worst-case bias,

$$\max_{\|f\| \le \|f_0\|} |B(W; f)| = \|f_0\| \max_{\|f\| \le 1} B(W; f),$$

where the equality holds because $B(W; \alpha f) = \alpha B(W; f)$ is degree-1 homogeneous and $\|\alpha f\| = |\alpha| \, \|f\|$ is degree-1 positively homogeneous and symmetric. Clearly, it only matters that $\|f_0\| < \infty$ and the particular finite value of it does not change which W minimizes the above. In light of this, we define the *worst-case bias* as

$$\mathfrak{B}(W; \mathcal{F}) = \max_{\|f\| \le 1} B(W; f).$$

Since $\sum_{i=1}^{n} (-1)^{T_i+1} W_i = 0$, we have that $B(W; f)$ is invariant to constant shifts to f, i.e., $B(W; f) = B(W; f + c)$, where $c \in \mathbb{R}$ represents a constant function $x \mapsto c$. To eliminate this irrelevant mode of $\mathcal{F}$, we can just consider the quotient space $\mathcal{F}/\mathbb{R}$, which consists of the equivalence classes $[f] = \{f + c : c \in \mathbb{R}\}$ endowed with the norm $\|[f]\| = \min_{c \in \mathbb{R}} \|f + c\|$. Note that by construction, $B(W; [f]) = B(W, f)$ is well defined. For brevity, we will simply refer to $\mathcal{F}$ and $\|\cdot\|$ when we mean $\mathcal{F}/\mathbb{R}$ and the corresponding norm.

We will consider spaces $(\mathcal{F}, \|\cdot\|)$ that satisfy the following conditions:

Assumption 2. The space $\mathcal{F}$ is a Banach space.

Assumption 3. For each $W \in \mathcal{W}$, $f \mapsto B(W; f)$ is a continuous mapping $\mathcal{F} \to \mathbb{R}$.

Since $B(W, f)$ is also linear in f, these assumptions imply that, for each W, the operator $B(W, \cdot)$ is in the continuous dual space of $\mathcal{F}$. Hence,

$$\mathfrak{B}(W; \mathcal{F}) = \|B(W; \cdot)\|_*$$

is *precisely* the dual norm of the bias, where the dual norm of a continuous linear operator A on a Banach space with norm $\|\cdot\|$ is $\|A\|_* = \sup_{\|u\| \le 1} A(u)$. This also guarantees that $\mathfrak{B}(W; \mathcal{F})$ is finite and well-defined.

Definition 1. A weighting (or matching) method $W(T, X)$ is said to be *worst-case bias minimizing (WCBM)* if for some $\mathcal{W}$ and $(\mathcal{F}, \|\cdot\|)$ satisfying Assumptions 2 and 3 we have

$$W(T, X) \in \arg\min_{W \in \mathcal{W}} \mathfrak{B}(W; \mathcal{F}) \neq \mathcal{W}.$$

Let $\mathfrak{B}_{\min}(\mathcal{F}) = \min_{W \in \mathcal{W}} \mathfrak{B}(W; \mathcal{F})$ be the optimal value. Clearly, if a weighting method $W(T, X)$ is WCBM with $(\mathcal{F}, \|\cdot\|)$ and $\mathcal{W}$ then the bias of $\hat{\tau}_W$ is bounded by

$$|B(W; f_0)| \le \|f_0\| \, \mathfrak{B}_{\min}(\mathcal{F}).$$

5 Existing methods as WCBM

Surprisingly, a great many methods for causal inference that are standard in practice are also in fact WCBM. On the one hand, this interpretation gets at the core of the structural motivations behind many of these methods (e.g., "if you believe the conditional expectation is Lipschitz and nothing more then you should pairwise match") and allows one to choose a method appropriate to one's beliefs about problem structure. On the other hand, these results provide motivation that WCBM is the *right* framework in which to think about weighting and matching for causal inference and this motivates us to consider new WCBM methods in Sec. 6.

5.1 One-to-one matching

One-to-one (pairwise) matching is by far the most common matching method. In one-to-one matching, each treated subject is paired with exactly one control subject so that the sum of pairwise distances is minimized as measured by some distance metric $\delta(x, x')$ on $\mathcal{X}$ [31]. Usually, the Mahalanobis metric is used:

$$\delta(x, x') = \sqrt{(x - x')\hat{\Sigma}^{-1}(x - x'},$$

where $\hat{\Sigma}$ is the pooled sample covariance matrix. One-to-one matching can be done either without replacement (each control subject used at most once) or with replacement (each control subject could be reused and matched to two or more treated subjects). The estimate of SATT is the average pairwise differences of outcomes. This estimator is exactly $\hat{\tau}_W$ where the weight on control subject i is $1/n_1$ times the number of times subject i was matched, i.e., the matched control sample is the (multi-)set of control subjects that got matched to treated subjects.

One-to-one matching is WCBM.

Theorem 2. *One-to-one matching with pairwise distance metric $\delta(x, x')$ with replacement and without replacement are both WCBM with*

- $\|f\| = \sup_{x \neq x'} \frac{f(x)-f(x')}{\delta(x,x')}$*, the Lipschitz constant of f;*
- $\mathcal{F} = \{f : \|f\| < \infty\}$*;*
- $\mathcal{W}_0$ *is either* $\mathcal{W}_0^{\textit{nonnegative}}$ *or* $\mathcal{W}_0^{\textit{w/ rep.}}$ *(with $n_0' = n_1$) if with replacement; and*
- $\mathcal{W}_0$ *is either* $\left\{W_{\mathcal{T}_0} \in \mathcal{W}_0^{\textit{nonnegative}} : n_1 W_i \leq 1 \,\forall i\right\}$ *or* $\mathcal{W}_0^{\textit{w/o rep.}}$ *(with $n_0' = n_1$) if without replacement.*

Remark 1. Note that even if the weights are not restricted to be multiples of $1/n_1$, the *optimal* unrestricted weights will end up to be multiples of $1/n_1$ regardless. That is, the optimal weighting is optimal matching for Lipschitz functions.

Remark 2. Note that $(\mathcal{F}, \|\cdot\|)$ is *not* a Banach space. In particular, constant functions have zero Lipschitz constant. However, as required, $\mathcal{F}/\mathbb{R}$ is a Banach space and evaluation differences are continuous because they are bounded by the magnitude.

Remark 3. Algorithmically, one-to-one matching with replacement amounts to finding the control subject of minimal distance to each treated subject in a greedy manner. One-ton-one matching without replacement amounts to minimum-sum-of-distances bipartite matching with unbalanced parts, which is easily solved by the Ford-Fulkerson algorithm [10].

An alternative to optimal pairwise matching is caliper matching whereby we only match subjects that are within a distance δ_0 from one another. This method is also WCBM.

Corollary 3. *Caliper matching with pairwise distance metric $\delta(x, x')$ and threshold δ_0 (if feasible) with replacement and without replacement are both WCBM with*

- $\|f\| = \sup_{x \neq x'} \frac{f(x)-f(x')}{\max\{\delta_0,\, \delta(x,x')\}}$*;*
- $\mathcal{F} = \{f : \|f\| < \infty\}$*;*
- $\mathcal{W}_0$ *is either* $\mathcal{W}_0^{\textit{nonnegative}}$ *or* $\mathcal{W}_0^{\textit{w/ rep.}}$ *(with $n_0' = n_1$) if with replacement; and*
- $\mathcal{W}_0$ *is either* $\left\{W_{\mathcal{T}_0} \in \mathcal{W}_0^{\textit{nonnegative}} : n_1 W_i \leq 1 \,\forall i\right\}$ *or* $\mathcal{W}_0^{\textit{w/o rep.}}$ *(with $n_0' = n_1$) if without replacement.*

5.2 Coarsened exact matching

CEM [19] is a weighting method whereby one coarsens the covariates into a few (M) strata via a coarsening function $C : \mathcal{X} \to \{1, \dots, M\}$, and then one matches exactly within each stratum. For example, if there are 5 treated subjects and 3 control subjects in a given stratum then each of the 3 control subjects is given weight proportional to $5/3$, whereas if there were 0 treated subject the weights would be 0. The case of a stratum containing only treated subjects is not allowed (no extrapolation). ([20] suggests that in this case one estimates the "feasible average treatment effect on the treated," meaning to modify the sample of interest from the treated sample to the subset that has good matches.) Under Assumption 1, lack of any overlap is rare for large n.

Theorem 4. *CEM with coarsening function $C : \mathcal{X} \to \{1, \dots, M\}$ is WCBM with*

- $\mathcal{F} = \left\{f : \left|f^{-1}(C^{-1}(j))\right| = 1 \;\forall j = 1, \dots, M\right\}$*, i.e., piece-wise constant on the coarsening partitions;*
- $\|f\| = \sup_{x \in \mathcal{X}} |f(x)|$ *for $f \in \mathcal{F}$, otherwise ∞; and*
- $\mathcal{W}_0$ *is either* $\mathcal{W}_0^{\textit{general}}$ *or* $\mathcal{W}_0^{\textit{nonnegative}}$*,*

assuming that each partition that contains a treatment subject also contains a control subject (no extrapolation).

5.3 Mean-matched sampling

Very often, practitioners will evaluate the quality of a matched control sample by measuring the Mahalanobis distance between the matched control sample and the treated sample:

$$M_V(W) = \left\| V^{-1/2} \left(\tfrac{1}{n_1} \textstyle\sum_{i \in \mathcal{T}_1} X_i - \sum_{i \in \mathcal{T}_0} W_i X_i \right) \right\|_2,$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ and V is some positive definite matrix usually taken to be $V = \hat{\Sigma}$, the pooled sample covariance matrix of X. This distance is a rotated 2-norm between the sample means. Mean-matched sampling finds a matched control sample of a prescribed size to minimize this distance.

Theorem 5. *Mean-matched sampling n_0' subjects (with or without replacement) from the control set is WCBM with*

- $\mathcal{F} = \left\{x \mapsto \beta_0 + \beta^T x : \beta \in \mathbb{R}^d\right\}$*;*
- $\|x \mapsto \beta_0 + \beta^T x\| = \sqrt{\beta^T V \beta + \beta_0^2}$ *and $\|f\| = \infty$ otherwise; and*
- $\mathcal{W}_0$ *is either* $\mathcal{W}_0^{\textit{w/ rep}}$ *or* $\mathcal{W}_0^{\textit{w/o rep}}$*, respectively.*

Remark 4. Since finite, the space $(\mathcal{F}, \|\cdot\|)$ is always a Banach space and evaluations (and hence their differences) are always continuous. See Thms. 5.33 and 5.35 of [18].

6 Kernel WCBM methods

In the previous section we saw that a variety of standard methods for causal inference are WCBM. Each was recovered using a different form of structure on the conditional expectations of outcomes. In this section we develop a range of new WCBM based on kernels and their corresponding reproducing kernel Hilbert spaces (RKHS). Kernels are standard in machine learning (ML) as ways to generalize the structure of learned conditional expectation functions, like classifiers or regressors [37]. Kernels have many applications in statistics [3, 13, 45]. The same way kernels are used to generalize the structure of learned functions in ML, we can use these to generalize the structure of f_0. This will lead to new methods for causal inference that are potentially very powerful.

A Hilbert space is an inner-product space such that the norm induced by the inner product, $\|f\|^2 = \langle f, f\rangle$, yields a Banach space. An RKHS $\mathcal{F}$ is a Hilbert space of functions for which, for every $x \in \mathcal{X}$, the map $f \mapsto f(x)$ is a continuous mapping [3]. Continuity and the Riesz representation theorem imply that for each $x \in \mathcal{X}$ there is $\mathcal{K}(x, \cdot) \in \mathcal{F}$ such that $\langle \mathcal{K}(x, \cdot), f(\cdot)\rangle = f(x)$ for every $f \in \mathcal{F}$. The symmetric map $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called the reproducing kernel of $\mathcal{F}$. The name is motivated by the fact that $\mathcal{F} = \mathrm{closure}\left(\mathrm{span}\left\{\mathcal{K}(x, \cdot) \ : \ x \in \mathcal{X}\right\}\right)$. Thus $\mathcal{K}$ fully characterizes $\mathcal{F}$. Prominent examples of kernels for $\mathcal{X} \subset \mathbb{R}^d$ are:

(i) The polynomial kernel $\mathcal{K}_s(x, x') = (1 + x^T x'/s)^s$, whose RKHS spans the finite-dimensional space of all polynomials of degree up to s.

(ii) The exponential kernel $\mathcal{K}(x, x') = e^{x^T x'}$, the infinite-dimensional limit of the polynomial kernel.

(iii) The Gaussian kernel $\mathcal{K}(x, x') = e^{-\left\|x-x'\right\|^2}$. The corresponding RKHS is infinite-dimensional [43].

For $X \in \mathcal{X}^n$ and a kernel $\mathcal{K}$, the Gram matrix is $K_{ij} = \mathcal{K}(X_i, X_j)$, which is always positive semi-definite (PSD). Generally, we normalize the covariate data before putting it in a kernel so that the sample has zero sample mean and identity pooled sample covariance

Some kernels have a special property, known as universality, that allows them to approximate any continuous function arbitrarily well. Both the Gaussian and exponential kernels are universal [41].

Definition 2. For $\mathcal{X}$ compact Hausdorff, a kernel is *universal* if for any continuous function $g : \mathcal{X} \to \mathbb{R}$ and $\epsilon > 0$, there exists $f \in \mathcal{F}$ in the corresponding RKHS such that $\sup_{x \in \mathcal{X}} |f(x) - g(x)| \leq \epsilon$.

As we will see in Sec. 6.1, universality is one way to guarantee model-free consistency.

Note that any RKHS $\mathcal{F}$ satisfies Assumptions 2 and 3. As such it gives rise to WCBM matching and weighting methods.

Theorem 6. *Let $\mathcal{F}$ be an RKHS with kernel $\mathcal{K}$. Let K be the Gram matrix on X. Then,*

$$\mathfrak{B}(W; \mathcal{F}) = \Big(\tfrac{1}{n_1^2} e_{n_1}^T K_{\mathcal{T}_1, \mathcal{T}_1} e_{n_1} + W_{\mathcal{T}_0}^T K_{\mathcal{T}_0 \mathcal{T}_0} W_{\mathcal{T}_0} - \tfrac{2}{n_1} e_{n_1}^T K_{\mathcal{T}_1, \mathcal{T}_0} W_{\mathcal{T}_0}\Big)^{1/2}.$$

Remark 5. If $W_{\mathcal{T}_0} \in \{0, 1/n_0'\}^{\mathcal{T}_0}$ then $\mathfrak{B}(W; \mathcal{F})$ is exactly the kernel *maximum mean discrepancy* (MMD) statistic between the treated sample and the matched control sample. Kernel MMD is a common test statistic in two-sample goodness-of-fit testing [13, 38]. We can interpret minimizing this discrepancy as trying to make the two samples appear to come from the exact same distribution.

Next, we review the various possible methods this can give rise to. We will see that these kernel methods can offer superior inferential power.

In the following, we let $k_0 = K_{\mathcal{T}_0 \mathcal{T}_1} e_{n_1}/n_1$.

Kernel weighting with general weights

For general unconstrained weighting, we can find the optimal weights in closed form. A basic application of Lagrange multipliers to Theorem 6 yields

$$\mathrm{argmin}_{W_{\mathcal{T}_0} \in \mathcal{W}_0^{\text{general}}} \mathfrak{B}(W; \mathcal{F}) = \left\{ u + \tfrac{1 - u^T e_{n_0}}{v^T e_{n_0}} e_{n_0} : \begin{array}{c} K_{\mathcal{T}_0 \mathcal{T}_0} u = k_0 \\ K_{\mathcal{T}_0 \mathcal{T}_0} v = e_{n_0} \end{array} \right\}.$$

If $K_{\mathcal{T}_0 \mathcal{T}_0}$ is invertible, this consists of a single point. Generally, the Gram matrix for the Gaussian and exponential kernels is invertible with probability one.

Kernel weighting with nonnegative weights

For nonnegative weights, we can formulate a linearly-constrainted convex-quadratic optimization problem to find the optimal weights:

$$\mathrm{argmin}_{W_{\mathcal{T}_0} \in \mathcal{W}_0^{\text{nonnegative}}} \mathfrak{B}(W; \mathcal{F}) = \mathrm{argmin}_{W \in \mathbb{R}_+^{n_0} : e_{n_0}^T W = 1} \left(W K_{\mathcal{T}_0 \mathcal{T}_0} W - 2 k_0^T W\right).$$

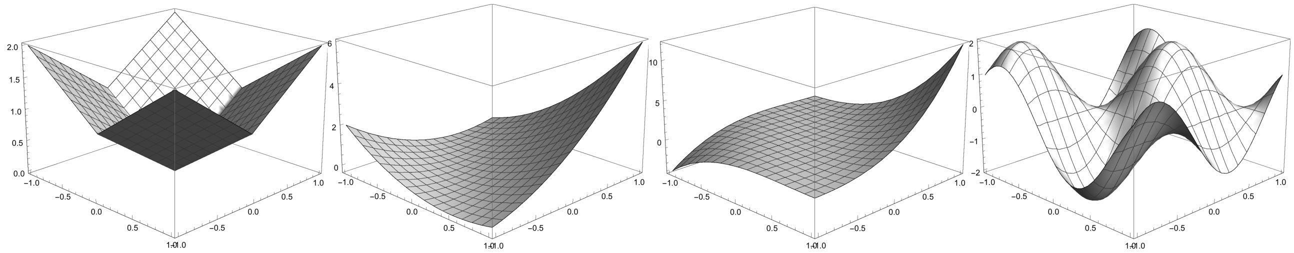
This problem can be solved in polynomial time with interior point methods [4] and is amenable to solution with off-the-shelf solvers like Gurobi.

Kernel matching with replacement

For matching with replacement, we can formulate a linear-integer-constrainted convex-quadratic optimization problem to find the optimal weights:

$$\mathrm{argmin}_{W_{\mathcal{T}_0} \in \mathcal{W}_0^{\text{w/ rep.}}} \mathfrak{B}(W; \mathcal{F}) = \frac{1}{n_0'} \underset{W' \in \mathbb{Z}^{n_0} : e_{n_0}^T W' = n_0'}{\mathrm{argmin}} \left(\tfrac{1}{n_0'} W' K_{\mathcal{T}_0 \mathcal{T}_0} W' - 2 k_0^T W'\right),$$

Figure 1: Various effect functions in Section 7



(a) ℓ_1 norm (b) Quadratic (c) Cubic (d) Sinusoidal

where we used the change of variables $W' = n_0' W_{\mathcal{T}_0}$. This problem is NP-hard (reducible to number partitioning for $\operatorname{rank}(K_{\mathcal{T}_0\mathcal{T}_0}) = 1$), but it is also amenable to solution by off-the-shelf integer programming solvers like Gurobi.

Kernel matching without replacement

For matching without replacement, we can formulate a linear-integer-constrainted convex-quadratic optimization problem to find the optimal weights:

$$\operatorname{argmin}_{W_{\mathcal{T}_0} \in \mathcal{W}_0^{\text{w/o rep.}}} \mathfrak{B}(W; \mathcal{F}) =$$
$$\frac{1}{n_0'} \operatorname*{argmin}_{W' \in \{0,1\}^{n_0} : e_{n_0}^T W' = n_0'} \left(\frac{1}{n_0'} W' K_{\mathcal{T}_0\mathcal{T}_0} W' - 2k_0^T W' \right).$$

Again, the problem is generally "hard" but can be solved in practice using off-the-shelf integer programming solvers.

6.1 Consistency

Next, we express conditions for our kernel estimators to have bias converging to zero. That is, despite the confounding in the data, we can match on X to achieve zero bias.

Definition 3. A Banach space is said to be *B-convex* if there exists $N \in \mathbb{N}$ and $\eta < N$ such that for every $g_1, \ldots, g_N$ with $\|g_i\| \leq 1 \ \forall i$ there exists a choice of signs so that $\|\pm g_1 \pm \cdots \pm g_N\| \leq \eta$.

It is easy to verify that all the Banach spaces so far considered are B-convex. In particular, every Hilbert space or finite-dimensional Banach space is B-convex (see [23] Ch. 9). We use this condition to characterize consistency.

Theorem 7. *Suppose Assumption 1 holds and that*

(i) *the subjects $i = 1, 2, \ldots$ form an iid process;*

(ii) *for each n, W is WCBM with $(\mathcal{F}, \|\cdot\|)$, $\mathcal{W}$ such that $\mathcal{W}_0 \supseteq \left\{W_{\mathcal{T}_0} \in \{0, 1/n_0'\}^{\mathcal{T}_0} : n_0' \in \mathbb{N}, \underline{n}_0' \leq n_0' \leq n_0\right\}$ for some fixed $\underline{n}_0' \geq 1$;*

(iii) *$f_0 \in \operatorname{Closure}_\infty(\mathcal{F})$, i.e., $\forall \epsilon > 0, \exists g_0 \in \mathcal{F} : \sup_{x \in \mathcal{X}} |f_0(x) - g_0(x)| \leq \epsilon$; and*

(iv) *either*

(a) *$\mathcal{F}$ is B-convex and*

$$\mathbb{E}\left[\max\nolimits_{\|f\| \leq 1} \left(f(X_1) - \mathbb{E}\left[f(X_1) \middle| T = 1\right]\right)^2 \middle| T = 1\right] < \infty$$

or

(b) *$\mathcal{F}$ is a Hilbert space and*

$$\mathbb{E}\left[\max\nolimits_{\|f\| \leq 1} \left(f(X_1) - \mathbb{E}\left[f(X_1) \middle| T = 1\right]\right) \middle| T = 1\right] < \infty.$$

Then,

$$\mathbb{E}\left[\hat{\tau}_W - \mathrm{SATT} \middle| X, T\right] \longrightarrow 0 \ \textit{almost surely, as} \ n \to \infty.$$

Moreover, if $\operatorname{Var}\left(Y_1 \middle| X_1\right)$ *is bounded and* $\max_i W_i \to 0$, *then*

$$\hat{\tau}_W - \mathrm{SATT} \longrightarrow 0 \ \textit{in probability, as} \ n \to \infty.$$

Remark 6. One way to satisfy condition (iii) is to have $f_0 \in \mathcal{F}$, i.e., to make the correct structural assumption. Universal kernels, however, always satisfy condition (iii) whenever f_0 is continuous.

Remark 7. Note that the result is quite strong: for almost all realization of subjects, the bias is eventually zero. This is stronger than the average over subjects (a coarser notion of bias) being zero. In particular, as shown, under some additional conditions, we get convergence of the estimation error to zero in probability.

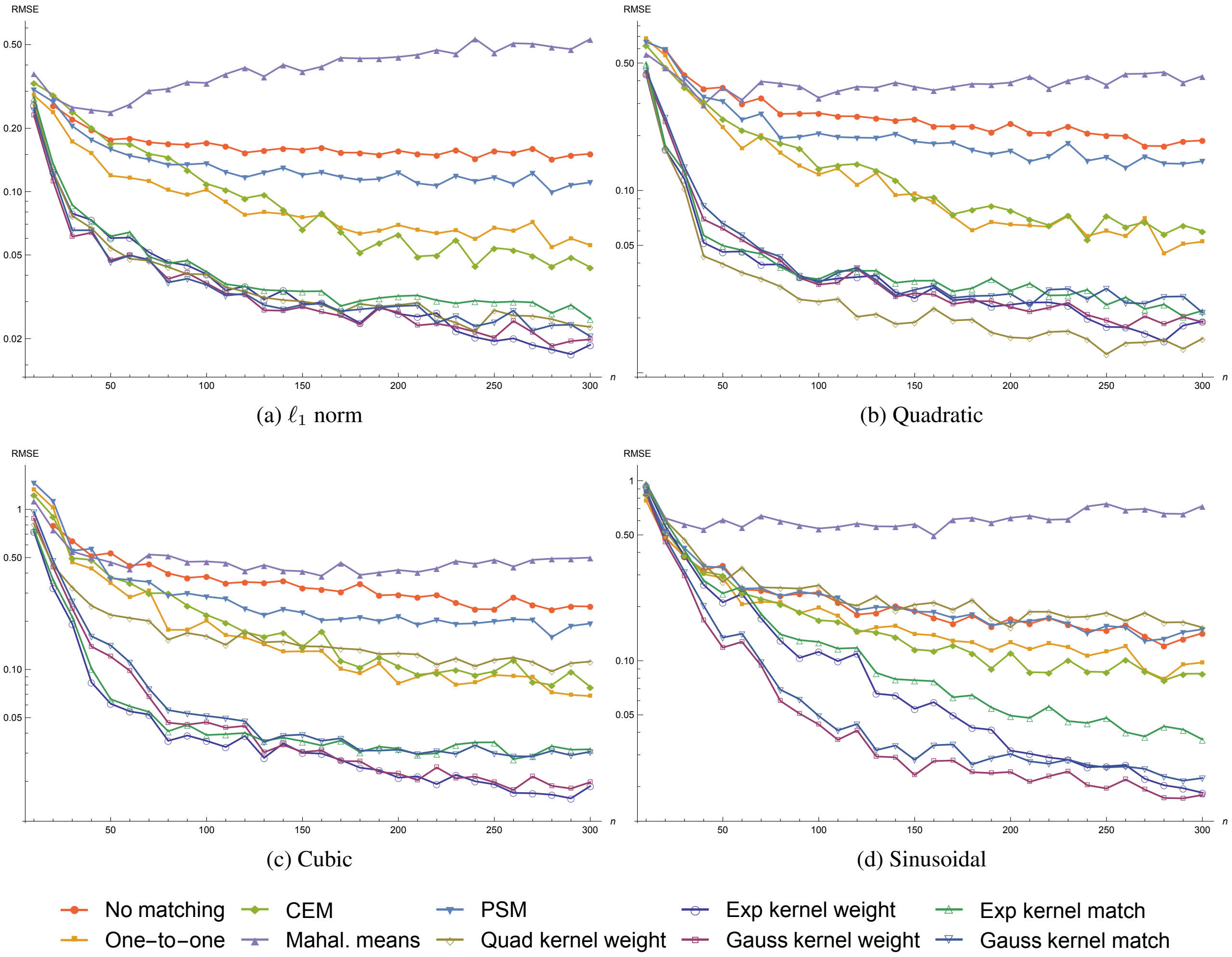
Remark 8. If $Y_1 | X_1$ is homoscedastic then $\operatorname{Var}\left(Y_1 \middle| X_1\right)$ is trivially bounded (constant). Matching estimators achieve $\max_i W_i \to 0$ if we let the size of the matched control sample grow. For weighting, a constraint can be added.

7 Numerical experiments

In this section, we study the comparative efficiency of various causal estimators, including our new kernel estimators.

Consider the following fictitious observational study with one treatment and control. Subjects are drawn at random

Figure 2: RMSE (on log scale) of various causal estimators for various effect functions in Section 7



(a) ℓ_1 norm

(b) Quadratic

(c) Cubic

(d) Sinusoidal

from a population. For each subject we observe a two-dimensional vector of covariates $X_i \in \mathbb{R}^2$. In the population, these are distributed as uniform on $[-1, 1]^2$. Each subject has either received treatment or control and we observe T_i. In the population, T_i is distributed as Bernoulli with probability $0.8/\left(1+\sqrt{2}\,\|X_i\|_2\right)$, which ranges $0.27 \sim 0.8$.

The potential outcomes are distributed as

$$Y_i(0) = f_0(X_i) + \epsilon_{0i},\; Y_i(1) = f_1(X_i) + \epsilon_{1i},$$

where $\epsilon_{0i}, \epsilon_{1i} \sim \mathcal{N}(0, 0.1)$ is independent noise. We focus on the case of small residual noise (variance not explained by X_i) so to tease out the comparative efficiency in matching X (if residual noise is big, any method that only matches on X will do badly). We let f_1 be any function whatsoever. We consider a variety of possible cases for f_0:

(a) ℓ_1 norm: $f_0(x) = |x_1| + |x_2|$;

(b) quadratic: $f_0(x) = (x_1 + x_2) + (x_1 + x_2)^2$;

(c) cubic: $f_0(x) = (x_1 + x_2)^2 + (x_1 + x_2)^3$;

(d) sin: $f_0(x) = \sin(\pi(x_1 + x_2)) + \cos(\pi(x_1 - x_2))$.

These are shown in Figure 1.

For each $n = 10, 20, \ldots, 300$, we produce 100 replicates. For each experiment we consider a variety of estimators:

(a) No matching: we take the whole control sample to be the matched sample ($W_i = 1/n_0$);

(b) One-to-one: we match n_1 control subjects using optimal bipartite matching on the matrix of pairwise Mahalanobis distances between treated and control subjects;

(c) CEM: we find the largest $b \geq 0$ such that coarsening each of the covariates into even bins $\{[-1, -1 + 2^{b-1}), \ldots, [1 - 2^{b-1}, 1]\}$ leaves no box (product of two bins) that contains only treated subjects, then we perform exact matching within each box;

(d) Mahal. means: we match n_1 control subjects with replacement to minimize the Mahalanobis distance between the means of the two samples;

(e) PSM: we match n_1 control subjects using propensity score matching by fitting a logistic regression to impute propensity scores and doing optimal bipartite matching on imputed scores;

(f) Quad kernel weight: we use nonnegative kernel weighting with the quadratic kernel;

(g) Exp kernel weight: we use nonnegative kernel weighting with the exponential kernel;

(h) Gauss kernel weight: we use nonnegative kernel weighting with the Gaussian kernel;

(i) Exp kernel match: we match n_1 control subjects with replacement using kernel matching with the exponential kernel; and

(j) Gauss kernel match: we match n_1 control subjects with replacement using kernel matching with the Gaussian kernel.

We use Gurobi v6.5 (`www.gurobi.com`) to solve all quadratic and integer optimization problems. For each estimator, we compute $\hat{\tau}_W - \mathrm{SATT}$. Then, we measure the RMSE over the 100 replicates, $\mathrm{RMSE} = (\hat{\mathbb{E}}_{100}\left[(\hat{\tau}_W - \mathrm{SATT})^2\right])^{1/2}$. We plot the results in Figure 2. Note the log scale.

The results clearly show the power of our approach. In each case, every one of our exponential- or Gaussian-kernel-based estimators outperforms standard causal estimators by an order of magnitude (base 10). The advantage is particularly noticeable in smaller samples and for our kernel weighting methods. This can be explained by the fact that it can be difficult to find a good control pair for every treated subject in small samples, and similarly it can be difficult to have a fine enough coarsening of the data without creating a stratum that only has treated subjects. At the same time, by *optimizing* the mismatch as characterized by the dual norm of the bias one can achieve small mismatch with even small samples. This is in agreement with the observation made by [21] that one only requires a small sample to have a very small objective in a multi-criterion partitioning problem.

Another observation is that matching based on parametric models can be fragile. This can be seen here for PSM, which is based on a misspecified logistic model, and also for estimators that match on X itself. We also see that mean-matched sampling does very poorly in every example, even doing worse than no matching. Indeed, matching the means only makes sense if the effect is *purely linear*. A linear model assumption is very fragile and even small violations can trip up mean-matched sampling. Similarly, matching per the quadratic kernel depends on an assumption of quadratic effect. Indeed, the estimator based on the quadratic kernel does the best of all estimators when the effect is quadratic (panel b). However, unlike linear, a quadratic model is generally more robust as quadratics can better approximate a wider range of functions. Accordingly, we see that the estimator based on the quadratic kernel has reasonable performance even when the effect is not quadratic (panels a and c), while extreme violations trip it up (panel d).

Overall, the universal kernels (exponential and Gaussian) seem to do the best by far. They appear to provide a good balance between generality of model with efficiency of balancing. They are general enough so that we can ensure consistency even if the true effect is not in the corresponding RKHS. And, fully optimizing mismatch as measured by the dual norm of the bias in their RKHS can lead to small objective value even for moderate n.

8 Conclusion

We presented a novel framework for matching and weighting estimators for causal inference from observational data. The framework is based on minimizing the dual norm of the bias operator with respect to a space of possible conditional expectation functions. Many existing methods common in practice appear to fit this framework. We developed new, kernel-based estimators using the framework and showed they satisfy consistency. Our new estimators prove exceedingly successful in a numerical experiment.

References

[1] R. Ahuja, T. Magnanti, and J. Orlin. *Network flows: theory, algorithms, and applications*. Prentice Hall, Upper Saddle River, 1993.

[2] A. Beck. A convexity condition in Banach spaces and the strong law of large numbers. *Proc. Amer. Math. Soc.*, 13(2):329–334, 1962.

[3] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, 2004.

[4] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[5] Y.-X. Chen and W.-J. Zhu. Note on the strong law of large numbers in a Hilbert space. *Gen. Math.*, 19(3):11–18, 2011.

[6] W. G. Cochran. *Planning and analysis of observational studies*. John Wiley & Sons, New York, 1983.

[7] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 2009.

[8] M. R. Elliott. Model averaging methods for weight

trimming. *Journal of official statistics*, 24(4):517, 2008.

[9] R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, 1925.

[10] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canadian journal of Mathematics*, 8(3):399–404, 1956.

[11] D. A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.

[12] A. S. Goldberger. Structural equation methods in the social sciences. *Econometrica*, 1972.

[13] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, 2006.

[14] X. S. Gu and P. R. Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420, 1993.

[15] K. Hirano and G. W. Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3-4):259–278, 2001.

[16] D. E. Ho, K. Imai, G. King, and E. A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236, 2007.

[17] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. 2009.

[18] J. K. Hunter and B. Nachtergaele. *Applied Analysis*. World Scientific, 2001.

[19] S. M. Iacus, G. King, and G. Porro. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, page mpr013, 2011.

[20] S. M. Iacus, G. King, and G. Porro. A theory of statistical inference for matching methods in applied causal research. 2015.

[21] N. Kallus. Optimal a priori balance in the design of controlled experiments. 2014.

[22] G. King and R. Nielsen. Why propensity scores should not be used for matching. 2015.

[23] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

[24] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, New York, 1998.

[25] J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2000.

[26] J. Pearl. Remarks on the method of propensity score. *Statistics in Medicine*, 28(9):1415–1416, 2009.

[27] J. Pearl. The causal foundation of structural equation modeling. In R. Hoyle, editor, *Handbook of structural equation modeling*, pages 68–91. Sage, Newbury Park, 2012.

[28] J. Pearl et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[29] J. Peters, J. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. 2012.

[30] G. Pisier. Martingales in banach spaces (in connection with type and cotype). 2011.

[31] P. R. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989.

[32] P. R. Rosenbaum. *Observational studies*. Springer, New York, 2002.

[33] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[34] H. L. Royden. *Real Analysis*. Prentice Hall, 1988.

[35] D. B. Rubin. Comments on "randomization analysis of experimental data". *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[36] D. B. Rubin. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28(9):1420–1423, 2009.

[37] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[38] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statis.*, 41(5):2263–2291, 2013.

[39] J. S. Sekhon. The neyman-rubin model of causal inference and estimation via matching methods, 2008.

[40] I. Shrier. Propensity scores. *Statistics in Medicine*, 28(8):1317–1318, 2009.

[41] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. 2010.

[42] O. Stegle, D. Janzing, K. Zhang, J. M. Mooij, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, pages 1687–1695, 2010.

[43] I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inform. Theory*, 52(10):4635–4643, 2006.

[44] E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[45] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. 2012.

9 Appendix A: Proofs

Proof of Theorem 1. Let us write SATT as

$$\mathrm{SATT} = \tfrac{1}{n_1}\sum_{i\in\mathcal{T}_1} Y_i - \tfrac{1}{n_1}\sum_{i\in\mathcal{T}_1} Y_i(0).$$

It is then clear that SATT differs from $\hat\tau_W$ only in the second term, that is,

$$\begin{aligned}\hat\tau - \mathrm{SATT} &= \tfrac{1}{n_1}\sum_{i\in\mathcal{T}_1} Y_i(0) - \sum_{i\in\mathcal{T}_0} W_iY_i(0)\\ &= \sum_{i=1}^n (-1)^{T_i+1}W_iY_i(0)\\ &= \sum_{i=1}^n(-1)^{T_i+1}W_if_0(X_i) + \sum_{i=1}^n(-1)^{T_i+1}W_i\epsilon_i,\end{aligned}$$

where we recognize the last term as $E(W)$. For each term of $E(W)$ we have

$$\begin{aligned}&\mathbb{E}\left[(-1)^{T_i+1}W_i\epsilon_i\middle|X,T\right]\\ &\quad=(-1)^{T_i+1}W_i\left(\mathbb{E}\left[Y_i(0)|X,T\right]-f_0(X_i)\right)\\ &\quad=(-1)^{T_i+1}W_i\left(\mathbb{E}\left[Y_i(0)|X\right]-f_0(X_i)\right)=0,\end{aligned}$$

where the first equality is by definition of ϵ_i and the fact that $W_i = W_i(X,T)$ and the second is by Assumption 1. □

Proof of Theorem 2. Let D be the distance matrix $D_{ii'} = \delta(X_i, X_{i'})$. For this choice of $(\mathcal{F}, \|\cdot\|)$, by linear optimization duality we get

$$\begin{array}{lll}\mathfrak{B}(W;\mathcal{F}) = \frac{1}{n_1}\sup\limits_{v_i-v_{i'}\le D_{ii'}\ \forall i,i'}\left(\sum\limits_{i\in\mathcal{T}_1}v_i - \sum\limits_{i\in\mathcal{T}_0}n_1W_iv_i\right) & & \\ \quad= \frac{1}{n_1}\min_S & \sum_{i,i'} D_{ii'}S_{ii'} & \\ \qquad\qquad\text{s.t.} & S\in\mathbb{R}_+^{n\times n} & \\ & \sum_{i'=1}^n(S_{ii'}-S_{i'i}) = 1 & \forall i\in\mathcal{T}_1\\ & \sum_{i'=1}^n (S_{ii'}-S_{i'i}) = -n_1W_i & \forall i\in\mathcal{T}_0.\end{array}$$

This describes a min-cost network flow problem with sources $\mathcal{T}_1$ with inputs 1, sinks $\mathcal{T}_0$ with outputs W_i, edges between every two nodes with costs $D_{ii'}$ and without capacities. Consider any source $i\in\mathcal{T}_1$ and any sink $i'\in\mathcal{T}_0$ and any path $i, i_1, \dots, i_m, i'$. By the triangle inequality, $D_{ii'} \le D_{ii_1} + D_{i_1i_2} + \cdots + D_{i_mi'}$. Therefore, as there are no capacities, it is always preferable to send the flow from the sources to the sinks along the direct edges from $\mathcal{T}_1$ to $\mathcal{T}_0$. That is, we can eliminate all other edges and write

$$\begin{array}{lll}\mathfrak{B}(W;\mathcal{F}) = \frac{1}{n_1}\min_S & \sum_{i\in\mathcal{T}_1,\,i'\in\mathcal{T}_0} D_{ii'}S_{ii'} & \\ \qquad\qquad\text{s.t.} & S\in\mathbb{R}_+^{\mathcal{T}_1\times\mathcal{T}_0} & \\ & \sum_{i'\in\mathcal{T}_0} S_{ii'} = 1 & \forall i\in\mathcal{T}_1\\ & \sum_{i\in\mathcal{T}_1} S_{ii'} = n_1W_i & \forall i'\in\mathcal{T}_0.\end{array}$$

In the case of with replacement and $\mathcal{W}_0 = \mathcal{W}_0^{\text{nonnegative}}$, using the transformation $W_i' = n_1W_i$, we get

$$\begin{array}{lll}\min\limits_{W\in\mathcal{W}}\mathfrak{B}(W;\mathcal{F}) & & \\ \quad= \frac{1}{n_1}\min\limits_{S,W'} & \sum\limits_{i\in\mathcal{T}_1,\,i'\in\mathcal{T}_0} D_{ii'}S_{ii'} & \\ \qquad\text{s.t.} & S\in\mathbb{R}_+^{\mathcal{T}_1\times\mathcal{T}_0} & \\ & W_i'\in\mathbb{R}_+^{\mathcal{T}_0} & \\ & \sum_{i\in\mathcal{T}_0} W_i' = n_1 & \\ & \sum_{i'\in\mathcal{T}_0} S_{ii'} = 1 & \forall i\in\mathcal{T}_1\\ & \sum_{i\in\mathcal{T}_1} S_{ii'} - W_i' = 0 & \forall i'\in\mathcal{T}_0.\end{array}$$

This describes a min-cost netwrok flow problem with sources $\mathcal{T}_1$ with inputs 1; nodes $\mathcal{T}_0$ with 0 exogenous flow; one sink with output n_1; edges from each $i\in\mathcal{T}_1$ to each $i'\in\mathcal{T}_0$ with flow variable $S_{ii'}$, cost $D_{ii'}$, and without capacity; and edges from each $i\in\mathcal{T}_0$ to the sink with flow variable W_i' and without cost or capacity. Because all data is integer, the optimal solution of $W' = n_1W$ is integer (see [1]). Hence, since $\mathcal{W}_0^{\text{w/ rep.}} \subseteq \mathbb{Z}/n_1$, the solution is the same when we restrict to $\mathcal{W}_0 = \mathcal{W}_0^{\text{w/ rep.}}$. This solution (in terms of W') is equal to sending the whole input 1 from each source in $\mathcal{T}_1$ to the node in $\mathcal{T}_0$ with smallest distance and from there routing this flow to the sink, which corresponds exactly to one-to-one matching with replacement.

In the case of no replacement and for $\mathcal{W}_0 = \{W\in\mathcal{W}_0^{\text{nonnegative}} : n_1W_i \le 1\,\forall i\}$, using the transformation $W_i' = n_1W_i$, we get

$$\begin{array}{lll}\min\limits_{W\in\mathcal{W}}\mathfrak{B}(W;\mathcal{F}) & & \\ \quad= \frac{1}{n_1}\min\limits_{S,W'} & \sum\limits_{i\in\mathcal{T}_1,\,i'\in\mathcal{T}_0} D_{ii'}S_{ii'} & \\ \qquad\text{s.t.} & S\in\mathbb{R}_+^{\mathcal{T}_1\times\mathcal{T}_0} & \\ & W_i'\in\mathbb{R}_+^{\mathcal{T}_0} & \\ & \sum_{i\in\mathcal{T}_0} W_i' = n_1 & \\ & W_i' \le 1\quad \forall i\in\mathcal{T}_0 & \\ & \sum_{i'\in\mathcal{T}_0} S_{ii'} = 1 & \forall i\in\mathcal{T}_1\\ & \sum_{i\in\mathcal{T}_1} S_{ii'} - W_i' = 0 & \forall i'\in\mathcal{T}_0.\end{array}$$

This describes the same min-cost netwrok flow problem except that the edges from each $i\in\mathcal{T}_0$ to the sink have a capacity of 1. Because all data is integer, the optimal solution of S and $W' = n_1W$ is integer (see [1]). Hence, since $\mathcal{W}_0^{\text{w/o rep.}} \subseteq \mathbb{Z}/n_1$, the solution is the same when we restrict to $\mathcal{W}_0 = \mathcal{W}_0^{\text{w/o rep.}}$. The optimal $S_{ii'}$ is integer and so, by $\sum_{i'\in\mathcal{T}_0} S_{ii'} = 1$, for each $i\in\mathcal{T}_1$ there is exactly one $i'\in\mathcal{T}_0$ with $S_{ii'} = 1$ and all others are zero. $S_{ii'} = 1$ denotes matching i with i'. The optimal W_i' is integral and so, by $W_i' \le 1$, $W_i' \in \{0,1\}$. Hence, for each $i\in\mathcal{T}_0$, $\sum_{i'\in\mathcal{T}_1} S_{ii'} \in \{0,1\}$ so we only use node i at most once. The cost of S is exactly the sum of pairwise distances in the match. Hence, the optimal solution corresponds exactly to one-to-one matching without replacement. □

Proof of Corollary 3. Apply Theorem 2 with the metric $\delta'(x,x') = \mathbb{I}_{[x\neq x']}\max\{\delta(x,x'),\,\delta_0\}$. □

Proof of Theorem 4. This choice of space leads to

$$\mathfrak{B}(W;\mathcal{F}) = \sum_{j=1}^{M} \left| \tfrac{1}{n_1} \sum_{i\in\mathcal{T}_1} \mathbb{I}_{[C(X_i)=j]} - \sum_{i\in\mathcal{T}_0} W_i \mathbb{I}_{[C(X_i)=j]} \right|.$$

That is, the worst-case f assigns ±1 to each partition in order to make the difference of values in that partition be nonnegative. Then clearly the optimal choice of $W \in \mathbb{R}^{\mathcal{T}_0}$ is to make each of these absolute values equal zero. This happens exactly when, for each $i \in \mathcal{T}_0$,

$$\begin{aligned} W_i &= \tfrac{1}{n_1} \tfrac{\left|i'\in\mathcal{T}_1 : C(X_{i'})=C(X_i)\right|}{\left|i'\in\mathcal{T}_0 : C(X_{i'})=C(X_i)\right|} \\ &= \tfrac{1}{n_1} \tfrac{\text{num treatment subjects in same partition as } i}{\text{num control subjects in same partition as } i}, \end{aligned}$$

where $0/0 = 0$ and we never encounter dividing a positive integer by 0 due to the no-extrapolation assumption. Because the weight is nonnegative, the solution is unchanged when restricting to nonnegative weights. □

Proof of Theorem 5. By duality of norms,

$$\begin{aligned} \mathfrak{B}(W;\mathcal{F}) &= \sup_{\beta^T V \beta \le 1} \beta^T \left(\tfrac{1}{n_1} \sum_{i\in\mathcal{T}_1} X_i - \sum_{i\in\mathcal{T}_0} W_i X_i \right) \\ &= M_V(W). \end{aligned}$$

The optimal W minimizes this discrepancy over subsamples from control with the allowable size. □

Proof of Theorem 6. We have

$$\begin{aligned} &\mathfrak{B}^2(W;\mathcal{F}) = \max_{\|f\|\le1} \left(\textstyle\sum_{i=1}^n (-1)^{T_i+1} W_i f(X_i)\right)^2 \\ &= \left\langle \textstyle\sum_{i=1}^n (-1)^{T_i+1} W_i \mathcal{K}(X_i,\cdot), \sum_{i=1}^n (-1)^{T_i+1} W_i \mathcal{K}(X_i,\cdot) \right\rangle \\ &= \textstyle\sum_{i,j=1}^n (-1)^{T_i+T_j} W_i W_j K_{ij}, \end{aligned}$$

which when written in block form gives rise to the result. □

Proof of Theorem 7. First we show $\mathfrak{B}_{\min}(\mathcal{F}) \to 0$ a.s. by showing that we can construct a feasible $\tilde{W}$ such that $\mathfrak{B}(\tilde{W};\mathcal{F}) \to 0$ a.s. Let $p(x) = \mathbb{P}\left(T=1 \middle| X=x\right)$. By Assumption 1, $0 < p(X) < 1$ a.s. So there exists $\alpha > 0$ such that $q(x) = \alpha p(x)/(1-p(x))$ is a.s. in $(0,1)$. For each i, let $\tilde{W}'_i \in \{0,1\}$ be Bernoulli with probability $q(X_i)$. Then we have that $X_i \big| T=0, \tilde{W}'_i = 1$ is distributed as $X_i \big| T = 1$. Let $n'_0 = \sum_{j\in\mathcal{T}_0} \tilde{W}'_j$ and note that $n'_0 \ge \underline{n}'_0$ eventually a.s. For each $i \in \mathcal{T}_0$, set $\tilde{W}_i = \tilde{W}'_i / n'_0$. Let $\zeta(f) = \mathbb{E}\left[f(X_1)\middle| T=1\right]$, $\xi_i(f) = (T_i + \tilde{W}'_i)\left(f(X_i) - \zeta(f)\right)$. Let $A_0 = \frac{1}{n_0}\sum_{i\in\mathcal{T}_0} \xi_i$ and $A_1 = \frac{1}{n_1}\sum_{i\in\mathcal{T}_1}\xi_i$. Adding and subtracting ζ, we see $B(\tilde{W};f) = A_1(f) - (n_0/n'_0)A_0(f)$. By construction of $\tilde{W}'_i$, we see that $\mathbb{E}\left[\xi_i\right] = 0$ (i.e., Bochner integral). By (iv), $\|\xi\|_*$ has (a) second or (b) first moment. By (i), each ξ_i is independent. Therefore, by [2] for (iv)(a) (since B-convexity of $\mathcal{F}$ implies B-convexity of $\mathcal{F}^*$ [30]) or by [5] for (iv)(b), a law of large numbers holds yielding, a.s., $\|A_0\|_* \to 0$ and $\|A_1\|_* \to 0$. Since $(n_0/n'_0) \to \alpha\mathbb{E}\left[p(X_1)\right] < \infty$ a.s., we have that $\|B(\tilde{W};\cdot)\|_* \to 0$ a.s. Since $\tilde{W}$ is feasible, a.s.

$$\mathfrak{B}_{\min}(\mathcal{F}) = \mathfrak{B}(W;\mathcal{F}) \le \mathfrak{B}(\tilde{W};\mathcal{F}) = \|B(\tilde{W};\cdot)\|_* \to 0.$$

Fix $\epsilon > 0$. By (iii), $\exists g_0 \in \mathcal{F} : \sup_x |f_0(x) - g_0(x)| \le \epsilon/2$.

$$\begin{aligned} |B(W;f_0)| &\le |B(W;g_0)| + 2 \sup_{i=1,\dots,n} |f_0(X_i) - g_0(X_i)| \\ &\le \|g_0\|\,\mathfrak{B}(W;\mathcal{F}) + \epsilon = \|g_0\|\,\mathfrak{B}_{\min} + \epsilon \to \epsilon. \end{aligned}$$

Since true for any $\epsilon > 0$, $|B(W;f_0)| \to 0$ a.s. By Assumption 1 and Theorem 1, $\mathbb{E}\left[\hat{\tau}_W - \text{SATT}\middle| X,T\right] = B(W;f_0)$.

By Theorem 1, $\hat{\tau}_W - \text{SATT} = B(W;f_0) + E(W)$. By the strong law of large numbers, $\frac{1}{n_1}\sum_{i\in\mathcal{T}_1}\epsilon_i \to 0$ a.s. By mean-independence of the residual,

$$\mathbb{E}\left[\left(\textstyle\sum_{i\in\mathcal{T}_0} W_i\epsilon_i\right)^2 \middle| X,T\right] = \textstyle\sum_{i\in\mathcal{T}_0} W_i^2 \operatorname{Var}\left(Y_i \middle| X_i\right).$$

By assumption $\operatorname{Var}\left(Y_i\middle|X_i\right) \le M < \infty$ so the above is bounded by $M(\max_i W_i)^2 \to 0$. Since both convergence in L_2 and convergence a.s. imply convergence in probability, we get the desired result. □