Journal of Machine Learning Research 21 (2020) 1-54 Submitted 2/19; Revised 3/20; Published 4/20

Generalized Optimal Matching Methods for Causal Inference

Nathan Kallus KALLUS@CORNELL.EDU
Department of Operations Research and Information Engineering and Cornell Tech
Cornell University
New York, NY 10044, USA

Editor: Joris Mooij

Abstract

We develop an encompassing framework for matching, covariate balancing, and doubly-robust methods for causal inference from observational data called generalized optimal matching (GOM). The framework is given by generalizing a new functional-analytical formulation of optimal matching, giving rise to the class of GOM methods, for which we provide a single unified theory to analyze tractability and consistency. Many commonly used existing methods are included in GOM and, using their GOM interpretation, can be extended to optimally and automatically trade off balance for variance and outperform their standard counterparts. As a subclass, GOM gives rise to kernel optimal matching (KOM), which, as supported by new theoretical and empirical results, is notable for combining many of the positive properties of other methods in one. KOM, which is solved as a linearly-constrained convex-quadratic optimization problem, inherits both the interpretability and model-free consistency of matching but can also achieve the $\sqrt{n}$-consistency of well-specified regression and the bias reduction and robustness of doubly robust methods. In settings of limited overlap, KOM enables a very transparent method for interval estimation for partial identification and robust coverage. We demonstrate this in examples with both synthetic and real data.

Keywords: Causal inference, optimal covariate balance, embeddings, matching, convex optimization.

1. Introduction

In causal inference, matching is the pursuit of comparability between samples that differ in systematic ways due to selection (often self-selection) by way of subsampling or re-weighting the samples (Stuart, 2010). Optimal matching (Rosenbaum, 1989), wherein each treated unit is matched to one or more control units (for estimating the effect on the treated) to minimize some objective (such as sum) in the list of within-match pairwise distances so to optimize comparability,[1] as implemented in the popular *R* package `optmatch`, is arguably one of the most commonly used methods for causal inference on treatment effects, whether used as an estimator or as a preprocessing step before regression analysis (Ho et al., 2007).

1. The term "optimal matching" has also been used in other contexts such as full matching (Rosenbaum, 1991) and optimal near-fine balance (Zubizarreta, 2012), but the most common usage by far, which we follow here, refers to matching made on one-to-one, one-to-many, or many-to-many basis using network flow and bipartite approaches, such as optimal bipartite (one-to-one) matching with weights equal to covariate vector distances.

Since the introduction of optimal matching, a variety of other methods for matching on covariates have been developed, including coarsened exact matching (Iacus et al., 2011a), genetic matching (Diamond and Sekhon, 2013), combining optimal matching with near-fine balance on one stratification (Yang et al., 2012), and using integer programming to match sample mean vectors for simultaneous near-fine balance on multiple stratifications (Zubizarreta, 2012). These have largely been developed independently and ad hoc given distinct motivations and definitions for what is "balance" exactly and how to improve it relative to no matching, optimally or non-optimally. There are also a variety of other methods for causal inference on treatment effects such as regression analysis (Lin, 2013), propensity score matching and weighting (Rosenbaum and Rubin, 1983), and doubly robust methods that combine the latter two (Robins et al., 1994).

In this paper, we develop an encompassing framework and theory for matching and weighting methods and related methods for causal inference that reveal the connections and motivations behind these various existing methods and, moreover, give rise to new and improved ones. We begin by providing a functional analytical characterization of optimal matching as a weighting method that minimizes worst-case conditional mean squared error given the observed data and given assumptions on (a) the space of feasible conditional expectation functions, (b) the space of feasible weights, and (c) the magnitude of residual variance. By generalizing the lattermost, we develop a new optimal matching method that correctly and automatically accounts for the balance-variance trade-off inherent in matching and by doing so can reduce effect estimation error. By generalizing all three and using functional analysis and modern optimization, we develop a new class of generalized optimal matching (GOM) methods that construct matched samples or redistributions of the units to eliminate imbalances. This also gives rise to the interpretation of *imbalance* as the *dual norm of bias*. It turns out that many existing methods are included in GOM, including nearest-neighbor matching, one-to-one matching, optimal caliper matching, coarsened exact matching, various near-fine balance approaches, and linear regression adjustment. Moreover, using the lens of GOM many of these too are extended to new methods that judiciously and automatically trade off balance for variance and that outperform their standard matching counterparts. We provide theory on both tractability and consistency that applies generally to GOM methods.

Finally, as a subclass of GOM, we develop kernel optimal matching (KOM), which is particularly notable for combining the interpretability and potential use as preprocessing of matching methods (Ho et al., 2007), the non-parametric nature and model-free consistency of optimal matching (Rosenbaum, 1989; Abadie and Imbens, 2006), the $\sqrt{n}$-consistency of well-specified regression-based estimators (Lin, 2013), the robustness (Robins et al., 1994) of augmented inverse propensity weight estimators, the careful selection of matched sample size of monotonic imbalance bounding methods (Iacus et al., 2011b), and the model-selection flexibility of Gaussian-process regression (Williams and Rasmussen, 2006). We show that KOM can be interpreted as Bayesian optimal in a certain sense, that it is computationally tractable, and that it is consistent. We discuss how to tune the hyperparameters of KOM and demonstrate the efficacy of doing so. KOM also allows for a transparent way to bound any irreducible biases due to a lack of overlap between the control and treated populations, which leads to robust interval estimates that can partially identify effects in a highly interpretable manner. We develop the augmented kernel weighted estimator and

establish robustness and bias reduction guarantees for it related to those of the augmented inverse propensity weighted estimator. Furthermore, we establish similar guarantees for KOM used as a preprocessing step before linear regression, rigorously establishing that it reduces model dependence and yields fast estimation under a well-specified model without ceding model-free consistency under misspecification. We study the practical usefulness of KOM by applying the new methods developed to a semi-simulated case study using real data and find that KOM offers significant benefits in estimation error and also in robustness to practical issues like limited overlap and lack of model specification.

The paper proceeds as follows. In Sec. 2, we setup the problem and provide a reinterpretation of optimal matching from a functional analytical lens. Based on this, we define GOM in Sec. 3. Specifically, we generalize the notion of balance in Sec. 3.1, discuss its trade off with variance in Sec. 3.2, define GOM precisely in Sec. 3.3, establish computational tractability in Sec. 3.4, review the many existing methods GOM encompasses in Sec. 3.5, and provide general consistency guarantees in Sec. 3.6. We then present the sub-class of KOM methods in Sec. 4, discuss hyperparameter selection in Sec. 4.1, present consistency guarantees that are stronger than the general GOM results in Sec. 4.2, discuss KOM-weighted regression adjustment in Sec. 4.3, extend KOM to semi-kernel optimal matching in Sec. 4.4, discuss kernel selection in Sec. 4.5, provide an empirical study using data from the Infant Health and Development Program in Sec. 4.6, and provide recommendations for practice in Sec. 4.7. In Sec. 5 we provide detail on how GOM generalizes many existing methods, and in Sec. 6 we conclude. Algorithm listings are given in Appendix A. Inference with KOM is discussed in Appendix B. Connections to and generalization of Equal-Percent Bias Reduction are discussed in Appendix C. Finally, all proofs are given in Appendix D.

2. Re-interpreting Optimal Matching

In this section we present the first building blocks toward generalizing optimal matching. We set up the causal estimation problem and provide a bias-variance decomposition of error. Through a new functional analytical lens on optimal matching, we uncover it as a specific case of finding weights that minimize worst-case error, but only under zero residual variance of outcomes given covariates. Our first generalization is to consider non-zero residual variance, giving rise to a balance-variance Pareto-efficient version of optimal matching and to a method that automatically chooses the exchange between balance and variance.

2.1. Setting

The observed data consists of n independent and identically distributed (iid) observations $\{(X_i, T_i, Y_i) : i = 1, \ldots, n\}$ of the variables (X, T, Y), where $X \in \mathcal{X}$ denotes baseline covariates, $T \in \{0, 1\}$ treatment assignment, and $Y \in \mathbb{R}$ outcome. The space $\mathcal{X}$ is general; assumptions about it will be specified as necessary. For $t = 0, 1$, we let $\mathcal{T}_t = \{i : T_i = t\}$ and $n_t = |\mathcal{T}_t|$. We also let $T_{1:n} = (T_1, \ldots, T_n)$ and $X_{1:n} = (X_1, \ldots, X_n)$ denote all the observed treatment assignments and baseline covariates, respectively. Using Neyman-Rubin potential outcome notation (Imbens and Rubin, 2015, Ch. 2), we let $Y_i(0)$, $Y_i(1)$ be the real-valued potential outcomes for unit i and assume the stable unit treatment value assumption (Rubin,

1980) holds. We let $Y_i = Y_i(T_i)$, capturing consistency and non-interference. We define

$$f_0(x) = \mathbb{E}\left[Y(0) \mid X = x\right], \quad \epsilon_i = Y_i(0) - f_0(X_i), \quad \sigma_i^2 = \mathrm{Var}\left(Y_i(0) \mid X_i\right).$$

We consider estimating the *sample average treatment effect on the treated*:

$$\mathrm{SATT} = \tfrac{1}{n_1}\textstyle\sum_{i\in\mathcal{T}_1}(Y_i(1) - Y_i(0)) = \overline{Y}_{\mathcal{T}_1}(1) - \overline{Y}_{\mathcal{T}_1}(0),$$

where $\overline{Y}_{\mathcal{T}_t}(s) = \frac{1}{n_t}\sum_{i\in\mathcal{T}_t} Y(s)$ is the average outcome of treatment s in the t-treated sample. As $\overline{Y}_{\mathcal{T}_1}(1)$ is observed, we consider estimators of the form

$$\hat{\tau} = \overline{Y}_{\mathcal{T}_1}(1) - \hat{\overline{Y}}_{\mathcal{T}_1}(0)$$

for some choice of $\hat{\overline{Y}}_{\mathcal{T}_1}(0)$. We will focus on *weighting* estimators $\hat{\overline{Y}}_{\mathcal{T}_1}(0) = \sum_{i\in\mathcal{T}_0} W_i Y_i$ given weights $W \in \mathbb{R}^{\mathcal{T}_0}$. We consider *honest* weights that only depend on the observed $X_{1:n}, T_{1:n}$ and not on observed outcome data, that is, $W = W(X_{1:n}, T_{1:n})$. The resulting estimator has the form

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

Here honesty simply refers to the fact that, conditioned on $X_{1:n}, T_{1:n}$, we are focusing on *linear* estimators. An alternative weighting estimator, which we call the augmented estimator, can be derived as a generalization of the doubly-robust augmented inverse propensity weighting (AIPW) estimator (Robins et al., 1994; Robins, 1999; Scharfstein et al., 1999; Robins et al., 1994):

$$\hat{\tau}_{W,\hat{f}_0} = \tfrac{1}{n_1}\textstyle\sum_{i\in\mathcal{T}_1}(Y_i - \hat{f}_0(X_i)) - \sum_{i\in\mathcal{T}_0} W_i(Y_i - \hat{f}_0(X_i)), \tag{2}$$

where $\hat{f}_0(x)$ is a regression estimator for $f_0(x)$. The standard AIPW for SATT would be given by $W_{\hat{p},i} = \frac{\hat{p}(X_i)}{n_1(1-\hat{p}(X_i))}$ where $\hat{p}(x)$ is a binary regression estimator for $\mathbb{P}\left(T = 1 \mid X = x\right)$.

Given a data set, we measure the risk of a weighting estimator as its conditional mean squared error (CMSE), conditioned on all the observed data upon which the weights depend as honest weights:

$$\mathrm{CMSE}(\hat{\tau}) = \mathbb{E}\left[(\hat{\tau} - \mathrm{SATT})^2 \mid X_{1:n}, T_{1:n}\right].$$

Note that the CMSE is a function of the sample $X_{1:n}, T_{1:n}$. Correspondingly, our optimization analysis will be conditional on the $X_{1:n}, T_{1:n}$ sample, while consistency and inference analysis will consider the randomness of this sample.

When choosing weights W, one may restrict to a certain space of allowable weights $\mathcal{W}$. Throughout, we consider only permutation symmetric sets, satisfying $P\mathcal{W} = \mathcal{W}$ for all permutation matrices $P \in \mathbb{R}^{\mathcal{T}_0\times\mathcal{T}_0}$. For example, $\mathcal{W}^{\text{general}} = \mathbb{R}^{\mathcal{T}_0}$ allows *all* weights; $\mathcal{W}^{\text{simplex}} = \left\{W \geq 0 : \sum_{i\in\mathcal{T}_0} W_i = 1\right\}$ restricts to weights that give a probability measure, preserving the unit of analysis and ensuring no extrapolation in estimating $\overline{Y}_{\mathcal{T}_1}(0)$; $\mathcal{W}^{b\text{-simplex}} = \mathcal{W}^{\text{simplex}} \cap [0,b]^{\mathcal{T}_0}$ further bounds how much weight we can put on a single unit; $\mathcal{W}^{n_0'\text{-multisubset}} = \mathcal{W}^{\text{simplex}} \cap \{0, 1/n_0', 2/n_0', \dots\}^{\mathcal{T}_0}$ limits us to integer-multiple weights that exactly correspond to sub-sampling a multisubset of controls of cardinality n_0'; $\mathcal{W}^{n_0'\text{-subset}} = \mathcal{W}^{1/n_0'\text{-simplex}} \cap \mathcal{W}^{n_0'\text{-multisubset}}$ corresponds to sub-sampling a usual subset of cardinality n_0'; and $\mathcal{W}^{\text{multisubsets}} =$

$\cup_{n_0'=1}^{n_0} \mathcal{W}^{n_0'\text{-multisubset}}$ and $\mathcal{W}^{\text{subsets}} = \cup_{n_0'=1}^{n_0} \mathcal{W}^{n_0'\text{-subset}}$ correspond to sub-sampling any multisubset or subset, respectively. We have the inclusions:

$$\mathcal{W}^{\text{subsets}} \subseteq \mathcal{W}^{\text{multisubsets}} \subseteq \mathcal{W}^{\text{simplex}} \subseteq \mathcal{W}^{\text{general}}. \tag{3}$$

A standing assumption is that of *weak mean-ignorability*:

Assumption 2.1 *Conditioned on X, $Y(0)$ is mean-independent of T:* $\mathbb{E}[Y(0) \mid T, X] = \mathbb{E}[Y(0) \mid X]$.

A second assumption, which we will relax in the context of partial identification, is *overlap.*

Assumption 2.2 $\mathbb{P}(T = 0) \in (0, 1)$ *and* $\mathbb{P}(T = 0 \mid X)$ *is bounded away from 0.*

2.2. Decomposing the Conditional Mean Squared Error

In this section we decompose the CMSE of estimators of the form in eq. (1) into a bias term and a variance term. Let us define

$$\begin{aligned} B(W; f) &= \tfrac{1}{n_1} \textstyle\sum_{i \in \mathcal{T}_1} f(X_i) - \sum_{i \in \mathcal{T}_0} W_i f(X_i) \\ V^2(W; \sigma^2_{1:n}) &= \textstyle\sum_{i \in \mathcal{T}_0} W_i^2 \sigma_i^2 + \frac{1}{n_1^2} \sum_{i \in \mathcal{T}_1} \sigma_i^2 \\ E^2(W; f, \sigma^2_{1:n}) &= B^2(W; f) + V^2(W; \sigma^2_{1:n}) \end{aligned}$$

Proposition 1 *Under Asn. 2.1,*

$$\mathbb{E}[\hat{\tau}_W - \text{SATT} \mid X_{1:n}, T_{1:n}] = B(W; f_0), \quad \text{CMSE}(\hat{\tau}_W) = E^2(W; f_0, \sigma^2_{1:n}).$$

(Note our use of f_0 as the true conditional expectation function of $Y(0)$ and f as a generic function-valued variable in the space of all functions $\mathcal{X} \to \mathbb{R}$.)

The above provides a decomposition of the risk of $\hat{\tau}_W$ into a (conditional) bias term and a (conditional) variance term, which must be balanced to minimize overall risk. The first term, $B(W; f_0)$, is exactly the conditional bias of $\hat{\tau}_W$. We refer to $V^2(W; f_0)$ as the variance term of the error.[2] More generally, if the units are not independent, the proof makes clear that Prop. 1 holds with the variance term $(W, e_{n_1}/n_1)^T \Sigma (W, e_{n_1}/n_1)$ where e_{n_1} is the vector of all ones of length n_1 and Σ is the conditional covariance matrix.

An analogous result holds for the augmented estimator when $\hat{f}_0$ is fit across split folds.

Corollary 2 *Under Asn. 2.1, if $\hat{f}_0 \perp\!\!\!\perp Y_{1:n} \mid X_{1:n}, T_{1:n}$ then*

$$\mathbb{E}[\hat{\tau}_{W,f_0} - \text{SATT} \mid X_{1:n}, T_{1:n}] = B(W; f_0 - \hat{f}_0), \qquad \text{CMSE}(\hat{\tau}_{W,f_0}) = E^2(W; f_0 - \hat{f}_0, \sigma^2_{1:n})$$

2. The conditional variance of $\hat{\tau}_W$ actually differs from $V^2(W; \sigma^2_{1:n})$ by exactly $\frac{1}{n_1^2} \sum_{i \in \mathcal{T}_1} (\text{Var}(Y_i(1) \mid X_i) - \text{Var}(Y_i(0) \mid X_i))$, which accounts for the conditional variance of SATT and its covariance with $\hat{\tau}_W$. Note this difference is constant in W and so it does not matter whether the estimand we consider is SATT or $\mathbb{E}[\text{SATT} \mid X_{1:n}, T_{1:n}]$. Since we do not condition on potential outcomes, $\mathbb{E}[\text{SATT} \mid X_{1:n}, T_{1:n}]$ may be a more natural estimand; by the above observation, however, it makes no difference which is considered the estimand and all the results remain the same for either.

2.3. Re-interpreting Optimal Matching

In this section, we provide an interpretation of optimal matching as minimizing worst-case CMSE. We consider two forms of optimal matching: nearest neighbor matching (NNM) and optimal one-to-one matching (1:1M). In both, each treated unit is matched to one control unit to minimize the sum of distances between matches as measured by a given extended pseudo-metric $\delta(X_i, X_j)$.[3] NNM allows for *replacement* of control units whereas 1:1M does not. In the end, the weight W_i assigned to a control unit $i \in \mathcal{T}_0$ is equal to $1/n_1$ times the number of times it has been matched. So, under 1:1M, W_i is capped at $1/n_1$ and the result is equivalent to constructing a subset of cardinality n_1, where all $n_0 - n_1$ unmatched control units have been pruned away. Under NNM, the result is equivalent to a *multi*-subset of the control sample of cardinality n_1.

Next, consider an alternative perspective. We seek weights W that depend only on data $X_{1:n}, T_{1:n}$ and that minimize the resulting CMSE. The CMSE depends on unknowns: f_0 and $\sigma^2_{1:n}$. In order to get a handle on the CMSE, we make assumptions about these unknowns. First, we assume (implausibly) that X_i is completely predictive of $Y_i(0)$ so that $\sigma_i^2 = 0$. Second, we assume that f_0 is a Lipschitz continuous function with respect to δ. That is,[4]

$$\exists \gamma \geq 0 : \|f_0\|_{\mathrm{Lip}(\delta)} \leq \gamma \quad \text{where} \quad \|f\|_{\mathrm{Lip}(\delta)} := \sup_{x \neq x'} \tfrac{f(x)-f(x')}{\delta(x,x')} \leq \gamma.$$

Assuming nothing else, we may seek W to minimize the worst-case CMSE. If we limit ourselves to simplex weights $\mathcal{W} = \mathcal{W}^{\mathrm{simplex}}$, the next theorem, adapted from Kallus (2017, Theorem 2), shows that this is precisely equivalent to optimal matching.

Theorem 3 *Fix a pseudo-metric $\delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Then, for any $\gamma > 0$, NNM and 1:1M are equivalent to*

$$W \in \mathrm{argmin}_{W \in \mathcal{W}} \sup_{\|f\|_{Lip(\delta)} \leq \gamma} \left\{ E^2(W; f, \mathbf{0}) = B^2(W; f) \right\}, \tag{4}$$

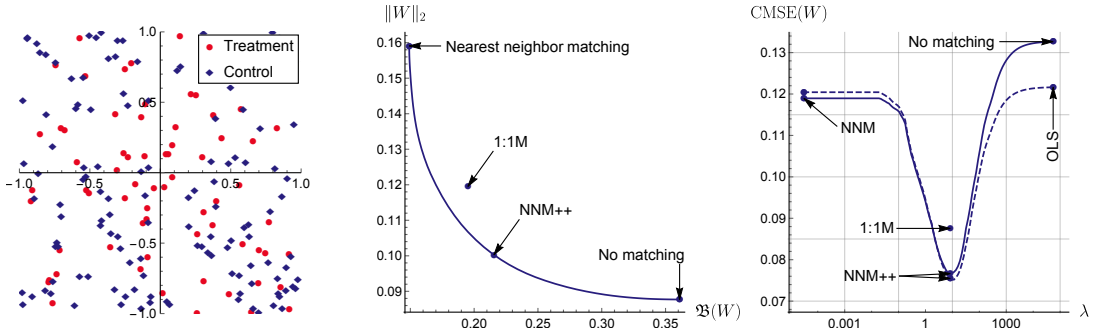
where $\mathcal{W} = \mathcal{W}^{simplex}$ for NNM and $\mathcal{W} = \mathcal{W}^{1/n_1\text{-}simplex}$ for 1:1M.

Therefore, optimal matching is indeed optimal in a minimax CMSE sense, given the assumptions and restrictions made. (Alternatively, without restricting $\sigma_i^2 = 0$, it minimizes just the worst-case bias.) That is, the above theorem relates optimal matching – an *a priori* choice of re-weighting based on data $X_{1:n}, T_{1:n}$ – to the *a posteriori* error of estimating a causal treatment effect and says that this choice minimizes the worst-case error over all γ-Lipschitz functions. Surprisingly, we need not restrict to selecting each control unit an integer multiple number of times – optimal matching, which results in a subset or multisubset of the control sample, minimizes this error among all continuous weights in the (bounded) simplex.

The above reinterpretation is closely related to the Rubinstein-Kantorovich theorem that establishes the dual forms of the Wasserstein metric (Kantorovich and Rubinstein, 1958). The Wasserstein metric is an example of an integral probability metric (IPM; Müller, 1997),

3. Compared to a metric, an extended pseudo-metric may also assign zero or infinity distance to distinct elements.
4. Note that since δ may be a pseudo-metric, $\|f_0\|_{\mathrm{Lip}(\delta)} < \infty$ implies $f(x) = f(x')\ \forall x, x' : \delta(x, x') = 0$.

Figure 1: The Balance-Variance Trade-off in Optimal Matching



(a) $X_{1:n}$ (b) Balance-variance (c) CMSE

which are distance metrics between probability distributions given by the maximum moment discrepancy of functions in a given class. The Wasserstein metric is given by the class of 1-Lipschitz functions and thus eq. (4) shows that optimal matching minimizes the Wasserstein metric between the empirical distributions of the treated sample and the re-weighted control sample. Other examples of IPMs are the kernel maximum mean discrepancy (MMD; Gretton et al., 2006) and the total variation distance.

This reinterpretation of optimal matching involved three critical choices: a restriction on the conditional expectation f_0, a restriction on the space weights $\mathcal{W}$, and a restriction on the magnitude of residual variance $\sigma^2_{1:n}$. In this paper, we consider different such choices that lead to methods that *generalize* optimal matching.

3. Generalizing Optimal Matching

In this section we consider generalizing the restrictions and assumptions that made optimal matching equivalent to minimizing worst-case error. Doing so recovers other common matching methods and other causal estimation methods, as well as give rise to new matching methods such as KOM.

3.1. Generalizing Balance

Balance between the control and treatment samples can be understood as the extent to which they are comparable. In estimating SATT, we want the samples to be comparable on their values of f_0 so that the bias due to the systematic differences between the samples is minimal. When we re-weight the control sample by W, the absolute discrepancy in values of f_0 is $|B(W; f_0)|$. As seen in the preceding section, by minimizing this quantity over all possible realizations of Lipschitz functions, optimal matching is seeking the best possible balance over this class of functions. We now generalize this functional restriction, leading to more general balance metrics called bias-dual-norm balance metrics (Kallus, 2017), which also take the form of some IPM between the empirical distributions of the treated sample and the re-weighted control sample.

Since the bias depends on f_0 but we do not know f_0, we consider guarding against any reasonable realization of f_0. Bias is linear in f_0, *i.e.*, $B(W;\alpha f+\alpha' f')=\alpha B(W;f)+\alpha' B(W;f')$. So, we must limit the "size" of f_0. In particular, we consider the bias relative to some extended magnitude $\|f\|\in[0,\infty]$ of f_0 that is absolutely homogeneous, *i.e.*, $\|\alpha f\|=|\alpha|\,\|f\|$ (where $|\alpha|\,\infty=\infty$), and satisfies the triangle inequality (where $\infty\leq\infty$). This allows us to generalize the notion of balance for optimal matching. Letting $0/0=0$, we redefine imbalance more generally as

$$\mathfrak{B}(W;\|\cdot\|)=\sup_{f:\mathcal{X}\to\mathbb{R}} B(W;f)/\|f\|=\sup_{\|f\|\leq 1} B(W;f), \tag{5}$$

where the last equality is due to the homogeneity of $B(W;\cdot)$ and $\|\cdot\|$. The last equality also shows that $\mathfrak{B}(W;\|\cdot\|)$ is of the form of an IPM between the treated sample and reweighted control sample given by the class functions in the unit ball of $\|\cdot\|$. To ensure that $\mathfrak{B}(W;\|\cdot\|)$ is well-defined, we restrict our attention only to certain magnitude functions. We require that $B(W;\cdot)$ is bounded with respect to $\|\cdot\|$, *i.e.*, $\forall W\in\mathcal{W}\ \exists M_W>0: B(W;f)\leq M_W\|f\|$.[5] Then $\{f:\|f\|<\infty\}$ is a semi-normed[6] vector space and $B(W;\cdot)$ is a well-defined, continuous, linear operator on the Banach completion of the quotient space $\{f:\|f\|<\infty\}/\{f:\|f\|=0\}$, *i.e.*, $B(W;\cdot)$ is in its dual space. (See Ledoux and Talagrand, 1991; Royden, 1988 for Banach spaces.) In particular, $\mathfrak{B}(W;\|\cdot\|)$ is precisely the *dual norm of the bias as an operator on conditional expectation functions*, which is necessarily finite and well-defined for any $W\in\mathcal{W}$:

$$\mathfrak{B}(W;\|\cdot\|)=\|B(W;\ \cdot\ )\|_*<\infty.$$

Definition 4 *Given $\mathcal{W}$ and $\|\cdot\|:[\mathcal{X}\to\mathbb{R}]\to[0,\infty]$ such that $\|\cdot\|$ is absolutely homogeneous and satisfies the triangle inequality and $\forall W\in\mathcal{W}\ \exists M_W>0: |B(W;f)|\leq M_W\|f\|$,*[7] *$\mathfrak{B}(W;\|\cdot\|)$ is called a* bias-dual-norm (BDN) imbalance metric *on $\mathcal{W}$.*

3.2. The Balance-Variance Trade-off

Even if the control and treatment samples are made completely comparable, there is inherent error to the estimation of outcomes in each sample. Just a few controls may provide the best matches, and hence the least bias. But, if σ_i^2 are nonzero, then averaging the outcome in these few units has higher variance than averaging more units by, *e.g.*, finding a few more but perhaps less good matches. In one extreme, if σ_i^2 are zero, there is no added variance and we best use the best matches (*e.g.*, reuse controls with replacement). In the other extreme, we conceive of σ_i^2 being so large, that we do not care about the bias due to imbalance and we would prefer to do *no matching* on the samples so to minimize variance (assuming homoskedasticity) and estimate SATT as the simple mean difference of the raw treated and control samples.

5. This is necessary: were $B(W;\cdot)$ not bounded for some $W\in\mathcal{W}$ then for any $M>0$ we would have some f with $B(W;f)>M\|f\|$ so that indeed $\mathfrak{B}(W;\|\cdot\|)=\infty$ is not defined. Alternatively, we can let $\mathfrak{B}(W;\|\cdot\|)=\infty$ and treat this as a constraint on the feasibility W; the resulting feasible set $\mathcal{W}$ satisfies the condition by construction. Note also that M_W may also depend on $X_{1:n},T_{1:n}$.
6. Compared to a norm, a semi-norm may assign zero magnitude to non-zeros.
7. Note that this condition is relative to $\mathcal{W}$. For example, for $\|\cdot\|=\|\cdot\|_{\mathrm{Lip}(\delta)}$, the condition does hold for $\mathcal{W}=\mathcal{W}^{\mathrm{simplex}}$ but does *not* hold for $\mathcal{W}=\mathcal{W}^{\mathrm{general}}$. So the balance metric in optimal matching is *not* valid for general non-simplex weights.

This trade off between bias and variance is well understood in matching (Zubizarreta, 2015; Chan et al., 2016; Zhao, 2016; Iacus et al., 2011b). The most common approach to this trade off in optimal matching is to disallow replacement (1:1M instead of NNM) and to increase the number of matches (1:kM instead of 1:1M) (Stuart, 2010, §3.1.2). But given an explicit understanding of balance as bounding bias, these are only heuristic and need not be on the Pareto-efficient frontier of achievable balance and variance.

Per Thm. 3, NNM is given by optimizing only balance and ignoring variance. Some approaches like 1:1M seek to alleviate this by forcing a more even distribution of weights. However, per Prop. 1, given an imbalance metric, the best way to trade off balance and variance for minimal error is by directly regularizing imbalance by the sum of squared weights. This suggests that the right way to trade off balance and variance in optimal matching is to consider eq. (4) with $\sigma^2_{1:n} \neq \mathbf{0}$. Plugging $\sigma^2_i = \lambda\gamma^2$ for $\lambda \geq 0$ into eq. (4), we refer to the result as Balance-Variance Efficient Nearest Neighbor Matching (BVE-NNM) because it generates the Pareto-efficient frontier of achievable variance and (matching) balance and, for some λ, it gives the minimax optimal weights. We will revisit BVE-NNM in Sec. 5 and develop NNM++, which automatically selects λ using cross-validation. For now, we explore the balance-variance trade-off in an example.

In general, moving beyond optimal matching toward GOM, Prop. 1 provides an explicit form of the total estimation risk in terms of these competing objectives and suggests that the best choice lies somewhere in between focusing solely on balance or solely on variance, where balance can be understood more broadly than sum of matched-pair distances.

Example 1 *Let* $X \sim \text{Unif}[-1,1]^2$, $\mathbb{P}(T=1 \mid X) = 0.95/(1+\frac{3}{\sqrt{2}}\|X\|_2)$. *Fix a draw of* $X_{1:n}, T_{1:n}$ *with* $n = 200$. *We plot the resulting draw, which has* $n_0 = 130, n_1 = 70$, *in Fig. 1a. For a range of* λ *we compute the resulting BVE-NNM weights using the Mahalanobis distance* $\delta(x,x') = (x-x')\hat{\Sigma}_0^{-1}(x-x')$ *where* $\hat{\Sigma}_0$ *is the sample covariance of* $X \mid T=0$. *We plot the resulting space of achievable balance and variance in Fig. 1b. In one extreme* ($\lambda = 0$) *we have NNM and in the other* ($\lambda = \infty$) *we have no matching. Since 1:1M minimizes the same balance criterion (sum of pair distances), we can plot the balance and variance it achieves on the same axes. As intended, 1:1M achieves a trade off between the two extremes, but it is not actually on the Pareto-efficient frontier since it does not trade these off in the* optimal *way. Next, let* $Y(0) \mid X \sim \mathcal{N}(\|X\|_2^2 - e^T X/2, \sqrt{3})$. *In Fig. 1c, varying* λ, *we plot the resulting CMSE of* $\hat{\tau}_W$ *(solid) and* $\hat{\tau}_{W,\hat{f}_0}$ *(dashed) for* $\hat{f}_0$ *given by ordinary least squares (OLS). Since OLS has in-sample residuals summing to zero,* $\hat{\tau}_{W,\hat{f}_0}$ *for* $\lambda = \infty$ *corresponds to simple OLS regression adjustment. We see that tuning* λ *correctly can amount to a significant improvement in CMSE.*

3.3. Generalized Optimal Matching Methods

Optimal matching minimized the worst-case squared error given certain restrictions on f_0, $\sigma^2_{1:n}$, and $\mathcal{W}$. Generalizing these restrictions, we can consider a whole range of generalized optimal matching methods that minimize CMSE by trading off variance to new, generalized notions of balance.

Definition 5 *Given* $\mathcal{W}$, $\|\cdot\|$ *satisfying the assumptions of Def. 4 and* $\lambda \in [0,\infty]$, *the generalized optimal matching method* $\text{GOM}(\mathcal{W}, \|\cdot\|, \lambda)$ *is given by the weights* W *that solve*

$$\min_{W\in\mathcal{W}} \left\{ \mathfrak{E}^2(W; \|\cdot\|, \lambda) := \mathfrak{B}^2(W; \|\cdot\|) + \lambda \|W\|_2^2 \right\}. \tag{6}$$

We let $\mathfrak{E}^2_{\min}(\mathcal{W}, \|\cdot\|, \lambda)$ *denote the value of this minimum. If* $\lambda = \infty$*, then we define* W *as a minimizer of the first term in eq.* (6) *over the minimizers of the second term.*

In other words, per Prop. 1, fixing $X_{1:n}, T_{1:n}$, the SATT estimator given by $\mathrm{GOM}(\mathcal{W}^{\text{general}}, \|\cdot\|, \lambda)$ is the one given by letting $\widehat{\overline{Y}}_{\mathcal{T}_1}(0)$ be the minimax (with respect to squared error risk) linear estimator constrained in $\{\sum_{i\in\mathcal{T}_0} W_i Y_i : W \in \mathcal{W}\}$ for the target $L(f_{1:n}) = \frac{1}{n_1}\sum_{i\in\mathcal{T}_1} f_i$, which is a linear operator on the possible realizations $f_{1:n} \in \mathcal{F}_{1:n} = \{(f(X_1), \dots, f(X_n)) : \|f\| \le \gamma\} \subseteq \mathbb{R}^n$, given observations $Y_i = f_i + \epsilon_i$ for $i \in \mathcal{T}_0$, if we assume that $\sigma_i^2 = \lambda\gamma^2$. In particular, Prop. 1 shows that such minimax linear estimators are exactly those that minimize a BDN imbalance metric plus a 2-norm regularization on weights. More generally, if we have knowledge of heteroskedasticity or even if the units are not independent, we would take the second term to be $W^T\Lambda W$ for some positive semi-definite Λ. In this paper, we focus only on $\Lambda = \lambda I$ for the sake of simplicity. Our consistency results, nonetheless, will apply under heteroskedasticity even when we use a single λ.

Thm. 3 established that NNM is equivalent to $\mathrm{GOM}(\mathcal{W}^{\text{simplex}}, \|\cdot\|_{\mathrm{Lip}(\delta)}, 0)$ and 1:1M is equivalent to $\mathrm{GOM}(\mathcal{W}^{1/n_1\text{-simplex}}, \|\cdot\|_{\mathrm{Lip}(\delta)}, 0)$. BVE-NNM is given by $\mathrm{GOM}(\mathcal{W}^{\text{simplex}}, \|\cdot\|_{\mathrm{Lip}(\delta)}, \lambda)$. Similarly, for any $\|\cdot\|$, no matching is given by $\mathrm{GOM}(\mathcal{W}, \|\cdot\|, \infty)$ for any $\mathcal{W} \ni (1/n_0, \dots, 1/n_0)$, examples of which include $\mathcal{W}^{\text{simplex}}$, $\mathcal{W}^{\text{multisubsets}}$, and $\mathcal{W}^{\text{subsets}}$.

It follows by Prop. 1 that GOM leads to a bound on the CMSE. Define

$$\|[f]\| = \inf_{g: B(W;g)=0\ \forall W\in\mathcal{W}} \|f+g\|,$$

which acts on the quotient space that eliminates degrees of freedom that are irrelevant to $B(W; f)$. For example, when $\mathcal{W} \subseteq \mathcal{W}^{\text{simplex}}$, this includes all constant shifts. Note $\|[f]\|$ is always smaller than $\|f\|$.

Corollary 6 *Suppose* $\sigma^2 \ge \sigma_i^2$ *and* $\gamma \ge \|[f_0]\|$*. Let* $\lambda = \sigma^2/\gamma^2$ *and let* W *be given by* $\mathrm{GOM}(\mathcal{W}, \|\cdot\|, \lambda)$*. Then*

$$\mathrm{CMSE}(\hat{\tau}_W) \le \gamma^2(\mathfrak{E}^2_{\min}(\mathcal{W}, \|\cdot\|, \lambda) + \lambda/n_1).$$

And, if $\hat{f}_0 \perp\!\!\!\perp Y_{1:n} \mid X_{1:n}, T_{1:n}$ *and* $\gamma \ge \|[f_0 - \hat{f}_0]\|$ *, then*

$$\mathrm{CMSE}(\hat{\tau}_{W,\hat{f}_0}) \le \gamma^2(\mathfrak{E}^2_{\min}(\mathcal{W}, \|\cdot\|, \lambda) + \lambda/n_1).$$

In particular, note that GOM controls a bound on the CMSE and not the CMSE itself, which is unknowable. That is, it is a minimax approach, obtaining the best-possible control on the range of possible CMSE values by a linear estimator, for each given $X_{1:n}, T_{1:n}$.

For subset-based matching, the balance-variance Pareto-efficient frontier given by varying λ is given by solely-balance-optimizing fixed-sized subsets.

Proposition 7 *Given* $\|\cdot\|$ *and* $\lambda \in [0, \infty]$*, there exists* $n(\lambda) \in \{1, \dots, n_0\}$ *such that* $\mathrm{GOM}(\mathcal{W}^{subsets}, \|\cdot\|, \lambda)$ *is equivalent to* $\mathrm{GOM}(\mathcal{W}^{n(\lambda)\text{-}subset}, \|\cdot\|, 0)$.

In particular, to compute $\mathrm{GOM}(\mathcal{W}^{\text{subsets}}, \|\cdot\|, \lambda)$ we may search over $\mathrm{GOM}(\mathcal{W}^{n_0'\text{-subset}}, \|\cdot\|, 0)$ for $n_0' \in \{1, \ldots, n_0\}$ and pick the one that minimizes $\mathfrak{E}(W; \|\cdot\|, \lambda)$. Note that the converse is not true: there may be some cardinalities that are *not* on the Pareto-efficient frontier of balance-variance efficient subsets. An example of this will be seen in Ex. 4. We also have the following relationship between optimal fixed-cardinality multisubsets and subsets:

Proposition 8 *Given $\|\cdot\|$, $\lambda \in [0, \infty]$ and $n_0' \in \{1, \ldots, n_0\}$, the following are equivalent:* $\mathrm{GOM}(\mathcal{W}^{n_0'\text{-}mutlisubset}, \|\cdot\|, \infty)$, $\mathrm{GOM}(\mathcal{W}^{n_0'\text{-}subset}, \|\cdot\|, 0)$, *and* $\mathrm{GOM}(\mathcal{W}^{n_0'\text{-}subset}, \|\cdot\|, \lambda)$.

3.4. Tractability

GOM is given by an optimization problem, which begs the question of when is it computationally tractable. We can first establish that the objective is always convex.

Proposition 9 *Given any $\mathcal{W}$, $\|\cdot\|$ satisfying the assumptions of Def. 4 and $\lambda \geq 0$, $\mathfrak{E}^2(W; \|\cdot\|, \lambda)$ is convex in W.*

This means that if $\mathcal{W} = \mathcal{W}^{\text{simplex}}$ then problem (6) is convex. Indeed, we can show that we can solve it in polynomial time.

Proposition 10 *Given an evaluation oracle for $\mathfrak{B}(W; \|\cdot\|)$, we can solve problem* (6) *for $\mathcal{W} = \mathcal{W}^{simplex}$ up to ϵ precision in time and oracle calls polynomial in $n, \log(1/\epsilon)$.*

In all cases we consider, $\mathfrak{B}(W; \|\cdot\|)$ is easy to evaluate. Moreover, in all cases we consider with $\mathcal{W} = \mathcal{W}^{\text{simplex}}$, we will in fact be able to formulate problem (6) as a linearly-constrained convex-quadratic optimization problem, which are not only polynomially-time solvable but also easily solved in practice using off-the-shelf solvers like Gurobi (`www.gurobi.com`), which we use in all numerics in this paper to solve such problems in tens to hundreds of milliseconds on a personal laptop computer. This includes the case of kernel optimal matching, which we introduce in Sec. 4.

If $\mathcal{W} = \mathcal{W}^{n_0'\text{-subset}}$ then, by Prop. 8, problem (6) is equivalent to a convex-objective binary optimization problem:

$$\min_{U \in \{0,1\}^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} U_i = n_0'} \mathfrak{B}(U/n_0'; \|\cdot\|). \tag{7}$$

Unlike simplex weights, this problem is *not* polynomial-time solvable.

If $\mathcal{W} = \mathcal{W}^{\text{subsets}}$ then Prop. 7 shows that problem (6) is equivalent to searching over the solutions $U_{n_0'}$ to problem (7) for $n_0' \in \{1, \ldots, n_0\}$ and picking the one with minimal $\mathfrak{B}(U_{n_0'}/n_0'; \|\cdot\|) + \lambda/n_0'$.

In all cases we consider in this paper with $\mathcal{W} = \mathcal{W}^{n_0'\text{-subset}}$ or $\mathcal{W} = \mathcal{W}^{\text{subsets}}$, we will be able to formulate problem (6) as, respectively, a single or a series of either binary quadratic or mixed-integer-linear optimization problem(s). These problems, generally hard in the sense of being NP-hard, can be solved for many practical sizes of n also by Gurobi. In fact, we solve these problems too in our numerical examples.

Table 1: Existing Matching Methods as GOM

Method	GOM with $\|\cdot\| =$	$\mathcal{W} =$	See
1:1M	$\|\cdot\|_{\mathrm{Lip}(\delta)}$	$\mathcal{W}^{1/n_1\text{-simplex}}$	Thm. 1
NNM	$\|\cdot\|_{\mathrm{Lip}(\delta)}$	$\mathcal{W}^{\text{simplex}}$	Thm. 1
Optimal caliper matching	$\|\cdot\|_{\partial(\hat{\mu}_n,\delta)}$	$\mathcal{W}^{1/n_1\text{-simplex}}$	Prop. 22
Coarsened exact matching	$\|\cdot\|_{L_\infty(C)}$	$\mathcal{W}^{\text{simplex}}$	Prop. 23
Mean-matching and fine balance	$\|\cdot\|_{2-\mathrm{lin}}$	$\mathcal{W}^{\text{subsets}}$	Prop. 24
Combined pair- and mean-matching	$\|\cdot\|_{\mathrm{Lip}(\delta)} \oplus_\rho \|\cdot\|_{2-\mathrm{lin}}$	$\mathcal{W}^{\text{subsets}}$	Prop. 25
Regression adjustment	$\|\cdot\|_{2-\mathrm{lin}}$	$\mathcal{W}^{\text{general}}$	Props. 26, 27

3.5. Existing Matching Methods as GOM

A surprising fact is that many matching methods commonly used in practice – not just NNM and 1:1M – are actually also GOM. Optimal-caliper matching (OCM), which finds a control match for each treated unit so to minimize the size of a caliper that would contain all pair distances, is GOM with respect to an averaged form of the Lipschitz norm. Coarsened exact matching (CEM) (Iacus et al., 2011a) is GOM with respect to the L_∞ norm on piecewise linear functions. In addition, various matching and weighting methods that use mean matching, near-fine balance, and combinations thereof with pair matching (Zubizarreta, 2012; Greenberg, 1953; Rubin, 1973; Rosenbaum et al., 2007; Yang et al., 2012; Bertsimas et al., 2015) are also GOM with norms given by certain parametric spaces and their direct sums with Lipschitz spaces.

Some of these results arise from observations in Kallus (2017), which studies method that minimize only a BDN imbalance objective *without* the regularization as in GOM. In particular, like NNM and 1:1M, many of the above are GOM with $\lambda = 0$. By automatically selecting λ using hyperparameter estimation, we can develop extensions of these methods, such as NNM++ and CEM++, that automatically and optimally trade off balance and variance and successfully reduce overall estimation error. Finally, regression adjustment methods are also GOM, revealing a close connection to matching but also a nuanced but important difference in the handling of extrapolation.

For the sake of a more fluid presentation we defer the full presentation of all these results to Sec. 5. We summarize the results in in Table 1, showing how a wide range of existing methods to achieve covariate balance all arise from the GOM framework. Before proving these and discussing extensions to consider balance-variance tradeoffs, we focus first on presenting a new, unified analysis of all GOM method and on developing KOM as a special class of GOM.

3.6. Consistency

In this section we characterize conditions for GOM to lead to consistent estimation. The conditions include "correct specification" by requiring that $\|[f_0]\| < \infty$. We also need the following technical condition on $\|\cdot\|$ for consistency. All magnitudes that we consider in this paper satisfy this condition.

Definition 11 $\|\cdot\|$ *is* B-convex *if there is* $N \in \mathbb{N}, \eta < N$ *such that for any* $\|g_1\|, \ldots, \|g_N\| \leq 1$ *there is a choice of signs so that* $\|\pm g_1 \pm \cdots \pm g_N\| \leq \eta$.

Theorem 12 *Suppose Asns. 2.1 and 2.2 hold and that*

(i) *for each* n, W *is given by* $\mathrm{GOM}(\mathcal{W}, \|\cdot\|, \lambda_n)$,

(ii) $\mathcal{W}$, $\|\cdot\|$ *satisfy the conditions of Def. 4,*

(iii) $\lambda_n \in [\underline{\lambda}, \overline{\lambda}] \subset (0, \infty)$,

(iv) $\mathcal{W}^{subsets} \subseteq \mathcal{W}$,

(v) $\|\cdot\|$ *is* B*-convex,*

(vi) $\mathbb{E}[\sup_{\|f\|\leq 1} (f(X_1) - f(X_2))^2 \mid T_1 = 1, T_2 = 1] < \infty$,

(vii) $\mathrm{Var}(Y(0) \mid X)$ *is almost surely bounded, and*

(viii) $\|[f_0]\| < \infty$.

Then, $\hat{\tau}_W - \mathrm{SATT} = o_p(1)$.

Condition (iv) is satisfied for subset, mutlisubset, and simplex matching. Condition (vi) requires bounded moments: for either near-fine balance or for optimal matching ($\|\cdot\|_{\mathrm{Lip}(\delta)}$) with Euclidean distances, the condition is satisfied if covariates have second moments. Condition (viii) requires correct specification of the outcome model. For example, for near-fine balance, expected potential outcomes have to be *additive* in the factors (*i.e.*, linear). We will relax this in the case of KOM and prove model-free consistency.

Note that, letting $\mathrm{PATT} = \mathbb{E}[Y(1) - Y(0) \mid T = 1]$, under the above conditions, central limit and Slutsky's theorems imply $\mathrm{SATT} - \mathrm{PATT} = O_p(1/\sqrt{n})$. Therefore, all consistency and rate results extend to estimating PATT.

The results of Thm. 12 can be extended to the augmented estimator:

Theorem 13 *Suppose all assumptions of Thm. 12 except (viii) hold, that* $\hat{f}_0 \perp\!\!\!\perp Y_{1:n} \mid X_{1:n}, T_{1:n}$, *and that* $(\mathbb{E}[(\hat{f}_0(X) - \tilde{f}_0(X))^2])^{1/2} = O(1/\sqrt{n})$ *for some fixed* $\tilde{f}_0$. *Then the following three results hold:*

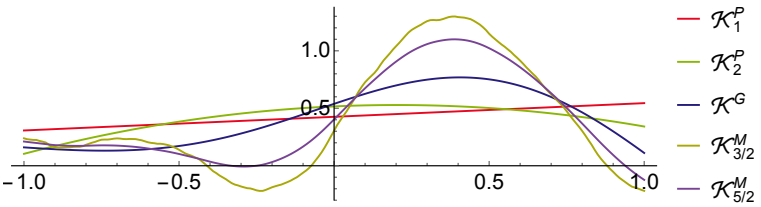
(a) *If* $\tilde{f}_0 = f_0$*:* $\hat{\tau}_{W,\hat{f}_0} - \mathrm{SATT} = o_p(1)$.

(b) *If* $\|[\tilde{f}_0]\|, \|[f_0]\| < \infty$*:* $\hat{\tau}_{W,\hat{f}_0} - \mathrm{SATT} = o_p(1)$.

(c) *If* $\|[f_0]\| < \infty, \|[\hat{f}_0]\| = O_p(1)$*:* $\hat{\tau}_{W,\hat{f}_0} - \mathrm{SATT} = o_p(1)$.

The consistency results for $\hat{\tau}_W$ and $\hat{\tau}_{W,\hat{f}_0}$ are stronger in the case of KOM, which we discuss next.

Figure 2: Random functions drawn from a Gaussian process



4. Kernel Optimal Matching

In this section we develop kernel optimal matching (KOM) methods, which are given by GOM using a reproducing kernel Hilbert space (RKHS). Kernels methods are standard in machine learning as ways to generalize the structure of learned conditional expectation functions, like classifiers or regressors (Scholkopf and Smola, 2001). Kernels have many applications in statistics, applied and theoretical (Berlinet and Thomas-Agnan, 2004; Gretton et al., 2006; Zhang et al., 2012; Kallus, 2018).

An RKHS $\mathcal{F}$ is a Hilbert space of functions $\mathcal{X} \to \mathbb{R}$ such that, for any x, $E_x : f \in \mathcal{F} \mapsto f(x)$ is a bounded operator. That is, $\mathcal{F}$ is a inner product space that is closed with respect to the norm the inner product defines, $\|f\| = \langle f, f\rangle$, and $|f(x)| \le M_x\|f\|$ for some M_x for every x. Since, for every x, E_x is linear and bounded, it has a representer, that is, it can equivalently be written as an inner product, $E_x f = \langle \mathcal{K}(x,\cdot), f\rangle$ for some $\mathcal{K}(x,\cdot) \in \mathcal{F}$, an element of $\mathcal{F}$ defined for every x. The resulting bivariate function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semidefinite (PSD) kernel, meaning that for any $m, x_1, \dots, x_m$ the Gram matrix $K_{ij} = \mathcal{K}(x_i, x_j)$ (which is real square symmetric) is PSD (has nonnegative eigenvalues). The closure of the span of $\{\mathcal{K}(x,\cdot) : x \in \mathcal{X}\}$ reproduces $\mathcal{F}$, hence the name. Conversely, the span of every PSD kernel can be uniquely completed to form an RKHS and hence induces a norm $\|\cdot\|$ (Berlinet and Thomas-Agnan, 2004). PSD kernels also describe the covariance of Gaussian processes (GP): we say that $f \sim \mathcal{GP}(\mu, \mathcal{K})$ if for any $m, x_1, \dots, x_m$, $(f(x_1), \dots, f(x_m))$ are jointly normal and $\mathbb{E}f(x_i) = \mu(x_i)$, $\mathrm{Cov}(f(x_i), f(x_j)) = \mathcal{K}(x_i, x_j)$.[8]

Popular examples of kernels on $\mathbb{R}^d$ are polynomial $\mathcal{K}_\nu^{\mathrm{P}}(x,x') = (1 + \frac{x^T x'}{\nu})^\nu$, exponential $\mathcal{K}^{\mathrm{E}}(x,x') = e^{x^T x'}$, Gaussian $\mathcal{K}^{\mathrm{G}}(x,x') = e^{-\frac{1}{2}\|x-x'\|^2}$, and Matérn $\mathcal{K}_\nu^{\mathrm{M}}(x,x') = \frac{(\sqrt{2\nu}\|x-x'\|)^\nu}{2^{\nu-1}\Gamma(\nu)}\,\mathrm{BK}_\nu(\sqrt{2\nu}\|x-x'\|)$ where BK_ν is a modified Bessel function of the second kind. For $s = \nu + d/2 \in \mathbb{N}_0$, the Matérn kernel induces a norm that is equivalent to the Sobolev norm of order s given by $\|f\|_{H^s}^2 = \sum_{\alpha\in\mathbb{N}_0^d : \|\alpha\|_1 \le s} \int_{\mathbb{R}^d} (D^\alpha f)^2$ (Wendland, 2004, Cor. 10.13). More generally, any differential norm $\|f\|^2 = \sum_{\alpha\in\mathbb{N}_0^d} a_{\|\alpha\|_1} \int_{\mathbb{R}^d} (D^\alpha f)^2$ with $a_0 > 0, a_{\lceil (d+1)/2\rceil} > 0$ also corresponds to an RKHS norm (Williams and Rasmussen, 2006, Sec. 6.2.1) (we treat the case of purely differential regularization, $a_0 = 0$, in Sec. 4.4).

8. Note the difference between "kernel" as the reproducer of an RKHS or covariance operator of a GP and "kernel" as the integral kernel used in non-parametric smoothing methods such as kernel density estimation, Nadaraya-Watson kernel regression, and local linear regression. In particular, in the context of matching, Heckman et al. (1997) interpret 1:k NNM as a regression adjustment using k-nearest-neighbor regression to impute control potential outcomes and suggest to replace this regression with a kernel regression. This is wholly distinct from KOM and, in particular, is not GOM and does not minimize worst-case CMSE. Instead, it directly imputes potential outcomes using kernel regression.

Fig. 2 displays random draws (on the same realization path) of functions $\mathbb{R} \to \mathbb{R}$ from the Gaussian processes with mean zero and the kernels above. Generally, we either normalize covariate data before putting it in a kernel so that the control sample has zero sample mean and identity covariance or we just fit a rescaling matrix to the data (see Sec. 4.5 for details). The IPM with respect to the unit ball of an RKHS is the MMD (Gretton et al., 2006).

RKHS norms always satisfy the conditions of Def. 4 by definition. We call the resulting GOM method, kernel optimal matching (KOM).

Definition 14 *Given a PSD kernel $\mathcal{K}$ on $\mathcal{X}$, the kernel optimal matching* $\mathrm{KOM}(\mathcal{W}, \mathcal{K}, \lambda)$ *is given by* $\mathrm{GOM}(\mathcal{W}, \|\cdot\|, \lambda)$ *where $\|\cdot\|$ is the RKHS norm induced by $\mathcal{K}$. We also overload our notation: $\mathfrak{B}(W; \mathcal{K}) = \mathfrak{B}(W; \|\cdot\|)$, $\mathfrak{E}(W; \mathcal{K}, \lambda) = \mathfrak{E}(W; \|\cdot\|, \lambda)$.*

Compared to the space of Lipschitz function, RKHSs impose more structure but can still be extremely general. Lipschitz functions are not only infinite dimensional, they are also non-separable, *i.e.*, no countable subset is dense. They have too little structure to be practically useful without immense amounts of data or very little covariates, as we will see empirically. In contrast, RKHSs are much more flexible: they are always separable and may be chosen to be as general as needed and may still approximate functions arbitrarily well, as we will see in Sec. 4.2. Additionally, they easily adapt to the data, as we will see in Sec. 4.5. These properties and the strong theoretical guarantees that hold specifically for this class, which we present in the following sections make KOM a particularly appealing subclass of GOM and practically useful matching method.

KOM is generally given by a quadratic optimization problem.

Proposition 15 *Let $\mathcal{K}$ be a PSD kernel and let $K_{ij} = \mathcal{K}(X_i, X_j)$. Then,* $\mathrm{KOM}(\mathcal{W}, \mathcal{K}, \lambda)$ *is given by the optimization problem*

$$\min_{W \in \mathcal{W}} \ \tfrac{1}{n_1^2} e_{n_1}^T K_{\mathcal{T}_1, \mathcal{T}_1} e_{n_1} - \tfrac{2}{n_1} e_{n_1}^T K_{\mathcal{T}_1 \mathcal{T}_0} W + W^T (K_{\mathcal{T}_0 \mathcal{T}_0} + \lambda I) W. \tag{8}$$

Problem (8) has a convex-quadratic objective. When $\mathcal{W} = \mathcal{W}^{\mathrm{simplex}}$, the problem is a linearly-constrained quadratic optimization problem, which is polynomially time solvable and easily computed with off-the-shelf solvers. For subset-based matching $\mathrm{KOM}(\mathcal{W}^{\mathrm{subsets}}, \mathcal{K}, \lambda)$ is given by the following convex-quadratic binary optimization problem for some $n(\lambda)$, as a consequence of Props. 7 and 8:

$$\min_{U \in \{0,1\}^{\mathcal{T}_0} : \sum_{i \in \mathcal{T}_0} U_i = n(\lambda)} \tfrac{1}{n_1^2} e_{n_1}^T K_{\mathcal{T}_1, \mathcal{T}_1} e_{n_1} - \tfrac{2}{n_1 n(\lambda)} e_{n_1}^T K_{\mathcal{T}_1 \mathcal{T}_0} U + \tfrac{1}{n(\lambda)^2} U^T K_{\mathcal{T}_0 \mathcal{T}_0} U.$$

Unless otherwise noted, when referring to KOM, we mean on the simplex.

In Cor. 6, we saw that GOM immediately leads to a bound on CMSE. For the case of KOM, we can also interpret it as Bayesian optimal, exactly minimizing the posterior CMSE of $\hat{\tau}_W$, rather than merely bounding it.

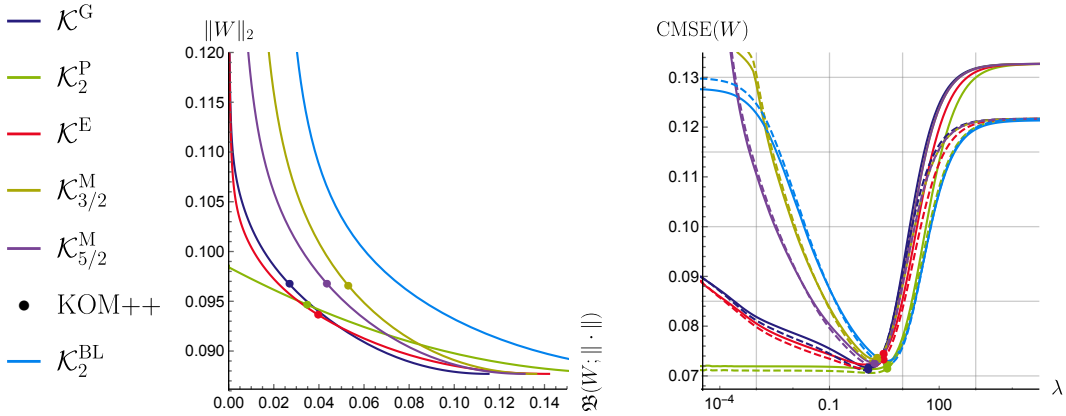
Proposition 16 *Let $\mathcal{K}$ be a PSD kernel, $c \in \mathbb{R}$, $\gamma^2, \sigma^2 \geq 0$, $\lambda = \sigma^2/\gamma^2$. Suppose our prior is that $f_0 \sim \mathcal{GP}(c, \gamma^2 \mathcal{K})$ and $Y_i(0) \sim \mathcal{N}(f_0(X_i), \sigma^2)$. Then*

$$\mathbb{E}[(\hat{\tau}_W - \mathrm{SATT})^2 \mid X_{1:n}, T_{1:n}] = \gamma^2 (\mathfrak{E}(W; \mathcal{K}, \lambda) + \lambda / n_1).$$

If instead our prior is $f_0 \sim \mathcal{GP}(\hat{f}_0, \gamma^2 \mathcal{K})$

$$\mathbb{E}[(\hat{\tau}^2_{W, \hat{f}_0} - \mathrm{SATT})^2 \mid X_{1:n}, T_{1:n}] = \gamma^2 (\mathfrak{E}(W; \mathcal{K}, \lambda) + \lambda / n_1).$$

Figure 3: The Balance-Variance Trade-off in KOM



(a) Balance-variance (b) CMSE

In the Appendix C, we discuss the equal-percent bias reduction (Rubin, 1976a, EPBR) properties of KOM and new kernel-based generalizations thereof.

4.1. Automatic Selection of λ

An important question that remains is how to choose λ. Using the Bayesian interpretation of KOM given by Prop. 16, we treat the choice of λ as hyperparameter estimation problem for the prior and employ an empirical Bayes approach.

Given a kernel $\mathcal{K}$, we postulate $f_0 \sim \mathcal{GP}(c, \gamma^2\mathcal{K})$, $Y_i \sim \mathcal{N}(f_0(X_i), \sigma^2)$, where we set $c = \overline{Y}_{\mathcal{T}_0}$. Given this model, we can ask what is the likelihood of the data for a given assignment to γ^2, σ^2. This is known as the *marginal likelihood* as it marginalizes over the actual regression function f_0 rather than asking what is the likelihood under a particular choice thereof (as in MLE). It is straightforward to show that the negative log marginal likelihood of the control data given the prior parameters γ^2, σ^2 is

$$\begin{aligned}\ell(\gamma^2,\sigma^2) &= -\log \mathbb{P}\left(Y_{\mathcal{T}_0} \mid X_{\mathcal{T}_0}, \gamma^2, \sigma^2\right) \\ &= \tfrac{1}{2}(Y_{\mathcal{T}_0} - \overline{Y}_{\mathcal{T}_0})^T(\gamma^2 K + \sigma^2 I)^{-1}(Y_{\mathcal{T}_0} - \overline{Y}_{\mathcal{T}_0}) + \tfrac{1}{2}\log|\gamma^2 K + \sigma^2 I| + \tfrac{n_0 \log(2\pi)}{2},\end{aligned}$$

where K is the Gram matrix on $\mathcal{T}_0$. Choosing $\hat{\gamma}^2, \hat{\sigma}^2$ to minimize this quantity, we let $\hat{\lambda} = \hat{\sigma}^2/\hat{\gamma}^2$. We give the name KOM++ to KOM with $\lambda = \hat{\lambda}$.

In a strict sense, KOM++ does not produce "honest" weights because $\hat{\lambda}$ depends on $Y_{\mathcal{T}_0}$. However, as we only extract a single parameter $\hat{\lambda}$, we guard against data mining, as seen by the resulting low CMSE in the next example and in our investigation of its performance on real data in Sec. 4.6.

Example 2 *Let us revisit Ex. 1 to study KOM. We consider KOM with the Gaussian, quadratic, exponential, Matérn $\nu = 3/2$, and Matérn $\nu = 5/2$ kernels. We plot the resulting balance-variance landscape in Fig. 3a. Note that the horizontal axis is different across the curves and so the curves are not immediately comparable. We plot the resulting CMSE in*

Fig. 3b. In both figures, we point out the result of KOM++, which chooses λ by marginal likelihood and which appears to perform well across all kernels. We include SKOM with the second-order Beppo-Levi kernel $\mathcal{K}_2^{BL}$ as detailed in Sec. 4.4.

4.2. Consistency

We now characterize conditions for KOM to lead to consistent estimation. Under correct specification, we can guarantee $\sqrt{n}$-consistency. Under incorrect specification but with a C_0-universal kernel, we can still ensure consistency. A C_0-universal kernel, defined below, is one that can arbitrarily approximate compactly-supported continuous functions in L_∞. The Gaussian and Matérn kernels are C_0-universal and the exponential kernel is C_0-universal on compact spaces (Sriperumbudur et al., 2010).

Definition 17 *A PSD kernel $\mathcal{K}$ on a Hausdorff $\mathcal{X}$ (*e.g.*, $\mathbb{R}^d$) is* C_0-universal *if, for any continuous function $g : \mathcal{X} \to \mathbb{R}$ with compact support (*i.e.*, for some C compact, $\{x : g(x) \neq 0\} \subseteq C$) and $\eta > 0$, there exists $m, \alpha_1, x_1, \ldots, \alpha_m, x_m$ such that $\sup_{x\in\mathcal{X}} |\sum_{j=1}^m \alpha_i \mathcal{K}(x_j, x) - g(x)| \leq \eta$.*

Theorem 18 *Suppose Asns. 2.1 and 2.2 hold and that*

(i) *for each n, W is given by* $\mathrm{KOM}(\mathcal{W}, \mathcal{K}, \lambda_n)$,

(ii) $\lambda_n \in [\underline{\lambda}, \overline{\lambda}] \subset (0, \infty)$,

(iii) $\mathcal{W}^{subsets} \subseteq \mathcal{W}$,

(iv) $\mathbb{E}[\mathcal{K}(X, X) \mid T = 1] < \infty$, *and*

(v) $\mathrm{Var}(Y(0) \mid X)$ *is almost surely bounded.*

Then the following two results hold:

(a) *If $\|[f_0]\| < \infty$:* $\hat{\tau}_W - \mathrm{SATT} = O_p(n^{-1/2})$.

(b) *If $\mathcal{K}$ is C_0-universal:* $\hat{\tau}_W - \mathrm{SATT} = o_p(1)$.

As before, condition (iii) is satisfied for subset, mutlisubset, and simplex matching. Condition (iv) is trivially satisfied for any bounded kernel ($\mathcal{K}(x, x) \leq M$), such as the Gaussian and Matérn kernels. Condition (v) is generally weak and, in particular, is satisfied under homoskedasticity. Case (a) is the case of a well-specified model, even if $\mathcal{K}$ induces an infinite-dimensional RKHS (all C_0-universal kernels do). For example, while the exponential kernel is infinite dimensional and C_0-universal on compact spaces, polynomial functions (*e.g.*, linear) have finite norm in its induced RKHS. Moreover, under the common semiparameteric specification where f_0 is assumed to be Sobolev (*e.g.*, be square-integrable so that $\mathrm{Var}(Y(0)) < \infty$ and have square-integrable derivatives of degrees up to $\lceil (d+1)/2 \rceil$), it is well-specified by the Matérn kernel. Case (b) is the case of a misspecified model, wherein a C_0-universal kernel still guarantees model-free consistency.

The results can be extended to the augmented estimator.

Theorem 19 *Suppose the conditions of Thm. 18 hold and that* $\hat{f}_0 \perp\!\!\!\perp Y_{1:n} \mid X_{1:n}, T_{1:n}$. *Then the following five results hold:*

(a) If $\|[\hat{f}_0 - f_0]\| = o_p(1)$*:* $\quad \hat{\tau}_{W,\hat{f}_0} - \mathrm{SATT} = \frac{1}{n_1}\sum_{i\in\mathcal{T}_1}\epsilon_i - \sum_{i\in\mathcal{T}_0} W_i\epsilon_i + o_p(n^{-1/2})$.

(b) If $\|[f_0]\| < \infty, \|[\hat{f}_0]\| = O_p(1)$*:* $\quad \hat{\tau}_{W,\hat{f}_0} - \mathrm{SATT} = O_p(n^{-1/2})$.

If $(\mathbb{E}[(\hat{f}_0(X) - \tilde{f}_0(X))^2])^{1/2} = O(r(n))$ *for* $r(n) = o(1)$ *and*

(c) If $\tilde{f}_0 = f_0$*:* $\quad \hat{\tau}_{W,\hat{f}_0} - \mathrm{SATT} = O_p(r(n) + n^{-1/2})$.

(d) If $\mathcal{K}$ *is* C_0*-universal:* $\quad \hat{\tau}_{W,\hat{f}_0} - \mathrm{SATT} = o_p(1)$.

(e) If $\|[\tilde{f}_0]\|, \|[f_0]\| < \infty$*:* $\quad \hat{\tau}_{W,\hat{f}_0} - \mathrm{SATT} = O_p(r(n) + n^{-1/2})$.

Case (c) says that $\hat{f}_0$ is weakly L_2-consistent with rate $r(n)$. The condition holds with a parametric $r(n) = n^{-1/2}$, for example, if f_0 is linear and $\hat{f}_0$ is given by OLS on $X_{\mathcal{T}_0}, Y_{\mathcal{T}_0}$. Cases (d), (e), and (b) say that when $\hat{f}_0$ may be inconsistent, kernel weights can correct for the error, with the situation varying on whether the regression estimator is itself in the RKHS. Thus, one useful case is when $\hat{f}_0$ is given by a parametric regression and we use KOM with a C_0-universal kernel: then we get a parametric rate when the model is correct but do not sacrifice consistency when it is not. Case (a) shows that if $\hat{f}_0$ is consistent in the RKHS norm then we are only left with the irreducible error that involves only the residuals ϵ_i, which of course X cannot control for. Example of such cases include when $\hat{f}_0$ is given by a well-specified kernel ridge regression and we use KOM with the same kernel or when its given by any nonparametric regression and its derivatives up to $\lceil (d+1)/2 \rceil$ are weakly consistent and we use KOM with the Matérn kernel.

For the case where $\hat{f}_0$ is given by kernel ridge regression and W by KOM with $\mathcal{W}^{\text{general}}$ (see Sec. 5.6) and both use the same kernel and λ, we get a closed form for the augmented KOM estimator:

$$\tfrac{1}{n_1} e_{n_1}^T (Y_{\mathcal{T}_1} - K_{\mathcal{T}_1\mathcal{T}_0}(K_{\mathcal{T}_0\mathcal{T}_0} + \lambda I)^{-1}(K_{\mathcal{T}_0\mathcal{T}_0} + 2\lambda I)(K_{\mathcal{T}_0\mathcal{T}_0} + \lambda I)^{-1} Y_{\mathcal{T}_0}).$$

This estimator essentially debiases the kernel ridge regression adjustment – bias which is unavoidable in nonparametric kernel ridge regression (*e.g.*, universal kernel). In the parametric case (rank of $K_{\mathcal{T}_0\mathcal{T}_0}$ is bounded), we can set $\lambda = 0$ and recover plain OLS adjustment as it is already unbiased.

In a related alternative usage of augmented estimators, Athey et al. (2016) recently showed that in a setting with a *correctly specified* but *high-dimensional* parametric (linear) model, fast, consistent estimation is possible using $\hat{\tau}_{W,\hat{f}_0}$ with $\hat{f}_0$ given by LASSO (Tibshirani, 1996) and W given by the equivalent of $\mathrm{GOM}(\mathcal{W}^{\text{simplex}}, \|\cdot\|_{\text{1-lin}}, \lambda)$.

4.3. Kernel Matching to Reduce Model Dependence

A popular use of matching, as implemented in the popular *R* package `MatchIt`, is as preprocessing before regression analysis, in which case matching is commonly understood to reduce model dependence (Ho et al., 2007). Similar in spirit to double robustness in the face

of potential model misspecification, this is understood commonly and in Ho et al. (2007) as pruning unmatched control subjects before a linear regression-based treatment effect estimation. More generally, however, we can consider any nonnegative weights W, whether subset or simplex weights. This leads to the following weighted least squares estimator:

$$\hat{\tau}_{\mathrm{WLS}(W)} = \operatorname{argmin}_{\tau\in\mathbb{R}} \min_{\alpha\in\mathbb{R},\beta_1,\beta_2\in\mathbb{R}^d} \sum_{i=1}^n (T_i/n_1 + (1-T_i)W_i) \\ \times (Y_i - \alpha - \tau T_i - \beta_1^T X_i - \beta_2^T (X_i - \overline{X}_{\mathcal{T}_1})T_i)^2$$

When W is given by KOM, we can show that this procedure indeed achieves the desired robustness: consistency without model dependence and parametric rates when the model is correctly specified.

Theorem 20 *Suppose the conditions of Thm. 18 hold and that $\mathcal{K}$ is C_0-universal, $\mathcal{X}$ bounded, and $\mathbb{E}[XX^T \mid T=1]$ non-singular. Then,*

(a) Regardless of f_0: $\hat{\tau}_{\mathrm{WLS}(W)} - \mathrm{SATT} = o_p(1)$.

(b) If $\exists \alpha_0, \beta_0$ s.t. $f_0(x) = \alpha_0 + \beta_0^T x$: $\hat{\tau}_{\mathrm{WLS}(W)} - \mathrm{SATT} = O_p(n^{-1/2})$.

4.4. Semi-Kernel Optimal Matching

We next extend KOM to the semiparametric case with unconstrained parametric part, where we combine both a parametric exact balance criterion such as exact matching of means with a non-parametric criterion such as that of KOM. A notable example will include matching against all functions with square-integrable Hessians, as in smoothing splines (Friedman et al., 2001, Sec. 5.7). First, we define *conditionally* PSD kernels.

For a class of functions $\mathcal{G} \subseteq [\mathcal{X} \to \mathbb{R}]$, a $\mathcal{G}$-conditionally PSD kernel on $\mathcal{X} \subset \mathbb{R}^d$ is a symmetric function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies $\sum_{i=1}^m v_i v_j \mathcal{K}(x_i, x_j) \geq 0$ for every $m, x_1, \ldots, x_m, v_1, \ldots, v_m$ satisfying $\sum_{i=1}^n v_i g(x_i) = 0$ for all $g \in \mathcal{G}$. For example, $\{0\}$-conditionally PSD kernels are just the PSD kernels. Given a $\mathcal{G}$-conditionally PSD kernel $\mathcal{K}$, we can define a corresponding magnitude:

$$\|f\|^2 = \inf \left\{ \sum_{i,j=1}^{\infty} \alpha_i \alpha_j \mathcal{K}(x_i, x_j) : \begin{array}{c} f = g + \sum_{i=1}^{\infty} \alpha_i \mathcal{K}(x_i, \cdot), \\ g \in \mathcal{G}, \sum_{i=1}^{\infty} \alpha_i^2 \mathcal{K}(x_i, x_i) < \infty, \\ \sum_{i=1}^{\infty} \alpha_i g'(x_i) = 0 \ \forall g' \in \mathcal{G} \end{array} \right\}. \tag{9}$$

If $\mathcal{K}$ is $\mathcal{G}$-conditionally PSD, then we refer to $\mathrm{GOM}(\mathcal{W}, \|\cdot\|, \lambda)$ with $\|\cdot\|$ as in eq. (9) as $\mathrm{SKOM}(\mathcal{W}, \mathcal{K}, \lambda)$, abbreviating SKOM for semi-kernel optimal matching and treating $\mathcal{G}$ as implicit in $\mathcal{K}$.

An important example is smooth functions on $\mathbb{R}^d$. Let $\mathcal{G}$ be all polynomials of degree at most $\nu - 1$: $\mathcal{G}_\nu^{\mathrm{poly}} = \mathrm{span}\{x^\alpha : \alpha \in \mathbb{N}_0^d, \|\alpha\|_1 \leq \nu - 1\}$. Then, For $\nu > d/2$, the Beppo-Levi kernel $\mathcal{K}_\nu^{\mathrm{BL}}(x, x') = \kappa_\nu(\|x - x'\|_2)$ is $\mathcal{G}_\nu^{\mathrm{poly}}$-conditionally PSD, where $\kappa_\nu(0) = 0$ and $\kappa_\nu(u) = (-1)^{\nu + (d-2)/2} u^{2\nu - d} \log(u)$ for d even and $\kappa_{d,\nu}(u) = u^{2\nu - d}$ for d odd. The Beppo-Levi kernel's corresponding magnitude in eq. (9) is equivalent to the square-integral of the ν^{th} derivatives (Wendland, 2004, Prop. 10.39): $\|f\|_{\mathrm{BL}}^2 = \sum_{\alpha \in \mathbb{N}_0^d, \|\alpha\|_1 = \nu} \binom{\nu}{\alpha} \int_{\mathbb{R}^d} (D^\alpha f)^2$.

Two important cases are cubic and thin-plate splines. In $d = 1$, the cubic spline kernel $\mathcal{K}_2^{\mathrm{BL}}(x, x') = |x - x'|^3$ is conditionally PSD with respect to all linear functions $\mathcal{G}^{\mathrm{lin}} = \mathcal{G}_2^{\mathrm{poly}}$.

In $d = 2$, the thin-plate spline kernel $\mathcal{K}_2^{\mathrm{BL}}(x, x') = \|x - x'\|^2 \log(\|x - x'\|)$ is also $\mathcal{G}^{\mathrm{lin}}$-conditionally PSD. In either case, the magnitude in eq. (9) is equivalent to the roughness of f, or square integral of the Hessian:

$$\|f\|^2_{\mathrm{Roughness}} = \textstyle\int_{\mathbb{R}^d} \|\nabla^2 f\|^2_{\mathrm{Frobenius}}.$$

Thus, $\mathrm{SKOM}(\mathcal{W}, \mathcal{K}_2^{\mathrm{BL}}, \lambda)$ seeks weights to balance *all* smooth functions. In particular, as the linear part of f is completely unconstrained since linear functions have zero Hessian, it will, if possible, *exactly* balance all linear functions, *i.e.*, it will exactly match the sample means.

Like KOM, SKOM admits a solution as a convex-quadratic-objective optimization problem.

Proposition 21 *Let $\mathcal{G} = \mathrm{span}\{g_1, \dots, g_m\}$, let $\mathcal{K}$ be a $\mathcal{G}$-conditionally PSD kernel, let $K_{ij} = \mathcal{K}(X_i, X_j)$, let $G_{ij} = g_i(X_j)$, and let $N \in \mathbb{R}^{n \times k}$ have columns forming a basis for the null space of G. Then $\mathrm{SKOM}(\mathcal{W}, \mathcal{K}, \lambda)$ is given by $W = N_{\mathcal{T}_0} U$ with U given by the optimization problem*

$$\begin{array}{rl} \min & U^T N^T (K + \lambda I_{\mathcal{T}_0}) N U \\ s.t. & U \in \mathbb{R}^k,\ N_{\mathcal{T}_0} U \in \mathcal{W},\ N_{\mathcal{T}_1} U = -e_{n_1}/n_1 \end{array} \tag{10}$$

where $(I_{\mathcal{T}_0})_{ij} = \mathbb{I}[i = j \in \mathcal{T}_0]$. Moreover, $N^T K N$ is PSD matrix so that the quadratic objective is convex.

Note that in the case of $\mathcal{W} = \mathcal{W}^{\mathrm{simplex}}$, if a constant function is in $\mathcal{G}$ as in the case of smooth functions, then the constraint $\sum_{i=1}^n W_i = 1$ is redundant in problem (10) as it is already enforced by the other constraints. In particular, the constraints necessarily imply $B(W; g) = 0$ for all $g \in \mathcal{G}$.

This also means that, in the case of SKOM over smooth functions with $\mathcal{W} = \mathcal{W}^{\mathrm{simplex}}$, there may exist *no* solution at all to the SKOM unless the treated sample mean $\overline{X}_{\mathcal{T}_1} = \frac{1}{n_1} \sum_{i \in \mathcal{T}_1} X_i$ is in the convex hull of the control sample $\mathrm{conv}\{X_i : i \in \mathcal{T}_0\}$. If it is, then SKOM will seek the weights that simultaneously match the means exactly without extrapolation and balance all smooth functions. If it is not, a solution will nonetheless exist if we instead use $\mathcal{W} = \mathcal{W}^{\mathrm{general}}$ (effectively fit a spline), but this allows extrapolation and is inadvisable. More appropriately, in the case where exactly matching means without extrapolation is not feasible, one should instead seek to achieve approximate matching without extrapolation by simply using standard KOM (which penalizes linear terms in f_0). First-order discrepancies can be emphasized by putting higher weight on linear functions by, *e.g.*, using a direct sum of a universal RKHS with an appropriately weighted linear RKHS.

4.5. Automatic Selection of $\mathcal{K}$

We can go further than just selecting λ in a data-driven manner for KOM and also use marginal likelihood to choose $\mathcal{K}$. Consider a parametrized family of kernels $\mathfrak{K} = \{\mathcal{K}_\theta(x, x') : \theta \in \Theta\}$. The most common example is parameterizing the length-scale of the Gaussian kernel: $\left\{\mathcal{K}_\theta(x, x') = \mathcal{K}^{\mathrm{G}}(x/\theta, y/\theta) : \theta > 0\right\}$. But we can easily conceive of more complex structures such as fitting a rescaling matrix for any kernel:

Table 2: MSE of various estimators in Ex. 3

	$\hat{\tau}_W$	$\hat{\tau}_{W,\hat{f}_0}$	$\hat{\tau}_{\mathrm{WLS}(W)}$
KOM++ ARD $\mathcal{K}_2^P$	0.028481	0.028698	0.028685
KOM++ ARD $\mathcal{K}_{3/2}^M$	0.028886	0.029165	0.029182
KOM++ ARD $\mathcal{K}_{5/2}^M$	0.028983	0.029279	0.029323
KOM++ ARD $\mathcal{K}^G$	0.029033	0.029288	0.029316
KOM++ ARD $\mathcal{K}^E$	0.029072	0.029188	0.029128
KOM++ $\mathcal{K}^G$	0.029783	0.029834	0.029856
KOM++ $\mathcal{K}_{5/2}^M$	0.029895	0.029935	0.029956
KOM++ $\mathcal{K}_{3/2}^M$	0.029944	0.029980	0.030001
KOM++ $\mathcal{K}_2^P$	0.030391	0.030471	0.030543
IPW	0.033168	0.033146	0.033126
PSLLM	0.033379	0.033399	–
No matching	0.034188	0.032925	0.032925
1:1M	0.034283	0.034769	0.034797
PS1:1M	0.038382	0.037646	0.037914
CEM++	0.039811	0.039611	0.039533
CEM	0.040418	0.040228	0.040073
NNM++	0.042890	0.043184	0.043511
NNM	0.047071	0.047399	0.047695
LLM	0.081411	0.081411	–

$\left\{\mathcal{K}_\theta(x,x') = \mathcal{K}(\theta x, \theta y) : \theta \in \Theta \subseteq \mathbb{R}^{d\times d}\right\}$, where Θ can be restricted to diagonal matrices in order to rescale each covariate (known as automatic relevance detection, ARD), can be unrestricted to fit a full covariance structure, or can be restricted to matrices with only $d' < d$ rows in order to find a projection onto a lower dimensional space. Additionally, we can consider mixtures of kernels, $\{\theta\mathcal{K}_1 + (1-\theta)\mathcal{K}_2 : \theta \in [0,1], \mathcal{K}_1 \in \mathfrak{K}_1, \mathcal{K}_2 \in \mathfrak{K}_2\}$, and more complex structures like the spectral mixture kernel (Wilson and Adams, 2013).

It is easy to see that given any such parameterized kernel $\mathcal{K}_\theta$, the negative log marginal likelihood is simply given by the parametrized Gram matrix:

$$\ell(\theta,\gamma^2,\sigma^2) = \tfrac{1}{2}(Y_{\mathcal{T}_0} - \overline{Y}_{\mathcal{T}_0})^T(\gamma^2 K_\theta + \sigma^2 I)^{-1}(Y_{\mathcal{T}_0} - \overline{Y}_{\mathcal{T}_0}) + \tfrac{1}{2}\log|\gamma^2 K_\theta + \sigma^2 I| + \tfrac{n_0\log(2\pi)}{2},$$

where $K_{\theta,i,j} = \mathcal{K}_\theta(X_i, X_j)$ for $i,j \in \mathcal{T}_0$. As before, we can optimize this over $\theta, \gamma^2, \sigma^2$ jointly to select both $\mathcal{K}$ and λ for KOM.

Note that if it is the case that for any $\theta \in \Theta$ and unitary matrix U we have $\mathcal{K}_\theta(Ux, Ux') = \mathcal{K}_{\theta'}(x,x')$ for some $\theta' \in \Theta$, then KOM after marginal likelihood is affinely invariant. For example, this is the case when we parametrize either an unrestricted low-dimensional projection or full covariance matrix for the kernel. This means that it is not necessary to preprocess the data to make the sample covariance identity by studentization.

Example 3 *We revise the setup of Ex. 1 with higher dimensions and data size: we let* $n = 500$, $X \sim \mathrm{Unif}[-1,1]^5$, $\mathbb{P}(T=1 \mid X) = 0.95/(1 + \frac{3}{\sqrt{5}}\|X\|_2)$, *and* $Y(0) \mid X \sim \mathcal{N}(X_1^2 + X_2^2 - X_1/2 - X_2/2, \sqrt{3})$, *so that there are three redundant covariates. We consider all weighting methods in the preceding examples along with the following additional*

methods: KOM++ with ARD; inverse propensity weights (including AIPW when used in the augmented estimator) and propensity score matching (PS1:1M) with propensities estimated by logistic regression; and local linear matching as proposed by Heckman et al. (1997) applied either to X (LLM) or to propensities (PSLLM). For CEM, we coarsen into the greatest number of levels per covariate while maintaining at least one control unit in each stratum with a treated unit. For local linear matching, we use triweight kernel on Mahalanobis distances and 50% span. For each method, we interpret the result as a set of weights and consider either the simple weighting estimator $\hat{\tau}_W$, the augmented estimator $\hat{\tau}_{W,\hat{f}_0}$ with $\hat{f}_0$ given by OLS, and the weighted least squares estimator $\hat{\tau}_{\mathrm{WLS}(W)}$ (except for local linear matching due to negative weights). We run 500 replications and tabulate the marginal *mean squared error (MSE) for estimating SATT in Tab. 2.*

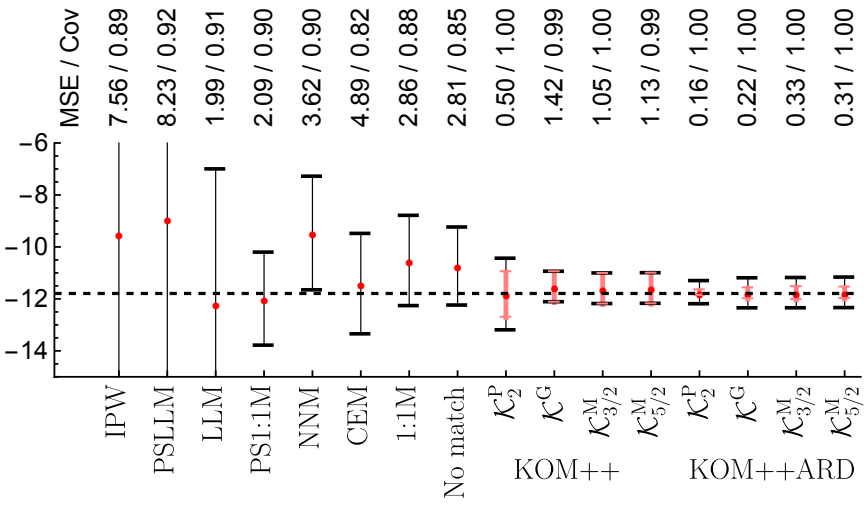
4.6. Infant Health and Development Program

We next consider evaluating the practical usefulness of KOM++ by studying data from the Infant Health and Development Program (IHDP). IHDP was a randomized experiment intended to measure the effect of a program consisting of child care and home visits from a trained provider on early child development (Brooks-Gunn et al., 1992), as measured through cognitive test scores.

To make an observational study from this data, we follow the construction of Hill (2012), where the subject of study is a child. We make one modification to further exacerbate overlap. Like Hill (2012), we remove all children with non-white mothers from the treatment group. To make overlap worse, we further remove all children with mothers aged 23 or younger from the treatment group and all children with mothers that are either white or aged 26 or older in the control group. Following Hill (2012), mother's age and race are omitted as covariates. In sum, the treatment group ($n_1 = 94$) consists only of children with older white mothers and the control group ($n_0 = 279$) consists only of children with younger nonwhite mothers, creating groups that are highly disparate in socioeconomic privilege. (The age cutoffs are near the mean and were chosen so to keep the data non-linearly-separable or else propensity score methods with scores estimated by logistic regression would be undefined.) Each unit is described by 25 covariates, 6 continuous and 19 binary, corresponding to measurements on the child (birth weight, etc.), measurements on the child's mother (smoked during pregnancy, etc.), and site. We generate outcomes $Y_i(0), Y_i(1)$ precisely as described by the *non-linear* response surface of Hill (2012) ("response surface B," which specified f_0 as the exponential of a linear function in covariates) conditioning on the coefficient on mother's age being zero.

We consider a range of methods: standard methods, KOM++ with various kernels, and KOM++ with ARD. The standard methods we consider include: inverse propensity weighting (IPW) and propensity score matching (PS1:1M) both using propensity scores estimated by logistic regression, NNM and 1:1M on the Mahalanobis distance, local linear matching (Heckman et al., 1997) on covariates (LLM) or propensities (PSLLM) using the triweight kernel on Mahalanobis distances with 50% span, and CEM on a coarsening chosen for maximal overlap. Coarsening on all variables (treating continuous variables as indicators for being above or below the mean) creates too many strata (2^{25}) and leaves only one treated unit in a stratum with at least one control unit. Instead, for CEM, we select half of the

Figure 4: Effect Estimation for the Infant Health and Development Program



covariates (13) to maximize the number of treated units that are in a stratum with at least one control unit after coarsening on only these covariates. Following the suggestion of Iacus et al. (2011a), we then proceed to prune all treated units in strata without overlap, leaving 69 units (only for CEM). We also omit NNM++ due to the high computational burden of its cross-validation procedure. For KOM++ we consider the quadratic ($\mathcal{K}_2^{\mathrm{P}}$), Gaussian ($\mathcal{K}^{\mathrm{G}}$), and Matérn ($\mathcal{K}_{3/2}^{\mathrm{M}}, \mathcal{K}_{5/2}^{\mathrm{M}}$) kernels and either using or not using ARD. We compute standard errors for all methods using the method of Imbens and Rubin (2015, §19.6). We construct confidence intervals by adding 1.96 standard errors to either the point estimate for all standard methods or to the interval estimate given by eq. (14) (see Appendix B) for KOM++, using the magnitude of f_0 as estimated by the marginal likelihood step.

We plot the results in Fig. 4. At the top, the figure lists the *marginal* mean-squared errors (MSE) $\mathbb{E}[(\hat{\tau} - \mathrm{SATT})^2]$ and the coverage of the 95% confidence intervals over 10,000 runs (note that SATT differs by run). Below the MSEs, the figure shows the results from one representative example run, showing SATT (dashed line), point estimates (red dots), confidence intervals (black bars), and, in the case of KOM++, interval estimates (pink bars). It is clear that among matching and weighting methods, KOM++ and, in particular, KOM++ when using ARD leads to significantly smaller error. As a weighting method, it can easily be combined with any regression technique by reweighting the training data or by reweighting the average of residuals.

The low overlap in this example leads IPW to produce extreme weights and suffer high MSE. In comparison, KOM++, while it cannot fix the unavoidable bias due to lack of overlap in and of itself, is able to maintain stable weights and small MSE by considering error directly while limiting extrapolation. 1:1M also provides rather stable weights but has a much harder time achieving good balance, leading to significantly higher MSE than KOM++. Through the lens of GOM we can identify two causes for this. On the one hand, 1:1M is only a heuristic way to trade off balance and variance: as seen in Ex. 1, it is not necessarily on the Pareto-efficient frontier. On the other, it is trying to balance far too much than is really necessary. One way to understand 1:1M's slow convergence

rate in $d \geq 2$ (Abadie and Imbens, 2006) is that finding good pairs becomes rapidly hard as dimension grows modestly. GOM offers another, functional-analytic perspective: 1:1M, which optimizes balance with respect to Lipschitz functions, is trying to balance far too much. Lipschitz functions are not only infinite dimensional, they are also non-separable so they have too little structure to be practically useful. In comparison, a C_0-universal RKHS, such as those given by the Gaussian or Matérn kernels, can still approximate any function arbitrarily well, but is still a separable space, admitting a countable orthonormal basis so that KOM++ is essentially balancing a countable number of moments. Essentially, this imposes enough structure so that good balance is actually achievable but not too much so that the resulting method remains fully non-parametric. Often, as in this example, even quadratic is enough, but it does not hurt too much to use the Gaussian or Matérn kernels and guarantee consistency without specification. Of course, if f_0 is truly extremely unsmooth, traditional matching (*i.e.*, optimizing against Lipschitz f_0) may be perform better, but this is generally not the case and would still not prevent consistency of KOM when using a C_0-universal RKHS. Adding ARD to KOM++ significantly improves its performance by learning the right representation of the data to balance, leading to lower MSEs.

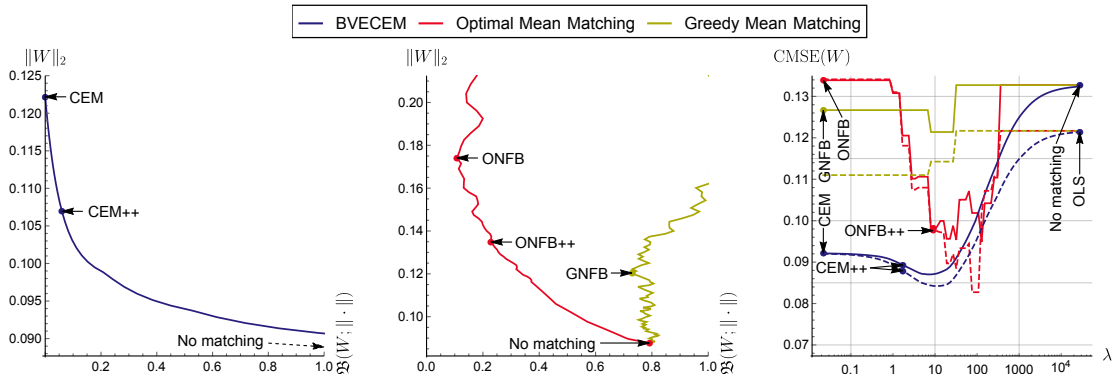
4.7. Recommendations for Practical Use

In sum the results, both theoretical and empirical, suggest that KOM++ can offer significant benefits in estimation precision and robustness. Using KOM with a C_0-universal kernel such as Gaussian or Matérn is non-parametric, just like optimal matching, and guarantees consistency regardless of model specification, which is particularly reassuring in an observational study. At the same time, the empirical results provide strong evidence that, when applied correctly using KOM++ with ARD for parameter tuning, using these nonparametric kernels incurs little to no deterioration in precision compared to using parametric kernels like quadratic when quadratic happens to be well-specified (which of course could not be known in practice). Therefore, a robust general-purpose recommendation for the use of KOM in practice is to use the Gaussian or Matérn kernel with KOM++ with ARD for parameter tuning. The choice of which C_0-universal kernel seems to matter little based on the results in Sec. 4.6 and Ex. 3.. In theory, the Matérn kernel only requires the existence of enough derivatives for "correct specification" and a speed up from $o_p(1)$ to $O_p(1/\sqrt{n})$. The Gaussian kernel requires more but it is more standard in practice and is simpler in form. In practice, such abstract notations of specification may have little relevance as both kernels yield very similar MSEs and what matters most is the reassuring blanket guarantee of model-free consistency, which is shared by both.

As a matching method, KOM is amenable to the same inference methods, such as that of (Abadie and Imbens, 2006), which was developed for optimal matching and generalized in (Imbens and Rubin, 2015, §19.6). It also gives rise to new inference methods based on the empirical Bayesian estimation of hyperparameter used in KOM++. These enable the construction of confidence intervals for KOM++ estimates.

More importantly, KOM provides an explicit bound on bias, which can be useful in instances with limited overlap where the bias can be significant or even ultimately irreducible. It is again advisable to use a C_0-universal kernel for constructing interval estimates using KOM as in eq. (14). The question for balance is whether the matched samples are com-

Figure 5: The Balance-Variance Trade-off in CEM and NFB



(a) Balance-variance (b) Balance-variance (c) CMSE

parable for the purpose of effect estimation. This means that evaluating balance just on differences of covariates means (as for example done on a Love plot) will not be helpful if effects are nonlinear. Evaluating balance using KOM with a nonparametric, C_0-universal kernel, however, will necessarily protect against any possible form of the effect. It will also be smaller than a similar bound for a pair-matching method because the imbalances it will leave will necessarily be huge, especially for data with more than just a couple dimensions. Correspondingly, the interval estimate produced by KOM with a C_0-universal kernel will be both precise and reliable in practice even when overlap is low so that both estimating effects and assessing specification is difficult.

5. Existing Matching Methods as GOM

As discussed in Sec. 3.5, many matching methods commonly used in practice – not just NNM and 1:1M – are GOM. In this section, we review these results and discuss practical implications, such as considering the balance-variance trade off. Like NNM and 1:1M, many of these methods are GOM with $\lambda = 0$, in which case they just minimize a bias-dual-norm balance metric (Kallus, 2017). Note that while very many existing methods can be seen as GOM, a few cannot, including full matching (Rosenbaum, 1991; Sävje et al., 2017), which focuses on finding strata composed of multiple units to be used for stratification-based estimation and inference (on the other hand, given these strata, the matching is equivalent to exact matching, which weights control units in each stratum by the inverse fraction of control units in the stratum, which is GOM with respect to L_∞).

5.1. Nearest-Neighbor Matching

Per Thm. 3, NNM is equivalent to GOM($\mathcal{W}^{\text{simplex}}, \|\cdot\|_{\text{Lip}(\delta)}, 0$). Minimizing CMSE when residual variances are not zero, however, would lead to GOM($\mathcal{W}^{\text{simplex}}$, $\|\cdot\|_{\text{Lip}(\delta)}, \lambda$), which refer to as BVE-NNM and which is given by the following linearly-

constrained convex-quadratic optimization problem:

$$\begin{array}{ll} \min & (\sum_{i\in\mathcal{T}_0,j\in\mathcal{T}_1}\delta(X_i,X_j)S_{ij})^2+\lambda\left\|W\right\|_2^2 \\ \text{s.t.} & W\in\mathbb{R}^{\mathcal{T}_0},\ S\in\mathbb{R}_+^{\mathcal{T}_0\times\mathcal{T}_1} \\ & \sum_{i\in\mathcal{T}_0}S_{ij}=1/n_1 \qquad\qquad \forall j\in\mathcal{T}_1 \\ & \sum_{j\in\mathcal{T}_1}S_{ij}=W_i \qquad\qquad \forall i\in\mathcal{T}_0 \end{array}$$

Ex. 1 illustrates the need to carefully weigh the imperative to balance in the face of variance in NNM but also begs the question of how should we appropriately tune the exchange rate λ. We present a approach we call NNM++ based on the interpretation of optimal matching as protecting against Lipschitz continuous functions and using cross validation for hyperparameter estimation. Assuming homoskedasticity, the hyperparameters of interest are the residual variance σ^2 and Lipschitz constant γ.

NNM++ proceed as follows. We consider regularization parameters $\Psi\subseteq[0,\infty)$ and m disjoint folds $\mathcal{T}_0=\mathcal{T}_0^{(1)}\sqcup\cdots\sqcup\mathcal{T}_0^{(m)}$. For each $\psi\in\Psi$ and validation fold $k=1,\dots,m$, we find $\hat{f}_0^{(k)}$ that minimizes the sum of squared errors in $\mathcal{T}_0\backslash\mathcal{T}_0^{(k)}$ regularized by ψ times the Lipschitz constant. Out of fold, a range of functions agrees with the fitted value $\hat{v}_i=\hat{f}_0^{(k)}(X_i)$, $\hat{\gamma}=\|\hat{f}_0^{(k)}\|_{\mathrm{Lip}(\delta)}$: $\hat{f}_0^{(k)}(x)\in[\min_{i\in\mathcal{T}_0\backslash\mathcal{T}_0^{(k)}}(\hat{v}_i+\hat{\gamma}\delta(X_i,x)),\max_{i\in\mathcal{T}_0\backslash\mathcal{T}_0^{(k)}}(\hat{v}_i-\hat{\gamma}\delta(X_i,x))]$. For the purposes of evaluating out-of-fold error, we use the point-wise midpoint of this interval. We select $\hat{\psi}$ with least out-of-fold mean squared error averaged over folds, let $\hat{\sigma}^2$ be this least error, and refit the $\hat{\psi}$ regularized problem on the whole $\mathcal{T}_0$ sample to estimate $\hat{\gamma}$. This cross-validation procedure is summarized in Alg. 1, listed in Appendix A. Note that we are not interested in a good fit of $\hat{f}_0$ – only a handle on the hyperparameter $\lambda=\sigma^2/\gamma^2$. Finally, we compute the BVE-NNM weights with $\hat{\lambda}=\hat{\sigma}^2/\hat{\gamma}^2$. In Ex. 1, NNM++ is shown using 10-fold cross validation in Figs. 1b and 1c.

Note that given only that f_0 is Lipschitz, is not generally possible to estimate its Lipschitz constant without bias given noisy observations. The above cross-validation procedure will necessarily shrink the estimate and have some downward bias. Moreover, NNM++ does not produce "honest" weights. It is also a very computationally intensive procedure, requiring solving many large quadratic optimization problems. We merely present NNM++ as one principled way to trade off balance and variance in optimal matching that seems to perform competitively.

An alternative approach to finding something on the Pareto-efficient frontier that is strictly "honest" is to compute 1:1M and find *anything* on the frontier that dominates it, *e.g.*, in Ex. 1, we may choose any point on the frontier that is below or to the left of 1:1M in Fig. 1b and improve upon it.

5.2. Optimal-Caliper Matching

A sibling of NNM is optimal-caliper matching (OCM), which selects matches with distances that all fit in the smallest possible single caliper. Allowing replacement, NNM is one of many OCM solutions. Without replacement, NNM and OCM differ. The next theorem shows that OCM is GOM with $\lambda=0$.

Proposition 22 *Fix a pseudo-metric $\delta:\mathcal{X}\times\mathcal{X}\to\mathbb{R}_+$. Let*

$$\|f\|_{\partial(\mu,\delta)}=\mathbb{E}_{x\sim\mu,x'\sim\mu}\left[\delta(x,x')^{-1}|f(x)-f(x')|\mid x\neq x'\right].$$

OCM with replacement is equivalent to $\mathrm{GOM}(\mathcal{W}^{simplex}, \|\cdot\|_{\partial(\hat{\mu}_n,\delta)}, 0)$*, where* $\hat{\mu}_n$ *is the empirical distribution of* X*. OCM without replacement is equivalent to* $\mathrm{GOM}(\mathcal{W}^{1/n_1\text{-}simplex}, \|\cdot\|_{\partial(\hat{\mu}_n,\delta)}, 0)$*.*

5.3. Coarsened Exact Matching

Given a coarsening $C : \mathcal{X} \to \{1, \ldots, J\}$ stratifying $\mathcal{X}$, CEM (Iacus et al., 2011a) minimizes the coarsened L_1 distance,

$$\sum_{j=1}^{M} \left| \frac{1}{n_1} \sum_{i \in \mathcal{T}_1} \mathbb{I}_{[C(X_i)=j]} - \sum_{i \in \mathcal{T}_0} W_i \mathbb{I}_{[C(X_i)=j]} \right|, \tag{11}$$

by simply equating the matched control distribution in each stratum by setting $W_i = n_1^{-1} |\{j \in \mathcal{T}_1 : [C(X_i) = C(X_j)]\}| / |\{j \in \mathcal{T}_0 : [C(X_i) = C(X_j)]\}|$.

CEM is also GOM with $\lambda = 0$ as the next result, adapted from Kallus (2017, Theorem 4), shows.

Proposition 23 *Fix a coarsening function* $C : \mathcal{X} \to \{1, \ldots, J\}$*. Let*

$$\|f\|_{L_p(C)} = \begin{cases} \|(\sup_{x \in \mathcal{C}^{-1}(j)} |f(x)|)_{j=1}^{J}\|_p & \left|f(C^{-1}(j))\right| = 1 \ \forall j, \\ \infty & \textit{otherwise.} \end{cases}$$

(I.e., the vector p-norm of the values taken by f *on the coarsened regions if* f *is piecewise constant.) CEM is equivalent to* $\mathrm{GOM}(\mathcal{W}^{simplex}, \|\cdot\|_{L_\infty(C)}, 0)$*.*

In practice, one often considers a sequence of nested coarsenings, each coarser than the previous, and chooses one to control balance and extreme weights. Instead, we can simply consider a Balance-Variance Efficient CEM (BVE-CEM) given by $\mathrm{GOM}(\mathcal{W}^{\mathrm{simplex}}, \|\cdot\|_{L_\infty(C)}, \lambda)$ for general λ. The BVE-CEM weights are given by the following optimization problem

$$\min_{W \in \mathcal{W}^{\mathrm{simplex}}} (\sum_{j=1}^{J} |\frac{1}{n_1} \sum_{i \in \mathcal{T}_1} \mathbb{I}_{[C(X_i)=j]} - \sum_{i \in \mathcal{T}_0} W_i \mathbb{I}_{[C(X_i)=j]}|)^2 + \lambda \|W\|_2^2 \tag{12}$$

Unlike CEM, the solution does not have a closed form. We can solve this optimization problem explicitly by considering all combinations of signs for the J absolute values. For each combination, a Lagrange multiplier argument yields an optimal solution. By further observing that we need only consider monotonic deviations from the usual CEM solution, we obtain Alg. 2, listed in Appendix A, which finds the BVE-CEM weights in $O(J^2)$ time.

We also consider a CEM++ variant given by estimating λ and using the corresponding BVE-CEM. Unlike BVE-NNM, here the class of functions $\{f : \|f\|_{L_\infty(C)} < \infty\}$ is generally "small."[9] Therefore, we sidestep a complicated validation scheme and simply estimate the parameters in-sample but use the one-standard-error rule (Friedman et al., 2001, §7.10) to carefully tune $\hat{\gamma}$. Set $\hat{\mu}_j = \sum_{i \in \mathcal{T}_0} \mathbb{I}_{[C(X_i)=j]} Y_i / \sum_{i \in \mathcal{T}_0} \mathbb{I}_{[C(X_i)=j]}$ and $\hat{\sigma}^2 = \frac{1}{n_0 - J} \sum_{i \in \mathcal{T}_0} (Y_i - \hat{\mu}_{C(X_i)})^2$. To estimate γ, we seek the smallest $\hat{\gamma} \geq 0$ such that the minimal error over functions with range $2\hat{\gamma}$ is no worse than $\hat{\sigma}^2$ plus its standard error $\frac{\hat{\sigma}^2 \sqrt{2}}{n_0 - J}$. (Note

9. Specifically, parametric with fewer parameters than data, $J < n_0$. In NNM++, the class of functions was not only infinite dimensional but also *non-separable*.

$\hat{\sigma}^2$ error is achieved by $2\hat{\gamma} = \max_j \hat{\mu}_j - \min_j \hat{\mu}_j$.) That is, we seek smallest $\hat{\gamma} \geq 0$ with $\frac{1}{n_0-J} \min_{c\in\mathbb{R}} \sum_{i\in\mathcal{T}_0} (Y_i - \max(\min(\hat{\mu}_j, c+\hat{\gamma}), c-\hat{\gamma}))^2 \leq \hat{\sigma}^2(1+\sqrt{2}/(n_0-J))$. We do this by using bisection search on $\hat{\gamma}$ and a nested golden section search on c. We refer to BVE-CEM with $\hat{\lambda} = \hat{\sigma}^2/\hat{\gamma}^2$ as CEM++.

5.4. Mean matching and near-fine balance

Suppose $\mathcal{X} \subseteq \mathbb{R}^d$ so X_i is vector-valued. Mean matching (Zubizarreta, 2012, 2015; Greenberg, 1953; Rubin, 1973; Bertsimas et al., 2015) are methods that find a subset of control units $\mathcal{T}_0{}' \subseteq \mathcal{T}_0$ to reduce the Mahalanobis distance between the sample means

$$M_V(\mathcal{T}_0{}') = \|V^{1/2}(\tfrac{1}{n_1}\textstyle\sum_{i\in\mathcal{T}_1} X_i - \tfrac{1}{|\mathcal{T}_0{}'|}\sum_{i\in\mathcal{T}_0{}'} X_i)\|_2.$$

We can consider optimal mean matching (OMM) as fully minimizing $M_V(\mathcal{T}_0{}')$ over all possible subsets $\mathcal{T}_0{}'$, which we show is GOM.

Proposition 24 *Suppose $\mathcal{X} \subseteq \mathbb{R}^d$. Let $V \in \mathbb{R}^{d\times d}$ be positive definite and*

$$\|f\|^2_{2\text{-}lin(V)} = \begin{cases} \alpha^2 + \beta^T V^{-1}\beta & f(x) = \alpha + \beta^T x, \\ \infty & \textit{otherwise.} \end{cases}$$

Then OMM is equivalent to $\mathrm{GOM}(\mathcal{W}^{subsets}, \|\cdot\|_{2\text{-}lin(V)}, 0)$.

Again, we may consider the more general $\mathrm{GOM}(\mathcal{W}^{\mathrm{subsets}}, \|\cdot\|_{2\text{-lin}(V)}, \lambda)$. As per Prop. 7 this can be written as the following convex-quadratic binary optimization problem for some $n(\lambda)$ and with $U_i = n(\lambda)W_i$:

$$\min_{U\in\{0,1\}^{\mathcal{T}_0}:\sum_{i\in\mathcal{T}_0} U_i = n(\lambda)} (\textstyle\sum_{i\in\mathcal{T}_1}\frac{X_i}{n_1} - \sum_{i\in\mathcal{T}_0}\frac{U_i X_i}{n(\lambda)})^T V (\sum_{i\in\mathcal{T}_1}\frac{X_i}{n_1} - \sum_{i\in\mathcal{T}_0}\frac{U_i X_i}{n(\lambda)}).$$

An alternative form of mean matching would be to minimize the ℓ_p distance between sample means. If we define $\|x \mapsto \alpha + \beta^T x\|_{p\text{-lin}(V)} = \|(\alpha, V^{-1/2}\beta)\|_p$ (and ∞ for all non-linear functions) then $\mathrm{GOM}(\mathcal{W}^{\mathrm{subsets}}, \|\cdot\|_{p\text{-lin}(I)}, \lambda)$ is given by the following convex binary optimization problem for some $n(\lambda)$

$$\min_{U\in\{0,1\}^{\mathcal{T}_0}:\sum_{i\in\mathcal{T}_0} U_i = n(\lambda)} \|\textstyle\sum_{i\in\mathcal{T}_1}\frac{X_i}{n_1} - \sum_{i\in\mathcal{T}_0}\frac{U_i X_i}{n(\lambda)}\|_{p'}, \tag{13}$$

where $1/p + 1/p' = 1$. This optimization problem is an integer *linear* optimization problem whenever $p' \in \{1, \infty\}$ (equivalently, $p \in \{1, \infty\}$).

If the covariates are 0-1 indicators (*e.g.*, if they are 2-level factors, if multilevel and encoded in unary as concatenated one-hot vectors, or if continuous and coarsened into multilevel factors and thus encoded), then the sample mean is simply the vector of sample proportions, *i.e.*, it is all marginal distributions of each multilevel factor. In this specific case, mean matching (especially when $p' = 1$, or equivalently $p = \infty$) is known as *near*-fine balance (Zubizarreta, 2012). In this case, we refer to $\mathrm{GOM}(\mathcal{W}^{\mathrm{subsets}}, \|\cdot\|_{\infty\text{-lin}(I)}, 0)$ as optimal near-fine balance (ONFB). When the marginal sample distributions can be made exactly equal, the resulting allocation is known as fine balance (Rosenbaum et al., 2007).

Zubizarreta (2015) considers the finding the simplex weights that minimize the variance of the set of weights subject to balance constraints on mean discrepancies. Because the

weights are in the simplex, their variance is equal to their squared Euclidean norm up to a constant. And, by convex duality, constraining mean discrepancies is the same as penalizing them. Therefore, Zubizarreta (2015) is equivalent to $\mathrm{GOM}(\mathcal{W}^{\text{simplex}}, \|\cdot\|_{1\text{-lin}(\mathrm{diag}(v))}, \lambda)$ for some $v \in \mathbb{R}^d_+$, $\lambda \geq 0$.

We can also consider a variant given by estimating λ for automatic subset size selection. Assuming $d < n_0$, the function class is "small," so we estimate λ in-sample. We let $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$ be given by OLS regression on $\{(X_i, Y_i) : i \in \mathcal{T}_0\}$ and set $\hat{\lambda} = \hat{\sigma}^2 / \|\hat{\beta}\|_p^2$. We refer to $\mathrm{GOM}(\mathcal{W}^{\text{subsets}}, \|\cdot\|_{\infty\text{-lin}}, \hat{\lambda})$ as ONFB++.

5.5. Mixed objectives

Methods such as Yang et al. (2012); Zubizarreta (2012) seek to minimize both the sum of pairwise distances and a discrepancy in means. These are also GOM.

Proposition 25 *Let $\mathcal{W}$, $\|\cdot\|_A$, $\|\cdot\|_B$, and $\rho > 0$ be given. Then*

$$\begin{aligned} &\mathfrak{B}(W; \|\cdot\|_{\|\cdot\|_A \oplus_\rho \|\cdot\|_B}) = \mathfrak{B}(W; \|\cdot\|_A) + \rho \mathfrak{B}(W; \|\cdot\|_B) \\ \textit{where} \quad &\|f\|_{\|\cdot\|_A \oplus_\rho \|\cdot\|_B} = \inf_{f_A + f_B = f} \max\left\{\|f_A\|_A, \|f_B\|_B / \rho\right\} \end{aligned}$$

Therefore, for example, $\mathrm{GOM}(\mathcal{W}^{\text{subsets}}, \|\cdot\|_{\|\cdot\|_{\mathrm{Lip}(\delta)} \oplus_\rho \|\cdot\|_{\infty\text{-lin}}}, \lambda)$ is given by the following integer linear optimization problem (for some $n(\lambda)$):

$$\begin{aligned} \min \quad & \tfrac{1}{n(\lambda)} \textstyle\sum_{i \in \mathcal{T}_0, j \in \mathcal{T}_1} \delta(X_i, X_j) S_{ij} + \tfrac{\rho}{n(\lambda)} \sum_{k=1}^d D_k \\ \text{s.t.} \quad & U \in \{0,1\}^{\mathcal{T}_0}, S \in \mathbb{R}_+^{\mathcal{T}_0 \times \mathcal{T}_1}, D \in \mathbb{R}^d \\ & \textstyle\sum_{i \in \mathcal{T}_0} U_i = n(\lambda) \\ & \textstyle\sum_{i \in \mathcal{T}_0} S_{ij} = n(\lambda)/n_1 && \forall j \in \mathcal{T}_1 \\ & \textstyle\sum_{j \in \mathcal{T}_1} S_{ij} = U_i && \forall i \in \mathcal{T}_1 \\ & D_k \geq \tfrac{n(\lambda)}{n_1} \textstyle\sum_{i \in \mathcal{T}_1} X_{ik} - \sum_{i \in \mathcal{T}_0} U_i X_{ik} && \forall k = 1, \ldots, d \\ & D_k \geq \textstyle\sum_{i \in \mathcal{T}_0} U_i X_{ik} - \tfrac{n(\lambda)}{n_1} \sum_{i \in \mathcal{T}_1} X_{ik} && \forall k = 1, \ldots, d \end{aligned}$$

For $n(\lambda) = n_0'$ chosen a priori, this is one of the problems considered by Zubizarreta (2012). Note our formulation has only n_0 discrete variables as we need not constrain S to be integral because the matching polytope is totally unimodular (Ahuja et al., 1993).

Similarly, given a coarsening function $C : \mathcal{X} \to \{1, \ldots, J\}$, (Yang et al., 2012, eq. (2)) is given by $\mathrm{GOM}(\mathcal{W}^{n_1\text{-subset}}, \|\cdot\|_{\|\cdot\|_{\mathrm{Lip}(\delta)} \oplus_\rho \|\cdot\|_{L_\infty(C)}}, 0)$, (Yang et al., 2012, eq. (3)) is given by $\mathrm{GOM}(\mathcal{W}^{n_1\text{-subset}}, \|\cdot\|_{\|\cdot\|_{\mathrm{Lip}(\delta)} \oplus_\rho \|\cdot\|_{L_1(C)}}, 0)$, and (Yang et al., 2012, eq. (4)) is given by $\mathrm{GOM}(\mathcal{W}^{n_1\text{-subset}}, \|\cdot\|_{\|\cdot\|_{\mathrm{Lip}(\delta)} \oplus_\rho \|\cdot\|_{L_2(C)}}, 0)$,[10] all for $\rho > 0$ sufficiently large.

Example 4 *Let us revisit Ex. 1 to study coarsened exact and near-fine balance matching. We coarsen each of the two covariates into an 8-level factor encoding its marginal octile in the sample.*[11] *We consider BVE-CEM with all $J = 64$ strata and plot the achievable balance-variance landscape in Fig. 5a. We point out both CEM and CEM++. No matching is in*

10. Using the vector 2-norm scaled by $(\sum_{i \in \mathcal{T}_1} \mathbb{I}_{[C(X_i)=j]})_{j=1}^J$.

11. In particular, we chose the greatest number ℓ of levels such that for the resulting ℓ^2 strata, every stratum with a treated unit had at least one control unit.

the far right, outside the plot area. We also consider mean-matching ($p = \infty, p' = 1$) on the resulting 16-dimensional vector, which corresponds to near-fine balance on the two coarsened covariates, and plot the balance-variance landscape in Fig. 5b. The red curve in Fig. 5b is the balance-variance achieved by $\mathrm{GOM}(\mathcal{W}^{n_0'\text{-}subset}, \|\cdot\|_{\infty\text{-}lin}, 0)$ *for* $n_0' \in \{1, \dots, n_0\}$ *($n_0' \leq 22$ is outside the plot area). ONFB is the leftmost point and has $n_0' = 33$. Note that not all points on the curve are on the Pareto-efficient frontier given by* $\mathrm{GOM}(\mathcal{W}^{subsets}, \|\cdot\|_{\infty\text{-}lin}, \lambda)$ *for $\lambda \in [0, \infty]$. In particular, any $n_0' < 33$ cannot be on the frontier (showing the converse of Prop. 7 is false). We also point out ONFB++ and no matching in the plot. The yellow curve in Fig. 5b is a commonly used greedy heuristic for mean-matching whereby, starting from an empty subset, we incrementally add the unused control unit that would minimize the mean-matching objective. We point out GNFB, given by choosing the point along the greedy path with minimal mean-matching objective. It is the leftmost point on the curve. We plot the resulting CMSE of $\hat{\tau}_W$ (solid) and $\hat{\tau}_{W,\hat{f}_0}$ (dashed) in Fig. 5c corresponding only to points on the Pareto-efficient frontier of each curve. The need to tune λ and consider the variance objective is clear. Even GNFB beats ONFB because the unintended larger matched set induced by sub-optimality. Correctly tuning λ, ONFB++ improves on both. Similarly, CEM++ improves on CEM.*

5.6. Regression as GOM

An alternative to matching is regression adjustment via OLS with interaction terms (Lin, 2013):[12]

$$\hat{\tau}_{\mathrm{OLS}} = \operatorname*{argmin}_{\tau \in \mathbb{R}} \min_{\alpha \in \mathbb{R}, \beta_1, \beta_2 \in \mathbb{R}^d} \sum_{i=1}^{n} \left(Y_i - \alpha - \tau T_i - \beta_1^T X_i - \beta_2^T (X_i - \overline{X}_{\mathcal{T}_1}) T_i\right)^2,$$

where $\overline{X}_{\mathcal{T}_1} = \frac{1}{n_1} \sum_{i \in \mathcal{T}_1} X_i$ is the treated sample mean vector. Surprisingly, this is exactly equivalent to an *unrestricted* version of mean-matching.

Proposition 26 *Let V positive definite be given and let W be given by* $\mathrm{GOM}(\mathcal{W}^{general}, \|\cdot\|_{2\text{-}lin(V)}, 0)$. *Then $\hat{\tau}_W = \hat{\tau}_{OLS}$.*

We actually prove this as a corollary of a more general result about the ridge-regression version of the regression adjustment with interaction terms:

$$\hat{\tau}_{\lambda\text{-ridge}} = \operatorname*{argmin}_{\tau \in \mathbb{R}} \min_{\alpha \in \mathbb{R}, \beta_1, \beta_2 \in \mathbb{R}^d} \sum_{i=1}^{n} \left(Y_i - \alpha - \tau T_i - \beta_1^T X_i - \beta_2^T (X_i - \overline{X}_{\mathcal{T}_1}) T_i\right)^2 + \lambda \|\beta_1\|_2^2 + \lambda \alpha^2.$$

Proposition 27 *Let W be given by* $\mathrm{GOM}(\mathcal{W}^{general}, \|\cdot\|_{2\text{-}lin}, \lambda)$ *for $\lambda \geq 0$. Then $\hat{\tau}_W = \hat{\tau}_{\lambda\text{-}ridge}$.*

12. Note that using the interaction term $(X_i - \overline{X}_{\mathcal{T}_1})T_i$ corresponds to estimating the effect on the treated whereas using the interaction term $(X_i - \overline{X})T_i$, as it appears in Lin (2013), corresponds to estimating the overall average effect. Note also that by no means does Lin (2013) recommend regression adjustment in observational settings; he instead studies the experimental setting, where model-agnostic consistency is assured.

These results reveal a very close connection between matching and regression adjustment and expands the scope of existing connections (Qingyuan and Daniel, 2017; Athey et al., 2016). But, there are several nuanced but important differences between regression adjustment and regular mean-matching (*i.e.*, with simplex or subset weights). Matching with simplex or subset weights results in a distribution over the sample units and thus is more interpretable, preserving the unit of analysis. This also allows certain randomization-based inferences. Moreover, whereas linear regression is subject to dangerous extrapolation (King and Zeng, 2006; Ho et al., 2007), matching with weights in the simplex (corresponding to convex combinations), including subsets, inherently prohibits extrapolation. OLS and mean-matching may coincide only if *exact* fine balance is feasible.

Proposition 27 can be easily generalized to an RKHS norm where it would recover regression adjustment via kernel ridge regression. More generally, recalling that GOM is a minimax linear estimator with coefficients constrained as $W \in \mathcal{W}$ (see Section 3.3), $\text{GOM}(\mathcal{W}^{\text{general}}, \|\cdot\|, \lambda)$ for any $\|\cdot\|$ is the *unconstrained* minimax linear estimator. Since the minimax linear estimator is also the minimax affine estimator, Donoho (1994, Corollary 1) gives that this estimator has worst-case CMSE (*i.e.*, $\mathfrak{E}^2_{\min}(\mathcal{W}^{\text{general}}, \|\cdot\|, \lambda)$) at most 5/4 of the general (non-linear) minimax risk (*i.e.*, the minimal worst-case CMSE where $\widehat{\overline{Y}}_{\mathcal{T}_1}(0)$ may be any measurable function of $Y_{\mathcal{T}_0}$, where the function may depend on $X_{1:n}, T_{1:n}$), although the general minimax risk may depend on the distribution of ϵ_i beyond σ_i^2. .

6. Conclusion

In this paper, we presented an encompassing framework and theory for matching methods for causal inference, arising from generalizations of a new functional analytical interpretation of optimal matching. On the one hand, this framework revealed a unifying thread between and provided a unified theoretical analysis for a variety of commonly used methods, including both matching and regression methods. This in turn lead to new extensions to methods subsumed in this framework that appropriately and automatically adjust the balance-variance trade-off inherent in matching revealed by the theory developed. These extensions lead to benefits in estimation error relative to their standard counterparts.

On the other hand, this framework lead to the development of a new class of matching methods based on kernels. The new methods, called KOM, were shown to have some of the more appealing properties of the different methods in common use, as supported by specialized theory developed. In particular, KOM yields either a distribution over or a subset of the control units preserving the unit of analysis and avoiding extrapolation, KOM has favorable consistency properties yielding parametric-rate estimation under correct specification and model-free consistency regardless thereof, KOM has favorable robustness properties when used in an augmented weighted estimator, KOM has similarly favorable robustness properties when used as preprocessing before regression, KOM++ judiciously and automatically weighs balance in the face of variance, and KOM allows for flexible model selection via empirical Bayes methods. These properties make KOM a particularly apt tool for causal inference. Beyond SATT, KOM may be used to estimate a variety of causal estimands, including SATE, ATE on a target population, and continuous treatments, as demonstrated in extensions of the present work (Kallus et al., 2018; Kallus and Santacatterina, 2019b,a).

References

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Alberto Abadie and Guido W Imbens. Estimation of the conditional variance in paired experiments. *Annales d'Economie et de Statistique*, pages 175–187, 2008.

Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, 2011.

RK Ahuja, TL Magnanti, and JB Orlin. *Network flows: theory, algorithms, and applications*. Prentice Hall, Upper Saddle River, 1993.

Susan Athey, Guido W Imbens, and Stefan Wager. Efficient inference of average treatment effects in high dimensions via approximate residual balancing. *arXiv preprint arXiv:1604.07125*, 2016.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, 2004.

Dimitris Bertsimas, Mac Johnson, and Nathan Kallus. The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 63(4):868–876, 2015.

Stephen Poythress Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

Jeanne Brooks-Gunn, Fong-ruey Liaw, and Pamela Kato Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3):350–359, 1992.

Kwun Chuen Gary Chan, Sheung Chi Phillip Yam, and Zheng Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700, 2016.

Richard Crump, V Joseph Hotz, Guido Imbens, and Oscar Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand, 2006.

Alexis Diamond and Jasjeet S Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

David L Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, pages 238–270, 1994.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer, Berlin, 2001.

Daniel P Giesy. On a convexity condition in normed linear spaces. *Transactions of the American Mathematical Society*, 125(1):114–146, 1966.

Paul W Goldberg, Christopher KI Williams, and Christopher M Bishop. Regression with input-dependent noise: A gaussian process treatment. In *Advances in neural information processing systems*, pages 493–499, 1998.

Bernard G Greenberg. The use of analysis of covariance and balancing in analytical surveys*. *American Journal of Public Health and the Nations Health*, 43(6_Pt_1):692–699, 1953.

Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2006.

M Grotschel, L Lovasz, and A Schrijver. *Geometric algorithms and combinatorial optimization*. Springer, New York, 1993.

James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654, 1997.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 2012.

Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236, 2007.

Stefano M Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, page mpr013, 2011a.

Stefano M Iacus, Gary King, and Giuseppe Porro. Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, 106 (493):345–361, 2011b.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Nathan Kallus. A framework for optimal matching for causal inference. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Nathan Kallus. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):85–112, 2018.

Nathan Kallus and Michele Santacatterina. Kernel optimal orthogonality weighting: A balancing approach to estimating effects of continuous treatments. *arXiv preprint arXiv:1910.11972*, 2019a.

Nathan Kallus and Michele Santacatterina. Optimal estimation of generalized average treatment effects using kernel optimal matching. *arXiv preprint arXiv:1908.04748*, 2019b.

Nathan Kallus, Brenton Pennicooke, and Michele Santacatterina. More robust estimation of sample average treatment effects using kernel optimal matching in an observational study of spine surgical interventions. *arXiv preprint arXiv:1811.04274*, 2018.

Leonid Vasilevich Kantorovich and G Sh Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958.

Gary King and Langche Zeng. The dangers of extreme counterfactuals. *Political Analysis*, 14(2):131–159, 2006.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedmans critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

Bernard Maurey. Type, cotype and k-convexity. *Handbook of the geometry of Banach spaces*, 2:1299–1332, 2003.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443, 1997.

G Pisier. Sur les espaces qui ne contiennent pas de ℓ_n^1 uniformément. *C. R. Acad. Sci. Paris Sér.*, 277:991–994, 1973.

Zhao Qingyuan and Percival Daniel. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1), 2017.

James M Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, pages 6–10, 1999.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

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

Paul R Rosenbaum. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):597–610, 1991.

Paul R Rosenbaum. *Design of Observational Studies*. Springer, New York, 2010.

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

Paul R Rosenbaum, Richard N Ross, and Jeffrey H Silber. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*, 102(477):75–83, 2007.

Halsey L Royden. *Real Analysis*. Prentice Hall, 1988.

Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.

Donald B Rubin. Multivariate matching methods that are equal percent bias reducing, i: Some examples. *Biometrics*, pages 109–120, 1976a.

Donald B Rubin. Multivariate matching methods that are equal percent bias reducing, ii: Maximums on bias reduction for fixed sample sizes. *Biometrics*, pages 121–132, 1976b.

Donald B Rubin. Comments on "randomization analysis of experimental data: The fisher randomization test comment". *Journal of the American Statistical Association*, 75(371): 591–593, 1980.

Donald B Rubin and Neal Thomas. Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics*, pages 1079–1093, 1992.

Fredrik Sävje, Michael J. Higgins, and Jasjeet S. Sekhon. Generalized full matching and extrapolation of the results from a large-scale voter mobilization experiment. *arXiv preprint arXiv:1202.3757*, 2017.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *arXiv preprint arXiv:1003.0887*, 2010.

Elizabeth 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.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT Press, Cambridge, MA, 2006.

Andrew Gordon Wilson and Ryan Prescott Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

Dan Yang, Dylan S Small, Jeffrey H Silber, and Paul R Rosenbaum. Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics*, 68(2):628–636, 2012.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions. *arXiv preprint arXiv:1601.05890*, 2016.

José R Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107 (500):1360–1371, 2012.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

Appendix A. Algorithm listings

ALGORITHM 1: Cross-Validation Estimation for NNM++

input: Control data $Y_{\mathcal{T}_0}$, distance matrix $D \in \mathbb{R}^{\mathcal{T}_0 \times \mathcal{T}_0}$, regularizer grid $\Psi \subseteq \mathbb{R}_+$, and number of folds m.

Randomly split the control data into disjoint folds $\mathcal{T}_0 = \mathcal{T}_0^{(1)} \sqcup \cdots \sqcup \mathcal{T}_0^{(m)}$.

for $k \in \{1, \ldots, m\}$, $\psi \in \Psi$: **do**

Solve $\begin{array}{ll} \min_{\hat{v},\hat{\gamma}} & \frac{1}{|i \in \mathcal{T}_0 \setminus \mathcal{T}_0^{(k)}|} \sum_{i \in \mathcal{T}_0 \setminus \mathcal{T}_0^{(k)}} (\hat{v}_i - y_i)^2 + \psi \hat{\gamma} \\ \text{s.t.} & \hat{v}_i - v_j \leq \hat{\gamma} D_{ij} \quad \forall i, j \in \mathcal{T}_0 \setminus \mathcal{T}_0^{(k)} \end{array}$.

Set $\hat{Y}_i = \frac{1}{2}(\min_{j \in \mathcal{T}_0 \setminus \mathcal{T}_0^{(k)}} (\hat{v}_j + \hat{\gamma} D_{ij}) + \max_{j \in \mathcal{T}_0 \setminus \mathcal{T}_0^{(k)}} (\hat{v}_j - \hat{\gamma} D_{ij}))$.

Set $\hat{\sigma}^2_{k,\psi} = \sum_{i \in \mathcal{T}_0^{(k)}} (Y_i - \hat{Y}_i)^2 / (|\mathcal{T}_0^{(k)}| - 1)$.

end for

Set $\hat{\sigma}^2_\psi = \frac{1}{m} \sum_{k=1}^m \hat{\sigma}^2_{k,\psi}$ for $\psi \in \Psi$, $\hat{\psi} = \operatorname{argmin}_{\psi \in \Psi} \hat{\sigma}^2_\psi$ and $\hat{\sigma}^2 = \min_{\psi \in \Psi} \hat{\sigma}^2_\psi$.

Solve $\begin{array}{ll} \min_{\hat{v},\hat{\gamma}} & \frac{1}{|\mathcal{T}_0|} \sum_{i \in \mathcal{T}_0} (\hat{v}_i - y_i)^2 + \hat{\psi} \hat{\gamma} \\ \text{s.t.} & \hat{v}_i - v_j \leq \hat{\gamma} D_{ij} \quad \forall i, j \in \mathcal{T}_0 \end{array}$.

output: $\hat{\lambda} = \hat{\sigma}^2 / \hat{\gamma}^2$.

ALGORITHM 2: BVE-CEM (solves eq. (12))

input: Data $X_{1:n}, T_{1:n}$, coarsening function $C : \mathcal{X} \to \{1, \ldots, J\}$, and exchange λ.

Let $n_{tj} = \sum_{i \in \mathcal{T}_t} \mathbb{I}_{[C(X_i)=j]}$ for $t = 0, 1$, $j = 1, \ldots, J$.

Let $q_j = n_{1j}/(n_1 n_{0j})$ and sort $q_{j_1} \leq \cdots \leq q_{j_J}$ ($q_{j_0} = -\infty$, $q_{j_{J+1}} = \infty$).

Set $v^* = \infty$.

for $J_+ = 0, \ldots, J$, $J_- = 0, \ldots, J - J_+$ **do**

$\begin{array}{rl} \text{Set } n_{0+} & = \sum_{k=1}^{J_+} n_{0 j_{J+1-k}},\ n_{0-} = \sum_{k=1}^{J_-} n_{0 j_k},\ r = \sum_{k=J_-+1}^{J+1-J_+} n_{0 j_k} q_{j_k}, \\ r_\Delta & = \sum_{k=1}^{J_+} n_{0 j_{J+1-k}} q_{0 j_{J+1-k}} - \sum_{k=1}^{J_-} n_{0 j_k} q_{j_k},\ r_2 = \sum_{k=J_-+1}^{J-J_+} n_{0 j_k} q_{j_k}^2, \\ w_+ & = \frac{2n_{0-}(1-r+r_\Delta)+\lambda(1-r)}{4n_{0+}n_{0-}+\lambda(n_{0+}+n_{0-})},\ w_- = \frac{2n_{0+}(1-r-r_\Delta)+\lambda(1-r)}{4n_{0+}n_{0-}+\lambda(n_{0+}+n_{0-})},\ \text{and} \\ v & = (\sum_{k=1}^{J_+} n_{0 j_{J+1-k}} \left| q_{j_{J+1-k}} - w_+ \right| + \sum_{k=1}^{J_-} n_{0 j_k} |q_{j_k} - w_-|)^2 \\ & \quad + \lambda(n_{0p} w_+^2 + n_{0m} w_-^2 + r_2). \end{array}$

if $v < v^*$ **then**

$$\text{Set } W_i = \begin{cases} 1/n_1 & i \in \mathcal{T}_1, \\ w_- & i \in \mathcal{T}_0,\ q_{C(X_i)} \leq q_{j_{J_-}}, \\ w_+ & i \in \mathcal{T}_0,\ q_{C(X_i)} \geq q_{j_{J+1-J_+}}, \\ q_{C(X_i)} & \text{otherwise.} \end{cases}$$

end if

end for

output: W.

Appendix B. Inference and Partial Identification with KOM

In order to conduct inference on the value of SATT using KOM, it is important to develop appropriate standard errors or other confidence intervals for $\hat{\tau}_W$. There are several options for estimating standard errors. One general-purpose option is the bootstrap. In applying the bootstrap to KOM, we re-optimize the weights for each bootstrap sample and record the

resulting estimator to produce the bootstrap distribution (rather than, say, using a weighting function precomputed at the onset on the complete dataset). Quantile, studentized, and $\mathrm{BC_A}$ bootstrap intervals are possible choices (Efron and Tibshirani, 1994). Another general-purpose option is to use the estimate $\hat{\tau}_{\mathrm{WLS}(W)}$ and employ the corresponding robust sandwich (Huber-White) standard errors.

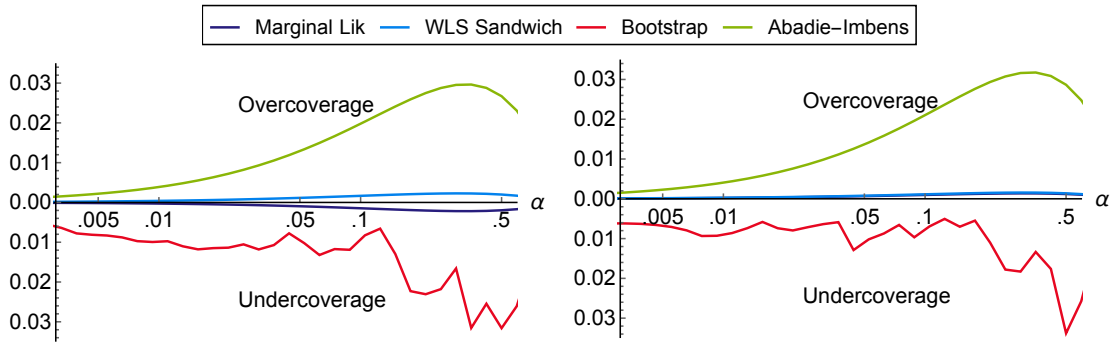
However, more specialized procedures are possible. One particularly appealing nature of matching is the transparent structure of the data (Rosenbaum, 2010, Ch. 6): it preserves the unit of analysis since the result, like the raw control sample, is still a valid distribution over the control units, whether it is a subset, a multisubset with duplicates, or any redistribution that is nonnegative and sums to one. This is preserved in KOM and enables the use of similar inferential methods as used in standard matching that interpret the data as a weighted sample.[13] In particular, since the weighted estimator $\hat{\tau}_W$ for SATT exactly matches the criteria presented in (Imbens and Rubin, 2015, §19.8), one approach to compute standard errors is to use the within-treatment-group matching techniques developed in (Abadie and Imbens, 2008, 2011, 2006) to estimate residual variances. In the specific case of KOM++, one can also use the marginal likelihood estimate of the residual variance to produce a standard error based on Prop. 1.[14] The next example explores how to use these methods to produce confidence intervals for KOM. Then, we will see how the interpretable nature of KOM can also allow us to account for unavoidable imbalances that lead to deceptive point estimates and instead produce more honest interval estimates.

Example 5 *We revisit Ex. 2 to look at (finite-sample) inference on SATT using KOM++ based on each of the confidence intervals above. We consider the situation of a constant effect* $Y(1)-Y(0)=\tau$ *and plot the desired significance* α *against the difference of the actual coverage and* $1-\alpha$ *in Fig. 6 for two examples: quadratic and Matérn. Coverage is computed keeping* $X_{1:n}, T_{1:n}$ *fixed. A conservative confidence interval corresponds to a point above the horizontal axis. For the method of (Imbens and Rubin, 2015, §19.6), we use a single match. Given an estimate* $\hat{\tau}$*, standard error estimate* $\hat{s}$*, and desired significance* α*, we construct the confidence region* $\hat{\tau} \pm \Phi^{-1}(1-\alpha/2)\hat{s}$*. For the bootstrap, we construct the confidence interval as the interval between the* $\alpha/2$ *and* $1-\alpha/2$ *quantiles of the bootstrap distribution over 1000 re-samples.*

All above confidence interval methods discard the conditional bias term in the error of $\hat{\tau}_W$. To quote Imbens and Rubin (2015), "with a sufficiently flexible estimator, this term will generally be small," meaning asymptotically insignificant compared to standard error. Indeed, in the above example, bias is small and estimated standard errors alone achieve approximately valid confidence intervals. However, in settings where overlap is limited, the bias may be in fact be significant. In the extreme case of no overlap (Asn. 2.2 does not hold), causal effects may be unidentifiable and bias may be unavoidable even asymptotically. In settings of low overlap, standard inverse propensity weighting approaches lead to very large weights and high variance and, in the extreme case of no overlap, they lead to infinite weights and provide no insights into average causal effects (unless we change the target of

13. Moreover, KOM can even be restricted to produce subset weights if such are desired. KOM, however, does not preserve the property of producing finitely-many coarsened strata of units as in such methods as full matching (Rosenbaum, 1991).
14. Marginal likelihood can also be used to estimate heteroskedastic noise (Goldberg et al., 1998).

Figure 6: Inference with KOM++



(a) KOM++ $\mathcal{K}_2^{\mathrm{P}}$ (b) KOM++ $\mathcal{K}_{5/2}^{\mathrm{M}}$

estimation as in Crump et al., 2006). Similarly, matching methods will fail to find good matches and a significant bias will remain.

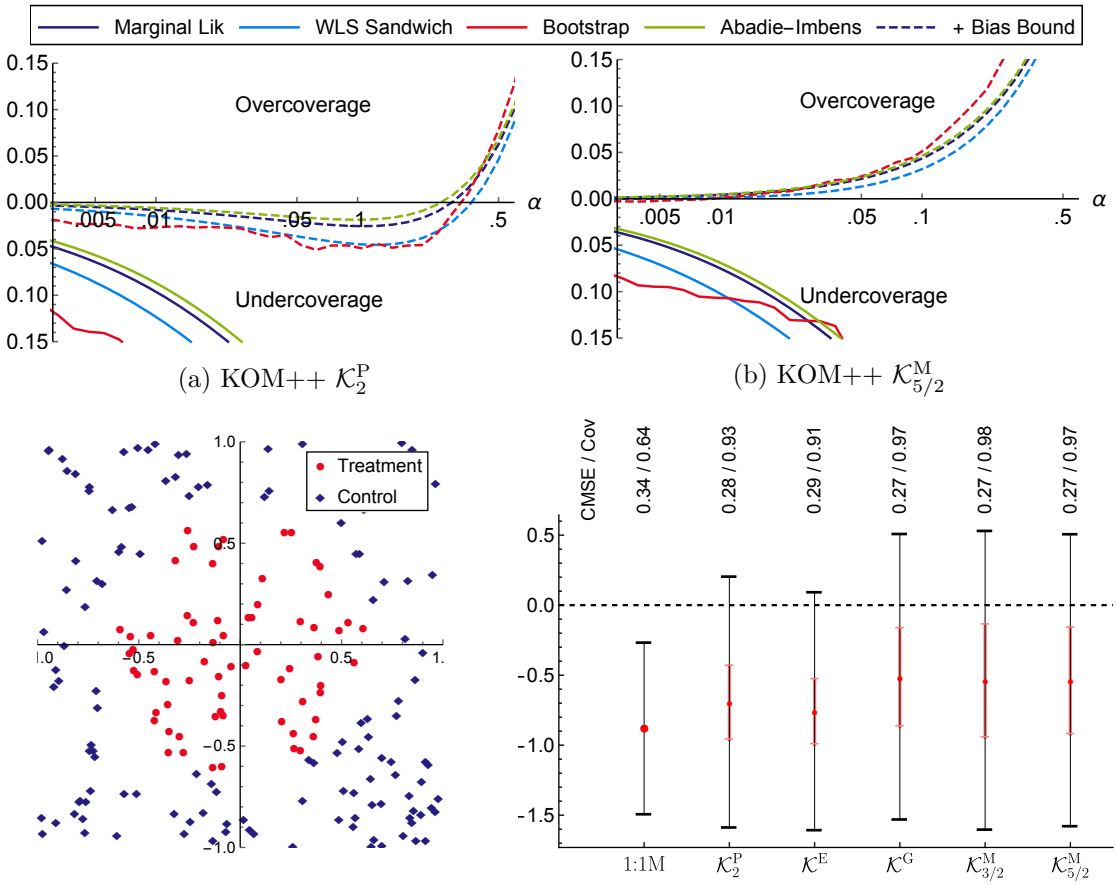
KOM opens the door to the possibility of partial identification of causal effects in the absence of overlap by bounding or approximately bounding the bias. With KOM, we can obtain an explicit bound on the bias: it is bounded by $\|[f_0]\|\mathfrak{B}(W;\mathcal{K})$. We may have an a priori bound on $\|[f_0]\| \leq \hat{\gamma}$. For example, using the Beppo-Levi kernel presented in the next section, this can take the form of an a priori bound on the roughness of f_0. Alternatively, in the case of KOM++, we may take a data-driven approach and rely on the marginal likelihood estimate $\hat{\gamma}$ of $\|[f_0]\|$. That is, in the absence of overlap, we assume we can still judge the complexity of f_0 on the treated population of X, if not its actual values, by observing its values with noise on the control population of X. In either case, assuming that $\|[f_0]\| < \infty$, we can obtain an *interval estimate* of the treatment effect:

$$\hat{\mathfrak{T}}_W = [\hat{\tau}_W - \hat{\gamma}\mathfrak{B}(W;\mathcal{K}),\ \hat{\tau}_W + \hat{\gamma}\mathfrak{B}(W;\mathcal{K})]. \tag{14}$$

This interval accounts for the possible bias precisely in terms of the covariate imbalances left in the re-weighted samples and in the extent to which f_0 could, in the worst case, depend on these imbalances and induce the worst bias. If this characterization of f_0 is valid (*i.e.*, $\|[f_0]\| \leq \hat{\gamma}$) then this interval contains $\mathrm{SATT} + \frac{1}{n_1}\sum_{i\in\mathcal{T}_1}\epsilon_i - \sum_{i\in\mathcal{T}_0} W_i\epsilon_i$, as a trivial consequence of Prop. 1. To the interval estimate in eq. (14), we can further add standard errors to account for the residual variance in ϵ_i and produce a robust confidence interval that provides coverage even in cases of limited overlap.

That KOM preserves the unit of analysis is critical. Using negative weights that do not necessarily sum to one, one can always make perfectly zero any imbalance metric, including that used by KOM – the result is essentially equivalent to regression (see Sec. 5.6). This, however, requires extreme extrapolation and the sense in which this is a "perfect match" is deceptive: in the absence of overlap and parametric specification, identification is simply impossible. Instead, KOM avoids extrapolation by preserving the unit of analysis and restricts only to a valid distribution over the controls (potentially, if so restricted, to a proper subset of the control sample). The imbalances between this distribution of controls and the empirical distribution of treated units are transparent and easily red off from the

Figure 7: Partial Identification with KOM++



(a) KOM++ $\mathcal{K}_2^{\mathrm{P}}$

(b) KOM++ $\mathcal{K}_{5/2}^{\mathrm{M}}$

(c) $X_{1:n}, T_{1:n}$

(d) 95% Confidence Intervals

result of the KOM optimization problem. This enables us to construct an honest and robust interval for the effect, without extrapolating beyond what we actually observe.

Example 6 *We repeat Ex. 5 but change the distribution of covariates to eliminate overlap completely. Instead of drawing T as given by $\mathbb{P}(T=1 \mid X)$ in Ex. 1, we fix $T_i = 1$ whenever this given propensity function is greater than 0.4 and otherwise $T_i = 0$. This yields the draw in Fig. 7c, which has $n_0 = 133, n_1 = 67$. We repeat Ex. 5 either as before (solid lines) or by instead adding the confidence terms to the endpoints of the interval estimate in eq. (14) using the $\hat{\gamma}$ estimate from the application of KOM++ (dashed lines) and plot the coverage in Figs. 7a and 7b. In Fig. 7d, we compare the point estimate given by 1:1M (red dot) for SATT (dashed line) to the interval estimate given by KOM++ (pink intervals), both surrounded by a confidence interval given by the standard error of (Imbens and Rubin, 2015, §19.6). The CMSE (above plot) of the KOM++ point estimates is not a substantial improvement over 1:1M – little can be done in the face of such an extreme lack of overlap – but the robust confidence intervals that arise from accounting for the non-vanishing bias help in achieving correct coverage (above plot) in the face of lack of overlap.*

Appendix C. Connections to and Generalization of Equal-Percent Bias Reduction

Equal-percent bias reduction (EPBR) (Rubin, 1976a,b) is a property of matching methods stipulating that, on average, the reduction in discrepancy in the mean vector be equal across all the the covariates relative to doing no matching at all. This is the case if and only if the matching reduces imbalance for any linear function of the covariates. In this section we discuss connections to KOM as well as non-linear generalizations of EPBR.

For the sake of exposition, we define EPBR somewhat differently and explain the connection below.

Definition 28 *For $\mathcal{X} \subseteq \mathbb{R}^d$, a matching method W is* linearly EPBR *relative to W', written $W \preceq_{\textit{lin-EPBR}} W'$, if $|\mathbb{E}[B(W; f)]| \leq |\mathbb{E}[B(W'; f)]|$ for any $f(x) = \beta^T x + \alpha$, $\beta \in \mathbb{R}^d$, $\alpha \in \mathbb{R}$.*

Usually, we compare relative to the no-matching weights $W_i^{(0)} = 1/n_0$. The definition of EPBR in Rubin (1976a) is equivalent to saying in our definition that W is comparable to $W^{(0)}$ in lin-EPBR, *i.e.*, either $W \preceq_{\text{lin-EPBR}} W^{(0)}$ or $W^{(0)} \preceq_{\text{lin-EPBR}} W$. This equivalence is proven in the following. The definition in Rubin (1976a) is stated in terms of the equivalent proportionality statement below, relative to $W^{(0)}$ and without magnitude restrictions on α.

Proposition 29 *Suppose $\mathcal{X} \subseteq \mathbb{R}^d$. Then, $W \preceq_{\textit{lin-EPBR}} W'$ if and only if $\mathbb{E}[\frac{1}{n_1}\sum_{i\in\mathcal{T}_1} X_i - \sum_{i\in\mathcal{T}_0} W_i X_i] = \alpha\mathbb{E}[\frac{1}{n_1}\sum_{i\in\mathcal{T}_1} X_i - \sum_{i\in\mathcal{T}_0} W_i' X_j]$ for some $\alpha \in [-1, 1]$ as vectors in $\mathbb{R}^d$.*

In Rubin and Thomas (1992), the authors show that in the special case of proportionally ellipsoidal distributions, affinely invariant matching methods are EPBR. We restate and reprove the result in terms of our definitions.

Definition 30 *For $\mathcal{X} \subseteq \mathbb{R}^d$, a matching method W is* affinely invariant *if $W(X_{1:n}, T_{1:n}) = W(X_{1:n}A^T + \mathbf{1}_n a^T, T_{1:n})$ for all non-singular $A \in \mathbb{R}^{d\times d}$ and $a \in \mathbb{R}^d$ (*i.e., *$x \mapsto Ax + a$ is applied to each data point separately).*

Definition 31 *Two random vectors Z and Z' are* proportionally ellipsoidal *if there exists PSD matrix Σ, $\alpha, \alpha' \in \mathbb{R}_+$, and a characteristic function $\phi : \mathbb{R} \to \mathbb{C}$ such that, for any v, $v^T Z$ and $v^T Z'$ have characteristic functions $e^{iv^T\mathbb{E}[Z]t}\phi(\alpha v^T\Sigma v t^2)$ and $e^{iv^T\mathbb{E}[Z']t}\phi(\alpha' v^T\Sigma v t^2)$, respectively.*

Proposition 32 *If $X \mid T = 0$, $X \mid T = 1$ are proportionally ellipsoidal and W is affinely invariant, then W and $W^{(0)}$ are lin-EPBR comparable.*

Methods based on the Mahalanobis distance are affinely invariant. KOM with a fitted scaling matrix is also affinely invariant as discussed in 4.5. All unitarily invariant matching methods can be made to be affinely invariant by preprocessing the data. This procedure is detailed in Alg. 3. KOM is unitarily invariant if the kernel is unitarily invariant, including all kernels studied in this paper. With the exception of CEM, all methods we have studied have been unitarily invariant.

Definition 33 *For $\mathcal{X} \subseteq \mathbb{R}^d$, a matching method W is* unitarily invariant *if $W(X_{1:n}, T_{1:n}) = W(X_{1:n}A^T, T_{1:n})$ for all unitary $A = A^{-T} \in \mathbb{R}^{d\times d}$. Equivalently, W is* unitarily invariant *if it only depends on $X_{1:n}X_{1:n}^T, T_{1:n}$.*

Proposition 34 *Suppose $\mathcal{X} \subseteq \mathbb{R}^d$. If W is unitarily invariant then Alg. 3 produces an affinely invariant weighting method.*

At the same time, affinely invariant methods only make sense if we have more datapoints than the dimension of the data, since the dimension of the data is precisely the dimension of the space of linear functions on the data.

Proposition 35 *Suppose $\mathcal{X} \subseteq \mathbb{R}^d$. If W is affinely invariant then W is constant over all data $X_{1:n}$ that is affinely independent.*[15]

In other words, if $n \leq d$, then any affinely invariant matching method does nothing useful at all because it is (generically) invariant to the data.

The above exposition establishes the connection of KOM to EPBR and the benefits it bestows. If we either fit a scaling matrix by marginal likelihood or studentize the data, then KOM is affinely invariant and hence EPBR for proportionally ellipsoidal data, *i.e.*, has uniform improvement over all linear outcomes. However, one of the main attractions of KOM is in dealing with *non-linear* outcomes. We next present a direct generalization of EPBR to non-linear outcomes, allowing us to characterize when matching methods can have performance guarantees over families of *non-linear* outcomes. We recreate the analogous lin-EPBR results for the non-linear version.

Definition 36 *A matching method W is* $\mathcal{F}$-EPBR *relative to W', written $W \preceq_{\mathcal{F}\text{-}EPBR} W'$, if $|\mathbb{E}\left[B(W; f)\right]| \leq |\mathbb{E}\left[B(W'; f)\right]|$ for every $f \in \mathcal{F}$.*

Proposition 37 *Let $\mathcal{F}$ be a linear subspace of the functions $\mathcal{X} \to \mathbb{R}$ under pointwise addition and scaling. Then, $W \preceq_{\mathcal{F}\text{-}EPBR} W'$ if and only if there exists $\alpha \in [-1, 1]$ such that, as operators on $\mathcal{F}$, $\mathbb{E}\left[B(W; \cdot)\right] = \alpha\mathbb{E}\left[B(W'; \cdot)\right]$.*

Definition 38 *Let $\mathcal{K}$ be a PSD kernel on $\mathcal{X}$ and let $\mathcal{F}$ be its RKHS. The $\mathcal{X}$-valued random variables Z and Z' are* proportionally $\mathcal{K}$-ellipsoidal *if there exist $\mu, \mu' \in \mathcal{F}$, PSD compact $C \in \mathcal{F}\otimes\mathcal{F}$, $\alpha, \alpha' \in \mathbb{R}_+$, and a characteristic function $\phi : \mathbb{R} \to \mathbb{C}$ such that, for every $x \in \mathcal{X}$, $K(Z, x)$ and $K(Z', x)$ are a real random variables distributed with characteristic functions $e^{i\mu(x)t}\phi(C(x, x)t^2)$, $e^{i\mu'(x)t}\phi(\alpha C(x, x)t^2)$ respectively.*

For example, proportionally ellipsoidal is equivalent to proportionally $\mathcal{K}$-ellipsoidal with $\mathcal{K}(x, x') = x^T x'$.

Definition 39 *Let $\mathcal{K}$ be a PSD kernel on $\mathcal{X}$ and let $\mathcal{F}$ be its RKHS. A matching method W is* $\mathcal{K}$-affinely invariant *if $\exists W' : \mathcal{F}^n \times \{0, 1\}^n \to \mathcal{W}$ such that for any bounded non-singular $A \in \mathcal{F} \otimes \mathcal{F}$ and $a \in \mathcal{F}$, we have $W(X_{1:n}, T_{1:n}) = W'(\{A\mathcal{K}(X_i, \cdot) + a\}_{i=1}^n, T_{1:n})$.*

15. $X_{1:n}$ are said to be affinely independent if $X_{2:n} - X_1$ are linearly independent.

ALGORITHM 3: Affine Invariance by Studentization

input: Data $X_{1:n}, T_{1:n}$, a matching method $W(X_{1:n}, T_{1:n})$.
Let $\hat{\mu} = \frac{1}{n_0}\sum_{i\in\mathcal{T}_0} X_i$, $\widehat{\Sigma} = \frac{1}{n_0-1}\sum_{i\in\mathcal{T}_0}(X_i-\hat{\mu})(X_i-\hat{\mu})^T$.
Eigen-decompose $\widehat{\Sigma} = U\,\mathrm{Diag}(\tau_1,\ldots,\tau_d)U^T$.
Set $\widehat{\Sigma}^{\dagger/2} = U\,\mathrm{Diag}(\mathbb{I}_{[\tau_1\neq 0]}\tau_1^{-1/2},\ldots,\mathbb{I}_{[\tau_d\neq 0]}\tau_d^{-1/2})U^T$.
output: $W((X_{1:n}-\hat{\mu})\widehat{\Sigma}^{\dagger/2}, T_{1:n})$.

ALGORITHM 4: $\mathcal{K}$-Affine Invariance by $\mathcal{K}$-Studentization

input: Data $X_{1:n}, T_{1:n}$, a PSD kernel $\mathcal{K}$, a $\mathcal{K}$-unitarily invariant $W(X,T)$.
Let $K_{ij} = \mathcal{K}(X_i, X_j)$, $E^{(0)}_{ij} = \mathbb{I}_{[j\in\mathcal{T}_0]}/n_0$.
Set $K^C = (I-E^{(0)})K(I-E^{(0)})^T$, $M_{ij} = \sum_{l\in\mathcal{T}_0} K^C_{il}K^C_{jl}$.
Compute the pseudo-inverse $M^\dagger$ and let $\overline{K} = K^C M^\dagger K^C$.
output: $W'(\overline{K}, T_{1:n})$.

Proposition 40 *Let $\mathcal{K}$ be a PSD kernel on $\mathcal{X}$ and let $\mathcal{F}$ be its RKHS. If $X \mid T=0$, $X \mid T=1$ are proportionally $\mathcal{K}$-ellipsoidal and W is affinely invariant then W is $\mathcal{F}$-EPBR relative to $W^{(0)}$.*

Definition 41 *Let $\mathcal{K}: \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a PSD kernel. W is* $\mathcal{K}$-unitarily invariant *if it depends on the data via its Gram matrix,* i.e., *$\exists W' : \mathcal{S}_+^{n\times n}\times\{0,1\}^n \to \mathcal{W}$ such that $W(X_{1:n}, T_{1:n}) = W'((\mathcal{K}(X_i,X_j))_{i,j=1}^n, T)$.*

For example, KOM is $\mathcal{K}$-unitarily invariant.

Proposition 42 *If W' is $\mathcal{K}$-unitarily invariant then Alg. 4 produces a $\mathcal{K}$-affinely invariant weighting method.*

However, there are limits to $\mathcal{K}$-affine invariance. The following shows that $\mathcal{K}$-affine invariance only makes sense for non-universal kernels since all C_0-universal kernels are strictly positive definite (has Gram matrix that is positive definite whenever all datapoints are distinct).

Proposition 43 *Suppose $\mathcal{K}$ is strictly positive definite and W is $\mathcal{K}$-affinely invariant. Then, W is constant over all $X_{1:n}$ that are distinct.*

Appendix D. Proofs

Proof [Proof of Prop. 1] Define

$$\Xi(W) = \tfrac{1}{n_1}\textstyle\sum_{i\in\mathcal{T}_1}\epsilon_i - \sum_{i\in\mathcal{T}_0} W_i\epsilon_i. \tag{15}$$

Rewrite $\mathrm{SATT} = \frac{1}{n_1}\sum_{i\in\mathcal{T}_1} Y_i - \frac{1}{n_1}\sum_{i\in\mathcal{T}_1} Y_i(0)$, it is clear SATT differs from $\hat{\tau}_W$ only in the second term so that, letting $W_i = 1/n_1$ for $i\in\mathcal{T}_1$,

$$\begin{aligned}\hat{\tau} - \mathrm{SATT} &= \tfrac{1}{n_1}\textstyle\sum_{i\in\mathcal{T}_1} Y_i(0) - \sum_{i\in\mathcal{T}_0} W_i Y_i(0) = B(W; f_0) + \Xi(W)\\ &= \textstyle\sum_{i=1}^n (-1)^{T_i+1} W_i f_0(X_i) + \sum_{i=1}^n (-1)^{T_i+1} W_i\epsilon_i,\end{aligned}$$

For each i, we have

$$\mathbb{E}[\epsilon_i \mid X_{1:n}, T_{1:n}] = \mathbb{E}[Y_i(0) \mid X_i, T_i] - f_0(X_i) = \mathbb{E}[Y_i(0) \mid X_i] - f_0(X_i) = 0,$$

where the first equality is by definition of ϵ_i and the second by Asn. 2.1. Since $W_i = W_i(X_{1:n}, T_{1:n})$, we have $\mathbb{E}[(-1)^{T_i+1} W_i \epsilon_i \mid X_{1:n}, T_{1:n}] = 0$. For each i, j,

$$\begin{aligned}
\mathbb{E}\left[(-1)^{T_i+T_j} W_i W_j \epsilon_i \epsilon_j \mid X_{1:n}, T_{1:n}\right] &= (-1)^{T_i+T_j} W_i W_j \mathbb{E}\left[\epsilon_i \epsilon_j \mid X_{1:n}, T_{1:n}\right] \\
&= W_i W_j \operatorname{Cov}(Y_i(0), Y_j(0)) = \mathbb{I}_{[i=j]} W_i^2 \sigma_i^2, \\
\mathbb{E}[(-1)^{T_i+T_j} W_i W_j \epsilon_i f_0(X_j) \mid X_{1:n}, T_{1:n}] &= 0,
\end{aligned}$$

completing the proof. ■

Proof [Proof of Thm. 3] Let D be the distance matrix $D_{ii'} = \delta(X_i, X_{i'})$. By the definition of the Lipschitz norm and linear optimization duality we get,

$$\begin{aligned}
\mathfrak{B}(W; \|\cdot\|_{\mathrm{Lip}(\delta)}) = \sup_{v_i - v_{i'} \le D_{ii'}} & \tfrac{1}{n_1} \textstyle\sum_{i \in \mathcal{T}_1} v_1 - \sum_{i \in \mathcal{T}_0} W_i v_i \\
= \min_{S \in \mathbb{R}_+^{n \times n}} & \textstyle\sum_{i,i'} D_{ii'} S_{ii'} \\
\text{s.t.} \quad & \textstyle\sum_{i'=1}^n (S_{ii'} - S_{i'i}) = 1/n_1 \quad \forall i \in \mathcal{T}_1 \\
& \textstyle\sum_{i'=1}^n (S_{ii'} - S_{i'i}) = -W_i \quad \forall i \in \mathcal{T}_0.
\end{aligned}$$

This describes a min-cost flow problem with sources $\mathcal{T}_1$ with inputs $1/n_1$, 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 $i, i_1, \ldots, i_m, i'$. By the triangle inequality, $D_{ii'} \le D_{ii_1} + D_{i_1 i_2} + \cdots + D_{i_m i'}$. 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

$$\begin{aligned}
\mathfrak{B}(W; \|\cdot\|_{\mathrm{Lip}(\delta)}) = \min_{S \in \mathbb{R}_+^{\mathcal{T}_1 \times \mathcal{T}_0}} & \textstyle\sum_{i \in \mathcal{T}_1, i' \in \mathcal{T}_0} D_{ii'} S_{ii'} \\
\text{s.t.} \quad & \textstyle\sum_{i' \in \mathcal{T}_0} S_{ii'} = 1/n_1 \quad \forall i \in \mathcal{T}_1 \\
& \textstyle\sum_{i' \in \mathcal{T}_1} S_{i'i} = W_i \quad \forall i \in \mathcal{T}_0.
\end{aligned}$$

For the case of NNM, using the transformation $W_i' = n_1 W_i$, we get

$$\begin{aligned}
\min_{W \in \mathcal{W}^{\mathrm{simplex}}} \mathfrak{B}(W; \|\cdot\|_{\mathrm{Lip}(\delta)}) = \tfrac{1}{n_1} \min_{S, W'} & \textstyle\sum_{i \in \mathcal{T}_1, i' \in \mathcal{T}_0} D_{ii'} S_{ii'} \\
\text{s.t.} \quad & S \in \mathbb{R}_+^{\mathcal{T}_1 \times \mathcal{T}_0},\ W' \in \mathbb{R}_+^{\mathcal{T}_0} \\
& \textstyle\sum_{i \in \mathcal{T}_0} W_i' = n_1 \\
& \textstyle\sum_{i' \in \mathcal{T}_0} S_{ii'} = 1 \qquad \forall i \in \mathcal{T}_1 \\
& \textstyle\sum_{i \in \mathcal{T}_1} S_{ii'} - W_{i'}' = 0 \qquad \forall i' \in \mathcal{T}_0.
\end{aligned}$$

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 is integer (see Ahuja et al., 1993). 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 NNM.

For the case of 1:1M, using the same transformation, we get

$$\begin{aligned}\min_{W\in\mathcal{W}^{1/n_1\text{-simplex}}} \mathfrak{B}(W;\|\cdot\|_{\mathrm{Lip}(\delta)}) = \tfrac{1}{n_1}\min_{S,W'} \quad & \textstyle\sum_{i\in\mathcal{T}_1,\, i'\in\mathcal{T}_0} D_{ii'}S_{ii'} \\ \text{s.t.} \quad & S\in\mathbb{R}_+^{\mathcal{T}_1\times\mathcal{T}_0},\ W'_i\in\mathbb{R}_+^{\mathcal{T}_0} \\ & \textstyle\sum_{i\in\mathcal{T}_0} W'_i = n_1 \\ & W'_i \le 1 \quad \forall i\in\mathcal{T}_0 \\ & \textstyle\sum_{i'\in\mathcal{T}_0} S_{ii'} = 1 \qquad \forall i\in\mathcal{T}_1 \\ & \textstyle\sum_{i\in\mathcal{T}_1} S_{ii'} - W'_i = 0 \quad \forall i'\in\mathcal{T}_0.\end{aligned}$$

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 is integer. Since the optimal $S_{ii'}$ is integer, 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 1:1M. ■

Proof [Proof of Cor. 6] Let W be given by $\mathrm{GOM}(\mathcal{W},\|\cdot\|,\lambda)$. Then $V^2(W;\sigma^2_{1:n}) \le \sigma^2(\|W\|_2+1/n_1) = \gamma^2\lambda(\|W\|_2+1/n_1)$ and $B^2(W;f_0)\le \inf_{g:B(W;g)=0\ \forall W\in\mathcal{W}} B^2(W;f_0+g) \le \gamma\mathfrak{B}(W;\|\cdot\|)$. Finally, apply Prop. 1. ■

Proof [Proof of Prop. 7] Let W be given by $\mathrm{GOM}(\mathcal{W}^{\text{subsets}},\|\cdot\|,\lambda)$ and let $n(\lambda)=\|W\|_0 = |\{i\in\mathcal{T}_0 : W^*_i>0\}|$. Then $W\in\mathcal{W}^{n(\lambda)\text{-subset}}$ so W is also given by $\mathrm{GOM}(\mathcal{W}^{n(\lambda)\text{-subset}},\|\cdot\|,\lambda)$. However, $\|W'\|_2^2 = 1/n(\lambda)$ for all $W'\in\mathcal{W}^{n(\lambda)\text{-subset}}$, so the variance term is constant among this space of weights. Therefore, W is given by $\mathrm{GOM}(\mathcal{W}^{n(\lambda)\text{-subset}},\|\cdot\|,0)$. ■

Proof [Proof of Prop. 8] Note $\mathrm{argmin}_{W\in\mathcal{W}^{n(\lambda)\text{-multisubset}}}\|W\|_2^2 = \mathcal{W}^{n(\lambda)\text{-subset}}$. So by definition of GOM for $\lambda=\infty$ we get the equivalence between the first and the second. The equivalence between the second and third was argued in the proof of Prop. 7. ■

Proof [Proof of Prop. 9] This follows because $\|W\|_2^2$ is convex, $\mathfrak{B}(W;\|\cdot\|)$ is nonnegative and is the supremum of affine functions in W (*i.e.*, $B(\lambda W+(1-\lambda)W';f) = \lambda B(W;f) + (1-\lambda)B(W';f)$) and hence convex, and the square is convex and nondecreasing on the nonnegative line. ■

Proof [Proof of Prop. 10] Define $F(W) = \sup_{\|f\|\leq 1} \sum_{i=1}^n (2T_i - 1)W_i f(X_i)$ for $W \in \mathbb{R}^n$, and rewrite problem (6) redundantly as

$$\begin{aligned}
\min\ & r + \lambda s \\
\text{s.t. } & W \in \mathbb{R}^n,\, r \in \mathbb{R}, s \in \mathbb{R}, t \in \mathbb{R} && (16)\\
& W_{\mathcal{T}_0} \in \mathcal{W}^{\text{simplex}} && (17)\\
& W_{\mathcal{T}_1} = e_{n_1}/n_1 && (18)\\
& \|W\|_2^2 \leq s && (19)\\
& t^2 \leq r && (20)\\
& F(W) \leq t,\, \|W\|_2 \leq 1 && (21)
\end{aligned}$$

Since each of (16)-(21) are convex, by (Grotschel et al., 1993, Thm. 4.2.2) we reduce the ϵ-optimization problem to ϵ-separation on each of (16)-(21), which is immediately trivial for (16)-(20) in polynomial time in n. By (Grotschel et al., 1993, Thms. 4.2.5, 4.2.7) we reduce ϵ-separation on (21) to ϵ-violation, which by binary search reduces ϵ-violation on $S_t = \left\{W \in \mathbb{R}^{TC} : F(W) \leq v, \|W\|_2 \leq 1\right\}$ for fixed $t \geq 0$. If $t = 0$, then $S_t = \{0\}$ and we are done. Otherwise, letting $E_i : f \mapsto f(X_i)$ and $M_i = \|[E_i]\|_* < \infty$, note that $F(W)$ is continuous since

$$\begin{aligned}
F(W) - F(W') &= \sup_{\|f\|\leq 1} B(W_{\mathcal{T}_0}; f) - \sup_{\|f\|\leq 1} B(W'_{\mathcal{T}_0}; f)\\
&\leq \sup_{\|f\|\leq 1} B(W_{\mathcal{T}_0} - W'_{\mathcal{T}_0}; f)\\
&\leq \sup_{\|f\|\leq 1} \|W_{\mathcal{T}_0} - W'_{\mathcal{T}_0}\|_2 (\textstyle\sum_{i=1}^n f(X_i))^{1/2}\\
&\leq \|W_{\mathcal{T}_0} - W'_{\mathcal{T}_0}\|_2 \|M\|_2
\end{aligned}$$

Therefore, since $F(0) = 0$, we have $\{\|W\|_2 \leq t/\|M\|_2\} \subset S_t \subset \{\|W\|_2 \leq 1\}$. Since we can check membership by evaluating $\|W\|_2$ and $F(w)$, by (Grotschel et al., 1993, Thm. 4.3.2) we have an ϵ-violation algorithm. ■

For the next few proofs we use the following lemma.

Lemma 44 *For random variables $Z_n \geq 0$ and any sub-sigma algebra $\mathcal{G}$, $\mathbb{E}[Z_n \mid \mathcal{G}] = O_p(1) \implies Z_n = O_p(1)$ and $\mathbb{E}[Z_n \mid \mathcal{G}] = o_p(1) \implies Z_n = o_p(1)$.*

Proof [Proof of Lemma 44] Suppose $\mathbb{E}[Z_n \mid \mathcal{G}] = O_p(1)$. Let $\nu > 0$ be given. Then $\mathbb{E}[Z_n \mid \mathcal{G}] = O_p(1)$ says that there exist N, M such that $\mathbb{P}(\mathbb{E}[Z_n \mid \mathcal{G}] > M) \leq \nu/2$ for all $n \geq N$. Let $M_0 = \max\{M, 2/\nu\}$. Then, for all $n \geq N$,

$$\begin{aligned}
\mathbb{P}(Z_n > M_0^2) &= \mathbb{P}(Z_n > M_0^2, \mathbb{E}[Z_n \mid \mathcal{G}] > M_0) + \mathbb{P}(Z_n > M_0^2, \mathbb{E}[Z_n \mid \mathcal{G}] \leq M_0)\\
&= \mathbb{P}(Z_n > M_0^2, \mathbb{E}[Z_n \mid \mathcal{G}] > M_0) + \mathbb{E}[\mathbb{P}(Z_n > M_0^2 \mid \mathcal{G}) \mathbb{I}_{[\mathbb{E}[Z_n|\mathcal{G}] \leq M_0]}]\\
&\leq \nu/2 + \mathbb{E}\left[\frac{\mathbb{E}[Z_n \mid \mathcal{G}]}{M_0^2} \mathbb{I}_{[\mathbb{E}[Z_n|\mathcal{G}] \leq M_0]}\right] \leq \nu/2 + 1/M_0 \leq \nu.
\end{aligned}$$

Now suppose $\mathbb{E}[Z_n \mid \mathcal{G}] = o_p(1)$. Let $\eta > 0, \nu > 0$ be given. Let N be such that $\mathbb{P}(\mathbb{E}[Z_n \mid \mathcal{G}] > \nu\eta/2) \leq \nu/2$. Then for all $n \geq N$:

$$\begin{aligned}\mathbb{P}(Z_n > \eta) &= \mathbb{P}(Z_n > \eta, \mathbb{E}[Z_n \mid \mathcal{G}] > \eta\nu/2) + \mathbb{P}(Z_n > \eta, \mathbb{E}[Z_n \mid \mathcal{G}] \leq \eta\nu/2) \\ &= \mathbb{P}(Z_n > \eta, \mathbb{E}[Z_n \mid \mathcal{G}] > \eta\nu/2) + \mathbb{E}[\mathbb{P}(Z_n > \eta \mid \mathcal{G})\mathbb{I}_{[\mathbb{E}[Z_n|\mathcal{G}] \leq \eta\nu/2]}] \\ &\leq \nu/2 + \mathbb{E}\left[\frac{\mathbb{E}[Z_n \mid \mathcal{G}]}{\eta}\mathbb{I}_{[\mathbb{E}[Z_n|\mathcal{G}] \leq \eta\nu/2]}\right] \leq \nu/2 + \nu/2 \leq \nu.\end{aligned}$$

■

Proof [Proof of Thm. 12] Let $p(x) = \mathbb{P}(T = 1 \mid X = x)$, $p_0 = \mathbb{P}(T = 0)$, $p_1 = \mathbb{P}(T = 1)$. By Asn. 2.2, 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 1 if $T_i = 1$ and otherwise Bernoulli with probability $q(X_i)$ (where the draw for i is fixed over n). Then we have that $X_i \mid T_i = 0, \tilde{W}'_i = 1$ is distributed identically as $X_i \mid T_i = 1$. Let $n'_0 = \sum_{i \in \mathcal{T}_0} \tilde{W}'_i$ and for each $i \in \mathcal{T}_0$ set $\tilde{W}_i = \tilde{W}'_i/n'_0$. For $i \in \mathcal{T}_1$, set $\tilde{W}_i = 1/n_1$. Let $X_i^{(1)}$ be new, independent replicates distributed as $X_i^{(1)} \sim (X \mid T = 1)$. Let $\xi_i(f) = \tilde{W}'_i(f(X_i) - f(X_i^{(1)}))$ and $\tilde{\xi}_i(f) = \tilde{W}'_i(f(X_i) - \mathbb{E}f(X_i^{(1)}))$. By construction of $\tilde{W}'_i$ and ξ, we see that $\mathbb{E}[\xi_i(f) \mid T_{1:n}] = 0$ for every f. And, by (vi), $\mathbb{E}[\|\xi\|_*^2 \mid T_{1:n}] \leq M < \infty$. Also, each ξ_i is independent. Let $A_t = \frac{1}{n_t}\sum_{i \in \mathcal{T}_t} \tilde{\xi}_i$. Then, by adding and subtracting $\mathbb{E}f(X_i^{(1)})$, we get

$$\begin{aligned}\mathfrak{B}(\tilde{W}; \|\cdot\|) &= \sup_{\|f\| \leq 1}(A_1(f) + \tfrac{n_0}{n'_0}A_0(f)) \\ &\leq \sup_{\|f\| \leq 1} A_1(f) + \tfrac{n_0}{n'_0}\sup_{\|f\| \leq 1} A_0(f) = \|A_1\|_* + \tfrac{n_0}{n'_0}\|A_0\|_*.\end{aligned}$$

Next, by Jensen's inequality, for each $t = 0, 1$, we have that

$$\mathbb{E}[\|A_t\|_*^2 \mid T_{1:n}] \leq \mathbb{E}[(\sup_{\|f\| \leq 1} \tfrac{1}{n_t}\textstyle\sum_{i \in \mathcal{T}_t} \xi_i(f))^2 \mid T_{1:n}].$$

Let ξ'_i be an identical and independent replicate of ξ_i, conditioned on T_i. Let ρ_i be iid Rademacher random variables independent of all else. Then, again by Jensen's inequality,

$$\begin{aligned}\mathbb{E}[\|A_t\|_*^2 \mid T_{1:n}] &\leq \mathbb{E}[(\sup_{\|f\| \leq 1} \tfrac{1}{n_t}\textstyle\sum_{i \in \mathcal{T}_t}(\xi_i(f) - \mathbb{E}[\xi'_i(f) \mid T_{1:n}]))^2 \mid T_{1:n}] \\ &\leq \mathbb{E}[(\sup_{\|f\| \leq 1} \tfrac{1}{n_t}\textstyle\sum_{i \in \mathcal{T}_t}(\xi_i(f) - \xi'_i(f)))^2 \mid T_{1:n}] \\ &= \mathbb{E}[(\sup_{\|f\| \leq 1} \tfrac{1}{n_t}\textstyle\sum_{i \in \mathcal{T}_t} \rho_i(\xi_i(f) - \xi'_i(f)))^2 \mid T_{1:n}] \\ &\leq 4\mathbb{E}[(\sup_{\|f\| \leq 1} \tfrac{1}{n_t}\textstyle\sum_{i \in \mathcal{T}_t} \rho_i\xi_i(f))^2 \mid T_{1:n}] \\ &= 4\mathbb{E}[\|\tfrac{1}{n_t}\textstyle\sum_{i \in \mathcal{T}_t} \rho_i\xi_i\|_*^2 \mid T_{1:n}]. \qquad (22)\end{aligned}$$

The B-convexity of $\mathcal{F}$ implies the B-convexity of $\mathcal{F}^*$ (Giesy, 1966; Maurey, 2003). Next, the B-convexity of $\mathcal{F}^*$ implies that it has a non-trivial Rademacher type $1 < p \leq 2$ (Pisier, 1973; Maurey, 2003). Consequently, by the definition of Rademacher type, there is some $C > 0$ such that

$$\begin{aligned}\mathbb{E}[\|\tfrac{1}{n_t}\textstyle\sum_{i \in \mathcal{T}_t} \rho_i\xi_i(f)\|_*^2 \mid T_{1:n}] &\leq \tfrac{C}{n_t^2}\mathbb{E}[(\textstyle\sum_{i \in \mathcal{T}_t} \|\xi_i\|_*^p)^{2/p} \mid T_{1:n}] \\ &\leq \tfrac{C}{n_t^2}\mathbb{E}[n_t^{2/p-1}\textstyle\sum_{i \in \mathcal{T}_t} \|\xi_i\|_*^2 \mid T_{1:n}] = \tfrac{MC}{n_t^{2-2/p}}.\end{aligned}$$

Hence, by iterated expectations and Markov's inequality, $\|A_t\|_* = O_p(n^{1/p-1})$. Moreover, $(n_0/n_0') \to \alpha/p_1 < \infty$ in probability as well as $\|\tilde{W}_{\mathcal{T}_0}\|_2^2 = 1/n_0' = O_p(1/n)$. Since $\tilde{W}_{\mathcal{T}_0}$ is feasible by (iv) and W is optimal,

$$\mathfrak{E}^2(W;\|\cdot\|,\lambda_n) \le \mathfrak{E}^2(\tilde{W}_{\mathcal{T}_0};\|\cdot\|,\lambda_n) \le \mathfrak{B}^2(\tilde{W}_{\mathcal{T}_0};\|\cdot\|) + \overline{\lambda}\|\tilde{W}_{\mathcal{T}_0}\|_2^2 = O_p(n^{1/p-1}).$$

By (viii), $\exists\gamma : \|[f_0]\| \le \gamma < \infty$. By (vii), $\exists\sigma^2 : \sigma_i^2 \le \sigma^2$. By Prop. 1,

$$\mathrm{CMSE}(\hat{\tau}_W) \le (\gamma^2 + \sigma^2/\underline{\lambda})\mathfrak{E}^2(W;\|\cdot\|,\lambda_n) + \sigma^2/n_1 = O_p(n^{1/p-1}).$$

Then, by Lemma 44, $\hat{\tau}_W - \mathrm{SATT} = O_p(n^{1/p-1}) = o_p(1)$. ■

Proof [Proof of Thm. 13]Recalling $\Xi(W)$ from eq. (15), we can write

$$\hat{\tau}_{W,\hat{f}_0} - \mathrm{SATT} = B(W;\tilde{f}_0 - \hat{f}_0) + B(W;f_0 - \tilde{f}_0) + \Xi(W),$$

From the proof of Thm. 12, we have $\mathfrak{B}(W;\|\cdot\|) = o_p(1)$, $\|W\|_2 = o_p(1)$, $\Xi^2(W) = o_p(1)$. Moreover,

$$\begin{aligned}|B(W;\tilde{f}_0 - \hat{f}_0)| &\le (\|W\|_2^2 + 1/n_1)^{1/2}(\textstyle\sum_{i=1}^n(\tilde{f}_0(X_i) - \hat{f}_0(X_i))^2)^{1/2} \\ &= o_p(1)O_p(1) = o_p(1).\end{aligned}$$

In case (a), $B(W;f_0 - \tilde{f}_0) = 0$ yields the result. In case (b), $\mathbb{E}|B(W;f_0 - \tilde{f}_0)| \le (\|f_0\| + \|\tilde{f}_0\|)\mathbb{E}\mathfrak{B}(W;\mathcal{K}) = O(n^{-1/2})$, yielding the result. In case (c), we write $\hat{\tau}_{W,\hat{f}_0} - \mathrm{SATT} = B(W;f_0 - \hat{f}_0) + \Xi(W)$ and $|B(W;f_0 - \hat{f}_0)| \le (\|[f_0]\| + \|[\hat{f}_0]\|)\mathfrak{B}(W;\mathcal{K}) = O_p(1)o_p(1) = o_p(1)$. ■

Proof [Proof of Prop. 15] Writing $W_i' = T_i/n_1 - (1-T_i)W_i$, by the representer property of $\mathcal{K}$ and by self-duality of Hilbert spaces,

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

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

Proof [Proof of Prop. 16] By Prop. 1, we have

$$\mathbb{E}\left[(\hat{\tau}_W - \mathrm{SATT})^2 \mid X,T,f_0\right] = B^2(W;f_0) + \sigma^2\|W\|_2^2 + \tfrac{\sigma^2}{n_1}.$$

Marginalizing over f_0 and writing $W_i' = T_i/n_1 + (1-T_i)W_i$, we get

$$\begin{aligned}\mathrm{CMSE}(\hat{\tau}_W) &= \textstyle\sum_{i,j=1}^n(-1)^{T_i+T_j}W_i'W_j'\mathbb{E}\left[f_0(X_i)f_0(X_j) \mid X_{1:n},T_{1:n}\right] + \sigma\left\|W'\right\|_2^2 \\ &= \textstyle\sum_{i,j=1}^n(-1)^{T_i+T_j}W_i'W_j'\gamma^2\mathcal{K}(X_i,X_j) + \sigma^2\left\|W'\right\|_2^2 \\ &= \gamma^2\mathfrak{B}(W;\mathcal{K}) + \sigma^2\|W'\|_2^2 = \gamma^2(\mathfrak{E}(W;\mathcal{K},\lambda) + \lambda/n_1),\end{aligned}$$

where the second equality is by the of Gaussian process prior and the third equality is due to Prop. 15. ■

Proof [Proof of Thm. 18] Repeat the proof of Thm. 12 up to eq. (22). By (iv), $M = \mathbb{E}[\|\xi_i\|_*^2 \mid T] \leq 4\mathbb{E}[\mathcal{K}(X,X) \mid T=1] < \infty$. Note that $\|\xi-\zeta\|_*^2 + \|\xi+\zeta\|_*^2 = 2\|\xi\|_*^2 + 2\|\zeta\|_*^2 + 2\langle\xi,\zeta\rangle - 2\langle\xi,\zeta\rangle = 2\|\xi\|_*^2 + 2\|\zeta\|_*^2$. By induction, $\sum_{\rho_i\in\{-1,+1\}^m} \|\sum_{i=1}^m \rho_i\xi_i\|_*^2 = 2^m \sum_{i=1}^m \|\xi_i\|_*^2$ (parallelogram law). Therefore, continuing from eq. (22), we have

$$\mathbb{E}[\|\tfrac{1}{n_t}\sum_{i\in\mathcal{T}_t}\rho_i\xi_i(f)\|_*^2 \mid T_{1:n}] = \tfrac{C}{n_t^2}\mathbb{E}[\sum_{i\in\mathcal{T}_t}\|\xi_i\|_*^2 \mid T_{1:n}] = \tfrac{MC}{n_t}.$$

We conclude that,

$$\mathfrak{E}^2(\tilde{W}_{\mathcal{T}_0};\mathcal{K},\lambda_n) = \|A_1 + (n_0/n_0')A_0\|_*^2 + \lambda_n\|\tilde{W}_{\mathcal{T}_0}\|_2^2 = O_p(1/n).$$

Because W is optimal and by (iii) $\tilde{W}_{\mathcal{T}_0}$ is feasible and $\lambda^{-1} \leq \underline{\lambda}^{-1}$, we then also have $\mathfrak{B}^2(W;\mathcal{K}) = O_p(1/n)$ and $\|W\|_2^2 = O_p(1/n)$. By (v), $\exists\sigma^2 : \sigma_i^2 \leq \sigma^2$. Therefore, $V^2(W;\sigma_{1:n}^2) \leq \sigma^2(\|W\|_2^2 + 1/n_1) = O_p(1/n)$ and hence by Lemma 44 $\Xi(W) = O_p(n^{-1/2})$.

In case (a), we have $B(W;f_0) \leq \|[f_0]\|\mathfrak{B}(W;\mathcal{K}) = O(n^{-1})$. Prop. 1 and Lemma 44 yield the result.

Now consider case (b). Fix $\eta > 0, \nu > 0$. Let $\tau = \sqrt{\nu\eta/3}/M$. Because $\mathfrak{B}(W;\mathcal{K}) = O_p(n^{-1/2}) = o_p(n^{-1/4})$ and $\|W\|_2 = O_p(n^{-1/2})$, there are M, N such that for all $n \geq N$ both $\mathbb{P}(n^{1/4}\mathfrak{B}(W;\mathcal{K}) > \sqrt{\eta}) \leq \nu/3$ and $\mathbb{P}(n^{1/2}(\|W\|_2 + 1/\sqrt{n_1}) > M\sqrt{\eta}) \leq \nu/3$. By existence of second moment, there is $g_0' = \sum_{i=1}^\ell \beta_i I_{S_i}$ with $(\mathbb{E}\left[(f_0(X) - g_0'(X))^2\right])^{1/2} \leq \tau/2$ where $I_S(x)$ are the simple functions $I_S(x) = \mathbb{I}[x \in S]$ for S measurable. Let $i = 1,\ldots,\ell$. Let $U_i \supset S_i$ open and $E_i \subseteq S_i$ compact be such that $\mathbb{P}(X \in U_i \backslash E_i) \leq \tau^2/(4\ell|\beta_i|)^2$. By Urysohn's lemma (Royden, 1988), there exists a continuous function h_i with support $C_i \subseteq U_i$ compact, $0 \leq h_i \leq 1$, and $h_i(x) = 1\,\forall x \in E_i$. Therefore, $(\mathbb{E}\left[(I_{S_i}(X) - h_i)^2\right])^{1/2} = (\mathbb{E}\left[(I_{S_i}(X) - h_i)^2\mathbb{I}[X \in U_i\backslash E_i]\right])^{1/2} \leq (\mathbb{P}(X \in U_i\backslash E_i))^{1/2} \leq \tau/(4\ell|\beta_i|)$. By C_0-universality, $\exists g_i = \sum_{j=1}^m \alpha_j\mathcal{K}(x_j,\cdot)$ such that $\sup_{x\in\mathcal{X}}|h_i(x) - g_i(x)| < \tau/(4\ell|\beta_i|)$. Because $\mathbb{E}\left[(h_i - g_i)^2\right] \leq \sup_{x\in\mathcal{X}}|h_i(x) - g_i(x)|^2$, we have $\sqrt{\mathbb{E}[(I_{S'}(X) - g_i)^2]} \leq \tau/(2\ell|\beta_i|)$. Let $g_0 = \sum_{i=1}^\ell \beta_i g_i$. Then $(\mathbb{E}\left[(f_0(X) - g_0(X))^2\right])^{1/2} \leq \tau/2 + \sum_{i=1}^\ell |\beta_i|\,\tau/(2\ell|\beta_i|) = \tau$ and $\|g_0\| < \infty$. Let $\delta_n = \sqrt{\frac{1}{n}\sum_{i=1}^n (f_0(X_i) - g_0(X_i))^2}$. Then, we have that

$$\begin{aligned}|B(W;f_0)| &\leq |B(W;g_0)| + |B(W;f_0 - g_0)| \\ &\leq \|[g_0]\|\,\mathfrak{B}(W;\mathcal{K}) + \sqrt{n}(\|W\|_2 + 1/\sqrt{n_1})\delta_n.\end{aligned}$$

Let $N' = \max\{N, 2\lceil\|[f_0]\|^4/\eta^2\rceil\}$. Then by union bound, for all $n \geq N'$, we have

$$\begin{aligned}\mathbb{P}(|B(W;f_0)| > \eta) &\leq \mathbb{P}(n^{-1/4}\|[g_0]\| > \sqrt{\eta}) + \mathbb{P}(n^{1/4}\mathfrak{B}(W;\mathcal{K}) > \sqrt{\eta}) \\ &\quad + \mathbb{P}(n^{1/2}(\|W\|_2 + 1/\sqrt{n_1}) > M\sqrt{\eta}) + \mathbb{P}(\delta_n > \sqrt{\eta}/M) \\ &\leq 0 + \nu/3 + \nu/3 + \nu/3 = \nu.\end{aligned}$$

Hence, $|B(W;f_0)| = o_p(1)$. Prop. 1 and Lemma 44 yield the result. ■

Proof [Proof of Thm. 19] From the proof of Thm. 18 we have $\mathfrak{B}^2(W;\mathcal{K}) = O_p(1/n), \|W\|_2^2 = O_p(1/n), \Xi(W) = O_p(n^{-1/2})$. In case (a), we have $|B(W; f_0 - \hat{f}_0)| \leq \|[f_0 - \hat{f}_0]\|\mathfrak{B}(W;\mathcal{K}) = o_p(n^{-1/2})$ and so $\hat{\tau}_{W,\hat{f}_0} - \text{SATT} = \Xi(W) + o_p(n^{-1/2})$, yielding the result. In case (b), we have $|B(W; f_0 - \hat{f}_0)| \leq (\|[f_0]\| + \|[\hat{f}_0]\|)\mathfrak{B}(W;\mathcal{K}) = O_p(1)O_p(n^{-1/2})$, yielding the result.

In the other cases, expand the error of $\hat{\tau}_{W,\hat{f}_0}$:

$$\begin{aligned}\hat{\tau}_{W,\hat{f}_0} - \text{SATT} &= B(W; \tilde{f}_0 - \hat{f}_0) + B(W; f_0 - \tilde{f}_0) + \Xi(W).\\ |B(W; \tilde{f}_0 - \hat{f}_0)| &\leq (\|W\|_2^2 + 1/n_1)^{1/2}(\textstyle\sum_{i=1}^n (\tilde{f}_0(X_i) - \hat{f}_0(X_i))^2)^{1/2}\\ &= O_p(n^{-1/2})O_p(1).\end{aligned}$$

In case (c), $B(W; f_0 - \tilde{f}_0) = 0$ yields the result. In case (d), we repeat the argument in the proof of Thm. 18 to show that both $|B(W; f_0)| \to 0$ and $|B(W; \tilde{f}_0)| \to 0$, yielding the result. In case (e), $|B(W; f_0 - \tilde{f}_0)| \leq (\|f_0\| + \|\tilde{f}_0\|)\mathfrak{B}(W;\mathcal{K}) = O_p(n^{-1/2})$, yielding the result. ■

Proof [Proof of Thm. 20] From the proof of Thm. 18, $\mathfrak{B}^2(W;\mathcal{K}) = O(1/n)$, $\|W\|_2^2 = O(1/n)$. From the proof of Prop. 27, for $\hat{f}_0(x) = \hat{\alpha} - \hat{\beta}^T x$ where $\hat{\alpha}, \hat{\beta} = \text{argmin}_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^d} \sum_{i\in\mathcal{T}_0} W_i(Y_i - \alpha - \beta^T X_i)^2$, we have $\hat{\tau}_{\text{WLS}(W)} = \frac{1}{n_1}\sum_{i\in\mathcal{T}_1}(Y_i - \hat{f}_0(X_i))$. Because least squares with intercept has zero in-sample bias, $\sum_{i\in\mathcal{T}_0} W_i \hat{f}_0(X_i) = \sum_{i\in\mathcal{T}_0} W_i Y_i$, so that by adding and substracting this term, we see that $\hat{\tau}_{\text{WLS}(W)} = \hat{\tau}_{W,\hat{f}_0} = \hat{\tau}_W - B(W; \hat{f}_0)$.

Let $\tilde{X}_i = (1, X_i)$, $\tilde{\beta} = (\hat{\alpha}, \hat{\beta}^T)^T$, $\hat{P} = \tilde{X}_{\mathcal{T}_0}^T W \tilde{X}_{\mathcal{T}_0} = \sum_{i\in\mathcal{T}_0} W_i \tilde{X}_i \tilde{X}_i^T$, $\hat{G} = \sum_{i\in\mathcal{T}_0} W_i f_0(X_i)\tilde{X}_i$, and $\hat{H} = \sum_{i\in\mathcal{T}_0} W_i \tilde{X}_i \epsilon_i$. Then $\tilde{\beta} = \hat{P}^{-1}(\hat{G} + \hat{H})$. Follow the argument in Thm. 18 for the case of $\mathcal{K}$ being C_0-universal to show $\hat{P} \to P$ in probability, where $P = \mathbb{E}[\tilde{X}\tilde{X}^T \mid T = 1]$. By the Schor complement, since $\mathbb{E}[XX^T \mid T = 1]$ is non-singular, P is also non-singular and hence we have $\hat{P}^{-1} \to P^{-1}$ in probability. Follow the argument in Thm. 18 for the case of $\mathcal{K}$ being C_0-universal to show $\hat{G} \to G = \mathbb{E}[f_0(X)\tilde{X} \mid T = 1]$ in probability. Moreover, letting $M > 1$ be such that $\sup_{x\in\mathcal{X}} \|x\|_2 \leq M$ by assumption, $\mathbb{E}[\|\hat{H}\|_2^2 \mid X_{1:n}, T_{1:n}] = \sum_{i\in\mathcal{T}_0} W_i^2\sigma_i^2\|\tilde{X}_i\|_2^2 \leq 2M^2\sigma^2\|W\|_2^2 = O_p(1/n)$ so that by Lemma 44 $\|\hat{H}\|_2^2 = O_p(1/n)$.

Consider case (a). By Thm. 18, $\hat{\tau}_W \to 0$ in probability. By the above, we have $\|\tilde{\beta}\|_2 = O_p(1)$. Let $\eta > 0, \rho > 0$ be given. Then there is $R > 0$ such that $\mathbb{P}(\|\tilde{\beta}\|_2 > R) \leq \rho/3$. Let M', N' be such that $\mathbb{P}(\sqrt{n}(\|W\|_2^2 + 1/n_1)^{1/2} > M') \leq \rho/3$ for all $n \geq N'$. Let $r = \eta/(8MM')$ and $\{\overline{\beta} : \|\overline{\beta} - \tilde{\beta}_1\|_2 \leq r\}, \ldots, \{\overline{\beta} : \|\overline{\beta} - \tilde{\beta}_\ell\|_2 \leq r\}$ be a finite cover of the compact $\{\overline{\beta} : \|\overline{\beta}\|_2 \leq R\}$. Let $f_{\overline{\beta}}(x) = \overline{\beta}^T(1, x)$. By C_0-universality and boundedness, $\exists g_k$ with $\|g_k\| < \infty$ and $\sup_{x\in\mathcal{X}} |g_k(x) - f_{\overline{\beta}}(x)| \leq \eta/(4M')$. Let $\Gamma = \max_{k=1,\ldots,\ell} \|g_k\|$ and let $N'' \geq N'$ be such that $\mathbb{P}(\mathfrak{B}(W;\mathcal{K}) > \eta/(2\Gamma)) \leq \rho/3$ for all $n \geq N''$. Note that $\sup_{x\in\mathcal{X}} |f_{\overline{\beta}}(x) - f_{\overline{\beta}'}(x)| \leq 2M\|\overline{\beta} - \overline{\beta}'\|_2$. Then we have that

$$\begin{aligned}&\sup_{\|\overline{\beta}\|\leq R} |B(W; f_{\overline{\beta}})|\\ &\quad\leq \sup_{\|\overline{\beta}\|\leq R} \min_k (|B(W; g_k)| + |B(W; f_{\overline{\beta}_k} - g_k)| + |B(W; f_{\overline{\beta}} - f_{\overline{\beta}_k})|)\\ &\quad\leq \Gamma\mathfrak{B}(W;\mathcal{K}) + \sqrt{n}(\|W\|_2^2 + 1/n_1)^{1/2}(\kappa + 2Mr)\\ &\quad= \Gamma\mathfrak{B}(W;\mathcal{K}) + \sqrt{n}(\|W\|_2^2 + 1/n_1)^{1/2}\eta/(2M')\end{aligned}$$

Finally, for all $n \geq N''$, we have

$$\begin{aligned}
&\mathbb{P}(|B(W;\hat{f}_0)| > \eta) \leq \mathbb{P}(\|\tilde{\beta}\| > R) + \mathbb{P}(|B(W;\hat{f}_0)| > \eta, \|\tilde{\beta}\| \leq R) \\
&\leq \rho/3 + \mathbb{P}(\sup_{\|\overline{\beta}\| \leq R} |B(W; f_{\overline{\beta}})| > \eta) \\
&\leq \rho/3 + \mathbb{P}(\mathfrak{B}(W;\mathcal{K}) > \eta/(2\Gamma)) + \mathbb{P}(\sqrt{n}(\|W\|_2^2 + 1/n_1)^{1/2} > \eta/(2M')) \leq \rho
\end{aligned}$$

which is eventually smaller than ρ. Since η, ρ were arbitrary we conclude that $B(W;\hat{f}_0) \to 0$ in probability so that $\hat{\tau}_{\mathrm{WLS}(W)} \to 0$ in probability.

Consider case (b). Let $\overline{\beta}_0 = (\alpha_0, \beta_0^T)^T$. Then $\tilde{\beta} - \overline{\beta}_0 = \hat{P}^{-1}\hat{H}$ so by the above $\tilde{\beta} - \overline{\beta}_0 = O_p(n^{-1/2})$. Noting that $\hat{\tau}_{\mathrm{WLS}(W)} - \mathrm{SATT} = \frac{1}{n_1}\sum_{i\in\mathcal{T}_1} \epsilon_i + (\overline{\beta}_0 - \tilde{\beta})^T(\frac{1}{n_1}\sum_{i\in\mathcal{T}_1} \tilde{X}_i)$ completes the proof. ■

Proof [Proof of Prop. 21] Writing $W_i' = (1-T_i)W_i - T_i/n_1$, $K' = K + \lambda I_{\mathcal{T}_0}$, we have that $\mathrm{SKOM}(\mathcal{W}, \mathcal{K}, \lambda)$ is given by

$$\begin{aligned}
&\min_{W\in\mathcal{W}} \sup_{\substack{\sum_{i,j=1}^\infty \alpha_i\alpha_j\mathcal{K}(x_i,x_j) \leq 1, \\ g\in\mathcal{G}, \sum_{i=1}^\infty \alpha_i^2\mathcal{K}(x_i,x_i) < \infty, \\ \sum_{i=1}^\infty \alpha_i g'(x_i) = 0 \ \forall g' \in \mathcal{G}}} \sum_{i=1}^n W_i' \left(g(X_i) + \sum_{j=1}^\infty \alpha_j \mathcal{K}(x_j, X_i) \right) + \lambda\|W\|_2^2 \\
&= \min_{W\in\mathcal{W}} \begin{cases} \infty & \exists g \in \mathcal{G} : \sum_{i=1}^n W_i' g(X_i) \neq 0 \\ W'^T K' W' & \forall g \in \mathcal{G} : \sum_{i=1}^n W_i' g(X_i) = 0 \end{cases} = \min_{\substack{W\in\mathcal{W}, \\ GW' = 0}} W'^T K' W'.
\end{aligned}$$

The result follows by writing $GW' = 0$ as $W' \in \mathrm{null}(G) = \mathrm{span}(N)$, which is in turn written as $W' = NU$ for a new variable $U \in \mathbb{R}^k$. ■

Proof [Proof of Prop. 22] Using similar arguments to the proof of Thm. 3, we get that $\mathfrak{B}(W; \|\cdot\|_{\partial(\hat{\mu}_n,\delta)})$ is equal (up to a scaling of $n(n-1)$) to

$$\begin{aligned}
\min_{S\in\mathbb{R}_+^{\mathcal{T}_1\times\mathcal{T}_0}, t\in\mathbb{R}_+} \quad & t \\
\text{s.t.} \quad & \textstyle\sum_{i'\in\mathcal{T}_0} S_{ii'} = 1/n_1 \quad \forall i \in \mathcal{T}_1 \\
& \textstyle\sum_{i'\in\mathcal{T}_1} S_{i'i} = W_i \quad \forall i \in \mathcal{T}_0 \\
& t \geq D_{ii'} S_{ii'} \quad \forall i \in \mathcal{T}_1, i' \in \mathcal{T}_0.
\end{aligned}$$

Hence, minimizing it over $\mathcal{W}^{\text{simplex}}$ or $\mathcal{W}^{1/n_1\text{-simplex}}$ we get the same network flow problems except with a bottleneck objective. The solution is still integer and gives the pair matching with minimal maximal pair distance, corresponding exactly to OCM with or without replacement. ■

Proof [Proof of Prop. 23] For f piecewise constant on the coarsening components let f_j denote its value on the j^{th} component. By writing $f(x) = \sum_{j=1}^M \mathbb{I}_{[C(x)=j]} f_j$ and exchanging sums we can rewrite $B(W;f)$ as

$$B(W;f) = \textstyle\sum_{j=1}^J f_j(\frac{1}{n_1}\sum_{i\in\mathcal{T}_1} \mathbb{I}_{[C(X_i)=j]} - \sum_{i\in\mathcal{T}_0} W_i \mathbb{I}_{[C(X_i)=j]}).$$

Under the constraint $|f_j| \leq 1 \; \forall j$, the maximizer of the above assigns ± 1 to each f_j in order to make the j^{th} term nonnegative. Hence,

$$\mathfrak{B}(W; \|\cdot\|_{L_\infty(C)}) = \textstyle\sum_{j=1}^{J} |\frac{1}{n_1} \sum_{i\in\mathcal{T}_1} \mathbb{I}_{[C(X_i)=j]} - \sum_{i\in\mathcal{T}_0} W_i \mathbb{I}_{[C(X_i)=j]}|,$$

which we recognize as the coarsened L_1 distance from eq. (11). ■

Proof [Proof of Prop. 24] Let $W \in \mathcal{W}^{\text{subsets}}$ and $\mathcal{T}_0' = \{i \in \mathcal{T}_0 : W_i > 0\}$. Then, by duality of Euclidean norms,

$$\mathfrak{B}(W; \|\cdot\|_{\text{2-lin}(V)}) = \sup_{\beta^T V \beta \leq 1} \beta^T (\tfrac{1}{n_1} \textstyle\sum_{i\in\mathcal{T}_1} X_i - \sum_{i\in\mathcal{T}_0} W_i X_i) = M_V(\mathcal{T}_0').$$

■

Proof [Proof of Prop. 25] By linearity, we have

$$\begin{aligned}
\mathfrak{B}(W; \|\cdot\|_{\|\cdot\|_{\text{A}} \oplus_\rho \|\cdot\|_{\text{B}}}) &= \sup_{\|f\|_{\|\cdot\|_{\text{A}} \oplus_\rho \|\cdot\|_{\text{B}}} \leq 1} B(W; f) \\
&= \sup_{\|f_A\|_{\text{A}} \leq 1, \|f_B\|_{\text{B}}/\rho \leq 1} B(W; f_A + f_B) \\
&= \mathfrak{B}(W; \|\cdot\|_{\text{A}}) + \rho \mathfrak{B}(W; \|\cdot\|_{\text{B}})
\end{aligned}$$

■

Proof [Proof of Prop. 26] Note that $\text{GOM}(\mathcal{W}^{\text{general}}, \|\cdot\|_{\text{2-lin}(V)}, 0)$ is the same as $\text{GOM}(\mathcal{W}^{\text{general}}, \|\cdot\|_{\text{2-lin}}, 0)$ applied to the data $\tilde{X}_i = V^{1/2} X_i$. However, applying OLS adjustment to data X_i or data $\tilde{X}_i$ is exactly the same because β_1, β_2 are unrestricted so we can make the transformation $\tilde{\beta}_1, \tilde{\beta}_2 = V^{-1/2}\beta_1, V^{-1/2}\beta_2$ without any effect except transforming $\tilde{X}_i$ to X_i. Finally, we apply Prop. 27 with $\lambda = 0$. ■

Proof [Proof of Prop. 27] We can rewrite the ridge-regression problem as

$$\begin{aligned}
&\min_{\tau,\alpha,\beta_1,\beta_2} \sum_{i=1}^{n} (Y_i - \alpha - \tau T_i - \beta_1^T X_i - \beta_2^T (X_i - \overline{X}_{\mathcal{T}_1}) T_i)^2 + \lambda \|\beta_1\|_2^2 + \lambda \alpha^2 \\
&= \min_{\tau,\alpha,\beta_1,\beta_2} \Bigg(\sum_{i\in\mathcal{T}_0} (Y_i - \alpha - \beta_1^T X_i)^2 + \lambda \|\beta_1\|_2^2 + \lambda \alpha^2 \\
&\qquad\qquad + \sum_{i\in\mathcal{T}_1} (Y_i - (\alpha + \tau - \beta_2^T \overline{X}_{\mathcal{T}_1}) - (\beta_1 + \beta_2)^T X_i)^2 \Bigg) \\
&= \min_{\alpha,\beta_1} \Big(\sum_{i\in\mathcal{T}_0} (Y_i - \alpha - \beta_1^T X_i)^2 + \lambda \|\beta_1\|_2^2 + \lambda \alpha^2 \Big) + \min_{\tilde{\alpha},\tilde{\beta}} \sum_{i\in\mathcal{T}_1} (Y_i - \tilde{\alpha} - \tilde{\beta}^T X_i)^2,
\end{aligned}$$

where we used the transformation $\tilde{\alpha} = \alpha + \tau - \beta_2^T \overline{X}_{\mathcal{T}_1}$, $\tilde{\beta} = \beta_1 + \beta_2$ and the fact that τ and β_2 are unrestricted to see that $\tilde{\alpha}, \tilde{\beta}$ are unrestricted. Because $\tilde{\alpha}$, $\tilde{\beta}$ solve an OLS problem with intercept on $\mathcal{T}_1$, the mean of in-sample residuals are zero, and therefore:

$$\begin{aligned}
0 = \tfrac{1}{n_1} \textstyle\sum_{i\in\mathcal{T}_1} (Y_i - \tilde{\alpha} - \tilde{\beta}^T X_i) &= \overline{Y}_{\mathcal{T}_1}(1) - \tilde{\alpha} - \beta_1^T \overline{X}_{\mathcal{T}_1} - \beta_2^T \overline{X}_{\mathcal{T}_1}, \\
\implies \tau_{\lambda\text{-ridge}} = \tilde{\alpha} - \alpha + \beta_2^T \overline{X}_{\mathcal{T}_1} &= \overline{Y}_{\mathcal{T}_1}(1) - \alpha - \beta_1^T \overline{X}_{\mathcal{T}_1} \\
&= \overline{Y}_{\mathcal{T}_1}(1) - (\alpha, \beta_1^T) \tilde{X}_{\mathcal{T}_1}^T \tfrac{e_{n_1}}{n_1},
\end{aligned}$$

where $\tilde{X}_i = (1, X_i)$ and $\tilde{X}_{\mathcal{T}_t}$ is the $n_t \times (d+1)$ matrix of these. Note α, β_1 solve ridge-regression on $\mathcal{T}_0$, so:

$$(\alpha, \beta_1^T)^T = (\tilde{X}_{\mathcal{T}_0}^T \tilde{X}_{\mathcal{T}_0} + \lambda I_{d+1})^{-1} \tilde{X}_{\mathcal{T}_0}^T Y_{\mathcal{T}_0}.$$

Therefore, letting $\tilde{W} = \frac{1}{n_1} \tilde{X}_{\mathcal{T}_0} (\tilde{X}_{\mathcal{T}_0}^T \tilde{X}_{\mathcal{T}_0} + \lambda I_{d+1})^{-1} \tilde{X}_{\mathcal{T}_1}^T e_{n_1}$, we must have

$$\tau_{\lambda\text{-ridge}} = \overline{Y}_{\mathcal{T}_1}(1) - \textstyle\sum_{i \in \mathcal{T}_0} \tilde{W}_i Y_i.$$

On the other hand, the weights W given by $\mathrm{GOM}(\mathcal{W}^{\mathrm{general}}, \|\cdot\|_{2\text{-lin}}, \lambda)$ minimize the following objective function:

$$\begin{aligned}
&\sup_{\alpha^2 + \|\beta\|_2^2 \le 1} \left(\tfrac{1}{n_1} \textstyle\sum_{i \in \mathcal{T}_1} (\alpha + \beta^T X_i) - \sum_{i \in \mathcal{T}_0} W_i (\alpha + \beta^T X_i) \right)^2 + \lambda \|W\|_2^2 \\
= &\sup_{\alpha^2 + \|\beta\|_2^2 \le 1} \left(\begin{pmatrix} W \\ -\frac{e_{n_1}}{n_1} \end{pmatrix}^T \tilde{X} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \right)^2 + \lambda \|W\|_2^2 = \left\| \tilde{X}^T \begin{pmatrix} W \\ -\frac{e_{n_1}}{n_1} \end{pmatrix} \right\|_2^2 + \lambda \|W\|_2^2 \\
= &\|\tilde{X}_{\mathcal{T}_0}^T W - \tilde{X}_{\mathcal{T}_1}^T \tfrac{e_{n_1}}{n_1}\|_2^2 + \lambda \|W\|_2^2 .
\end{aligned}$$

By first order optimality conditions on unrestricted W we have

$$0 = \tilde{X}_{\mathcal{T}_0}(\tilde{X}_{\mathcal{T}_0}^T W - \tilde{X}_{\mathcal{T}_1}^T \tfrac{e_{n_1}}{n_1}) + \lambda W \implies W = (\tilde{X}_{\mathcal{T}_0} \tilde{X}_{\mathcal{T}_0}^T + \lambda I_{n_0})^{-1} \tilde{X}_{\mathcal{T}_0} \tilde{X}_{\mathcal{T}_1}^T \tfrac{e_{n_1}}{n_1}.$$

Fix $\tilde{\lambda} > 0$. By applying the Sherman-Morrison-Woodbury (SMW) formula (Boyd and Vandenberghe, 2004, §C.4.3) *twice*, we have

$$\begin{aligned}
(\tilde{X}_{\mathcal{T}_0} \tilde{X}_{\mathcal{T}_0}^T + \tilde{\lambda} I_{n_0})^{-1} \tilde{X}_{\mathcal{T}_0} &\overset{\mathrm{SMW}}{=} (\tfrac{1}{\tilde{\lambda}} I_{n_0} - \tilde{X}_{\mathcal{T}_0} (\tfrac{1}{\tilde{\lambda}} \tilde{X}_{\mathcal{T}_0}^T \tilde{X}_{\mathcal{T}_0} + I_{d+1})^{-1} \tilde{X}_{\mathcal{T}_0}^T) \tilde{X}_{\mathcal{T}_0} \\
&= \tfrac{1}{\tilde{\lambda}} \tilde{X}_{\mathcal{T}_0} (I_{d+1} - (\tfrac{1}{\tilde{\lambda}} \tilde{X}_{\mathcal{T}_0}^T \tilde{X}_{\mathcal{T}_0} + I_{d+1})^{-1} \tfrac{1}{\tilde{\lambda}} \tilde{X}_{\mathcal{T}_0}^T \tilde{X}_{\mathcal{T}_0}) \\
&\overset{\mathrm{SMW}}{=} \tfrac{1}{\tilde{\lambda}} \tilde{X}_{\mathcal{T}_0} (\tfrac{1}{\tilde{\lambda}} \tilde{X}_{\mathcal{T}_0}^T \tilde{X}_{\mathcal{T}_0} + I_{d+1})^{-1} \\
&= \tilde{X}_{\mathcal{T}_0} (\tilde{X}_{\mathcal{T}_0}^T \tilde{X}_{\mathcal{T}_0} + \tilde{\lambda} I_{d+1})^{-1}.
\end{aligned}$$

If $\lambda > 0$ then set $\tilde{\lambda} = \lambda$. If $\lambda = 0$ and $\tilde{X}_{\mathcal{T}_0}^T \tilde{X}_{\mathcal{T}_0}$ is invertible then, by continuous transformation over the limit $\tilde{\lambda} \to 0$, the equation of the first to the last holds with $\tilde{\lambda} = 0$. In either case, this shows $W = \tilde{W}$, completing the proof. ■

Proof [Proof of Prop. 29] Let $\Delta = \mathbb{E}\left[\sum_{i \in \mathcal{T}_1} W_i X_i - \sum_{i \in \mathcal{T}_0} W_i X_i\right]$ and similarly define Δ'.

Suppose $\Delta = \alpha \Delta$ for $\alpha \in [-1, 1]$. Then, for any $f(x) = \beta^T x + \beta_0$, $|\mathbb{E}[B(W; f)]| = \left|\beta^T \Delta\right| = |\alpha| \left|\beta^T \Delta'\right| \le \left|\beta^T \Delta'\right| = |\mathbb{E}[B(W'; f)]|$.

Now, suppose W is linearly EPBR relative to W'. Then, for any β, we have that if $\beta^T \Delta' = 0$ then $B(W'; f) = 0$ for $f(x) = \beta^T x$, which by linear EPBR implies that $B(W; f) = 0$, which means that $\beta^T \Delta = 0$. In other words, $\mathrm{span}(\Delta')^\perp \subseteq \mathrm{span}(\Delta)^\perp$. Therefore, $\mathrm{span}(\Delta) \subseteq \mathrm{span}(\Delta')$, which exactly means that $\exists \alpha \in \mathbb{R}$ such that $\Delta = \alpha \Delta'$. If $\Delta' = 0$ then we can choose $\alpha = 0$. If $\Delta' \neq 0$ then there exists β such that $\left|\beta^T \Delta'\right| > 0$ and hence $|\alpha| = \left|\beta^T \Delta\right| / \left|\beta^T \Delta'\right| \le 1$ by linear EPBR. ■

Proof [Proof of Prop. 32] Fixing any $\mu \in \mathbb{R}^d$, $\mu \neq 0$, we can affinely transform the data so that $X \mid T = 0$ is spherical at zero (has zero mean, unit covariance, and its distribution is unitarily invariant) and $X \mid T = 1$ is distributed the same as $\mu + \alpha X \mid T = 0$ for some $\alpha \in \mathbb{R}_+$. For any affinely invariant W', we may assume this form for the data and any affinely invariant W' is also unitarily invariant so that by spherical symmetry, $\mathbb{E}\left[\sum_{i=1}^n (-1)^{T_i+1} W_i X_i\right] \in \text{span}(\mu)$. Both W and $W^{(0)}$ are affinely invariant. ■

Proof [Proof of Prop. 34] Fix X, T, $A \in \mathbb{R}^{d\times d}$, and $a \in \mathbb{R}^d$ with A non-singular. Let $\hat{\mu}$ and $\widehat{\Sigma}$ be defined as in Alg. 3 for X, T. Let $\hat{\mu}_A$ and $\widehat{\Sigma}_A$ be defined as in Alg. 3 for the transformed data $X_A = XA^T + \mathbf{1}_n a^T$, T. Then, $\hat{\mu}_A = A\hat{\mu} + a$ and $\widehat{\Sigma}_A = A\widehat{\Sigma}A^T$. The inner products produced by Alg. 3 on the transformed data are $((X_A)_i - \hat{\mu}_A)^T \widehat{\Sigma}_A^\dagger ((X_A)_j - \hat{\mu}_A) = (AX_i - A\hat{\mu})^T A^{-T} \widehat{\Sigma}^\dagger A^{-1}(AX_j - A\hat{\mu}) = (X_i - \hat{\mu})^T \widehat{\Sigma}^\dagger (X_j - \hat{\mu})$, which are the inner products produced by Alg. 3 on the untransformed data. ■

Proof [Proof of Prop. 35] If $X_{1:n}$ is affinely independent then it can be affinely mapped to the $(n-1)$-dimensional simplex, composed of 0 and the first $n-1$ unit vectors. If W is affinely invariant then it takes the same value on all affinely independent $X_{1:n}$ as it does on the $(n-1)$-dimensional simplex. ■

Proof [Proof of Prop. 37] The proof is the same as that of Prop. 29. ■

Proof [Proof of Prop. 40] The proof is the same as that of Prop. 32: without loss of generality we may assume that the embedded control data is spherical (ellipsoidal with zero mean and identity covariance operator and therefore spherically symmetric under unitary transformations) and that treated data is distributed like scaling and shifting the control data. ■

Proof [Proof of Prop. 42] The proof is the same as that of Prop. 34. ■

Proof [Proof of Prop. 43] Let $X_{1:n}$ and $X'_{1:n}$ each be a list of n distinct elements of $\mathcal{X}$. Since $\mathcal{K}$ is strictly positive definite, both $\{\mathcal{K}(X_i, \cdot) : i = 1, \ldots, n\}$ and $\{\mathcal{K}(X'_i, \cdot) : i = 1, \ldots, n\}$ are linearly independent sets of vectors. Therefore, there exists a bounded non-singular $A \in \mathcal{F} \otimes \mathcal{F}$ such that $AX_i = X'_i$. Since W is $\mathcal{K}$-affinely invariant, it is the same for $X_{1:n}$ and $X'_{1:n}$. ■