The Annals of Statistics
2019, Vol. 47, No. 2, 965–993
https://doi.org/10.1214/18-AOS1698

COVARIATE BALANCING PROPENSITY SCORE BY TAILORED LOSS FUNCTIONS[1]

BY QINGYUAN ZHAO

University of Pennsylvania

In observational studies, propensity scores are commonly estimated by maximum likelihood but may fail to balance high-dimensional pretreatment covariates even after specification search. We introduce a general framework that unifies and generalizes several recent proposals to improve covariate balance when designing an observational study. Instead of the likelihood function, we propose to optimize special loss functions—covariate balancing scoring rules (CBSR)—to estimate the propensity score. A CBSR is uniquely determined by the link function in the GLM and the estimand (a weighted average treatment effect). We show CBSR does not lose asymptotic efficiency in estimating the weighted average treatment effect compared to the Bernoulli likelihood, but CBSR is much more robust in finite samples. Borrowing tools developed in statistical learning, we propose practical strategies to balance covariate functions in rich function classes. This is useful to estimate the maximum bias of the inverse probability weighting (IPW) estimators and construct honest confidence intervals in finite samples. Lastly, we provide several numerical examples to demonstrate the tradeoff of bias and variance in the IPW-type estimators and the tradeoff in balancing different function classes of the covariates.

1. Introduction. To obtain causal relations from observational data, one crucial obstacle is that some pretreatment covariates are not balanced between the treatment groups. Exact matching, inexact matching and subclassification on raw covariates were first used by pioneers like Cochran [8, 9] and Rubin [42]. Later in the seminal work of Rosenbaum and Rubin [39], the propensity score, defined as the conditional probability of receiving treatment given the covariates, was established as a fundamental tool to adjust for imbalance in more than just a few covariates. Over the next three decades, numerous methods based on the propensity score have been proposed, most notably propensity score matching (e.g., [1, 41]), propensity score subclassification (e.g., [40]), and inverse probability weighting (e.g., [23, 38]); see Imbens [28], Lunceford and Davidian [34], Caliendo and Kopeinig [6], Stuart [47] for some comprehensive reviews.

Received March 2017; revised November 2017.

[1]The first version of this work was finished during the author's Ph.D. candidacy at Stanford University.

MSC2010 subject classifications. Primary 62P10; secondary 62C99.

Key words and phrases. Convex optimization, kernel method, inverse probability weighting, proper scoring rule, regularized regression, statistical decision theory.

With the rapidly increasing ability to collect high-dimensional covariates in the "big data" era (e.g., large number of covariates collected in health care claims data), propensity-score based methods often fail to produce satisfactory covariate balance [26]. In the meantime, numerical examples in Smith and Todd [46], Kang and Schafer [32] have demonstrated that the average treatment effect estimates can be highly sensitive to the working propensity score model. Conventionally, these two issues are handled by a specification search—the estimated propensity score is applicable only if it balances covariates well. A simple strategy is to gradually increase the model complexity by forward stepwise regression ([29], Sections 13.3–13.4), but as a numerical example below indicates, this has no guarantee to achieve sufficient covariate balance eventually.

More recently, several new methods were proposed to directly improve covariate balance in the design of an observational study, either by modifying the propensity score model [17, 27] or by directly constructing sample weights for the observations [7, 19, 21, 31, 53]. These methods have been shown to work very well empirically (particularly in smaller samples) and some asymptotic justifications were subsequently provided (e.g., [13, 52]).

In this paper, we will introduce a general framework that unifies and generalizes these proposals. The solution provided here is conceptually simple: in order to improve covariate balance of a propensity score model, one just needs to minimize, instead of the most widely used negative Bernoulli likelihood, a special loss function tailored to the estimand.

1.1. *A toy example.* To demonstrate the simplicity and effectiveness of the tailored loss function approach, we use the prominent simulation example of Kang and Schafer [32]. In this example, for each unit $i = 1, \ldots, n = 200$, suppose that $(Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})^T$ is independently distributed as $\mathrm{N}(0, I_4)$ and the true propensity scores are $p_i = \mathbb{P}(T_i = 1|\boldsymbol{Z}_i) = \operatorname{expit}(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.1Z_{i4})$ where $T_i \in \{0, 1\}$ is the treatment label. However, the observed covariates are nonlinear transformations of $\boldsymbol{Z}$: $X_{i1} = \exp(Z_{i1}/2)$, $X_{i2} = Z_{i2}/(1 + \exp(Z_{i1})) + 10$, $X_{i3} = (Z_{i1}Z_{i3}/25 + 0.6)^3$, $X_{i4} = (Z_{i2} + Z_{i4} + 20)^2$. To model the propensity score, we use a logistic model with some or all of $\{X_1, X_2, X_3, X_4, X_1^2, X_2^2, X_3^2, X_4^2\}$ as regressors. Using forward stepwise regression, two series of models are fitted using the Bernoulli likelihood and the loss function tailored for estimating the average treatment effect (ATE, see Section 3.1 for more detail). Inverse probability weights (IPW) are obtained from each fitted model and standardized differences of the regressors are used to measure covariate imbalance [41].

Figure 1 shows the paths of standardized difference for one realization of the simulation. A widely used criterion is that a standardized difference above 10% is unacceptable [3, 37], which is the dashed line in Figure 1. The left panel of Figure 1 uses the Bernoulli likelihood to fit and select logistic regression models. The standardized difference paths are not monotonically decreasing and never achieve

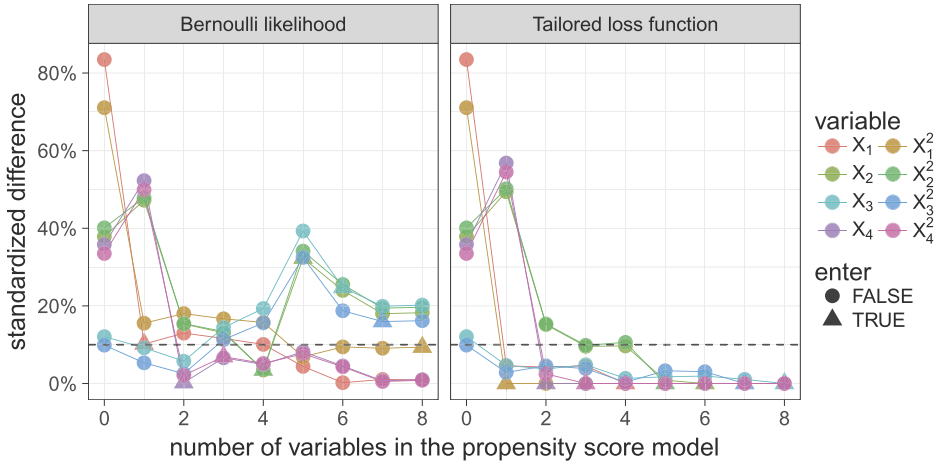

FIG. 1. *The tailored loss function proposed in this paper is much better than Bernoulli likelihood at reducing covariate imbalance. Propensity score is modeled by logistic regression and fitted by the tailored loss function or Bernoulli likelihood. Standardized difference is computed using inverse probability weighting (IPW) and pooled variance for the two treatment groups* [41]. *A standardized difference above* 10% *is often viewed unacceptable by many practitioners.*

the satisfactory level (10%) for all the regressors. In contrast, the right panel of Figure 1 uses the tailored loss function and all eight predictors are well balanced after three steps. In fact, as a feature of using the tailored loss function, all active regressors (variables in the selected model) are exactly balanced.

The toy example here is merely for presentation, but it clearly demonstrates that the proposed tailored loss function approach excels in balancing covariates. We will discuss some practical strategies that are more sophisticated than the forward stepwise regression in Section 4.

1.2. *Related work and our contribution.* The tailored loss function framework introduced here unifies a number of existing methods by exploring the (Lagrangian) duality of propensity scores and sample weights. Roughly speaking, the "moment condition" approaches advocated by Graham, De Xavier Pinto and Egel [17] and Imai and Ratkovic [27] correspond to the primal problem of minimizing the tailored loss over propensity score models, while the "empirical balancing" proposals (e.g., [19, 53]) correspond to the dual problem that solves some convex optimization problem over the sample weights subject to covariate balance constraints. The framework presented here is largely motivated by the aforementioned works. Part of the contribution of this paper is to bring together many pieces scattered in this literature—moment condition of estimating the propensity score, covariate balance, bias-variance tradeoff, different estimands, link function of a generalized linear model (the latter two are often overlooked)—and elucidate their roles in the design and analysis of an observational study.

A reader familiar with the development of this literature may recognize that many elements in the framework proposed here have already appeared in some previous works. Perhaps the closest approach is that of Imai and Ratkovic [27], as their covariate balancing moment conditions are essentially the first-order conditions of minimizing the tailored loss function. However, there are several benefits by taking a decision theoretic approach:

- We can visualize the tailored loss functions and obtain more insights. (See the Supplementary Material [51] for more detail.)
- We can understand when the moment equations have a unique solution by investigating convexity of the loss function.
- More importantly, we can use predictive algorithms developed in statistical learning to optimize covariate balance in high-dimensional problems and rich function classes. Moment constraints methods usually exactly balance several selected covariate functions but leave the others unattended. By regularizing the tailored loss function, the methods proposed in Section 4 can inexactly balance high-dimensional or even infinite-dimensional covariate functions. This usually results in more accurate estimates and more robust statistical inference.

Compared to the empirical balancing methods, the tailored loss function framework shows that they are essentially equivalent to certain models of propensity score. Asymptotic theory that are already established to propensity-score based estimators can now apply to empirical balancing methods. Our framework also allows the use of balancing weights in estimating more general estimands. For example, we can produce balancing weights to estimate the optimally weighted average treatment effect proposed by Crump et al. [10] that is more stable when there is limited overlap [33].

Last but not the least, we provide a novel approach to make honest, design-based and finite-sample inference for the weighted ATE. Instead of the improbable but commonly required assumption that the propensity score is correctly specified, the only major assumption we make is that the (unknown) true outcome regression function is in a given class. The function class can be high-dimensional and very rich. We give a Bayesian interpretation that underlies any design of an observational study and provide extensive numerical results to demonstrate the tradeoff in making different assumptions about the outcome regression function.

The next two sections are devoted to introducing the tailored loss functions. Section 4 proposes practical strategies motivated by statistical learning. Section 5 then considers some theoretical aspects about the tailored loss functions, including the dual interpretation, Bayesian interpretation, a new method for design-based inference and how to choose the tuning parameters. Section 6 uses numerical examples in two new settings to demonstrate the flexibility of the proposed framework and examine its empirical performance. Section 7 concludes the paper with some practical recommendations. Technical proofs are provided in the Supplementary Material [51].

2. Preliminaries on statistical decision theory. To start with, propensity score estimation can be viewed as a decision problem and this section introduces some terminologies in statistical decision theory. In a typical problem of making probabilistic forecast, the decision maker needs to pick an element as the prediction from $\mathcal{P}$, a convex class of probability measures on some general sample space Ω. For example, a weather forecaster needs to report the chance of rain tomorrow, so the sample space is $\Omega = \{\text{rain, no rain}\}$ and the prediction is a Bernoulli distribution. Propensity score is a (conditional) probability measure, but recall that the goal is to achieve satisfactory covariate balance rather than the best prediction of treatment assignment. At a high level, this is precisely the reason why we want to tailor the loss function when estimating the propensity score.

2.1. *Proper scoring rules.* At the core of statistical decision theory is the *scoring rule*, which can be any extended real-valued function $S : \mathcal{P} \times \Omega \to [-\infty, \infty]$ such that $S(P, \cdot)$ is $\mathcal{P}$-integrable for all $P \in \mathcal{P}$ [16]. If the decision is P and ω materializes, the decision maker's reward or utility is $S(P, \omega)$. An equivalent but more pessimistic terminology is *loss function*, which is just the negative scoring rule. These two terms will be used interchangeably in this paper.

If the outcome is probabilistic in nature and the actual probability distribution is Q, the expected score of forecasting P is

$$S(P, Q) = \int S(P, \omega) Q(d\omega).$$

To encourage honest decisions, the scoring rule S is generally required to be *proper*,

$$S(Q, Q) \geq S(P, Q) \qquad \forall P, Q \in \mathcal{P}. \tag{1}$$

The rule is called *strictly proper* if (1) holds with equality if and only if $P = Q$.

In observational studies, the sample space is commonly dichotomous $\Omega = \{0, 1\}$ (two treatment groups: 0 for control and 1 for treated), though there is no essential difficulty to extend the approach in this paper to $|\Omega| > 2$ (multiple treatments) or $\Omega \subset \mathbb{R}$ (continuous treatment). In the binary case, a probability distribution P can be characterized by a single parameter $0 \leq p \leq 1$, the probability of treatment. A classical result of Savage [45] asserts that every real-valued [except for possibly $S(0, 1) = \infty$ or $S(1, 0) = -\infty$] proper scoring rule S can be written as

$$S(p, 1) = G(p) + (1 - p)G'(p) = \int (1 - p)G''(p)\, dp + \text{const},$$

$$S(p, 0) = G(p) - pG'(p) = -\int pG''(p)\, dp + \text{const},$$

where $G : [0, 1] \to \mathbb{R}$ is a convex function and $G'(p)$ is a subgradient of G at the point $p \in [0, 1]$. When G is second-order differentiable, an equivalent but more

convenient representation is

$$\frac{\partial}{\partial p}S(p,t) = (t-p)G''(p), \qquad t = 0, 1. \tag{2}$$

Since the function G uniquely defines a scoring rule S, we shall call G a scoring rule as well.

A useful class of proper scoring rules is the following Beta family:

$$G''_{\alpha,\beta}(p) = p^{\alpha-1}(1-p)^{\beta-1}, \qquad -\infty < \alpha, \beta < \infty. \tag{3}$$

These scoring rules were first introduced by Buja, Stuetzle and Shen [5] to approximate the weighted misclassification loss by taking the limit $\alpha, \beta \to \infty$ and $\alpha/\beta \to c$. For example, if $c = 1$, the score $G_{\alpha,\beta}$ converges to the zero-one misclassification loss. Many important scoring rules belong to this family. For example, the Bernoulli log-likelihood function $S(p,t) = t\log p + (1-t)\log(1-p)$ corresponds to $\alpha = \beta = 0$, and the Brier score or the squared error loss $S(p,t) = -(t-p)^2$ corresponds to $\alpha = \beta = 1$. For our purpose of estimating propensity score, it will be shown later that the subfamily $-1 \le \alpha, \beta \le 0$ is particularly useful.

2.2. *Propensity score modeling by maximizing score.* Given i.i.d. observations $(\boldsymbol{X}_i, T_i) \in \mathbb{R}^d \times \{0,1\}$, $i = 1, 2, \ldots, n$ where T_i is the binary treatment assignment and $\boldsymbol{X}_i$ is a vector of d pretreatment covariates, we want to fit a model for the propensity score $p(\boldsymbol{X}) = \mathbb{P}(T = 1|\boldsymbol{X})$ in a prespecified family $\mathcal{P} = \{p_{\boldsymbol{\theta}}(\boldsymbol{X}) : \boldsymbol{\theta} \in \Theta\}$. Later on we will consider very rich model family, but for now let us focus on the generalized linear models with finite-dimensional regressors $\boldsymbol{\phi}(\boldsymbol{X}) = (\phi_1(\boldsymbol{X}), \ldots, \phi_m(\boldsymbol{X}))^T$ [35]

$$p_\theta(\boldsymbol{X}) = l^{-1}\big(f_\theta(\boldsymbol{X})\big) = l^{-1}\big(\theta^T\boldsymbol{\phi}(\boldsymbol{X})\big), \tag{4}$$

where l is the *link function*. In our framework, the tailored loss function is determined by the link function l (and the estimand). The most common choice is the logistic link

$$l(p) = \log\frac{p}{1-p}, \qquad l^{-1}(f) = \frac{e^f}{1+e^f}, \tag{5}$$

which will be used in all the numerical examples of this paper.

Given a strictly proper scoring rule S, the *maximum score (minimum loss) estimator* of θ is obtained by maximizing the average score

$$\hat{\theta}_n = \arg\max_\theta \mathcal{S}_n(\theta) = \frac{1}{n}\sum_{i=1}^n S\big(p_\theta(\boldsymbol{X}_i), T_i\big). \tag{6}$$

Notice that an affine transformation $S(p,t) \mapsto aS(p,t) + b(t)$ for any $a > 0$ and $-\infty < b(t) < \infty$ gives the same estimator $\hat{\boldsymbol{\theta}}_n$. Due to this reason, we will not differentiate between these equivalent scoring rules and use a single function $S(p,t)$ to represent all the equivalent ones.

When S is differentiable and assuming exchangeability of taking expectation and derivative, the maximizer of $\mathbb{E}[\mathcal{S}_n(\boldsymbol{\theta})]$, which is indeed $\boldsymbol{\theta}$ if the propensity score is correctly specified $p(\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x})$ (a property called Fisher consistency), is characterized by the following estimating equations:

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}\big[\mathcal{S}_n(\boldsymbol{\theta})\big] = \mathbb{E}\big[\nabla_{\boldsymbol{\theta}}\mathcal{S}_n(\boldsymbol{\theta})\big] = \mathbb{E}_{\boldsymbol{X},T}\big[\nabla_{\boldsymbol{\theta}} S\big(l^{-1}\big(\boldsymbol{\theta}^T\phi(\boldsymbol{X})\big), T\big)\big] = 0. \tag{7}$$

3. Tailoring the loss function.

3.1. *Covariate balancing scoring rules.* The tailored loss function framework is motivated by reinterpreting the first-order conditions (7) as covariate balancing constraints. Using the representation (2) and the inverse function theorem, we can rewrite (7) as

$$\mathbb{E}\big[\big(T - (1-T)\big)w(\boldsymbol{X}, T)\cdot\phi(\boldsymbol{X})\big] = 0, \tag{8}$$

where the weighting function

$$w(\boldsymbol{x}, t) = \frac{G''(p(\boldsymbol{x}))}{l'(p(\boldsymbol{x}))}\big[t\big(1 - p(\boldsymbol{x})\big) + (1-t)p(\boldsymbol{x})\big] \tag{9}$$

is determined by the scoring rule through G'' and the link function l. The maximum score estimator $\hat{\boldsymbol{\theta}}_n$ can be obtained by solving (8) with the expectation over the empirical distribution of $(\boldsymbol{X}, T)$ instead of the population. When the optimization problem (6) is strongly convex, the solution to (8) is also unique.

The next key observation is that every weighting function $w(\boldsymbol{x}, t)$ defines a weighted average treatment effect (ATE). To see this, we need to introduce the Neyman–Rubin causal model. Let $Y(t)$, $t = 0, 1$ be the potential outcomes and $Y = TY(1) + (1-T)Y(0)$ be the observed outcome. This paper assumes strong ignorability of treatment assignment [39], so the observational study is free of hidden bias.

ASSUMPTION 1. $T \perp\!\!\!\perp (Y(0), Y(1))|\boldsymbol{X}$, where $\perp\!\!\!\perp$ stands for (conditional) independence.

Let the observed outcomes be Y_i, $i = 1, \ldots, n$. Naturally, the weighted difference of Y_i,

$$\hat{\tau} = \sum_{i:T_i=1} w(\boldsymbol{X}_i, T_i)Y_i - \sum_{i:T_i=0} w(\boldsymbol{X}_i, T_i)Y_i, \tag{10}$$

estimates the following population parameter:

$$\tau_v = \mathbb{E}_{\boldsymbol{X},T,Y}\big\{\big(T - (1-T)\big)w(\boldsymbol{X}, T)Y\big\} = \mathbb{E}_{\boldsymbol{X},Y}\big[v(\boldsymbol{X})\big(Y(1) - Y(0)\big)\big],$$

TABLE 1

Correspondence of estimands, sample weighting functions and the covariate balancing scoring rules (corresponding to the logistic link) in the proposed Beta family. The estimand is a weighted average treatment effect $\tau_{\alpha,\beta} = \mathbb{E}[v_{\alpha,\beta}(\boldsymbol{X})(Y(1) - Y(0))]$ *and* $\tau^* = \tau_{\alpha,\beta}/\mathbb{E}[v_{\alpha,\beta}(\boldsymbol{X})]$

α	β	Estimand	$w(x,1)$	$w(x,0)$	$S(p,1)$	$S(p,0)$
-1	-1	$\tau = \tau^* = \mathbb{E}[Y(1) - Y(0)]$	$\frac{1}{p(x)}$	$\frac{1}{1-p(x)}$	$\log\frac{p}{1-p} - \frac{1}{p}$	$\log\frac{1-p}{p} - \frac{1}{1-p}$
-1	0	$\tau^* = \mathbb{E}[Y(1) - Y(0)\|T = 0]$	$\frac{1-p(x)}{p(x)}$	1	$-\frac{1}{p}$	$\log\frac{1-p}{p}$
0	-1	$\tau^* = \mathbb{E}[Y(1) - Y(0)\|T = 1]$	1	$\frac{p(x)}{1-p(x)}$	$\log\frac{p}{1-p}$	$-\frac{1}{1-p}$
0	0	$\tau = \mathbb{E}[p(\boldsymbol{X})(1 - p(\boldsymbol{X}))\cdot (Y(1) - Y(0))]$	$1 - p(x)$	$p(x)$	$\log p$	$\log(1-p)$

which is an (unnormalized) weighted average treatment effect. Here

$$
\begin{aligned}
v(\boldsymbol{X}) &= \mathbb{E}\big[T \cdot w(\boldsymbol{X}, 1)|\boldsymbol{X}\big] = \mathbb{E}\big[(1-T)\cdot w(\boldsymbol{X}, 0)|\boldsymbol{X}\big] \\
&= \frac{G''(p(\boldsymbol{X}))p(\boldsymbol{X})(1-p(\boldsymbol{X}))}{l'(p(\boldsymbol{X}))}.
\end{aligned} \tag{11}
$$

In practice, it is usually more meaningful to estimate the normalized version $\tau_w^* = \tau_w/\mathbb{E}[w(\boldsymbol{X})]$ by normalizing the weights $w_i = w(\boldsymbol{X}_i, T_i)$, $i = 1, \ldots, n$ separately among the treated and the control: $\hat{w}_i^* = \hat{w}_i / \sum_{j:T_j=T_i} \hat{w}_j$, $i = 1, \ldots, n$.

Table 1 shows that four mostly commonly used estimands, the average treatment effect (ATE), the average treatment effect on the control (ATC), the average treatment effect on the treated (ATT) and the optimally weighted average treatment effect (OWATE) under homoscedasticity [10], are weighted average treatment effects with

$$
v_{\alpha,\beta}(\boldsymbol{X}) = p(\boldsymbol{X})^{\alpha+1}\big(1 - p(\boldsymbol{X})\big)^{\beta+1} \tag{12}
$$

with different combinations of (α, β).

Therefore, in order to estimate $\tau_{\alpha,\beta} = \mathbb{E}[v_{\alpha,\beta}(\boldsymbol{X})(Y(1) - Y(0))]$, we just need to equate (11) with (12) and solve for G. The solution in general depends on the link function l. If the logistic link is used, it is easy to show that the solution belongs to the Beta family of scoring rules defined in (3). The loss functions corresponding to the four estimands are also listed in (12).

PROPOSITION 1. *Under Assumption 1, if l is the logistic link function, then $\tau_v = \tau_{\alpha,\beta}$ if $G = G_{\alpha,\beta}$.*

To use the framework developed here in practice, the user should "invert" the development in this section. First, the user should determine the estimand by its interpretation and whether there is insufficient covariate overlap (so OWATE may be desirable). Second, the user should decide on a link function (we recommend

logistic link). Lastly, the user can equate (11) with (12) or look up Table 1 to find the corresponding scoring rule.

The main advantage of using the "correct" scoring rule is that the weights will automatically balance the predictors $\phi(\boldsymbol{X})$. This is a direct consequence of the estimating equations (8) and is summarized in the next theorem. This is precisely the reason we call $G_{\alpha,\beta}$ or the corresponding $S_{\alpha,\beta}$ the *covariate balancing scoring rule* (CBSR) with respect to the estimand $\tau_{\alpha,\beta}$ and the logistic link function in this paper.

THEOREM 1. *If l is the logistic link function and $\hat{\theta}$ is obtained as in (6) by maximizing the CBSR corresponding to l and the estimand. Then the weights $\hat{w}_i$, $i = 1, \ldots, n$ computed by (9) exactly balance the sample regressors*

$$\sum_{i:T_i=1} \hat{w}_i \phi(\boldsymbol{X}_i) = \sum_{i:T_i=0} \hat{w}_i \phi(\boldsymbol{X}_i). \tag{13}$$

Furthermore, if the predictors include an intercept term [*i.e.*, 1 *is in the linear span of* $\phi(\boldsymbol{X})$], *then* $\hat{w}^*$ *also satisfies* (13).

Note that the Bernoulli likelihood ($\alpha = \beta = 0$) indeed corresponds to the estimand OWATE instead of the more commonly used ATE or ATT. This corresponds to the "overlap weights" recently proposed by Li, Morgan and Zaslavsky [33], where each observation's weight is proportional to the probability of being assigned to the opposite group. Theorem 3 of Li, Morgan and Zaslavsky [33] states that the "overlap weights" exactly balances the regressors when Bernoulli likelihood is used, which is a special case of our Theorem 1.

3.2. *Convexity.* To obtain covariate balancing propensity scores, one can solve the estimating equations (8) directly without using the tailored loss function. This is essentially the approach taken by Imai and Ratkovic [27], although it is unclear at this point that (13) has a unique solution. The first advantage of introducing the tailored loss functions is that some CBSR is strongly concave, so the solution to its first-order condition is always unique.

PROPOSITION 2. *Suppose the estimand is in the Beta-family equation (12) and let S be the CBSR corresponding to a link function l such that $v = v_{\alpha,\beta}$. Then the score functions $S(l^{-1}(f), 0)$ and $S(l^{-1}(f), 1)$ are both concave functions of $f \in \mathbb{R}$ if and only if $-1 \le \alpha, \beta \le 1$. Moreover, if $(\alpha, \beta) \neq (-1, 0)$, $S(l^{-1}(f), 0)$ is strongly concave; if $(\alpha, \beta) \neq (0, -1)$, $S(l^{-1}(f), 1)$ is strongly concave.*

Notice that the range of (α, β) in Proposition 2 includes the four estimands listed in Table 1. As a consequence, their corresponding score maximization problems can be solved very efficiently (e.g., by Newton's method). Motivated by this observation, in the next section we propose to fit propensity score models with more sophisticated strategies stemming from statistical learning.

4. Adaptive strategies. The generalized linear model considered in Section 3 amounts to a fixed low-dimensional model space

$$\mathcal{P}_{\mathrm{GLM}} = \{p(\boldsymbol{x}) = l^{-1}(f(\boldsymbol{x})) : f(\boldsymbol{x}) \in \operatorname{span}(\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x}))^T\}.$$

As mentioned previously in Section 1, in principle, we should not restrict to a single propensity score model as it can be misspecified. Propensity score is merely a nuisance parameter in estimating weighted ATEs. We shall see repeatedly in later sections that, in finite sample, it is more important to use flexible propensity score models that balance covariates well than to estimate the propensity score accurately. In this section, we incorporate machine learning methods in our framework to expand the model space.

4.1. *Forward stepwise regression.* To increase model complexity, perhaps the most straightforward approach is forward stepwise regression as illustrated earlier in Section 1. Instead of a fixed model space, forward stepwise gradually increases model complexity. Using the tailored loss functions in Section 3, active covariates are always exactly balanced and inactive covariates are usually well balanced also.

Motivated by this strategy, Hirano, Imbens and Ridder [24] studied the efficiency of the IPW estimator when the dimension of the regressors $\phi(\boldsymbol{x})$ is allowed to increase as the sample size n grows. Their renowned results claim that this *sieve* IPW estimator is semiparametrically efficient for estimating the weighted ATE. Here we show that the semiparametric efficiency still holds if the Bernoulli likelihood, the loss function that Hirano, Imbens and Ridder [24] used to estimate the propensity score, is replaced by the Beta family of scoring rules $G_{\alpha,\beta}$, $-1 \le \alpha, \beta \le 0$ in (3) or essentially any strongly concave scoring rule. This result is not too surprising as the propensity score is just a nuisance parameter whose estimation accuracy is of less importance in semiparametric inference. Conceptually, however, this result suggests that the investigator has the freedom to choose the loss function in estimating the propensity score and do not need to worry about loss of asymptotic efficiency. The advantages of using a tailored loss function are better accuracy in finite sample and more robustness against model misspecification, as detailed later in Section 5.

Let us briefly review the sieve logistic regression in Hirano, Imbens and Ridder [24]. For $m = 1, 2, \ldots$, let $\phi_m(\boldsymbol{x}) = (\varphi_{1m}(\boldsymbol{x}), \varphi_{2m}(\boldsymbol{x}), \ldots, \varphi_{mm}(\boldsymbol{x}))^T$ be a triangular array of orthogonal polynomials, which are obtained by orthogonalizing the power series: $\psi_{km}(\boldsymbol{x}) = \prod_{j=1}^{d} x_j^{\gamma_{kj}}$, where $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kd})^T$ is an d-dimensional multi-index of nonnegative integers and satisfies $\sum_{j=1}^{d} \gamma_{kj} \le \sum_{j=1}^{d} \gamma_{k+1,j}$. Let l be the logistic link function (5). Hirano, Imbens and Ridder [24] estimated the propensity score by maximizing the log-likelihood

$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} T_i \log\bigl(l^{-1}(\phi_m(\boldsymbol{X}_i)^T\theta)\bigr) + (1 - T_i) \log\bigl(1 - l^{-1}(\phi_m(\boldsymbol{X}_i)^T\theta)\bigr).$$

This is a special case of the score maximization problem (6) by setting $S = S_{0,0}$.

Theorem 2 below is an extension to the main theorem of Hirano, Imbens and Ridder [24]. Besides strong ignorability, the other technical assumptions in Hirano, Imbens and Ridder [24] are listed in the Supplementary Material [51]. Compared to the original theorem which always uses the maximum likelihood regardless of the estimand, the scoring rule is now tailored according to the estimand as described in Section 3.

THEOREM 2. *Suppose we use the Beta-family of covariate balancing scoring rules defined by equations* (2) *and* (3) *with* $-1 \le \alpha, \beta \le 0$ *and the logistic link* (5). *Under Assumption* 1 *and the technical assumptions in Hirano, Imbens and Ridder* [24], *if we choose suitable* m *growing with* n, *the propensity score weighting estimator* $\hat{\tau}_{\alpha,\beta}$ *and its normalized version* $\hat{\tau}^*_{\alpha,\beta}$ *are consistent for* $\tau_{\alpha,\beta}$ *and* $\tau^*_{\alpha,\beta}$. *Moreover, they reach the semiparametric efficiency bound for estimating* $\tau_{\alpha,\beta}$ *and* $\tau^*_{\alpha,\beta}$.

4.2. *Regularized regression.* In predictive modeling, stepwise regression is usually suboptimal especially if we have high-dimensional covariates (see, e.g., [20], Section 3). A more principled approach is to penalize the loss function

$$\hat{\theta}_\lambda = \arg\max_{p(\cdot)\in\mathcal{P}} \frac{1}{n}\sum_{i=1}^{n} S\big(p(\boldsymbol{X}_i), T_i\big) - \lambda J\big(p(\cdot)\big), \qquad p(\boldsymbol{x}) = l^{-1}\big(f(\boldsymbol{x})\big), \tag{14}$$

where $J(\cdot)$ is a regularization function that penalizes overly-complicated propensity score model $p(x)$ and the tuning parameter λ controls the degree of regularization. This estimator reduces to the optimum score estimator (6) when $\lambda = 0$.

The penalty term $J(\theta)$ should be chosen according to the model space $\mathcal{P}$ and the investigator's prior belief about the outcome regression function (see Section 5.4). In this paper, we consider three alternatives of model space and penalty:

1. Regularized GLM: the model space is the same generalized linear model $\mathcal{P}_{\text{GLM}}$ with potentially high-dimensional covariates, but the average score is penalized by the l_a-norm of $\boldsymbol{\theta}$, $J(p_{\boldsymbol{\theta}}) = \frac{1}{a}\sum_{k=1}^{m} |\theta_k|^a$ for some $a \ge 1$. Some typical choices are the l_1 norm $J(p_{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}\|_1$ (lasso) and the squared l_2 norm $J(p_{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}\|_2^2$ (ridge regression).
2. Reproducing kernel Hilbert space (RKHS): the model space is the RKHS generated by a given kernel K, $\mathcal{P}_{\text{RKHS}} = l^{-1}(\mathcal{H}_K)$, and the penalty is the corresponding norm of f, $J(p(\cdot)) = \|f\|_{\mathcal{H}_K}$.
3. Boosted trees: the model space is the additive trees: $\mathcal{P}_{k\text{-boost}} = \{p(\boldsymbol{x}) = l^{-1}(f(\boldsymbol{x})) : f = f_1 + f_2 + \cdots + f_m : f_k \in \mathcal{F}_{d\text{-tree}}, k = 1, 2, \ldots\}$, where $\mathcal{F}_{d\text{-tree}}$ is the space of step functions in the classification and regression tree (CART) with depth at most d [4]. This space is quite large and approximate fitting algorithms (boosting) must be used. There is no exact penalty function, but as noticed by Friedman, Hastie and Tibshirani [14] and illustrated later, boosting is closely related to the lasso penalty in regularized regression.

Since all the penalty terms considered here are convex, the regularized optimization problems can be solved very efficiently.

4.3. *RKHS regression*. Next we elaborate on the RKHS and boosting approaches since they might be foreign to researchers in causal inference. RKHS regression is a popular nonparametric method in machine learning that essentially extends the regularized GLM with ridge penalty to an infinite dimensional space [25, 48]. Let $\mathcal{H}_K$ be the RKHS generated by the kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which describes similarity between two vectors of pretreatment covariates. The RKHS model is most easily understood through the "feature map" interpretation. Suppose that K has an eigen-expansion $K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=1}^{\infty} c_k \phi_k(\boldsymbol{x}) \phi_k(\boldsymbol{x}')$ with $c_k \geq 0$, $\sum_{k=1}^{\infty} c_k^2 < \infty$. Elements of $\mathcal{H}_K$ have a series expansion

$$f(\boldsymbol{x}) = \sum_{k=1}^{\infty} \theta_k \phi_k(\boldsymbol{x}), \qquad \|f\|_{\mathcal{H}_K}^2 = \sum_{k=1}^{\infty} \theta_k^2 / c_k.$$

The eigenfunctions $\{\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots\}$ can be viewed as new regressors generated by the low-dimensional covariates $\boldsymbol{X}$. The standard generalized linear model (4) corresponds to a finite-dimensional linear reproducing kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=1}^{m} \phi_k(\boldsymbol{x}) \phi_k(\boldsymbol{x}')$, but in general the eigenfunctions (i.e., predictors) $\{\phi_k\}_{k=1}^{\infty}$ can be infinite-dimensional.

Although the parameter θ is potentially infinite-dimensional, the numerical problem (14) is computationally feasible via the "kernel trick" if the penalty is a function of the RKHS norm of $f(\cdot)$. The representer theorem (cf. [48]) states that the solution is indeed finite-dimensional and has the form $\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \hat{\gamma}_i K(\boldsymbol{x}, \boldsymbol{X}_i)$. Consequently, the optimization problem (14) can be solved with the n-dimension parameter vector $\boldsymbol{\gamma}$.

As a remark, the idea of using a kernel to describe similarity between covariate vectors is not entirely new to observational studies. However, most of the previous literature (e.g., [22]) uses kernel as a smoothing technique for propensity score estimation (similar to kernel density estimation) rather than generating a RKHS, although the kernel functions can be the same in principle.

4.4. *Boosting*. Boosting (particularly gradient boosting) can be viewed as a greedy algorithm of function approximation [15]. Let $\hat{f}$ be the current guess of f, then the next guess is given by the steepest gradient descent $\hat{f}_{\text{new}} = \hat{f} + \hat{\eta}\hat{h}$, where

$$\hat{h} = \arg \max_{h \in \mathcal{F}_{k\text{-tree}}} \frac{\partial}{\partial \eta} \frac{1}{n} \sum_{i=1}^{n} S_{\alpha,\beta}\big(l^{-1}\big(\hat{f}(\boldsymbol{X}_i) + \eta h(\boldsymbol{X}_i)\big), T_i\big) \quad \text{and} \tag{15}$$

$$\hat{\eta} = \arg \max_{\eta \geq 0} \frac{1}{n} \sum_{i=1}^{n} S_{\alpha,\beta}\big(l^{-1}\big(\hat{f}(\boldsymbol{X}_i) + \eta \hat{h}(\boldsymbol{X}_i)\big), T_i\big). \tag{16}$$

When using gradient boosting in predictive modeling, a practical advice is to not go fully along the gradient direction as it easily overfits the model. Friedman [15] introduced an tuning parameter $\nu > 0$ (usually much less than 1) and proposed to shrink each gradient update: $\hat{f}_{\text{new}} = \hat{f} + \nu \hat{\eta} \hat{h}$. Heuristically, this can be compared

with the difference between the forward stepwise regression which commits to the selected variables fully and the least angle regression or the lasso regression which moves an infinitesimal step forward each time [12]. We shall see in the next section that, in the context of propensity score estimation, boosting and lasso regression also share a similar dual interpretation.

4.5. *Adjustment by outcome regression*. So far we have only considered design-based estimators by building a propensity score model to weight the observations. Such estimators do not attempt to build models for the potential outcomes, $Y(0)$ and $Y(1)$. Design-based inference is arguably more straightforward as it attempts to mimic a randomized experiment by observation data. Nevertheless, in some applications it is reasonable to improve estimation accuracy by fitting outcome regression models.

Here we describe the augmented inverse probability weighting (AIPW) estimators of ATT and ATE. Let $g_0(\boldsymbol{X}) = \mathbb{E}[Y(0)|\boldsymbol{X}]$ and $g_1(\boldsymbol{X}) = \mathbb{E}[Y(1)|\boldsymbol{X}]$ be the true regression functions of the potential outcomes and $\hat{g}_0$ and $\hat{g}_1$ be the corresponding estimates. Let w^{ATT} and w^{ATE} be the weights obtained by maximizing CBSR ($S_{0,-1}$ for ATT and $S_{-1,-1}$ for ATE) with any of the above adaptive strategies. The AIPW estimators [38] are

$$\hat{\tau}^{\mathrm{ATT}}_{\mathrm{AIPW}} = \sum_{T_i=1} w_i^{\mathrm{ATT}}\big(Y_i - \hat{g}_0(\boldsymbol{X}_i)\big) - \sum_{T_i=0} w_i^{\mathrm{ATT}}\big(Y_i - \hat{g}_0(\boldsymbol{X}_i)\big) \quad \text{and}$$

$$\hat{\tau}^{\mathrm{ATE}}_{\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{g}_1(\boldsymbol{X}_i) - \hat{g}_0(\boldsymbol{X}_i)\big) + \sum_{T_i=1} w_i^{\mathrm{ATE}}\big(Y_i - \hat{g}_1(\boldsymbol{X}_i)\big) - \sum_{T_i=0} w_i^{\mathrm{ATE}}\big(Y_i - \hat{g}_0(\boldsymbol{X}_i)\big).$$

We will compare IPW and AIPW estimators in the numerical examples in Section 6.

5. Theoretical aspects. We have proposed a very general framework and several flexible methods to estimate the propensity score. Several important questions are left unsettled: if different loss functions are asymptotically equivalent as indicated by Theorem 2, why should we use the tailored loss functions in this paper (or any method listed in Section 1.2)? How should we choose among the adaptive strategies in Section 4? What is the bias-variance tradeoff in regularizing the propensity score model and how should we choose the regularization parameter λ in equation (14)? After fitting a propensity score model, how do we construct a confidence interval for the target parameter? This section addresses these questions through investigating the Lagrangian dual of the propensity score estimation problem.

5.1. *Covariate imbalance and bias.* As in Section 4.5, denote the true outcome regression functions by $g_t(\boldsymbol{X}) = \mathbb{E}[Y(t)|\boldsymbol{X}]$, $t = 0, 1$. Except for ATT, in this section we will only consider bias under the constant treatment effect model that $g_1(\boldsymbol{x}) = g_0(\boldsymbol{x}) + \tau^*$ for all $\boldsymbol{x}$. By definition, τ^* is also the (normalized) weighted average treatment effects.

Suppose $g_0(\boldsymbol{x})$ has the expansion $g_0(\boldsymbol{x}) = \sum_{k=1}^{\infty} \beta_k \phi_k(\boldsymbol{x})$ for some functions $\{\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots\}$. Let $\varepsilon_i = Y_i - g_{T_i}(\boldsymbol{X}_i)$ so $\mathbb{E}[\varepsilon_i|T_i, \boldsymbol{X}_i] = 0$. Given any weighting function $w(\boldsymbol{x}, t)$, $t = 0, 1$ on the sample [e.g., equations (11) and (12) with estimated propensity score] and denote $w_i = w(\boldsymbol{X}_i, T_i)$, the IPW-type estimator $\hat{\tau}^*$ defined in (10) has the decomposition

$$
\begin{aligned}
\hat{\tau}^* &= \tau^* + \left[\sum_{T_i=1} w_i g_0(\boldsymbol{X}_i) - \sum_{T_i=0} w_i g_0(\boldsymbol{X}_i)\right] + \left[\sum_{T_i=1} w_i \varepsilon_i - \sum_{T_i=0} w_i \varepsilon_i\right] \\
&= \sum_{k=1}^{\infty} \beta_k \cdot \left[\sum_{T_i=1} w_i \phi_k(\boldsymbol{X}_i) - \sum_{T_i=0} w_i \phi_k(\boldsymbol{X}_i)\right] + \left[\sum_{T_i=1} w_i \varepsilon_i - \sum_{T_i=0} w_i \varepsilon_i\right].
\end{aligned} \tag{17}
$$

The second term always has mean 0, so the bias of $\hat{\tau}^*$ is given by the first term (a fixed quantity conditional on $\{T_i, \boldsymbol{X}_i\}_{i=1}^n$), which is just the imbalance with respect to the covariate function $g_0(\boldsymbol{x})$. The second line decomposes the bias into the imbalance with respect to the basis functions $\{\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots\}$.

Equation (17) highlights the importance of covariate balance in reducing the bias of $\hat{\tau}^*$, especially if the propensity score model is misspecified. If the propensity score is modeled by a GLM with fixed regressors $\phi(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x}))$ and fitted by maximizing CBSR as in (6), an immediate corollary is the following.

THEOREM 3. *Under Assumption 1 and constant treatment effect that* $g_1(\boldsymbol{x}) \equiv g_0(\boldsymbol{x}) + \tau^*$ *for all* $\boldsymbol{x}$, *the estimator* $\hat{\tau}^*$ *obtained by maximizing CBSR with regressors* $\phi(\boldsymbol{x})$ *that include an intercept term is asymptotically unbiased if* $g_0(\boldsymbol{x})$ *is in the linear span of* $\{\phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x})\}$, *or more generally if* $\inf_{\boldsymbol{\eta}} \|g_0(\boldsymbol{x}) - \boldsymbol{\eta}^T \phi_m(\boldsymbol{x})\|_\infty \to 0$ *as* $n, m(n) \to \infty$.

The last condition says that $g_0(\boldsymbol{x})$ can be uniformly approximated by functions in the linear span of $\phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x})$ as $m \to \infty$. This holds under very mild assumption of g_0. For example, if the support of $\boldsymbol{X}$ is compact and $g_0(\boldsymbol{x})$ is continuous, the Weierstrass approximation theorem ensures that $g_0(\boldsymbol{x})$ can be uniformly approximated by polynomials.

Finally, we compare the results in Theorems 2 and 3. The main difference is that Theorem 2 uses *propensity score* models with increasing complexity, whereas Proposition 3 assumes uniform approximation for the *outcome regression* function. Since the unbiasedness in Proposition 3 does not make any assumption on the propensity score, the estimator $\hat{\tau}^*$ obtained by maximizing CBSR is more robust to the misspecified or overfitted propensity score model.

5.2. *Lagrangian duality.* In Section 1.2, we mentioned that the recently proposed "moment condition" approaches (e.g., [27]) and the "empirical balancing" approaches (e.g., [53]) can be unified under the framework proposed in this paper. We now elucidate this equivalence by exploring the Lagrangian dual of maximizing CBSR. First let us rewrite the score optimization problem (6) by introducing new variables f_i for each observation i:

$$\begin{aligned} &\underset{f,\theta}{\text{maximize}} && \frac{1}{n}\sum_{i=1}^{n} S\big(l^{-1}(f_i), T_i\big) \\ &\text{subject to} && f_i = \boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{X}_i), \qquad i = 1, \dots, n. \end{aligned} \tag{18}$$

Let the Lagrangian multiplier associated with the ith constraint be $(2T_i - 1)w_i/n$. By setting the partial derivatives of the Lagrangian equal to 0, we obtain

$$\frac{\partial Lag}{\partial \theta_k} = \frac{1}{n}\sum_{i=1}^{n}(2T_i - 1)w_i\phi_k(\boldsymbol{X}_i) = 0, \qquad k = 1, \dots, m, \tag{19}$$

$$\frac{\partial Lag}{\partial f_i} = \frac{1}{n}\left(\frac{\partial S(l^{-1}(f_i), T_i)}{\partial f_i} + (2T_i - 1)w_i\right) = 0, \qquad i = 1, \dots, n. \tag{20}$$

Equation (19) is the same as (13), meaning the optimal dual variables $\boldsymbol{w}$ balance the predictors $\phi_1, \dots, \phi_m$. Equation (20) determines $\boldsymbol{w}$ from f. By using the fact (2) and the scoring rule is CBSR, it turns out that $w_i = w(\boldsymbol{X}_i, T_i)$ is exactly the weights defined in (9), that is, the weights $\boldsymbol{w}$ in our estimator $\hat{\boldsymbol{\tau}}$ are dual variables of the CBSR-maximization problem (6).

Next we write down the Lagrangian dual problem of (18). In general, there is no explicit form because it is difficult to invert (9). However, in the particularly interesting cases $\alpha = 0$, $\beta = -1$ (ATT) and $\alpha = -1$, $\beta = -1$ (ATE), the dual problems are algebraically tractable. When $\alpha = 0$, $\beta = -1$ and if an intercept term is included in the GLM, (an equivalent form of) the dual problem is given by

$$\begin{aligned} &\underset{\boldsymbol{w}\geq 0}{\text{minimize}} && \sum_{i:T_i=0} w_i \log w_i \\ &\text{subject to} && \sum_{i:T_i=0} w_i\phi_k(\boldsymbol{X}_i) = \sum_{j:T_j=1} \phi_k(\boldsymbol{X}_j), \qquad k = 1, \dots, m. \end{aligned} \tag{21}$$

When $\alpha = \beta = -1$, the inverse probability weights are always greater than 1. The Lagrangian dual problem in this case is given by

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n}(w_i - 1)\log(w_i - 1) - w_i \\ &\text{subject to} && \sum_{i:T_i=0} w_i\phi_k(\boldsymbol{X}_i) = \sum_{j:T_j=1} w_j\phi_k(\boldsymbol{X}_j), \qquad k = 1, \dots, m. \\ & && w_i \geq 1, \qquad i = 1, \dots, n. \end{aligned} \tag{22}$$

The objective functions in (21) and (22) encourage the weights w to be close to uniform. They belong to a general distance measure $\sum_{i=1}^{n} D(w_i, v_i)$ in Deville and Särndal [11], where $D(\boldsymbol{w}, \boldsymbol{v})$ is a continuously differentiable and strongly convex function in $\boldsymbol{w}$ and achieves its minimum at $\boldsymbol{w} = \boldsymbol{v}$. When the estimand is ATT (or ATE), the target weight $\boldsymbol{v}$ is equal to 1 (or 2). Estimators of this kind are often called "calibration estimators" in survey sampling because the weighted sample averages are empirically calibrated to some known population averages.

All the previously proposed "empirical balancing" methods operate by solving convex optimization problems similar to (21) or (22). The maximum entropy problem (21) appeared first in Hainmueller [19] to estimate ATT and is called "entropy balancing." Zhao and Percival [52] used the primal-dual connection described above to show Entropy Balancing is doubly robust, a stronger property than Theorems 2 and 3. Unfortunately, the double robustness property does not extend to other estimands. Chan, Yam and Zhang [7] studied the calibration estimators with the general distance D and showed the estimator $\hat{\tau}$ is globally semiparametric efficient. When the estimand is ATE, Chan, Yam and Zhang [7] require the weighted sums of ϕ_k in (22) to be calibrated to $\sum_{i=1}^{n} \phi_k(\boldsymbol{X}_i)/n$ also. In view of Theorem 2, this extra calibration is not necessary for semiparametric efficiency. In an extension to entropy balancing, Hazlett [21] proposed to empirically balance kernel representers instead of fixed regressors. This corresponds to unregularized ($\lambda = 0$) RKHS regression introduced in Section 4.3. The unregularized problem is in general unfeasible, so Hazlett [21] had to tweak the objective to find a usable solution. Zubizarreta [53] proposed to solve a problem similar to (22) (the objective is replaced by the coefficient of variation of w and the exact balancing constraints are relaxed to tolerance level). Since that problem corresponds to use the unconventional link function $l(p) = 1/p$, Zubizarreta [53] needed to include the additional constraint that w is nonnegative to avoid model extrapolation.

5.3. *Inexact balance, multivariate two-sample test and bias-variance tradeoff.* When the CBSR maximization problem is regularized as in (14), its dual objective functions in (21) and (22) remain unchanged, but the covariate balancing constraints are no longer exact. Consider the regularized GLM approach in Section 4.2 with $J(p_{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}\|_a^a/a$ for some $a \geq 1$, the dual constraints are given by

$$\begin{aligned} &\left| \sum_{T_i=1} w_{\hat{\boldsymbol{\theta}}_\lambda}(\boldsymbol{X}_i, T_i)\phi_k(\boldsymbol{X}_i) - \sum_{T_i=0} w_{\hat{\boldsymbol{\theta}}_\lambda}(\boldsymbol{X}_i, T_i)\phi_k(\boldsymbol{X}_i) \right| \\ &\qquad \leq \lambda \cdot \left|(\hat{\boldsymbol{\theta}}_\lambda)_k\right|^{a-1}, \qquad k = 1, \ldots, m. \end{aligned} \tag{23}$$

The equality in (23) holds if $(\hat{\boldsymbol{\theta}}_\lambda)_k \neq 0$, which is generally true unless $a = 1$.

Following Section 5.1, if we assume constant treatment effect $\mathbb{E}[Y(1)|\boldsymbol{X}] \equiv \mathbb{E}[Y(0)|\boldsymbol{X}] + \tau^*$ and the outcome regression function is in the linear span of the

regressors $g_0(\boldsymbol{x}) = g_\beta(\boldsymbol{x}) = \sum_{k=1}^m \beta_k \phi_k(\boldsymbol{x})$, then the absolute bias of $\hat{\tau}_\lambda^*$ is

$$\left|\sum_{k=1}^{\infty} \beta_k \cdot \left[\sum_{T_i=1} w_i \phi_k(\boldsymbol{X}_i) - \sum_{T_i=0} w_i \phi_k(\boldsymbol{X}_i)\right]\right| \le \lambda \sum_{k=1}^{\infty} |\beta_k| \cdot |(\hat{\boldsymbol{\theta}}_\lambda)_k|^{a-1}$$
$$\le \lambda \|\beta\|_a \|\hat{\boldsymbol{\theta}}_\lambda\|_a^{a-1}.$$

The last inequality is due to Hölder's inequality and is tight. In other words, the dual constraints imply that

$$\sup_{\|\boldsymbol{\beta}\|_a \le 1} \left|\operatorname{bias}\left(\hat{\tau}_\lambda^*, g_{\boldsymbol{\beta}}\right)\right| = \lambda \|\hat{\boldsymbol{\theta}}_\lambda\|_a^{a-1}, \qquad a \ge 1. \tag{24}$$

The next proposition states that the right-hand side of the last equation is decreasing as the degree of regularization λ becomes smaller. This is consistent with the heuristic that the more we regularize the propensity score model, the more bias our estimator is.

PROPOSITION 3. *Given a strictly proper scoring rule S and a link function l such that $S(l^{-1}(f), t)$ is strongly concave and second-order differentiable in $f \in \mathbb{R}$ for $t = 0, 1$, let $\hat{\boldsymbol{\theta}}_\lambda$ be the solution to (14) with $J(p_{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}\|_a^a / a$ for a given $a \ge 1$. Then $\lambda \|\hat{\boldsymbol{\theta}}_\lambda\|_a^{a-1}$ is a strictly increasing function of $\lambda > 0$.*

The Lagrangian dual problems (21) and (22) highlight the bias-variance tradeoff when using CBSR to estimate the propensity score. The dual objective function measures the uniformity of w (closely related to the variance of $\hat{\tau}^*$) and the dual constraints bound the covariate imbalance of w [the minimax bias of $\hat{\tau}^*$ for $g_0(\boldsymbol{x}) = g_\eta(\boldsymbol{x}) = \sum_{k=1}^m \eta_k \phi_k(\boldsymbol{x})$ given $\|\boldsymbol{\eta}\|_a$]. The penalty parameter λ regulates this bias-variance tradeoff. When $\lambda \to 0$, the solution of (14) converges to the weights $\boldsymbol{w}$ that minimizes the $a/(a-1)$-norm of covariate imbalance. The limit of $r(\lambda)$ when $\lambda \to 0$ can be 0 or some positive value, depending on if the unregularized score maximization problem (6) is feasible or not. When $\lambda \to \infty$, the solution of (14) converges to uniform weights (i.e., no adjustment at all) whose estimator $\hat{\tau}^*$ has the smallest variance.

A particularly interesting case is the lasso penalty $J(p_{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}\|_1$. By (23), the maximum covariate imbalance is bounded by λ. Therefore, the approximate balancing weights proposed by Zubizarreta [53], Wang and Zubizarreta [49] can be viewed as putting weighted lasso penalty in propensity score estimation. Bounding the maximum covariate imbalance can be useful when the dimension of $\boldsymbol{X}$ is high; see Athey et al. [2].

The RKHS regression in Section 4.3 is a generalization to the regularized regression with potentially infinite-dimensional predictors and weighted l_2-norm penalty. The maximum bias under the sharp null is given by

$$\sup_{\|g_0\|_{\mathcal{H}_K} \le 1} \left|\operatorname{bias}\left(\hat{\tau}_\lambda^*, g_0\right)\right| \le \lambda \|\hat{f}_\lambda\|_{\mathcal{H}_K}. \tag{25}$$

The boosted trees in Section 4.4 does not have a dual problem since it is solved by a greedy algorithm. However, it shares a similar interpretation with the lasso-regularized GLM. With some algebra, the gradient direction in (15) can be shown to be

$$\hat{h} \propto \arg \max_{h \in \mathcal{F}_{k\text{-tree}}} \left[\sum_{T_i=1} w_i h(\boldsymbol{X}_i) - \sum_{T_i=0} w_i h(\boldsymbol{X}_i) \right].$$

That is, $\hat{h}$ is currently the most imbalanced k-tree. By taking a small gradient step in the direction $\hat{h}$, it reduces the covariate imbalance (bias) in this direction the fastest among all k-trees. To see this, when $k = 1$, maximum covariate imbalance among 1-trees is essentially the largest univariate Kolmogorov–Smirnov statistics. We illustrate this interpretation of boosting using the toy example in Section 1.1. Figure 2 plots the paths of Kolmogorov–Smirnov statistics as more trees are added to the propensity score model (the step size is $\nu = 0.1$). The behavior is similar to the lasso regularized CBSR-maximization (23) which reduces the largest univariate imbalance (instead of the largest Kolmogorov–Smirnov statistic).

As a final remark, the left-hand side of (24) or (25) indeed defines a distance metric between two probability distributions (the empirical distributions of the covariates over treatment and control). This distance is called *integral probability metric* [36] and has received increasing attention recently in the two-sample testing literature. In particular, a very successful multivariate two-sample test [18]

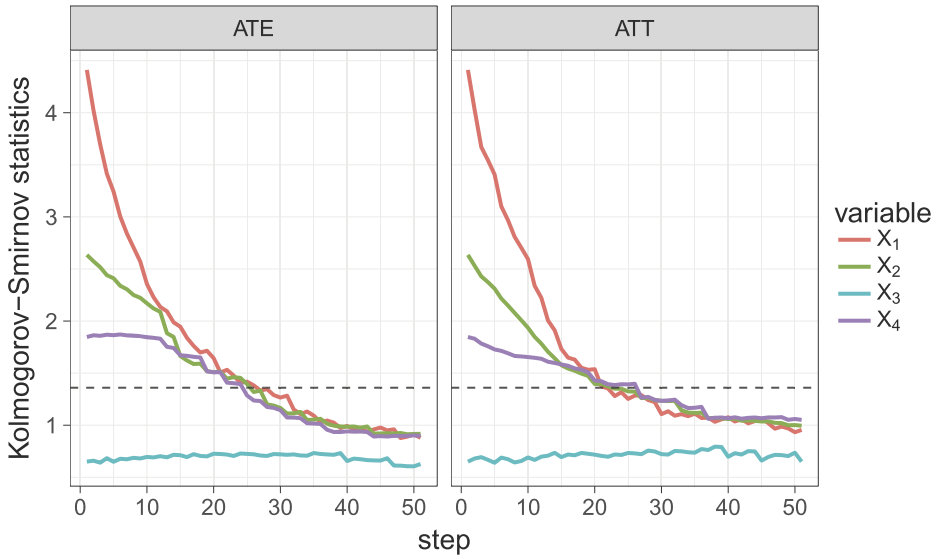

FIG. 2. *Boosting with* 1*-level trees is reducing the maximum Kolmogorov–Smirnov statistics. Two estimands* (*ATT and ATE*) *and their corresponding CBSR are considered. The dashed line is the upper* 0.05 *quantile of the asymptotic null distribution of the K-S statistic.*

uses the left-hand side of (25) as its test statistic. Here we have given an alternative statistical motivation of considering the integral probability metric.

5.4. *A Bayesian interpretation.* Besides the maximum bias interpretation (25), the RKHS model in Section 4.3 has another interesting Bayesian interpretation. Suppose the regression function g_0 is also random and generated from a Gaussian random field prior $g_0(\cdot) \sim \mathcal{G}(0, K)$ with mean function 0 and covariance function $K(\cdot, \cdot)$. Then the design MSE of $\hat{\tau}^*$ (conditional on $\{\boldsymbol{X}_i, T_i\}_{i=1}^n$) under constant treatment effect is given by

$$\begin{aligned}&\mathbb{E}_g\Bigg[\sum_{T_i=1} w_i g_0(\boldsymbol{X}_i) - \sum_{T_i=0} w_i g_0(\boldsymbol{X}_i)\Bigg]^2 + \sum_{i=1}^n w_i^2 \operatorname{Var}(Y_i|\boldsymbol{X}_i, T_i)\\ &\quad = \tilde{\boldsymbol{w}}^T K \tilde{\boldsymbol{w}} + \sum_{i=1}^n w_i^2 \operatorname{Var}(Y_i|\boldsymbol{X}_i, T_i),\end{aligned}$$

where $\tilde{w}_i = (2T_i - 1)w_i$, $i = 1, \ldots, n$ and (with some abuse of notation) K is the sample covariance matrix $K_{ij} = K(\boldsymbol{X}_i, \boldsymbol{X}_j)$, $i, j = 1, \ldots, n$. This is directly tied to the dual problem of CBSR maximization. For example, when the estimand is ATT and the link is logistic, using the "kernel trick" $\hat{f}(\boldsymbol{X}_i) = \boldsymbol{K}\hat{\boldsymbol{\gamma}}$ described in Section 4.3 it is not difficult to show the dual problem minimizes $\lambda \tilde{\boldsymbol{w}}^T \boldsymbol{K} \tilde{\boldsymbol{w}} + \sum_{i=1}^n w_i \log w_i$. Choosing different penalty parameters λ essentially amounts to different prior beliefs about the conditional variance $\operatorname{Var}(Y|\boldsymbol{X}, T)$. We will explore this Bayesian interpretation in a simulation example in Section 6.

In practice, optimally choosing the regularization parameter λ is essentially difficult as it requires prior knowledge about $\|g_0\|_{\mathcal{H}_K}$ and the conditional variance of Y (essentially the signal-to-noise ratio). Such difficulty exists in all previous approaches and we only attempt to provide a reasonable solution here. Our experience with the adaptive procedures in Section 4 is that once λ is reasonably small, the further reduction of maximum bias by decreasing λ becomes negligible in most cases. Our best recommendation is to plot the curve of the maximum bias versus λ, and then the user should use her best judgment based on prior knowledge about the outcome regression. To mitigate the problem of choosing λ, next we describe how to make valid statistical inference with an arbitrarily chosen λ.

5.5. *Design-based finite-sample inference.* When the treatment effect is not homogeneous, the derivation above no longer holds in general, although the bias-variance tradeoff is still expected if the effect is not too inhomogeneous. One exception is when the estimand is ATT. In this case, if the weights are normalized so $w_i = 1$ if $T_i = 1$, the finite sample bias of $\hat{\tau}^*$ is

$$\begin{aligned}&\Bigg[\sum_{T_i=1} w_i g_1(\boldsymbol{X}_i) - \sum_{T_i=0} w_i g_0(\boldsymbol{X}_i)\Bigg] - \frac{1}{n_1}\Bigg[\sum_{T_i=1} g_1(\boldsymbol{X}_i) - g_0(\boldsymbol{X}_i)\Bigg]\\ &\quad = \Bigg[\sum_{T_i=1} w_i g_0(\boldsymbol{X}_i) - \sum_{T_i=0} w_i g_0(\boldsymbol{X}_i)\Bigg].\end{aligned}$$

Therefore, the bias of $\hat{\tau}^*$ is only determined by how well $\boldsymbol{w}$ balances g_0. This fact was noticed in Zhao and Percival [52], Athey et al. [2], Kallus [31] and will be used to construct honest confidence interval for the ATT.

To derive design-based inference of weighted ATE, we assume strong ignorability (Assumption 1) and $Y_i \sim \mathrm{N}(g_{T_i}(\boldsymbol{X}_i), \sigma^2)$. The normality assumption is not essential when sample size is large, but the homoscedastic assumption is more difficult to relax. We assume the treatment effect is constant if the estimand is not ATT. The only other assumption we make is

ASSUMPTION 2. $g_0(\boldsymbol{x})$ is in a known RKHS $\mathcal{H}_K$.

Let the basis function of $\mathcal{H}_K$ be $\{\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots\}$ and $g_0(\boldsymbol{x}) = \sum_{k=1}^{\infty} \beta_k \phi_k(\boldsymbol{x})$. Suppose the propensity score is estimated by the RKHS regression described in Section 4.3. Then by the decomposition (17) and equation (25),

$$\left|\hat{\tau}^* - \tau^*\right| \preceq \lambda \|g_0\|_{\mathcal{H}_K} \|\hat{f}_\lambda\|_{\mathcal{H}_K} + \mathrm{N}\left(0, \sigma^2 \sum_{i=1}^{n} w_i^2\right), \tag{26}$$

where $\preceq$ means stochastically smaller. Therefore, if we can find an upper-$(\alpha/2)$ confidence limit for $\|g_0\|_{\mathcal{H}_K}$ [denoted by $\mathrm{CL}(\|g_0\|_{\mathcal{H}_K}, 1-\alpha/2)$] and a good estimate of σ (denoted by $\hat{\sigma}$), then a $(1-\alpha)$-confidence interval of τ^* is given by

$$\hat{\tau}^* \pm \left[\lambda \|\hat{f}_\lambda\|_{\mathcal{H}_K} \cdot \mathrm{CL}\left(\|g_0\|_{\mathcal{H}_K}, 1-\alpha/2\right) + \hat{\sigma} \|\boldsymbol{w}\|_2 z_{1-\alpha/2}\right], \tag{27}$$

where $z_{1-\alpha/2}$ is the upper-$(\alpha/2)$ quantile of the standard normal distribution. This inferential method can be further extended when an outcome regression adjustment is used (see Section 4.5) by replacing g_0 with $g_0 - \hat{g}_0$. Notice that in this case $\hat{g}_0$ and $\|\hat{g}_0\|_{\mathcal{H}_K}$ should be estimated using independent sample in order to maintain validity of (27).

Note that our Assumption 2 also covers the setting where $\boldsymbol{X}$ is high dimensional $(d \gg n)$ and $g_0(\boldsymbol{x}) = \sum_{k=1}^{d} \beta_k x_k$. In this case, estimating $\|g_0\|_{\mathcal{H}_K} = \|\boldsymbol{\beta}\|_2^2$ is of high interest in genetic heritability and we shall use a recent proposal by Janson, Foygel Barber and Candès [30] in our numeric example below. Estimating $\|g_0\|_{\mathcal{H}_K}$ when $\boldsymbol{X}$ is low dimensional can be done in a similar manner by weighting the coefficients.

Athey et al. [2] considered the inference of ATT when $\boldsymbol{X}$ is high dimensional, but a crucial assumption they require is that $\boldsymbol{\beta}$ is a very sparse vector so that g_0 can be accurately estimated by lasso regression. In this case, the maximum bias $\lambda \|\hat{g}_0 - g_0\|_{\mathcal{H}_K} \|\hat{f}_\lambda\|_{\mathcal{H}_K}$ is negligible if λ is carefully chosen. Our confidence interval above does not require the sparsity assumption since the procedure in Janson, Foygel Barber and Candès [30] does not need sparsity. Balancing functions in a kernel space is also considered in Hazlett [21] and Kallus [31], but they did not consider the statistical inference of weighted average treatment effects.

5.6. *Choosing the tuning parameters*. To use our framework, the user needs to choose three different parameters:

1. The estimand, indexed by α and β.
2. The function class $\mathcal{F}$ used to model the propensity score. All the functions in $\mathcal{F}$ are (approximately) balanced in the sample.
3. The regularization parameter λ.

In general, the estimand should be chosen according to the application. ATE ($\alpha = \beta = -1$) and ATT ($\alpha = 0$, $\beta = -1$) are the most common estimands, and there are certain benefits of using OWATE ($\alpha = \beta = 0$) [10, 33].

The function class $\mathcal{F}$ should be chosen according to prior knowledge of the outcome regression function $g_0(\boldsymbol{x})$. When the covariate dimension is low, we recommend using a universal kernel (such as a Gaussian or Laplacian kernel, see, e.g., Gretton et al. [18]) to estimate the propensity score, so the model space is dense in the space of continuous functions. Proposition 3 shows that this results in estimators with low bias. When the covariate dimension is high, more assumptions on the outcome regression function (such as linearity and sparsity) are often necessary.

The regularization parameter λ controls the bias-variance tradeoff of the propensity score weighting estimator. As shown in Section 5.4, the optimal choice of λ hinges on the signal-to-noise ratio of the outcome regression. In general, this is a very challenging and mostly unsolved open problem [50]. Here we propose several ways to choose λ in practice:

1. Choose the largest λ such that satisfactory covariate balance is achieved (e.g., if the standardized differences of prespecified covariate functions are no more than 10%). It is also useful to compare the coefficient of variation of the weights $\boldsymbol{w}$ at λ and at $\lambda \to 0$.
2. Estimate the signal-to-noise ratio of the outcome regression and choose the λ that minimizes the maximal mean squared error derived from equation (26). Note that this requires estimator of $\|g\|_{\mathcal{H}_K}$ and σ^2 and the design-based inference in Section 5.5 is no longer guaranteed to be valid because it uses the outcome data.
3. Choose λ by cross-validation. Cross-validation is routinely used in statistical learning to avoid overfitting [20]. Typically, the user first splits the data into K nonoverlapping groups, then cyclically use $K - 1$ groups to train the statistical model and use the 1 group left to validate the model. In predictive tasks, it is common to select the λ which minimizes the average loss in the validation groups. In our case, the loss function is a surrogate for covariate imbalance and it is more reasonable to directly minimize the covariate imbalance in the validation groups. In the simulations below, we will use the λ that minimizes the average norm of the gradient of the tailored loss function evaluated at the validation samples.

6. Numerical examples. This section provides two simulation examples to demonstrate the flexibility of the proposed framework.

6.1. *Simulation*: *Low-dimensional covariates*. To illustrate the bias-variance tradeoff in selecting model space and regularization parameter, in the following simulation we use a random regression function instead of a manually selected regression function to generate outcome observations. This is motivated by the Bayesian interpretation in Section 5.3. We believe this novel simulation design also better reflects the philosophy of design-based causal inference—the weights generated by the estimated propensity score should be robust against any reasonable outcome regression function.

In this simulation, we consider propensity score models fitted using six kernels: the Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sigma\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$ with $\sigma = 0.1$ or 0.3, the Laplace kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sigma\|\boldsymbol{x} - \boldsymbol{x}'\|)$ with $\sigma = 0.1$ or 0.3 and the polynomial kernel $k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T x' + 0.5)^d$ with $d = 1$ or 3. The sample size is $n = 1000$ and the covariates are generated by $\boldsymbol{X}_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \boldsymbol{I}_5)$. The true propensity score is a random function generated by $\operatorname{logit}(\mathbb{P}(T = 1|\boldsymbol{X} = \boldsymbol{x})) = f(\boldsymbol{x}) \sim \mathcal{G}(0, K(\cdot, \cdot))$ where the covariance function is either the polynomial kernel with degree 1 or the Gaussian kernel with $\sigma = 0.1$. Potential outcomes are generated from the sharp null model $Y_i(0) = Y_i(1) = g_0(\boldsymbol{X}_i) + \varepsilon_i$ where $g_0(\boldsymbol{X}_i)$ is a random function generated by the same Gaussian process with any of the six considered kernels and $\varepsilon_i \sim \mathrm{N}(0, 1)$. Note that for Gaussian and Laplace kernels, smaller σ indicates smoother random functions. For the polynomial kernels, a randomly generated function is just a linear or cubic function with random coefficients.

Along the regularization path of the RKHS regression described in Section 4.3, we consider three rules of choosing λ:

1. Stop early: choose the largest λ such that the coefficient of variation of $\boldsymbol{w}_\lambda$ is greater than 0.5 times the largest coefficient of variation in the regularization path.
2. Stop late: choose the λ with the largest coefficient of variation.
3. Cross validation as described in Section 5.6.

Table 2 reports the root mean squared error (RMSE) and average absolute bias over 100 simulations of the IPW estimators under different simulation settings and propensity score models. Some observations from this table:

1. There is no uniformly best kernel in fitting the propensity score. In particular, polynomial kernels perform poorly when g_0 is not a polynomial. Laplace kernel with $\sigma = 0.1$ performs relatively well in most of the simulation settings.
2. Under all settings, the lowest bias is always achieved when the fitting kernel is the same kernel that generates the outcome regression function g_0. This is expected from our Bayesian interpretation in Section 5.4.
3. When the fitting kernel is chosen correctly, the cross-validation rule of selecting λ has almost the same bias as "stop late" and usually smaller RMSE. When the fitting kernel is different from the kernel that generates g_0, cross-validation sometimes has even smaller bias.

TABLE 2

Simulation: *low-dimensional covariates*. *Outcome regression function* g_0 *and logit of propensity score* f *are randomly generated from some kennels*. *Three stopping criterion are considered*: *early*, *late and cross-validation* (*C-valid*). *Reported numbers are* 100 *times root mean squared error and average absolute bias* (*in the bracket*) *over* 1000 *simulations*. *In each block*, *numbers are bolded if no other method has smaller RMSE and smaller bias*

		$f \sim$ poly($d = 1$)			$f \sim$ gau($\sigma = 0.1$)		
g_0	**Fitting kernel**	**Early**	**Late**	**C-valid**	**Early**	**Late**	**C-valid**
lap($\sigma = 0.1$)	lap($\sigma = 0.1$)	2.3 (1.5)	**1.9 (0.7)**	**1.9 (0.7)**	2.0 (1.2)	1.7 (0.6)	1.7 (0.6)
	lap($\sigma = 0.3$)	2.4 (1.7)	1.8 (0.7)	1.8 (0.7)	2.0 (1.3)	1.7 (0.6)	**1.6 (0.6)**
	poly($d = 1$)	2.4 (1.2)	4.2 (1.6)	3.9 (1.5)	2.2 (1.3)	2.5 (1.1)	2.4 (1.1)
	poly($d = 3$)	2.1 (1.0)	4.0 (1.2)	3.2 (1.1)	2.0 (0.9)	3.4 (1.1)	2.7 (0.9)
	gau($\sigma = 0.1$)	2.0 (1.0)	3.3 (0.9)	3.0 (0.8)	1.7 (0.8)	2.7 (0.8)	2.5 (0.7)
	gau($\sigma = 0.3$)	2.4 (1.5)	2.3 (0.9)	2.2 (0.9)	1.9 (1.1)	2.0 (0.7)	1.9 (0.7)
lap($\sigma = 0.3$)	lap($\sigma = 0.1$)	3.5 (2.8)	3.0 (1.8)	2.9 (1.8)	3.4 (2.6)	2.8 (1.6)	2.7 (1.6)
	lap($\sigma = 0.3$)	3.8 (3.2)	**2.8 (1.8)**	**2.8 (1.8)**	3.5 (2.7)	**2.7 (1.5)**	**2.6 (1.6)**
	poly($d = 1$)	3.7 (2.6)	6.4 (3.8)	6.0 (3.6)	3.7 (2.9)	4.0 (2.8)	4.0 (2.7)
	poly($d = 3$)	3.4 (2.4)	5.6 (3.3)	4.6 (2.8)	3.2 (2.1)	5.4 (3.0)	4.3 (2.4)
	gau($\sigma = 0.1$)	3.1 (2.1)	4.5 (2.5)	4.1 (2.4)	2.9 (2.0)	4.2 (2.2)	3.8 (2.0)
	gau($\sigma = 0.3$)	3.6 (2.9)	3.3 (2.1)	3.2 (2.1)	3.2 (2.4)	3.1 (1.7)	3.0 (1.7)
poly($d = 1$)	lap($\sigma = 0.1$)	7.9 (7.0)	**2.3 (1.0)**	2.3 (1.1)	5.6 (4.7)	1.6 (0.4)	1.6 (0.4)
	lap($\sigma = 0.3$)	9.2 (8.4)	2.8 (1.6)	2.9 (1.8)	6.3 (5.5)	1.9 (0.7)	1.9 (0.8)
	poly($d = 1$)	4.8 (3.6)	2.7 (0.1)	**2.5 (0.1)**	5.1 (4.2)	**1.4 (0.0)**	**1.4 (0.0)**
	poly($d = 3$)	3.8 (2.6)	3.3 (0.7)	2.8 (0.8)	2.8 (1.7)	3.0 (0.7)	2.6 (0.8)
	gau($\sigma = 0.1$)	4.9 (3.8)	3.2 (0.8)	3.0 (0.9)	2.9 (1.9)	2.5 (0.4)	2.3 (0.4)
	gau($\sigma = 0.3$)	9.0 (8.1)	4.6 (3.2)	4.7 (3.4)	5.7 (4.7)	3.0 (1.5)	3.0 (1.6)
poly($d = 3$)	lap($\sigma = 0.1$)	5.7 (4.8)	3.9 (2.5)	3.9 (2.5)	6.1 (5.1)	3.6 (2.5)	3.6 (2.5)
	lap($\sigma = 0.3$)	6.1 (5.2)	4.0 (2.6)	4.0 (2.7)	6.3 (5.4)	3.7 (2.5)	3.7 (2.6)
	poly($d = 1$)	5.8 (4.4)	8.0 (5.3)	7.8 (5.1)	6.6 (5.6)	7.9 (6.7)	7.8 (6.5)
	poly($d = 3$)	**3.7 (2.4)**	**4.0 (1.5)**	3.6 (1.6)	3.5 (2.3)	3.6 (1.2)	**3.1 (1.2)**
	gau($\sigma = 0.1$)	4.6 (3.4)	4.4 (2.3)	4.2 (2.3)	4.7 (3.6)	3.9 (1.9)	3.7 (1.9)
	gau($\sigma = 0.3$)	6.2 (5.3)	4.6 (3.1)	4.7 (3.2)	6.0 (5.0)	4.1 (2.6)	4.1 (2.7)
gau($\sigma = 0.1$)	lap($\sigma = 0.1$)	5.0 (4.1)	**2.1 (0.9)**	**2.1 (0.9)**	4.7 (3.9)	**2.1 (0.9)**	**2.1 (0.9)**
	lap($\sigma = 0.3$)	5.5 (4.7)	2.1 (1.0)	2.2 (1.1)	5.0 (4.1)	2.1 (1.0)	2.1 (1.0)
	poly($d = 1$)	4.1 (2.8)	4.7 (2.3)	4.4 (2.1)	4.9 (4.0)	4.0 (2.6)	3.9 (2.6)
	poly($d = 3$)	3.1 (2.0)	4.1 (1.4)	3.4 (1.3)	3.2 (2.2)	3.7 (1.3)	3.1 (1.3)
	gau($\sigma = 0.1$)	3.0 (2.0)	**2.8 (0.6)**	**2.6 (0.6)**	3.1 (2.1)	2.5 (0.6)	**2.3 (0.6)**
	gau($\sigma = 0.3$)	4.9 (4.1)	2.5 (1.2)	2.5 (1.3)	4.3 (3.4)	2.4 (1.0)	2.3 (1.1)
gau($\sigma = 0.3$)	lap($\sigma = 0.1$)	4.5 (3.8)	3.6 (2.3)	3.6 (2.3)	4.9 (4.4)	3.5 (2.4)	3.5 (2.4)
	lap($\sigma = 0.3$)	4.7 (4.0)	3.3 (2.2)	3.3 (2.2)	5.0 (4.5)	3.3 (2.3)	3.3 (2.3)
	poly($d = 1$)	5.2 (4.0)	8.9 (6.0)	8.6 (5.9)	5.5 (4.8)	6.3 (5.0)	6.1 (4.9)
	poly($d = 3$)	4.3 (3.1)	7.0 (4.1)	5.7 (3.5)	4.3 (3.5)	6.4 (3.9)	5.2 (3.4)
	gau($\sigma = 0.1$)	3.8 (2.7)	5.1 (2.6)	4.7 (2.5)	4.0 (3.2)	4.7 (2.7)	4.3 (2.6)
	gau($\sigma = 0.3$)	4.0 (3.2)	3.2 (1.8)	**3.1 (1.8)**	4.3 (3.7)	3.3 (2.0)	**3.2 (2.0)**

4. Surprisingly, the kernel used to generate f (logit of the true propensity score) does not alter the qualitative conclusions above. Even if the "correct" kernel is used to fit the propensity score model, there is no guarantee that this better reduces the average bias than other "incorrect" kernels. For example, when f is simulated from poly$(d = 1)$, that is, f is a random linear function, using the linear propensity score model performs poorly unless g_0 is also a linear function.

6.2. *Simulation*: *High-dimensional covariates*. In our second example, we consider the case that the covariates $\boldsymbol{X}$ are high dimensional. In this simulation, the sample size $n = 1000$ and $\boldsymbol{X}_i \in \mathbb{R}^{100} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$ where $\Sigma_{ij} = 0.5^{|i-j|}$. The true propensity score is $\operatorname{logit}(\mathbb{P}(T_i = 1|\boldsymbol{X}_i)) = \rho \boldsymbol{X}_i^T \theta$ where $\rho = 1$ or 2, θ is a 100-dimensional vector whose first s_t entries are $1/\sqrt{s_t}$ and the rest are zero, and $s_t = 5$ or 100. The potential outcomes are generated from the sharp null model $Y_i(0) = Y_i(1) = \boldsymbol{X}_i^T \boldsymbol{\beta} + \varepsilon_i$, where the first s_y entries of β are $1/\sqrt{s_y}$ and the rest are zero, $s_y = 5, 20$ or 100, and ε_i is an independent Gaussian noise with standard deviation $\sigma = 5$.

In this simulation, the propensity score model is fitted by maximizing the CBSR corresponding to ATT ($\alpha = 0$, $\beta = -1$) with ridge penalty. The regularization parameter λ is chosen so that the coefficient of variation of the weights is just below 1. Three estimators are considered: the weighted difference estimator with no outcome adjustment (IPW), outcome adjustment fitted by the lasso (AIPW-L) and outcome adjustment fitted by the ridge regression (AIPW-R). The outcome regressions, either fitted by the lasso or the ridge penalty, are tuned by cross-validation.

Averaging over 1000 simulations, we report in Table 3 the root-mean-square error of the estimators (RMSE), the absolute bias (Bias), the estimated maximum bias as described in Section 5.5 which uses the EigenPrism method of Janson, Foygel Barber and Candès [30] to estimate $\|\boldsymbol{\beta}\|_2$ (Max Bias), coverage of the 95% confidence interval ignoring covariate imbalance as in Athey et al. [2] (CI), coverage of the honest 95%-confidence interval (27) (Honest CI) and the ratio of the length between the two confidence intervals (CI Ratio).

As shown in Table 3, outcome regression adjustment often improves estimation accuracy substantially. As expected, when $\boldsymbol{\theta}$ is dense ridge outcome regression performs better and when $\boldsymbol{\theta}$ is sparse lasso outcome regression performs better. In many settings, the actual bias is a substantial portion of the estimated maximum bias. Ignoring this bias in the construction of confidence interval can lead to serious undercoverage of the causal parameter, as indicated by the CI column in Table 3. Note that the sparsity assumption in Athey et al. [2] requires $s_y \ll \sqrt{n}/\log(d) \approx 6.9$, so the lack of coverage does not violate the theoretical results in Athey et al. [2] as the smallest s_y in this simulation is 5. Using the honest confidence interval derived in Section 5.5 ensures the desired coverage, although the confidence interval is about two to three times as wide and quite conservative as expected.

TABLE 3

Simulation: high-dimensional covariates. Reported values are the average RMSE, average absolute bias, average estimated maximum bias, coverage of the confidence interval ignoring bias due to inexact balance, coverage of the honest confidence interval proposed in Section 5.5 and the average ratio between the two confidence intervals over 1000 *simulations*

s_y	s_t	ρ	Method	RMSE	Bias	Max bias	CI	Honest CI	CI ratio
5	5	1	IPW	0.22	0.16	0.27	0.82	1	2.28
			AIPW-L	0.16	0.06	0.14	0.92	1	1.92
			AIPW-R	0.19	0.11	0.21	0.87	1	2.15
		2	IPW	0.48	0.45	0.56	0.24	1	3.25
			AIPW-L	0.29	0.22	0.33	0.72	1	2.71
			AIPW-R	0.41	0.36	0.42	0.40	1	3.02
	100	1	IPW	0.15	0.03	0.27	0.94	1	2.38
			AIPW-L	0.14	0.01	0.14	0.95	1	1.95
			AIPW-R	0.15	0.02	0.21	0.94	1	2.22
		2	IPW	0.18	0.10	0.59	0.89	1	3.60
			AIPW-L	0.16	0.03	0.31	0.93	1	2.73
			AIPW-R	0.18	0.07	0.44	0.90	1	3.27
20	5	1	IPW	0.17	0.08	0.28	0.90	1	2.38
			AIPW-L	0.16	0.05	0.22	0.91	1	2.24
			AIPW-R	0.16	0.06	0.21	0.92	1	2.23
		2	IPW	0.27	0.22	0.58	0.68	1	3.53
			AIPW-L	0.24	0.17	0.45	0.77	1	3.25
			AIPW-R	0.23	0.17	0.44	0.78	1	3.23
	100	1	IPW	0.16	0.07	0.28	0.91	1	2.38
			AIPW-L	0.16	0.05	0.21	0.92	1	2.22
			AIPW-R	0.16	0.06	0.21	0.92	1	2.21
		2	IPW	0.25	0.19	0.59	0.74	1	3.56
			AIPW-L	0.20	0.12	0.46	0.85	1	3.28
			AIPW-R	0.22	0.14	0.45	0.82	1	3.24
100	5	1	IPW	0.15	0.04	0.27	0.93	1	2.39
			AIPW-L	0.15	0.03	0.26	0.93	1	2.42
			AIPW-R	0.15	0.03	0.21	0.93	1	2.23
		2	IPW	0.18	0.09	0.58	0.90	1	3.60
			AIPW-L	0.18	0.08	0.55	0.89	1	3.63
			AIPW-R	0.18	0.07	0.45	0.91	1	3.27
	100	1	IPW	0.22	0.16	0.27	0.82	1	2.29
			AIPW-L	0.20	0.13	0.25	0.85	1	2.29
			AIPW-R	0.19	0.11	0.21	0.88	1	2.16
		2	IPW	0.48	0.45	0.59	0.23	1	3.29
			AIPW-L	0.45	0.41	0.48	0.33	1	3.16
			AIPW-R	0.41	0.37	0.43	0.43	1	3.01

7. Discussion. We have proposed a general method of obtaining covariate balancing propensity score which unifies many previous approaches. Our proposal is conceptually simple: the investigator just needs to tailor the loss function according to the link function and estimand. This offers great flexibility in incorporating adaptive strategies developed in statistical learning. We have given a through discourse on the dual interpretation of minimizing the tailored loss function, especially how regularization is linked to the bias-variance tradeoff in estimating the weighted average treatment effects. We provide honest inference that account for the bias incurred by inexact balance.

Throughout the paper we have taken an outright design perspective: without looking at the outcome data, the investigator tries to balance pre-treatment covariates as well as possible to mimic a randomized experiment, echoing the recommendations by Rubin [43, 44]. This allows us to give an interesting Bayesian interpretation of covariate balance: when checking covariate imbalance and deciding which propensity score model is "acceptable," the investigator is implicitly assuming a prior on the unknown outcome regression function.

Although the optimal choice of the regularization parameter λ depends on the signal-to-noise ratio of the outcome regression, simulation examples in Section 6 show that the λ selected by cross-validation usually results in a low biased estimator without overfitting the propensity score model. Thus we recommend selecting λ by cross-validation in practice. Moreover, when the covariate dimension is low we recommend using a universal kernel (particularly the Laplacian kernel) to estimate the propensity score. We encourage the user to try different kernels (e.g., Laplace kernel with different σ) as a secondary sensitivity analysis and report how the confidence interval changes with different modeling assumptions. When the covariate dimension is high, more assumptions on the outcome regression function $g_0(\boldsymbol{x})$ are necessary. One simplifying assumption is that $g_0(\boldsymbol{x})$ is linear, and it is tempting to further assume that $g_0(\boldsymbol{x})$ is sparse. However, our simulation results suggest that the user should be cautious about the inference assuming sparsity, as this modeling assumption essentially dismisses the bias due to inexact balance. Honest confidence interval can be constructed without assuming sparsity, but the interval can also be much wider if the covariate dimension is much larger than the sample size.

Acknowledgments. The author thanks Trevor Hastie, Hera Y. He, Dylan S. Small and three anonymous reviewers for valuable comments.

SUPPLEMENTARY MATERIAL

Supplement to "Covariate balancing propensity score by tailored loss functions" (DOI: 10.1214/18-AOS1698SUPP; .pdf). In this supplement we provide the detailed proof for the theoretical results and some graphical illustration of the Beta-family of scoring rules.

REFERENCES

[1] ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. MR2194325

[2] ATHEY, S., IMBENS, G. W., WAGER, S. et al. (2016). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 597–623.

[3] AUSTIN, P. C. and STUART, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat. Med.* **34** 3661–3679. MR3422140

[4] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

[5] BUJA, A., STUETZLE, W. and SHEN, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. Working Draft.

[6] CALIENDO, M. and KOPEINIG, S. (2008). Some practical guidance for the implementation of propensity score matching. *J. Econ. Surv.* **22** 31–72.

[7] CHAN, K. C. G., YAM, S. C. P. and ZHANG, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 673–700. MR3506798

[8] COCHRAN, W. G. (1953). Matching in analytical studies. *Am. J. Public Health Nation's Health* **43** 684–691.

[9] COCHRAN, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24** 295–313. MR0228136

[10] CRUMP, R., HOTZ, V. J., IMBENS, G. and MITNIK, O. (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical Report No. 330, National Bureau of Economic Research, Cambridge, MA.

[11] DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87** 376–382. MR1173804

[12] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

[13] FAN, J., IMAI, K., LIU, H., NING, Y. and YANG, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach. Technical Report, Princeton Univ., Princeton, NJ.

[14] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting. *Ann. Statist.* **28** 337–407. MR1790002

[15] FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328

[16] GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

[17] GRAHAM, B. S., DE XAVIER PINTO, C. C. and EGEL, D. (2012). Inverse probability tilting for moment condition model with missing data. *Rev. Econ. Stud.* **79** 1053–1079. MR2986390

[18] GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

[19] HAINMUELLER, J. (2011). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46. DOI:10.1093/pan/mpr025.

[20] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. MR2722294

[21] HAZLETT, C. (2016). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. Available at arXiv:1605.00155.

[22] HECKMAN, J. J., ICHIMURA, H. and TODD, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Rev. Econ. Stud.* **64** 605–654.

[23] HIRANO, K. and IMBENS, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Serv. Outcomes Res. Methodol.* **2** 259–278. DOI:10.1023/A:1020371312283.

[24] HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826

[25] HOFMANN, T., SCHÖLKOPF, B. and SMOLA, A. J. (2008). Kernel methods in machine learning. *Ann. Statist.* **36** 1171–1220. MR2418654

[26] IMAI, K., KING, G. and STUART, E. A. (2008). Misunderstanding between experimentalists and observationalists about causal inference. *J. Roy. Statist. Soc. Ser. A* **171** 481–502. MR2427345

[27] IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263. MR3153941

[28] IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

[29] IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951

[30] JANSON, L., FOYGEL BARBER, R. and CANDÈS, E. (2017). EigenPrism: Inference for high dimensional signal-to-noise ratios. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1037–1065. MR3689308

[31] KALLUS, N. (2016). Generalized optimal matching methods for causal inference. Available at arXiv:1612.08321.

[32] KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458

[33] LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2016). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. MR3803473

[34] LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **23** 2937–2960.

[35] MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London. MR3223057

[36] MÜLLER, A. (1997). Integral probability metrics and their generating classes of functions. *Adv. in Appl. Probab.* **29** 429–443. MR1450938

[37] NORMAND, S.-L. T., LANDRUM, M. B., GUADAGNOLI, E., AYANIAN, J. Z., RYAN, T. J., CLEARY, P. D. and MCNEIL, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *J. Clin. Epidemiol.* **54** 387–398.

[38] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

[39] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

[40] ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

[41] ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.
[42] RUBIN, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* 159–183.
[43] RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. MR2516795
[44] RUBIN, D. B. (2009). Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Stat. Med.* **28** 1420–1423. MR2724703
[45] SAVAGE, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783–801. MR0331571
[46] SMITH, J. A. and TODD, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *J. Econometrics* **125** 305–353. MR2143379
[47] STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812
[48] WAHBA, G. (1990). *Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442
[49] WANG, Y. and ZUBIZARRETA, J. R. (2017). Approximate balancing weights: Characterizations from a shrinkage estimation perspective. Available at arXiv:1705.00998.
[50] WONG, R. K. W. and CHAN, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika* **105** 199–213. MR3768874
[51] ZHAO, Q. (2018). Supplement to "Covariate balancing propensity score by tailored loss functions." DOI:10.1214/18-AOS1698SUPP.
[52] ZHAO, Q. and PERCIVAL, D. (2017). Entropy balancing is doubly robust. *J. Causal Inference* **5**.
[53] ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110** 910–922. MR3420672

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: qyzhao@wharton.upenn.edu