The Annals of Statistics
2020, Vol. 48, No. 2, 811–837
https://doi.org/10.1214/19-AOS1824
© Institute of Mathematical Statistics, 2020

MODEL-ASSISTED INFERENCE FOR TREATMENT EFFECTS USING REGULARIZED CALIBRATED ESTIMATION WITH HIGH-DIMENSIONAL DATA

BY ZHIQIANG TAN

Department of Statistics, Rutgers University, ztan@stat.rutgers.edu

Consider the problem of estimating average treatment effects when a large number of covariates are used to adjust for possible confounding through outcome regression and propensity score models. We develop new methods and theory to obtain not only doubly robust point estimators for average treatment effects, which remain consistent if either the propensity score model or the outcome regression model is correctly specified, but also model-assisted confidence intervals, which are valid when the propensity score model is correctly specified but the outcome model may be misspecified. With a linear outcome model, the confidence intervals are doubly robust, that is, being also valid when the outcome model is correctly specified but the propensity score model may be misspecified. Our methods involve regularized calibrated estimators with Lasso penalties but carefully chosen loss functions, for fitting propensity score and outcome regression models. We provide high-dimensional analysis to establish the desired properties of our methods under comparable sparsity conditions to previous results, which give valid confidence intervals when both the propensity score and outcome models are correctly specified. We present simulation studies and an empirical application which demonstrate advantages of the proposed methods compared with related methods based on regularized maximum likelihood estimation. The methods are implemented in the R package RCAL.

1. Introduction.

For observational studies, causal inference involves statistical modeling and estimation of population properties and associations from empirical data under structural assumptions (e.g., Tsiatis (2006)). In particular, as the main problem to be tackled in the paper, estimation of average treatment effects typically requires building and fitting outcome regression or propensity score models (e.g., Tan (2007)). The fitted outcome regression functions or propensity scores can then be used in various estimators for the average treatment effects, notably inverse probability weighted (IPW) estimators or augmented IPW estimators (Robins, Rotnitzky and Zhao (1994)).

A conventional approach for propensity score estimation (Rosenbaum and Rubin (1984)) involves fitting a propensity score model (often logistic regression) by maximum likelihood, check covariate balance, and then modify and refit the propensity score model until reasonable balance is achieved. However, this approach depends on ad hoc choices of what variables are included and whether nonlinear terms or interactions are used among others. The situation can be especially challenging when there are a large number of potentially confounding variables (or covariates) that need to be considered in outcome regression or propensity score models. In addition, another statistical issue is that uncertainty from the iterative process of model selection is complicated and often ignored in subsequent inference (i.e., confidence intervals or hypothesis testing) about treatment effects.

Received July 2018; revised January 2019.

MSC2010 subject classifications. Primary 62E20, 62F25, 62F35; secondary 62J05, 62J12.

Key words and phrases. Calibrated estimation, causal inference, doubly robust estimation, inverse probability weighting, Lasso penalty, model misspecification, propensity score, regularized M-estimation.

In this article, we develop new methods and theory for fitting logistic propensity score models and generalized linear outcome models, and then using the fitted values in augmented IPW estimators to estimate average treatment effects in high-dimensional settings where the number of covariate functions p is close to or even greater than the sample size n. There are two main elements in our approach. First, we employ regularized estimation with a Lasso penalty (Tibshirani (1996)) when fitting the outcome regression and propensity score models to deal with the large number of covariates under a sparsity assumption that only a small but unknown subset (relative to the sample size) of covariates are associated with nonzero coefficients in the propensity score and outcome regression models. Second, we carefully choose the loss functions for regularized estimation, different from least squares or maximum likelihood, such that the resulting augmented IPW estimator and Wald-type confidence intervals possess the following properties (G1) and at least one of (G2)–(G3) under suitable conditions:

(G1) The point estimator is doubly robust, that is, remains consistent if either the propensity score model or the outcome regression model is correctly specified.

(G2) The confidence intervals are valid if the propensity score model is correctly specified but the outcome regression model may be misspecified.

(G3) The confidence intervals are valid if the outcome regression model is correctly specified but the propensity score model may be misspecified.

If either property (G2) or (G3) is satisfied, then the confidence intervals are said to be model-assisted, borrowing the terminology from the survey literature (Särndal, Swensson and Wretman (1992)). If properties (G2)–(G3) are satisfied, then the confidence intervals are doubly robust.

Combining the two foregoing elements leads to a regularized calibrated estimator, denoted by $\hat{\gamma}^1_{\mathrm{RCAL}}$, for the coefficients in the propensity score model and a regularized weighted likelihood estimator, denoted by $\hat{\alpha}^1_{\mathrm{RWL}}$, for the coefficients in the outcome model within the treated subjects. See the loss functions in (12) and (14) or (53). The regularized calibrated estimator $\hat{\gamma}^1_{\mathrm{RCAL}}$ has recently been proposed in Tan (2017) as an alternative to the regularized maximum likelihood estimator for fitting logistic propensity score models, regardless of outcome regression models. As shown in Tan (2017), minimization of the underlying expected calibration loss implies reduction of not only the expected likelihood loss for logistic regression but also a measure of relative errors of limiting propensity scores that controls the mean squared errors of IPW estimators, when the propensity score model may be misspecified. In a complementary manner, our work here shows that $\hat{\gamma}^1_{\mathrm{RCAL}}$ can be used in conjunction with $\hat{\alpha}^1_{\mathrm{RWL}}$ to yield an augmented IPW estimator with valid confidence intervals if the propensity score model is correctly specified but the outcome regression model may be specified.

We provide high-dimensional analysis of the regularized weighted likelihood estimator $\hat{\alpha}^1_{\mathrm{RWL}}$ and the resulting augmented IPW estimator with possible model misspecification, while building on related results on convergence of $\hat{\gamma}^1_{\mathrm{RCAL}}$ to a target value $\bar{\gamma}^1_{\mathrm{CAL}}$ in Tan (2017). In particular, a new strategy is developed to tackle the technical issue that the weighted likelihood loss for $\hat{\alpha}^1_{\mathrm{RWL}}$ is defined depending on the estimator $\hat{\gamma}^1_{\mathrm{RCAL}}$. As a result, we obtain the convergence of $\hat{\alpha}^1_{\mathrm{RWL}}$ to a target value $\bar{\alpha}^1_{\mathrm{WL}}$ in the L_1 norm at the rate $(|S_\gamma| + |S_{\alpha^1}|)\{\log(p)/n\}^{1/2}$ and the associated Bregman divergence at the rate $(|S_\gamma| + |S_{\alpha^1}|)\log(p)/n$ under comparable conditions to those for high-dimensional analysis of standard Lasso estimators (e.g., Bühlmann and van de Geer (2011)), where $|S_\gamma|$ or $|S_{\alpha^1}|$ denotes the size of nonzero elements in $\bar{\gamma}^1_{\mathrm{CAL}}$ or respectively $\bar{\alpha}^1_{\mathrm{WL}}$. Furthermore, we establish an asymptotic expansion of the augmented IPW estimator based on $\hat{\gamma}^1_{\mathrm{RCAL}}$ and $\hat{\alpha}^1_{\mathrm{RWL}}$, and show that property (G1) is achieved provided $(|S_\gamma| + |S_{\alpha^1}|)(\log p)^{1/2} = o(n^{1/2})$ and property (G2) is achieved provided $(|S_\gamma| + |S_{\alpha^1}|)(\log p) = o(n^{1/2})$ with a nonlinear outcome model. With a linear outcome model, we obtain stronger results: property (G1) is achieved provided $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n)$ and both (G2) and (G3) are achieved provided $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$. These sparsity conditions are as weak as in previous works (e.g., Belloni, Chernozhukov and Hansen (2014); van de Geer et al. (2014)).

Related works. We compare and connect our work with related works in several areas. Nonpenalized calibrated estimation for propensity score models have been studied, sometimes independently (re)derived, in causal inference, missing-data problems and survey sampling (e.g., Folsom (1991); Tan (2010); Graham, De Xavier Pinto and Egel (2012); Hainmueller (2012); Imai and Ratkovic (2014); Kim and Haziza (2014); Vermeulen and Vansteelandt (2015); Chan, Yam and Zhang (2016)). The nonpenalized version of the estimator $\hat{\alpha}^1_{\mathrm{RWL}}$ for outcome regression models have also been proposed in Kim and Haziza (2014) and Vermeulen and Vansteelandt (2015), where one of the motivations is to circumvent the need of accounting for variation of such estimators of nuisance parameters, and hence simplify the computation of confidence intervals based on augmented IPW estimators. Our work generalizes these ideas to achieve statistical advantages in high-dimensional settings, where model-assisted or doubly robust confidence intervals would not be obtained without using regularized calibrated estimation. See Section 3.2 for further discussion.

For high-dimensional causal inference, Belloni, Chernozhukov and Hansen (2014) and Farrell (2015) employed augmented IPW estimators based on regularized maximum likelihood estimators in outcome regression and propensity score models, and obtained Wald-type confidence intervals that are valid when both the outcome regression and propensity score models are correctly specified, provided $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$ or refined rates depending on the product of $|S_\gamma|$ and $|S_{\alpha^1}|$; see Remark 10 later. Our main contribution is therefore to provide model-assisted or doubly robust confidence intervals for average treatment effects using differently configured augmented IPW estimators. Belloni, Chernozhukov and Hansen (2014) and Farrell (2015) also advocated post-Lasso refitting to potentially improve finite-sample performance.

Another related work is Athey, Imbens and Wager (2018), where valid confidence intervals are obtained for the sample treatment effect $n_1^{-1}\sum_{i:T_i=1}\{m_1^*(X_i) - m_0^*(X_i)\}$ (different from population treatment effects), if a linear outcome model is correctly specified. No propensity score model is explicitly used.

Our work is also connected to the literature of confidence intervals and hypothesis testing for a single or lower-dimensional coefficients in high-dimensional regression models (Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014)). Model-assisted inference does not seem to be addressed in these works, but can potentially be developed.

2. Setup. Suppose that $\{(Y_i, T_i, X_i) : i = 1, \dots, n\}$ are independent and identically distributed observations of (Y, T, X), where Y is an outcome variable, T is a treatment variable taking values 0 or 1, and X is a $d \times 1$ vector of measured covariates. In the potential outcomes framework (Splawa-Neyman (1990); Rubin (1974)), let (Y^0, Y^1) be potential outcomes that would be observed under treatment 0 or 1, respectively. By consistency, assume that Y is either Y^0 if $T = 0$ or Y^1 if $T = 1$, that is, $Y = (1 - T)Y^0 + TY^1$. There are two causal parameters commonly of interest: the average treatment effect (ATE), defined as $E(Y^1 - Y^0) = \mu^1 - \mu^0$ with $\mu^t = E(Y^t)$, and the average treatment effect on the treated (ATT), defined as $E(Y^1 - Y^0|T = 1) = \nu^1 - \nu^0$ with $\nu^t = E(Y^t|T = 1)$ for $t = 0, 1$. For concreteness, we mainly discuss estimation of μ^1 until Section 3.5 to discuss ATE and ATT.

Estimation of ATE is fundamentally a missing-data problem: only one potential outcome, Y_i^0 or Y_i^1, is observed and the other one is missing for each subject i. For identification of (μ^0, μ^1) and ATE, we make the following two assumptions throughout:

(i) Unconfoundedness: T and Y^0 and, respectively, T and Y^1 are conditionally independent given X (Rubin (1976));

(ii) Overlap: $0 < \pi^*(x) < 1$ for all x, where $\pi^*(x) = P(T = 1|X = x)$ is called the propensity score (PS) (Rosenbaum and Rubin (1983)).

Under these assumptions, (μ^0, μ^1) and ATE are often estimated by imposing additional modeling assumptions on the outcome regression function $m_t^*(X) = E(Y|T = t, X)$ or the propensity score $\pi^*(X) = P(T = 1|X)$.

Consider a conditional mean model for outcome regression (OR),

$$E(Y|T = 1, X) = m^1(X; \alpha^1) = \psi\{\alpha^{1\mathrm{T}} g^1(X)\}, \tag{1}$$

where $\psi(\cdot)$ is an (increasing) inverse link function, $g^1(x) = \{1, g_1^1(x), \dots, g_q^1(x)\}^{\mathrm{T}}$ is a vector of known functions such as $g^1(x) = (1, x^{\mathrm{T}})^{\mathrm{T}}$, and $\alpha^1 = (\alpha_0^1, \alpha_1^1, \dots, \alpha_q^1)^{\mathrm{T}}$ is a vector of unknown parameters. Throughout, superscript $^{\mathrm{T}}$ denotes a transpose, not the treatment variable T. Model (1) can be deduced from a generalized linear model with a canonical link (McCullagh and Nelder (1989)). Then the average negative log-(quasi-)likelihood function can be written (after dropping any dispersion parameter) as

$$\ell_{\mathrm{ML}}(\alpha^1) = \tilde{E}(T[-Y\alpha^{1\mathrm{T}} g^1(X) + \Psi\{\alpha^{1\mathrm{T}} g^1(X)\}]), \tag{2}$$

where $\Psi(u) = \int_0^u \psi(u')\,\mathrm{d}u'$, which is convex in u. Throughout, $\tilde{E}(\cdot)$ denotes the sample average. With high-dimensional data, a regularized maximum likelihood estimator, $\hat{\alpha}_{\mathrm{RML}}^1$, can be defined by minimizing the loss $\ell_{\mathrm{ML}}(\alpha^1)$ with the Lasso penalty (Tibshirani (1996)),

$$\ell_{\mathrm{RML}}(\alpha^1) = \ell_{\mathrm{ML}}(\alpha^1) + \lambda \|\alpha_{1:q}^1\|_1, \tag{3}$$

where $\|\cdot\|_1$ denotes the L_1 norm, $\alpha_{1:q}^1 = (\alpha_1^1, \dots, \alpha_q^1)^{\mathrm{T}}$ excluding α_0^1, and $\lambda \geq 0$ is a tuning parameter. The resulting estimator of μ^1 is then

$$\hat{\mu}_{\mathrm{OR}}^1 = \tilde{E}\{\hat{m}_{\mathrm{RML}}^1(X)\} = \frac{1}{n}\sum_{i=1}^n \hat{m}_{\mathrm{RML}}^1(X_i),$$

where $\hat{m}_{\mathrm{RML}}^1(X) = m^1(X; \hat{\alpha}_{\mathrm{RML}}^1)$, the fitted outcome regression function. Various theoretical results have been obtained on Lasso estimation in sparse, high-dimensional regression (e.g., Bühlmann and van de Geer (2011)). If model (1) is correctly specified, then it can be shown under suitable conditions that $\|\hat{\alpha}_{\mathrm{RML}}^1 - \alpha^{1*}\|_1 = O_p(1)\|\alpha^{1*}\|_0\{\log(q)/n\}^{1/2}$ and $\hat{\mu}_{\mathrm{OR}}^1 = \mu^1 + O_p(1)\{\|\alpha^{1*}\|_0 \log(q)/n\}^{1/2}$, where α^{1*} is the true value for model (1) such that $m_1^*(X) = m^1(X; \alpha^{1*})$ and $\|\alpha^{1*}\|_0$ is the size of nonzero elements in α^{1*}.

Alternatively, consider a propensity score (PS) model

$$P(T = 1|X) = \pi(X; \gamma) = \Pi\{\gamma^{\mathrm{T}} f(X)\}, \tag{4}$$

where $\Pi(\cdot)$ is an inverse link function, $f(x) = \{1, f_1(x), \dots, f_p(x)\}^{\mathrm{T}}$ is a vector of known functions such as $g^1(x) = (1, x^{\mathrm{T}})^{\mathrm{T}}$, and $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_p)^{\mathrm{T}}$ is a vector of unknown parameters. For concreteness, assume that model (4) is logistic regression with $\pi(X; \gamma) = [1 + \exp\{-\gamma^{\mathrm{T}} f(X)\}]^{-1}$, and hence the average negative log-likelihood function is

$$\ell_{\mathrm{ML}}(\gamma) = \tilde{E}[\log\{1 + \mathrm{e}^{\gamma^{\mathrm{T}} f(X)}\} - T\gamma^{\mathrm{T}} f(X)]. \tag{5}$$

To handle high-dimensional data, a Lasso penalized maximum likelihood estimator, $\hat{\gamma}_{\mathrm{RML}}$, is defined by minimizing the objective function

$$\ell_{\mathrm{RML}}(\gamma) = \ell_{\mathrm{ML}}(\gamma) + \lambda \|\gamma_{1:p}\|_1, \tag{6}$$

where $\gamma_{1:p} = (\gamma_1, \ldots, \gamma_p)^{\mathrm{T}}$ excluding γ_0, and $\lambda \geq 0$ is a tuning parameter. The fitted propensity score is then $\hat{\pi}_{\mathrm{RML}}(X) = \pi(X; \hat{\gamma}_{\mathrm{RML}})$. A (ratio) inverse probability weighted (IPW) estimator for μ^1 is

$$\hat{\mu}^1_{\mathrm{rIPW}}(\hat{\pi}_{\mathrm{RML}}) = \tilde{E}\left\{\frac{TY}{\hat{\pi}_{\mathrm{RML}}(X)}\right\} \Big/ \tilde{E}\left\{\frac{T}{\hat{\pi}_{\mathrm{RML}}(X)}\right\}.$$

From previous works (e.g., Bühlmann and van de Geer (2011)), if model (4) is correctly specified, then it can be shown under suitable conditions that $\|\hat{\gamma}_{\mathrm{RML}} - \gamma^*\|_1 = O_p(1)\|\gamma^*\|_0\{\log(p)/n\}^{1/2}$ and $\hat{\mu}^1_{\mathrm{rIPW}}(\hat{\pi}_{\mathrm{RML}}) = \mu^1 + O_p(1)\{\|\gamma^*\|_0 \log(p)/n\}^{1/2}$, where γ^* is the true value for model (4) such that $\pi^*(X) = \pi(X; \gamma^*)$ and $\|\gamma^*\|_0$ is the size of nonzero elements in γ^*.

To attain consistency for μ^1, the estimator $\hat{\mu}^1_{\mathrm{OR}}$ or $\hat{\mu}^1_{\mathrm{rIPW}}(\hat{\pi}_{\mathrm{RML}})$ relies on correct specification of OR model (1) or PS model (4), respectively. In contrast, there are doubly robust estimators depending on both OR and PS models in the augmented IPW form (Robins, Rotnitzky and Zhao (1994))

$$\hat{\mu}^1(\hat{m}^1, \hat{\pi}) = \tilde{E}\{\varphi(Y, T, X; \hat{m}^1, \hat{\pi})\},$$

where $\hat{m}^1(X)$ and $\hat{\pi}(X)$ are fitted values of $m_1^*(X)$ and $\pi^*(X)$ and

$$\varphi(Y, T, X; \hat{m}^1, \hat{\pi}) = \frac{TY}{\hat{\pi}(X)} - \left\{\frac{T}{\hat{\pi}(X)} - 1\right\}\hat{m}^1(X). \tag{7}$$

See Kang and Schafer (2007) and Tan (2010) for reviews in low-dimensional settings. Recently, in high-dimensional settings, Belloni, Chernozhukov and Hansen (2014) and Farrell (2015) studied the estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}_{\mathrm{RML}})$, using the fitted values $\hat{m}^1_{\mathrm{RML}}(X)$ and $\hat{\pi}_{\mathrm{RML}}(X)$ from Lasso penalized estimation or similar methods. Their results are mainly of two types. The first type shows double robustness: $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}_{\mathrm{RML}})$ remains consistent if either OR model (1) or PS model (4) is correctly specified. The second type establishes valid confidence intervals: $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}_{\mathrm{RML}})$ admits the usual influence function,

$$\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}_{\mathrm{RML}}) = \tilde{E}\{\varphi(Y, T, X; m^{1*}, \pi^*)\} + o_p(n^{-1/2}), \tag{8}$$

if both OR model (1) and PS model (4) are correctly specified. In general, the latter result requires a stronger sparsity condition than in consistency results only. For example, it is assumed that $\{\|\alpha^{1*}\|_0 + \|\gamma^*\|_0\}\log(p) = o(n^{1/2})$ in Belloni, Chernozhukov and Hansen (2014).

3. Theory and methods.

3.1. *Overview.* A limitation of existing high-dimensional methods discussed in Section 2 is that valid confidence intervals based on $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}_{\mathrm{RML}})$ is obtained only under the assumption that both OR model (1) and PS model (4) are correctly specified, even though the point estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}_{\mathrm{RML}})$ is doubly robust, that is, remains consistent if either OR model (1) or PS model (4) is correctly specified. To fill this gap, we develop new point estimators and confidence intervals for μ^1, depending on a propensity score model and an outcome regression model, such that properties (G1) and at least one of (G2)–(G3) are attained as described in Section 1.

To illustrate main ideas, consider a logistic propensity score model (4) and a linear outcome regression model,

$$E(Y|T = 1, X) = m^1(X; \alpha^1) = \alpha^{1\mathrm{T}} f(X), \tag{9}$$

that is, model (1) with the identity link and the vector of covariate functions $g^1(X)$ taken to be the same as $f(X)$ in model (4) (hence $q = p$). This condition can be satisfied possibly after enlarging model (1) or (4) to reach the same dimension. Our point estimator of μ^1 is

$$\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) = \tilde{E}\{\varphi(Y, T, X; \hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})\}, \tag{10}$$

where $\varphi(\cdot)$ is defined in (7), $\hat{\pi}^1_{\mathrm{RCAL}}(X) = \pi(X; \hat{\gamma}^1_{\mathrm{RCAL}})$, $\hat{m}^1_{\mathrm{RWL}}(X) = m^1(X; \hat{\alpha}^1_{\mathrm{RWL}})$, and $\hat{\gamma}^1_{\mathrm{RCAL}}$ and $\hat{\alpha}^1_{\mathrm{RWL}}$ are defined as follows. The estimator $\hat{\gamma}^1_{\mathrm{RCAL}}$ is a regularized calibrated estimator of γ from Tan (2017), defined as a minimizer of the Lasso penalized objective function,

$$\ell_{\mathrm{RCAL}}(\gamma) = \ell_{\mathrm{CAL}}(\gamma) + \lambda \|\gamma_{1:p}\|_1, \tag{11}$$

where $\lambda \geq 0$ is a tuning parameter and $\ell_{\mathrm{CAL}}(\gamma)$ is the calibration loss,

$$\ell_{\mathrm{CAL}}(\gamma) = \tilde{E}\{T\mathrm{e}^{-\gamma^{\mathrm{T}} f(X)} + (1-T)\gamma^{\mathrm{T}} f(X)\}. \tag{12}$$

The estimator $\hat{\alpha}^1_{\mathrm{RWL}}$ is a regularized weighted least-squares estimator of α^1, defined as a minimizer of

$$\ell_{\mathrm{RWL}}(\alpha^1; \hat{\gamma}^1_{\mathrm{RCAL}}) = \ell_{\mathrm{WL}}(\alpha^1; \hat{\gamma}^1_{\mathrm{RCAL}}) + \lambda \|\alpha^1_{1:p}\|_1, \tag{13}$$

where $\ell_{\mathrm{WL}}(\alpha^1; \hat{\gamma}^1_{\mathrm{RCAL}})$ is the weighted least-squares loss,

$$\ell_{\mathrm{WL}}(\alpha^1; \hat{\gamma}^1_{\mathrm{RCAL}}) = \tilde{E}\left[T \frac{1 - \hat{\pi}^1_{\mathrm{RCAL}}(X)}{\hat{\pi}^1_{\mathrm{RCAL}}(X)} \{Y - \alpha^{1\mathrm{T}} f(X)\}^2\right] \Big/ 2, \tag{14}$$

and $\lambda \geq 0$ is a tuning parameter. That is, the observations in the treated group are weighted by $\{1 - \hat{\pi}^1_{\mathrm{RCAL}}(X_i)\}/\hat{\pi}^1_{\mathrm{RCAL}}(X_i)$, which differs slightly from the commonly used inverse propensity score weight $1/\hat{\pi}^1_{\mathrm{RCAL}}(X_i)$.

There are simple and interesting interpretations of the preceding estimators. By the Karush–Kuhn–Tucker condition for minimizing (11), the fitted propensity score $\hat{\pi}^1_{\mathrm{RCAL}}(X)$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \frac{T_i}{\hat{\pi}^1_{\mathrm{RCAL}}(X_i)} = 1, \tag{15}$$

$$\frac{1}{n} \left| \sum_{i=1}^{n} \frac{T_i f_j(X_i)}{\hat{\pi}^1_{\mathrm{RCAL}}(X_i)} - \sum_{i=1}^{n} f_j(X_i) \right| \leq \lambda, \quad j = 1, \dots, p, \tag{16}$$

where equality holds in (16) for any j such that the jth estimate $(\hat{\gamma}^1_{\mathrm{RCAL}})_j$ is nonzero. Equation (15) shows that the inverse probability weights, $1/\hat{\pi}^1_{\mathrm{RCAL}}(X_i)$ with $T_i = 1$, sum to the sample size n, whereas equation (16) implies that the weighted average of each covariate $f_j(X_i)$ over the treated group may differ from the overall average of $f_j(X_i)$ by no more than λ. The Lasso penalty is used to induce the box constraints on the gradient of $\ell_{\mathrm{CAL}}(\gamma)$ instead of setting the gradient to 0.

By the Karush–Kuhn–Tucker condition for minimizing (13), the fitted outcome regression function $\hat{m}^1_{\mathrm{RWL}}(X)$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} T_i \frac{1 - \hat{\pi}^1_{\mathrm{RCAL}}(X_i)}{\hat{\pi}^1_{\mathrm{RCAL}}(X_i)} \{Y_i - \hat{m}^1_{\mathrm{RWL}}(X_i)\} = 0, \tag{17}$$

$$\frac{1}{n} \left| \sum_{i=1}^{n} T_i \frac{1 - \hat{\pi}^1_{\mathrm{RCAL}}(X_i)}{\hat{\pi}^1_{\mathrm{RCAL}}(X_i)} \{Y_i - \hat{m}^1_{\mathrm{RWL}}(X_i)\} f_j(X_i) \right| \leq \lambda, \quad j = 1, \dots, p, \tag{18}$$

where equality holds in (18) for any j such that the jth estimate $(\hat{\alpha}^1_{\mathrm{RWL}})_j$ is nonzero. Equation (17) implies that by simple calculation, the estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ can be recast as

$$
\begin{aligned}
\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) &= \tilde{E}\left[\hat{m}^1_{\mathrm{RWL}}(X) + \frac{T}{\hat{\pi}^1_{\mathrm{RCAL}}(X)}\{Y - \hat{m}^1_{\mathrm{RWL}}(X)\}\right] \\
&= \tilde{E}\{TY + (1-T)\hat{m}^1_{\mathrm{RWL}}(X)\},
\end{aligned}
\tag{19}
$$

which takes the form of linear prediction estimators known in the survey literature (e.g., Särndal, Swensson and Wretman (1992)): $\tilde{E}\{TY + (1-T)\hat{m}^1(X)\}$ for some fitted outcome regression function $\hat{m}^1(X)$. As a consequence, $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ always falls within the range of the observed outcomes $\{Y_i : T_i = 1, i = 1, \dots, n\}$ and the predicted values $\{\hat{m}^1_{\mathrm{RWL}}(X_i) : T_i = 0, i = 1, \dots, n\}$. This boundedness property is not satisfied by the estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}^1_{\mathrm{RML}})$.

We provide a high-dimensional analysis of the estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ in Section 3.3, allowing for possible model misspecification. Our main result shows that under suitable conditions, the estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ admits the asymptotic expansion

$$
\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) = \tilde{E}\{\varphi(Y, T, X; \bar{m}^1_{\mathrm{WL}}, \bar{\pi}^1_{\mathrm{CAL}})\} + o_p(n^{-1/2}),
\tag{20}
$$

where $\bar{\pi}^1_{\mathrm{CAL}}(X) = \pi(X; \bar{\gamma}^1_{\mathrm{CAL}})$, $\bar{m}^1_{\mathrm{WL}}(X) = m^1(X; \bar{\alpha}^1_{\mathrm{WL}})$ and $\bar{\gamma}^1_{\mathrm{CAL}}$ and $\bar{\alpha}^1_{\mathrm{WL}}$ are defined as follows. With possible model misspecification, the target value $\bar{\gamma}^1_{\mathrm{CAL}}$ is defined as a minimizer of the expected calibration loss

$$
E\{\ell_{\mathrm{CAL}}(\gamma)\} = E\{T\mathrm{e}^{-\gamma^{\mathrm{T}} f(X)} + (1-T)\gamma^{\mathrm{T}} f(X)\}.
$$

If model (4) is correctly specified, then $\bar{\pi}^1_{\mathrm{CAL}}(X) = \pi^*(X)$. Otherwise, $\bar{\pi}^1_{\mathrm{CAL}}(X)$ may differ from $\pi^*(X)$. The target value $\bar{\alpha}^1_{\mathrm{WL}}$ is defined as a minimizer of the expected loss

$$
E\{\ell_{\mathrm{WL}}(\alpha^1; \bar{\gamma}^1_{\mathrm{CAL}})\} = E\left[T\frac{1 - \bar{\pi}^1_{\mathrm{CAL}}(X)}{\bar{\pi}^1_{\mathrm{CAL}}(X)}\{Y - \alpha^{1\mathrm{T}} f(X)\}^2\right]/2.
$$

If model (9) is correctly specified, then $\bar{m}^1_{\mathrm{WL}}(X) = m^*_1(X)$. But $\bar{m}^1_{\mathrm{WL}}(X)$ may in general differ from $m^*_1(X)$. The following result can be deduced from Theorems 3 and 4. Suppose that the Lasso tuning parameter is specified as $\lambda = A^\dagger_0\{\log(p)/n\}^{1/2}$ for $\hat{\gamma}^1_{\mathrm{RCAL}}$ and $\lambda = A^\dagger_1\{\log(p)/n\}^{1/2}$ for $\hat{\alpha}^1_{\mathrm{RWL}}$, with some constants $A^\dagger_0$ and $A^\dagger_1$. Denote $S_\gamma = \{0\} \cup \{j : \bar{\gamma}^1_{\mathrm{CAL},j} \neq 0, j = 1, \dots, p\}$ and $S_{\alpha^1} = \{0\} \cup \{j : \bar{\alpha}^1_{\mathrm{WL},j} \neq 0, j = 1, \dots, p\}$.

PROPOSITION 1. *Suppose that Assumptions 1 and 2 hold as in Section 3.3, and* $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$. *Then for* $\hat{\gamma}^1_{\mathrm{RCAL}}$ *and* $\hat{\alpha}^1_{\mathrm{RWL}}$ *with sufficiently large constants* $A^\dagger_0$ *and* $A^\dagger_1$, *asymptotic expansion* (20) *is valid. Moreover, if either logistic PS model* (4) *or linear OR model* (9) *is correctly specified, then* $\bar{\pi}^1_{\mathrm{CAL}}(x) \equiv \pi^*(x)$ *or, respectively,* $\bar{m}^1_{\mathrm{WL}}(x) \equiv m^{1*}(x)$, *and the following results hold*:

(i) $n^{1/2}\{\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) - \mu^1\} \to_{\mathcal{D}} \mathrm{N}(0, V)$, *where* $V = \mathrm{var}\{\varphi(Y, T, X; \bar{m}^1_{\mathrm{WL}}, \bar{\pi}^1_{\mathrm{CAL}})\}$;

(ii) *a consistent estimator of V is*

$$
\hat{V} = \tilde{E}[\{\varphi(Y, T, X; \hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) - \hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})\}^2];
$$

(iii) *an asymptotic* $(1-c)$ *confidence interval for* μ^1 *is* $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) \pm z_{c/2}\sqrt{\hat{V}/n}$, *where* $z_{c/2}$ *is the* $(1 - c/2)$ *quantile of* $\mathrm{N}(0, 1)$.

That is, a doubly robust confidence interval for μ^1 *is obtained.*

REMARK 1. We discuss two implications of Proposition 1. First, the estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ is also locally efficient, that is, achieves the semiparametric efficiency bound, $n^{-1}\operatorname{var}\{\varphi(Y,T,X;m^{1*},\pi^*)\}$, for estimation of μ^1 (Hahn (1998)), when both models (4) and (9) are correctly specified, because in this case $\bar{\pi}^1_{\mathrm{CAL}}(x) \equiv \pi^*(x)$ and $\bar{m}^1_{\mathrm{WL}}(x) \equiv m^{1*}(x)$. Second, the results (i)–(iii) hold uniformly over all data-generating processes subject to the assumptions stated, which are much weaker than those needed for perfect model selection by Lasso-type methods (Bühlmann and van de Geer (2011)). Therefore, our method provides uniformly valid inference similarly as in Belloni, Chernozhukov and Hansen ((2014), Corollary 1).

3.2. *On construction of estimators.* We point out basic ideas underlying the construction of the estimators $\hat{\gamma}^1_{\mathrm{RCAL}}$ and $\hat{\alpha}^1_{\mathrm{RWL}}$ such that the estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ satisfies asymptotic expansion (20), even with model misspecification. The discussion is heuristic here, and formal theory is presented in Sections 3.3 and 3.4. In general, let $\hat{\alpha}^1$ be some estimator of α^1 in model (1), which is assumed to converge in probability to a limit $\bar{\alpha}^1$. Denote $\hat{m}^1(X) = m^1(X;\hat{\alpha}^1)$ and $\bar{m}^1(X) = m^1(X;\bar{\alpha}^1)$. Similarly, let $\hat{\gamma}$ be some estimator of γ in model (4), which is assumed to converge in probability to a limit $\bar{\gamma}$. Denote $\hat{\pi}(X) = \pi(X;\hat{\gamma})$ and $\bar{\pi}(X) = \pi(X;\bar{\gamma})$.

Consider a Taylor expansion of $\hat{\mu}^1(\hat{m}^1, \hat{\pi})$:

$$\hat{\mu}^1(\hat{m}^1,\hat{\pi}) = \hat{\mu}^1(\bar{m}^1,\bar{\pi}) + \Delta_1 + \Delta_2 + o_p(n^{-1/2}), \tag{21}$$

with

$$\Delta_1 = (\hat{\alpha}^1 - \bar{\alpha}^1)^{\mathrm{T}} \times \frac{\partial}{\partial \alpha^1} \tilde{E}\{\varphi(Y,T,X;\alpha^1,\gamma)\}\Big|_{(\alpha^1,\gamma)=(\bar{\alpha}^1,\bar{\gamma})},$$

$$\Delta_2 = (\hat{\gamma} - \bar{\gamma})^{\mathrm{T}} \times \frac{\partial}{\partial \gamma} \tilde{E}\{\varphi(Y,T,X;\alpha^1,\gamma)\}\Big|_{(\alpha^1,\gamma)=(\bar{\alpha}^1,\bar{\gamma})},$$

where $\varphi(Y,T,X;\alpha^1,\gamma) = \varphi(Y,T,X;m^1(\cdot;\alpha^1),\pi(\cdot;\gamma))$ from (7), and the remainder is taken to be $o_p(n^{-1/2})$ under suitable conditions. The term Δ_1 or Δ_2 represents part of the variation of $\hat{\mu}^1(\hat{m}^1,\hat{\pi})$ caused by the deviation of $\hat{m}^1(X)$ from the limit $\bar{m}^1(X)$ or, respectively, that of $\hat{\pi}(X)$ from the limit $\bar{\pi}(X)$. With model misspecification, the two terms, Δ_1 and Δ_2, are in general no smaller than $O_p(n^{-1/2})$, in low- and high-dimensional settings. This is because $\hat{\alpha}^1 - \bar{\alpha}^1$ and $\hat{\gamma} - \bar{\gamma}$ are at least $O_p(n^{-1/2})$ and the second terms in Δ_1 and Δ_2 may not vanish in probability. Therefore, in order that $\Delta_1 = o_p(n^{-1/2})$, $\Delta_2 = o_p(n^{-1/2})$, and (21) gives

$$\hat{\mu}^1(\hat{m}^1,\hat{\pi}) = \hat{\mu}^1(\bar{m}^1,\bar{\pi}) + o_p(n^{-1/2}), \tag{22}$$

it seems necessary that the second terms in Δ_1 and Δ_2 should be $o_p(1)$ (i.e., vanish in probability), and their population versions should satisfy

$$\frac{\partial}{\partial \alpha^1} E\{\varphi(Y,T,X;\alpha^1,\gamma)\}\Big|_{(\alpha^1,\gamma)=(\bar{\alpha}^1,\bar{\gamma})} = 0, \tag{23}$$

$$\frac{\partial}{\partial \gamma} E\{\varphi(Y,T,X;\alpha^1,\gamma)\}\Big|_{(\alpha^1,\gamma)=(\bar{\alpha}^1,\bar{\gamma})} = 0. \tag{24}$$

These conditions (23) and (24) are known to be sufficient for (22) to hold, under additional regularity conditions in low-dimensional settings (Kim and Haziza (2014); Vermeulen and Vansteelandt (2015)). As discussed below, such orthogonality conditions are also key to the construction of our regularized estimators $\hat{\gamma}^1_{\mathrm{RCAL}}$ and $\hat{\alpha}^1_{\mathrm{RWL}}$ in high-dimensional settings.

We now examine implications of (23)–(24) in the concrete situation of Section 3.1. For the linear OR model (9), Δ_1 reduces to

$$\Delta_1 = (\hat{\alpha}^1 - \bar{\alpha}^1)^{\mathrm{T}} \times \tilde{E}\left[\left\{1 - \frac{T}{\bar{\pi}(X)}\right\} f(X)\right]. \tag{25}$$

For the logistic PS model (4), Δ_2 reduces to

$$\Delta_2 = -(\hat{\gamma} - \bar{\gamma})^{\mathrm{T}} \times \tilde{E}\left[T \frac{1 - \bar{\pi}(X)}{\bar{\pi}(X)} \{Y - \bar{m}^1(X)\} f(X)\right]. \tag{26}$$

Then conditions (23) and (24) become

$$E\left[\left\{1 - \frac{T}{\bar{\pi}(X)}\right\} f(X)\right] = 0, \tag{27}$$

$$E\left[T \frac{1 - \bar{\pi}(X)}{\bar{\pi}(X)} \{Y - \bar{m}^1(X)\} f(X)\right] = 0. \tag{28}$$

The validity of conditions (27)–(28) depends on whether models (4) and (9) are correctly specified and how the estimators $(\hat{\alpha}^1, \hat{\gamma})$ are constructed.

If models (4) and (9) are correctly specified, then conditions (23)–(24) or (27)–(28) are satisfied for *any* consistent estimators $(\hat{\alpha}^1, \hat{\gamma})$ such that, in the limit, $m^1(X; \bar{\alpha}^1) = m^{*1}(X) = E(Y|T = 1, X)$ and $\pi(X; \bar{\gamma}) = \pi^*(X) = P(T = 1|X)$, including the regularized maximized likelihood estimators $(\hat{\alpha}^1_{\mathrm{RML}}, \hat{\gamma}_{\mathrm{RML}})$. In this case, conditions (23)–(24) can be expressed as

$$\frac{\partial}{\partial \alpha^1} E\{\varphi(Y, T, X; \alpha^1, \gamma)\}\bigg|_{(\alpha^1,\gamma)=(\alpha^{*1},\gamma^*)} = 0, \tag{29}$$

$$\frac{\partial}{\partial \gamma} E\{\varphi(Y, T, X; \alpha^1, \gamma)\}\bigg|_{(\alpha^1,\gamma)=(\alpha^{*1},\gamma^*)} = 0, \tag{30}$$

where α^{1*} and γ^* are the true values such that $m^1(X; \alpha^{1*}) = m^{1*}(X)$ and $\pi(X; \gamma^*) = \pi^*(X)$. While conditions (23)–(24) in general depend on the estimators $(\hat{\alpha}^1, \hat{\gamma})$ through their limits $(\bar{\alpha}^1, \bar{\gamma})$, such dependency is suppressed in (29)–(30) because $(\bar{\alpha}^1, \bar{\gamma})$ is assumed to coincide with (α^{1*}, γ^*) under correctly specified models. The identities (29) and (30) are known to hold for the AIPW estimating function $\varphi(\cdot)$ as a consequence of its double robustness (Robins and Rotnitzky (2001)). Moreover, the relationship (29)–(30) is called an orthogonality property and exploited by Belloni, Chernozhukov and Hansen (2014) to establish asymptotic expansion (8) for the estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}_{\mathrm{RML}})$, under correctly specified models in high-dimensional settings.

If, however, PS model (4) or OR model (9) is misspecified, then condition (27) or, respectively, (28) is in general violated, and hence asymptotic expansion (22) may no longer hold, for example, for the standard estimators $(\hat{\alpha}^1_{\mathrm{RML}}, \hat{\gamma}^1_{\mathrm{RML}})$. In this context, the estimators $\hat{\gamma}^1_{\mathrm{RCAL}}$ and $\hat{\alpha}^1_{\mathrm{RWL}}$ are constructed with loss functions $\ell_{\mathrm{CAL}}(\gamma)$ and $\ell_{\mathrm{WL}}(\alpha^1; \gamma)$ in (12) and (14) such that

$$\begin{aligned} \frac{\partial}{\partial \gamma} \ell_{\mathrm{CAL}}(\gamma) &= \frac{\partial}{\partial \alpha^1} \tilde{E}\{\varphi(Y, T, X; \alpha^1, \gamma)\} \\ &= \tilde{E}\left[\left\{1 - \frac{T}{\pi(X; \gamma)}\right\} f(X)\right], \end{aligned} \tag{31}$$

$$\begin{aligned} \frac{\partial}{\partial \alpha^1} \ell_{\mathrm{WL}}(\alpha^1; \gamma) &= \frac{\partial}{\partial \gamma} \tilde{E}\{\varphi(Y, T, X; \alpha^1, \gamma)\} \\ &= -\tilde{E}\left[T \frac{1 - \pi(X; \gamma)}{\pi(X; \gamma)} \{Y - \alpha^{1\mathrm{T}} f(X)\} f(X)\right]. \end{aligned} \tag{32}$$

As a result, conditions (27) and (28) are satisfied by $\bar{\pi}^1_{\mathrm{CAL}}(X) = \pi(X; \bar{\gamma}^1_{\mathrm{CAL}})$ and $\bar{m}^1_{\mathrm{WL}}(X) = m^1(X; \bar{\alpha}^1_{\mathrm{WL}})$, where $\bar{\gamma}^1_{\mathrm{CAL}}$ and $\bar{\alpha}^1_{\mathrm{WL}}$ are, respectively, the probability limits of $\hat{\gamma}^1_{\mathrm{RCAL}}$ and $\hat{\alpha}^1_{\mathrm{RWL}}$, defined as minimizers of the expected loss $E\{\ell_{\mathrm{CAL}}(\gamma)\}$ and $E\{\ell_{\mathrm{WL}}(\alpha^1; \bar{\gamma}^1_{\mathrm{CAL}})\}$. From (27) and (28), the second terms in (25) and (26) can be $O_p(\{\log(p)/n\}^{1/2})$ in the supremum norms under sub-Gaussian errors. Moreover, it can be shown that $\|\hat{\gamma}^1_{\mathrm{RCAL}} - \bar{\gamma}^1_{\mathrm{CAL}}\|_1 = O_p(1)|S_\gamma|\{\log(p)/n\}^{1/2}$ and $\|\hat{\alpha}^1_{\mathrm{RWL}} - \bar{\alpha}^1_{\mathrm{WL}}\|_1 = O_p(1)(|S_\gamma| + |S_{\alpha^1}|)\{\log(p)/n\}^{1/2}$, as demonstrated in Theorems 1 and 2 later. Consequently, the products Δ_1 and Δ_2 in (25) and (26) can be $O_p(1)(|S_\gamma| + |S_{\alpha^1}|)\log(p)/n$, which becomes $o_p(n^{-1/2})$, and hence asymptotic expansion (20) holds provided $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$ as stated in Proposition 1.

The estimator $\hat{\gamma}^1_{\mathrm{RCAL}}$ is called a regularized calibrated estimator of γ (Tan (2017)), because in the extreme case of $\lambda = 0$, equations (15)–(16) reduce to calibration equations, which can be traced to Folsom (1991) in the survey literature. Although such equations are intuitively appealing, the preceding discussion shows that $\hat{\gamma}^1_{\mathrm{RCAL}}$ can also be derived to reduce the variation associated with estimation of α^1 from linear OR model (9) for the estimator $\hat{\mu}^1(\hat{m}^1, \hat{\pi})$, when PS model (4) may be misspecified. Similarly, $\hat{\alpha}^1_{\mathrm{RWL}}$ is constructed to reduce the variation associated with estimation of γ from logistic PS model (4) for the estimator $\hat{\mu}^1(\hat{m}^1, \hat{\pi})$, when OR model (9) may be misspecified. By extending the meaning of calibrated estimation, we call $\hat{\alpha}^1_{\mathrm{RWL}}$ a regularized calibrated estimator of α^1 against model (4), as well as $\hat{\gamma}^1_{\mathrm{RCAL}}$ a regularized calibrated estimator of γ against model (9), when used to define $\hat{\mu}^1(\hat{m}^1, \hat{\pi})$.

REMARK 2. Conditions (23)–(24) were previously used by Kim and Haziza (2014) and Vermeulen and Vansteelandt (2015) to construct an augmented IPW estimator $\hat{\mu}^1(\hat{m}^1, \hat{\pi})$ for μ^1 in low-dimensional settings, where $(\hat{\alpha}^1, \hat{\gamma})$ are nonpenalized, defined by directly setting (31)–(32) to zero. One of their motivations is to achieve asymptotic expansion (22), and hence enable simple confidence intervals without the need of correcting for estimation of (α^1, γ). In the absence of (22), valid confidence intervals can still be derived in low-dimensional settings by invoking (21) and $n^{-1/2}$ asymptotic expansions for $\hat{\alpha}^1 - \bar{\alpha}^1$ and $\hat{\gamma} - \bar{\gamma}$ with usual influence functions (White (1982)), allowing for model misspecification. But this influence-function based approach is not applicable with high-dimensional data. Our work exploits conditions (23)–(24) to construct the regularized estimators $(\hat{\alpha}^1_{\mathrm{RWL}}, \hat{\gamma}^1_{\mathrm{RCAL}})$ and achieve asymptotic expansion (22) for $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$, so that valid confidence intervals for μ^1 can be obtained in high-dimensional settings.

REMARK 3. Our approach based on (23)–(24) and that of Belloni, Chernozhukov and Hansen (2014) and Chernozhukov et al. (2018) based on (29)–(30) can be compared as follows. Both approaches involve use of the doubly robust estimating function $\varphi(\cdot)$, but in different manners. Conditions (23)–(24) amount to requiring orthogonality to hold at whatever limits $(\bar{\alpha}^1, \bar{\gamma})$, which can be achieved by carefully choosing ("calibrating") the loss functions for the corresponding estimators. Conditions (29)–(30) can be seen as a special case of (23)–(24), with $(\bar{\alpha}^1, \bar{\gamma})$ at the true values, which can in general be achieved only by consistent estimators under correctly specified models.

3.3. *Theory with linear outcome regression.*

In this section, we assume that linear outcome model (9) is used together with logistic propensity score model (4), and develop theoretical results for the proposed estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ in high-dimensional settings. There are several technical issues we need to address in high-dimensional analysis, including how to handle the dependency of the estimator $\hat{\alpha}^1_{\mathrm{RWL}}$ on $\hat{\gamma}^1_{\mathrm{RCAL}}$, and what condition is required on the sparsity sizes of $\bar{\gamma}^1_{\mathrm{CAL}}$ and $\bar{\alpha}^1_{\mathrm{WL}}$.

First, we describe relevant results from Tan (2017) on the regularized calibrated estimator $\hat{\gamma}^1_{\mathrm{RCAL}}$ in model (4). The tuning parameter λ in (11) for defining $\hat{\gamma}^1_{\mathrm{RCAL}}$ is specified as $\lambda =$

$A_0\lambda_0$, with a constant $A_0 > 1$ and

$$\lambda_0 = C_1\sqrt{\log\{(1+p)/\epsilon\}/n},$$

where $C_1 > 0$ is a constant depending only on (C_0, B_0) from Assumption 1 below and $0 < \epsilon < 1$ is a tail probability for the error bound. For example, taking $\epsilon = 1/(1+p)$ gives $\lambda_0 = C_1\sqrt{2\log(1+p)/n}$, a familiar rate in high-dimensional analysis.

With possible model misspecification, the target value $\bar{\gamma}^1_{\mathrm{CAL}}$ is defined as a minimizer of the expected calibration loss $E\{\ell_{\mathrm{CAL}}(\gamma)\}$ as in Section 3.1. From a functional perspective, we write $\ell_{\mathrm{CAL}}(\gamma) = \kappa_{\mathrm{CAL}}(\gamma^{\mathrm T} f)$, where for a function $h(x)$,

$$\kappa_{\mathrm{CAL}}(h) = \tilde{E}\big[T\mathrm{e}^{-h(X)} + (1-T)h(X)\big].$$

As $\kappa_{\mathrm{CAL}}(h)$ is easily shown to be convex in h, the Bregman divergence associated with κ_{CAL} is defined such that for two functions $h(x)$ and $h'(x)$,

$$D_{\mathrm{CAL}}(h', h) = \kappa_{\mathrm{CAL}}(h') - \kappa_{\mathrm{CAL}}(h) - \langle \nabla\kappa_{\mathrm{CAL}}(h), h' - h\rangle,$$

where h is identified as a vector $(h_1, \ldots, h_n)$ with $h_i = h(X_i)$, and $\nabla\kappa_{\mathrm{CAL}}(h)$ denotes the gradient of $\kappa_{\mathrm{CAL}}(h)$ with respect to $(h_1, \ldots, h_n)$. The following result (Theorem 1) is restated from Tan ((2017), Corollary 2), where the convergence of $\hat{\gamma}^1_{\mathrm{RCAL}}$ to $\bar{\gamma}^1_{\mathrm{CAL}}$ is obtained in the L_1 norm $\|\hat{\gamma}^1_{\mathrm{RCAL}} - \bar{\gamma}^1_{\mathrm{CAL}}\|_1$ and the symmetrized Bregman divergence

$$D^{\dagger}_{\mathrm{CAL}}(\hat{h}^1_{\mathrm{RCAL}}, \bar{h}^1_{\mathrm{CAL}}) = D_{\mathrm{CAL}}(\hat{h}^1_{\mathrm{RCAL}}, \bar{h}^1_{\mathrm{CAL}}) + D_{\mathrm{CAL}}(\bar{h}^1_{\mathrm{CAL}}, \hat{h}^1_{\mathrm{RCAL}}),$$

where $\hat{h}^1_{\mathrm{RCAL}}(X) = \hat{\gamma}^{1\mathrm T}_{\mathrm{RCAL}} f(X)$ and $\bar{h}^1_{\mathrm{CAL}}(X) = \bar{\gamma}^{1\mathrm T}_{\mathrm{CAL}} f(X)$. See Lemma 7 in the Supplementary Material (Tan (2020)) for an explicit expression of $D^{\dagger}_{\mathrm{CAL}}$.

For a matrix Σ with row indices $\{0, 1, \ldots, k\}$, a compatibility condition (Bühlmann and van de Geer (2011)) is said to hold with a subset $S \in \{0, 1, \ldots, k\}$ and constants $\nu_0 > 0$ and $\xi_0 > 1$ if $\nu_0^2(\sum_{j\in S}|b_j|)^2 \le |S|(b^{\mathrm T}\Sigma b)$ for any vector $b = (b_0, b_1, \ldots, b_k)^{\mathrm T} \in \mathbb{R}^{1+k}$ satisfying

$$\sum_{j\notin S}|b_j| \le \xi_0 \sum_{j\in S}|b_j|. \tag{33}$$

Throughout, $|S|$ denotes the size of a set S. By the Cauchy–Schwarz inequality, this compatibility condition is implied by (hence weaker than) a restricted eigenvalue condition (Bickel, Ritov and Tsybakov (2009)) such that $\nu_0^2(\sum_{j\in S} b_j^2) \le b^{\mathrm T}\Sigma b$ for any vector $b \in \mathbb{R}^{1+k}$ satisfying (33).

ASSUMPTION 1. Suppose that the following conditions are satisfied:

(i) $\max_{j=0,1,\ldots,p}|f_j(X)| \le C_0$ almost surely for a constant $C_0 \ge 1$;

(ii) $\bar{h}^1_{\mathrm{CAL}}(X) \ge B_0$ almost surely for a constant $B_0 \in \mathbb{R}$, that is, $\pi(X; \bar{\gamma}^1_{\mathrm{CAL}})$ is bounded from below by $(1+\mathrm{e}^{-B_0})^{-1}$;

(iii) the compatibility condition holds for Σ_γ with the subset $S_\gamma = \{0\} \cup \{j : \bar{\gamma}^1_{\mathrm{CAL},j} \neq 0, j = 1, \ldots, p\}$ and some constants $\nu_0 > 0$ and $\xi_0 > 1$, where $\Sigma_\gamma = E[Tw(X; \bar{\gamma}^1_{\mathrm{CAL}}) f(X) \times f^{\mathrm T}(X)]$ is the Hessian of $E\{\ell_{\mathrm{CAL}}(\gamma)\}$ at $\gamma = \bar{\gamma}^1_{\mathrm{CAL}}$ and $w(X;\gamma) = \mathrm{e}^{-\gamma^{\mathrm T} f(X)}$;

(iv) $|S_\gamma|\lambda_0 \le \eta_0$ for a sufficiently small constant $\eta_0 > 0$, depending only on (A_0, C_0, ξ_0, ν_0).

THEOREM 1 (Tan (2017)). *Suppose that Assumption 1 holds. Then for $A_0 > (\xi_0+1)/(\xi_0-1)$, we have with probability at least $1-4\epsilon$,*

$$D^{\dagger}_{\mathrm{CAL}}(\hat{h}^1_{\mathrm{RCAL}}, \bar{h}^1_{\mathrm{CAL}}) + (A_0 - 1)\lambda_0\|\hat{\gamma}^1_{\mathrm{RCAL}} - \bar{\gamma}^1_{\mathrm{CAL}}\|_1 \le M_0|S_\gamma|\lambda_0^2, \tag{34}$$

where $M_0 > 0$ is a constant depending only on $(A_0, C_0, B_0, \xi_0, \nu_0, \eta_0)$.

REMARK 4. We provide comments about the conditions involved. First, Assumption 1(iii) can be justified from a compatibility condition for the Gram matrix $E\{f(X)f^{\mathrm{T}}(X)\}$ in conjunction with additional conditions such as for some constant $\tau_0 > 0$,

$$b^{\mathrm{T}}E\{f(X)f^{\mathrm{T}}(X)\}b \le (b^{\mathrm{T}}\Sigma_\gamma b)/\tau_0, \quad \forall b \in \mathbb{R}^{1+p}. \tag{35}$$

For example, (35) holds provided that $\pi^*(X)$ is bounded from below by a positive constant and $\pi(X;\bar\gamma^1_{\mathrm{CAL}})$ is bounded away from 1. But it is also possible that Assumption 1(iii) is satisfied even if (35) does not hold for any $\tau_0 > 0$. Therefore, $\pi(X;\bar\gamma^1_{\mathrm{CAL}})$ may not be bounded away from 1 under Assumption 1, although it is required to be bounded away from 0 by Assumption 1(ii). Second, Assumption 1(iv) can be relaxed to only require that $|S_\gamma|\lambda_0^2$ is sufficiently small, albeit under stronger conditions, for example, the variables $f_1(X),\dots,f_p(X)$ are jointly (not just marginally) sub-Gaussian (Huang and Zhang (2012); Negahban et al. (2012)). On the other hand, Assumption 1(iv) is already weaker than the sparsity condition, $|S_\gamma|\log(p) = o(n^{1/2})$, which is needed for obtaining valid confidence intervals for μ^1 from existing works (Belloni, Chernozhukov and Hansen (2014)) and our later results.

REMARK 5. For the Hessian Σ_γ, the weight $w(X;\bar\gamma^1_{\mathrm{CAL}})$ with $\bar\gamma^1_{\mathrm{CAL}}$ replaced by $\hat\gamma^1_{\mathrm{RCAL}}$ is identical to that used in the weighted least-square loss (14) to define $\hat\alpha^1_{\mathrm{RWL}}$, that is, $w(X;\hat\gamma^1_{\mathrm{RCAL}}) = \{1-\hat\pi^1_{\mathrm{RCAL}}(X)\}/\hat\pi^1_{\mathrm{RCAL}}(X)$. The Hessian of $\ell_{\mathrm{CAL}}(\gamma)$ at $\bar\gamma^1_{\mathrm{CAL}}$ is also the same as the Hessian of $\ell_{\mathrm{WL}}(\alpha^1;\bar\gamma^1_{\mathrm{CAL}})$ in α^1. This coincidence is a consequence of the pair of equations (31)–(32) satisfied by the loss functions $\ell_{\mathrm{CAL}}(\gamma)$ and $\ell_{\mathrm{WL}}(\alpha^1;\gamma)$.

Now we turn to the regularized weighted least-squares estimator $\hat\alpha^1_{\mathrm{RWL}}$. We develop a new strategy of inverting a quadratic inequality to address the dependency of $\hat\alpha^1_{\mathrm{RWL}}$ on $\hat\gamma^1_{\mathrm{RCAL}}$ and establish convergence of $\hat\alpha^1_{\mathrm{RWL}}$ under similar conditions as needed for Lasso penalized unweighted least-squares estimators in high-dimensional settings. The error bound obtained, however, depends on the sparsity size $|S_\gamma|$ and various constants in Assumption 1.

For theoretical analysis, the tuning parameter λ in (13) for defining $\hat\alpha^1_{\mathrm{RWL}}$ is specified as $\lambda = A_1\lambda_1$, with a constant $A_1 > 1$ and

$$\lambda_1 = \max\Big\{\lambda_0, \mathrm{e}^{-B_0}C_0\sqrt{8(D_0^2+D_1^2)}\sqrt{\log\{(1+p)/\epsilon\}/n}\Big\},$$

where $0<\epsilon<1$ is a tail probability for the error bound, (C_0,B_0) are from Assumption 1, and (D_0,D_1) are from Assumption 2 below. With possible model misspecification, the target value $\bar\alpha^1_{\mathrm{WL}}$ is defined as a minimizer of the expected loss $E\{\ell_{\mathrm{WL}}(\alpha^1;\bar\gamma^1_{\mathrm{CAL}})\}$ as in Section 3.1. The following result gives the convergence of $\hat\alpha^1_{\mathrm{RWL}}$ to $\bar\alpha^1_{\mathrm{WL}}$ in the L_1 norm $\|\hat\alpha^1_{\mathrm{RWL}}-\bar\alpha^1_{\mathrm{WL}}\|_1$ and the weighted (in-sample) prediction error defined as

$$Q_{\mathrm{WL}}(\hat m^1_{\mathrm{RWL}},\bar m^1_{\mathrm{WL}};\bar\gamma^1_{\mathrm{CAL}}) = \tilde E[Tw(X;\bar\gamma^1_{\mathrm{CAL}})\{\hat m^1_{\mathrm{RWL}}(X)-\bar m^1_{\mathrm{WL}}(X)\}^2], \tag{36}$$

where $\hat m^1_{\mathrm{RWL}}(X) = \hat\alpha^{1\mathrm{T}}_{\mathrm{RWL}}f(X)$ and $\bar m^1_{\mathrm{WL}}(X) = \bar\alpha^{1\mathrm{T}}_{\mathrm{WL}}f(X)$. In fact, $Q_{\mathrm{WL}}(\hat m^1_{\mathrm{RWL}},\bar m^1_{\mathrm{WL}};\bar\gamma^1_{\mathrm{CAL}})$ is the symmetrized Bregman divergence between $\hat m^1_{\mathrm{RWL}}(X)$ and $\bar m^1_{\mathrm{WL}}(X)$ associated with the loss $\kappa_{\mathrm{WL}}(h;\bar\gamma^1_{\mathrm{CAL}}) = \tilde E[Tw(X;\bar\gamma^1_{\mathrm{CAL}})\{Y-h(X)\}^2]/2$. See Section 3.4 for further discussion.

ASSUMPTION 2. Suppose that the following conditions are satisfied:

(i) $Y^1-\bar m^1_{\mathrm{WL}}(X)$ is uniformly sub-Gaussian given X: $D_0^2E(\exp[\{Y^1-\bar m^1_{\mathrm{WL}}(X)\}^2/D_0^2]-1|X) \le D_1^2$ for some positive constants (D_0,D_1);

(ii) the compatibility condition holds for Σ_γ with the subset $S_{\alpha^1} = \{0\} \cup \{j : \bar{\alpha}^1_{\mathrm{WL},j} \neq 0, j = 1, \ldots, p\}$ and some constants $\nu_1 > 0$ and $\xi_1 > 1$;
(iii) $(1+\xi_1)^2 \nu_1^{-2} |S_{\alpha^1}| \lambda_1 \le \eta_1$ for a constant $0 < \eta_1 < 1$.

THEOREM 2. *Suppose that linear outcome model* (9) *is used,* $A_0 > (\xi_0+1)/(\xi_0-1)$, $A_1 > (\xi_1+1)/(\xi_1-1)$ *and Assumptions* 1 *and* 2 *hold. If* $\log\{(1+p)/\epsilon\}/n \le 1$, *then we have with probability at least* $1-8\epsilon$,

$$\begin{aligned} Q_{\mathrm{WL}}(\hat{m}^1_{\mathrm{RWL}}, \bar{m}^1_{\mathrm{WL}}; \bar{\gamma}^1_{\mathrm{CAL}}) + \mathrm{e}^{\eta_{01}}(A_1-1)\lambda_1 \|\hat{\alpha}^1_{\mathrm{RWL}} - \bar{\alpha}^1_{\mathrm{WL}}\|_1 \\ \le \mathrm{e}^{4\eta_{01}} \xi_2^{-2} (M_{01}|S_\gamma|\lambda_0^2) + \mathrm{e}^{2\eta_{01}} \xi_3^2 (\nu_2^{-2} |S_{\alpha^1}| \lambda_1^2), \end{aligned} \tag{37}$$

where $\xi_2 = 1 - 2A_1/\{(\xi_1+1)(A_1-1)\} \in (0,1]$, $\xi_3 = (\xi_1+1)(A_1-1)$, *and* $\nu_2 = \nu_1(1-\eta_1)^{1/2}$, *depending only on* $(A_1, \xi_1, \nu_1, \eta_1)$, *and* $M_{01} = (D_0^2 + D_1^2)(\mathrm{e}^{\eta_{01}} M_0 + \eta_{02}) + (D_0^2 + D_0 D_1)\eta_{02}$, $\eta_{01} = (A_0-1)^{-1} M_0 \eta_0 C_0$, *and* $\eta_{02} = (A_0-1)^{-2} M_0^2 \eta_0$, *depending only on* $(A_0, C_0, B_0, \xi_0, \nu_0, \eta_0)$ *in Theorem* 1 *and* (D_0, D_1).

REMARK 6. Assumption 2(ii) is concerned about the same matrix Σ_γ as in Assumption 1(iii), but with the sparsity subset S_{α^1} from $\bar{\alpha}^1_{\mathrm{WL}}$ instead of S_γ from $\bar{\gamma}^1_{\mathrm{CAL}}$. The matrix Σ_γ is also the Hessian of the expected loss $E\{\ell_{\mathrm{WL}}(\alpha^1; \bar{\gamma}^1_{\mathrm{CAL}})\}$ at $\alpha^1 = \bar{\alpha}^1_{\mathrm{WL}}$, for reasons mentioned in Remark 5. Assumptions 2(ii)–(iii) are combined to derive a compatibility condition for the sample matrix $\tilde{\Sigma}_\gamma = \tilde{E}[Tw(X; \bar{\gamma}^1_{\mathrm{CAL}}) f(X) f^{\mathrm{T}}(X)]$. Assumption 2(iii) can be relaxed such that $|S_{\alpha^1}|\lambda_1^2$ is sufficiently small under further side conditions, but it is already weaker than the sparsity condition, $|S_{\alpha^1}|\log(p) = o(n^{1/2})$, later needed for valid confidence intervals for μ^1. Essentially, the conditions in Assumption 2 are comparable to those for high-dimensional analysis of standard Lasso estimators (e.g., Bühlmann and van de Geer (2011)).

REMARK 7. One of the key steps in our proof is to upper bound the product

$$(\hat{\alpha}^1_{\mathrm{RWL}} - \bar{\alpha}^1_{\mathrm{WL}})^{\mathrm{T}} \tilde{E}[Tw(X; \hat{\gamma}^1_{\mathrm{RCAL}})\{Y - \bar{m}^1_{\mathrm{WL}}(X)\} f(X)], \tag{38}$$

which involves the estimated weight $w(X; \hat{\gamma}^1_{\mathrm{RCAL}})$. If $\hat{\gamma}^1_{\mathrm{RCAL}}$ were replaced by $\bar{\gamma}^1_{\mathrm{CAL}}$, then it is standard to use the following bound:

$$\begin{aligned} (\hat{\alpha}^1_{\mathrm{RWL}} - \bar{\alpha}^1_{\mathrm{WL}})^{\mathrm{T}} \tilde{E}[Tw(X; \bar{\gamma}^1_{\mathrm{CAL}})\{Y - \bar{m}^1_{\mathrm{WL}}(X)\} f(X)] \\ \le \|\hat{\alpha}^1_{\mathrm{RWL}} - \bar{\alpha}^1_{\mathrm{WL}}\|_1 \times \|\tilde{E}[Tw(X; \bar{\gamma}^1_{\mathrm{CAL}})\{Y - \bar{m}^1_{\mathrm{WL}}(X)\} f(X)]\|_\infty. \end{aligned} \tag{39}$$

To handle the dependency on $\hat{\gamma}^1_{\mathrm{RCAL}}$, our strategy is to derive an upper bound of the difference between (38) and (39), depending on $Q_{\mathrm{WL}}(\hat{m}^1_{\mathrm{RWL}}, \bar{m}^1_{\mathrm{WL}}; \bar{\gamma}^1_{\mathrm{CAL}})$, which we seek to control. Carrying this bound leads to a quadratic inequality in $Q_{\mathrm{WL}}(\hat{m}^1_{\mathrm{RWL}}, \bar{m}^1_{\mathrm{WL}}; \bar{\gamma}^1_{\mathrm{CAL}})$, which can be inverted to obtain an explicit bound on $Q_{\mathrm{WL}}(\hat{m}^1_{\mathrm{RWL}}, \bar{m}^1_{\mathrm{WL}}; \bar{\gamma}^1_{\mathrm{CAL}})$. The resulting error bound (37) is of order $(|S_\gamma| + |S_{\alpha^1}|)\log(p)/n$, much sharper than could be obtained using other approaches, for example, directly bounding $\|\tilde{E}[Tw(X; \hat{\gamma}^1_{\mathrm{RCAL}})\{Y - \bar{m}^1_{\mathrm{WL}}(X)\} f(X)]\|_\infty$. There are also similar issues with estimated weights handled in van de Geer et al. (2014) and Belloni et al. (2018).

Finally, we study the proposed estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ for μ^1, depending on the regularized estimators $\hat{\gamma}^1_{\mathrm{RCAL}}$ and $\hat{\alpha}^1_{\mathrm{RWL}}$ from logistic propensity score model (4) and linear outcome regression model (9). The following result gives an error bound for $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$, allowing that both models (4) and (9) may be misspecified.

THEOREM 3. *Under the conditions of Theorem 2, if* $\log\{(1+p)/\epsilon\}/n \le 1$, *then we have with probability at least* $1-10\epsilon$,

$$\begin{aligned}(40)\qquad &|\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) - \hat{\mu}^1(\bar{m}^1_{\mathrm{WL}}, \bar{\pi}^1_{\mathrm{CAL}})| \\ &\quad \le M_{11}|S_\gamma|\lambda_0^2 + M_{12}|S_\gamma|\lambda_0\lambda_1 + M_{13}|S_{\alpha^1}|\lambda_0\lambda_1,\end{aligned}$$

where $M_{11} = M_{13} + \sqrt{D_0^2 + D_1^2}\mathrm{e}^{\eta_{01}}(\mathrm{e}^{\eta_{01}}M_0/2 + \eta_{02})$, $M_{12} = (A_0-1)^{-1}M_0$, $M_{13} = A_0(A_1-1)^{-1}M_1$, *and* M_1 *is a constant such that the right-hand side of* (37) *in Theorem 2 is upper bounded by* $\mathrm{e}^{\eta_{01}}M_1(|S_\gamma|\lambda_0\lambda_1 + |S_{\alpha^1}|\lambda_1^2)$.

Theorem 3 shows that $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ is doubly robust for μ^1 provided $(|S_\gamma| + |S_{\alpha^1}|)\lambda_1^2 = o(1)$, that is, $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n)$. In addition, Theorem 3 gives the $n^{-1/2}$ asymptotic expansion (20) provided $n^{1/2}(|S_\gamma| + |S_{\alpha^1}|)\lambda_1^2 = o(1)$, that is, $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$. To obtain valid confidence intervals for μ^1 via the Slutsky theorem, the following result gives the convergence of the variance estimator $\hat{V}$ to V, as defined in Proposition 1, allowing that both models (4) and (9) may be misspecified. For notational simplicity, denote $\hat{\varphi} = \varphi(T, Y, X; \hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ and $\hat{\varphi}_c = \hat{\varphi} - \hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ such that $\hat{V} = \tilde{E}(\hat{\varphi}_c^2)$. Similarly, denote $\bar{\varphi} = \varphi(T, Y, X; \bar{m}^1_{\mathrm{WL}}, \bar{\pi}^1_{\mathrm{CAL}})$ and $\bar{\varphi}_c = \bar{\varphi} - \hat{\mu}^1(\bar{m}^1_{\mathrm{WL}}, \bar{\pi}^1_{\mathrm{CAL}})$ such that $V = E(\bar{\varphi}_c^2)$.

THEOREM 4. *Under the conditions of Theorem 2, if* $\log\{(1+p)/\epsilon\}/n \le 1$, *then we have with probability at least* $1-10\epsilon$,

$$\begin{aligned}(41)\qquad |\tilde{E}(\hat{\varphi}_c^2 - \bar{\varphi}_c^2)| &\le 2M_{14}\{\tilde{E}(\bar{\varphi}_c^2)\}^{1/2}(|S_\gamma|\lambda_0 + |S_{\alpha^1}|\lambda_1) \\ &\quad + M_{14}(|S_\gamma|\lambda_0 + |S_{\alpha^1}|\lambda_1)^2,\end{aligned}$$

where M_{14} *is a positive constant depending only on* $(A_0, C_0, B_0, \xi_0, \nu_0, \eta_0)$ *in Theorem 1 and* $(A_1, D_0, D_1, \xi_1, \nu_1, \eta_1)$ *in Theorem 2. If, in addition, condition* (35) *holds, then we have with probability at least* $1-12\epsilon$,

$$\begin{aligned}(42)\qquad |\tilde{E}(\hat{\varphi}_c^2 - \bar{\varphi}_c^2)| &\le 2M_{15}\{\tilde{E}(\bar{\varphi}_c^2)\}^{1/2}(|S_\gamma|\lambda_0\lambda_1 + |S_{\alpha^1}|\lambda_1^2)^{1/2} \\ &\quad + M_{15}(|S_\gamma|\lambda_0\lambda_1 + |S_{\alpha^1}|\lambda_1^2),\end{aligned}$$

where M_{15} *is a positive constant, depending on* τ_0 *from* (35) *as well as* $(A_0, C_0, B_0, \xi_0, \nu_0, \eta_0)$ *and* $(A_1, D_0, D_1, \xi_1, \nu_1, \eta_1)$.

REMARK 8. Theorem 4 provides two rates of convergence for $\hat{V}$ under different conditions. Inequality (41) shows that $\hat{V}$ is a consistent estimator of V, that is, $\hat{V} - V = o_p(1)$, provided $(|S_\gamma| + |S_{\alpha^1}|)(\log p)^{1/2} = o(n^{1/2})$. Technically, consistency of $\hat{V}$ is sufficient for applying Slutsky theorem to establish confidence intervals for μ^1 in Proposition 1(iii). With additional condition (35), inequality (42) shows that $\hat{V}$ achieves the parametric rate of convergence, $\hat{V} - V = O_p(n^{-1/2})$, provided $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$.

REMARK 9. Combining Theorems 3–4 directly leads to Proposition 1, which gives doubly robust confidence intervals of μ^1. In addition, a broader interpretation of robust inference can be accommodated. All the results, Theorems 1–4, are developed to remain valid in the presence of misspecification of models (4) and (9), similarly as in classical theory of estimation with misspecified models (e.g., White (1982)). If both models (4) and (9) may be misspecified, then $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) \pm z_{c/2}\sqrt{\hat{V}/n}$ is an asymptotic $(1-c)$ confidence interval for the target value $\bar{\mu}^1 = E(\bar{\varphi})$, which in general differs from the true value μ^1. The *only* effect of models (4) and (9) being correctly specified is to ensure that $\bar{\mu}^1$ coincides with μ^1. Such robust estimation of variances is conceptually similar to the fact that White's (1980) sandwich variance estimator is valid for the least-squares estimator even when a linear model may be misspecified. For comparison, the standard estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}^1_{\mathrm{RML}})$ can also be shown to converge to a target value $\bar{\mu}^1_{\mathrm{ML}}$, which by double robustness, coincides with μ^1 if either model (4) or (9) is correctly specified, but otherwise differs from μ^1. As discussed in Section 3.2, asymptotic expansion (8) for $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}^1_{\mathrm{RML}})$ is in general invalid, and so are the associated confidence intervals for $\bar{\mu}^1_{\mathrm{ML}}$, unless both models (4) or (9) are correctly specified.

From the preceding discussion about model robustness, we obtain a direct extension of Proposition 1, allowing models (4) and (9) to be misspecified by some (small) bias terms, defined as

$$b_\gamma = \tilde{E}^{1/2}\big[\{1/\bar{\pi}^1_{\mathrm{CAL}}(X) - 1/\pi^*(X)\}^2\big],$$
$$b_{\alpha 1} = \tilde{E}^{1/2}\big[\{\bar{m}^1_{\mathrm{WL}}(X) - m^{1*}(X)\}^2\big].$$

These bias terms can be linked to those in Belloni, Chernozhukov and Hansen (2014) and Farrell (2015) by the following bounds (shown in the Supplement):

$$E(b_\gamma^2) \le O(1) \inf_\gamma E\big[\{1/\pi(X;\gamma) - 1/\pi^*(X)\}^2\big], \tag{43}$$

$$E(b_{\alpha 1}^2) \le O(1) \inf_{\alpha^1} E\big[\{m^1(X;\alpha^1) - m^{1*}(X)\}^2\big], \tag{44}$$

where $O(1)$ in each inequality depends only on constants (B_0, B_1) and (B_0^*, B_1^*) such that $\log\{\bar{\pi}^1_{\mathrm{CAL}}(X)/(1-\bar{\pi}^1_{\mathrm{CAL}}(X))\} \in [B_0, B_1]$ almost surely and $\log\{\pi^*(X)/(1-\pi^*(X))\} \in [B_0^*, B_1^*]$ almost surely.

PROPOSITION 2. *In addition to Assumptions 1 and 2, suppose that $\pi^*(X) \ge \mathrm{e}^{B_0^*}/(1+\mathrm{e}^{B_0^*})$ almost surely for a constant $B_0^* \in \mathbb{R}$. Then for sufficiently large constants A_0 and A_1,*

$$\begin{aligned} &|\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) - \hat{\mu}^1(\bar{m}^1_{\mathrm{WL}}, \pi^*)| \\ &\quad \le (|S_\gamma| + |S_{\alpha^1}|) O_p(\log(p)/n) + b_\gamma O_p(n^{-1/2}) + b_\gamma b_{\alpha 1}, \end{aligned} \tag{45}$$

$$\begin{aligned} &|\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) - \hat{\mu}^1(m^{1*}, \bar{\pi}^1_{\mathrm{CAL}})| \\ &\quad \le (|S_\gamma| + |S_{\alpha^1}|) O_p(\log(p)/n) + b_{\alpha 1} O_p(n^{-1/2}) + b_\gamma b_{\alpha 1}. \end{aligned} \tag{46}$$

Moreover, if $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$, $b_\gamma b_{\alpha 1} = o_p(n^{-1/2})$, and $b_\gamma = o_p(1)$ or $b_{\alpha 1} = o_p(1)$ (or both), then (i)–(iii) in Proposition 1 hold, with $\bar{\pi}^1_{\mathrm{CAL}}$ replaced by π^* or $\bar{m}^1_{\mathrm{WL}}$ replaced by m^{1*} (or both) in the definition of V.

REMARK 10. For $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}^1_{\mathrm{RML}})$ to admit asymptotic expansion (8), the rate condition in Farrell ((2015), Corrigendum) amounts to

$$\{|S_\gamma| \log^{\frac{3}{2}+\delta}(p)/n + b_\gamma^2\}^{\frac{1}{2}} \{|S_{\alpha^1}| \log^{\frac{3}{2}+\delta}(p)/n + b_{\alpha 1}^2\}^{\frac{1}{2}} = o_p(n^{-\frac{1}{2}}) \tag{47}$$

and $|S_{\alpha^1}| \log^{3/2+\delta}(p)/n + b_{\alpha 1}^2 = o_p(n^{-1/2})$ for some $\delta > 0$. The second condition may be dropped via cross-fitting in Chernozhukov et al. (2018). If b_γ^2 and $b_{\alpha 1}^2$ are of no larger order than, respectively, $|S_{\alpha^1}|\log^{3/2+\delta}(p)/n$ and $|S_{\alpha^1}|\log^{3/2+\delta}(p)/n$, then condition (47) reduces to $|S_\gamma||S_{\alpha^1}|\log^{3+2\delta}(p) = o(n)$, which, ignoring $\log(p)$ factors, is weaker than $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$: one of $|S_\gamma|$ and $|S_{\alpha^1}|$ can be large while the other is small enough with the product $|S_\gamma||S_{\alpha^1}|$ small. On the other hand, if b_γ or $b_{\alpha 1}$ is relatively large such that $|S_{\alpha^1}|b_\gamma^2$ or $|S_\gamma|b_{\alpha 1}^2$ is not $o_p(1)$, then condition (47) fails, but the right-hand side of (45) or (46) can be $o_p(n^{-1/2})$ provided $b_\gamma = o_p(1)$ or $b_{\alpha 1} = o_p(1)$, $b_\gamma b_{\alpha 1} = o_p(n^{-1/2})$, and $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$. This discussion shows the following comparison: $\hat{\mu}^1(\hat{m}^1_{\text{RML}}, \hat{\pi}^1_{\text{RML}})$ leads to valid confidence intervals under weaker sparsity conditions when both models (4) and (9) are nearly correct, but $\hat{\mu}^1(\hat{m}^1_{\text{RWL}}, \hat{\pi}^1_{\text{RCAL}})$ allows valid confidence intervals when either model (4) or (9) is nearly correct.

REMARK 11. Although our theory is developed in high-dimensional, parametric settings, there can be interesting implications in nonparametric settings where, for series estimation, $f_1(x), \ldots, f_p(x)$ are specified as basis functions of $x \in \mathbb{R}^d$ with the number of terms p growing with the sample size n. Denote by $(\hat{m}^1_{\text{RML}}, \hat{\pi}_{\text{RML}})$ and $(\hat{m}^1_{\text{RWL}}, \hat{\pi}^1_{\text{RCAL}})$ the corresponding series estimators. Assume that $m^{1*}(x)$ and $\pi^*(x)$ belong to Hölder classes with r_{m1}- and r_π-times continuous derivatives. By bias and variance formulas in series estimation (Newey (1997)), it might be expected from Proposition 2 that

$$\begin{aligned}
&|\hat{\mu}^1(\hat{m}^1_{\text{RWL}}, \hat{\pi}^1_{\text{RCAL}}) - \hat{\mu}^1(m^{1*}, \pi^*)| \\
&\quad \le \{p/n + (p^{-r_\pi/d} + p^{-r_{m1}/d})n^{-1/2} + p^{-(r_\pi + r_{m1})/d}\} O_p(1).
\end{aligned} \tag{48}$$

By choosing $p \propto n^{d/(r_\pi + r_{m1} + d)}$ to balance the two terms p/n and $p^{-(r_\pi + r_{m1})/d}$, the right-hand side of (48) is minimized as $O_p(n^{-(r_\pi + r_{m1})/(r_\pi + r_{m1} + d)})$, which becomes $o_p(n^{-1/2})$ if

$$\frac{r_\pi + r_{m1}}{r_\pi + r_{m1} + d} > \frac{1}{2} \quad \text{that is, } r_\pi + r_{m1} > d. \tag{49}$$

As the mean squared errors of $\hat{\pi}_{\text{RML}}$ and $\hat{m}^1_{\text{RML}}$ are of order $n^{-2r_\pi/(2r_\pi + d)}$ and $n^{-2r_{m1}/(2r_{m1} + d)}$ for suitable choices of p, respectively, it can be shown similarly as (47) that $\hat{\mu}^1(\hat{m}^1_{\text{RML}}, \hat{\pi}_{\text{RML}})$ satisfies asymptotic expansion (8) if $n^{-r_\pi/(2r_\pi + d)} n^{-r_{m1}/(2r_{m1} + d)} O_p(1) = o_p(n^{-1/2})$, that is,

$$\frac{r_\pi}{2r_\pi + d} + \frac{r_{m1}}{2r_{m1} + d} > \frac{1}{2}. \tag{50}$$

Condition (49) is weaker than (50), because $(r_\pi + r_{m1})/(r_\pi + r_{m1} + d) \ge r_\pi/(2r_\pi + d) + r_{m1}/(2r_{m1} + d)$ by Cauchy–Schwarz inequality, $2/(r_\pi + r_{m1} + d) \le 1/(2r_\pi + d) + 1/(2r_{m1} + d)$, where equality holds if and only if $r_\pi = r_{m1}$. On the other hand, under a weaker condition than (49), $r_\pi + r_{m1} > d/2$, an estimator of μ^1 is developed in Robins et al. (2017) to achieve asymptotic expansion in the form (8), using higher-order influence functions. We leave further investigation in this direction to future work.

3.4. *Theory with generalized linear outcome models.* In this section, we turn to the situation where a generalized linear model is used for outcome regression together with a logistic propensity score model, and develop appropriate methods and theory for obtaining confidence intervals for μ^1 in high-dimensional settings.

A technical complication compared with the situation of a linear outcome model in Section 3.3 is that the orthogonality conditions (23)–(24) with a nonlinear outcome model do not lead to as simple a pair of estimating functions as (31)–(32), which can be inverted to define loss functions in γ and α^1 sequentially. There are, however, different approaches that can be used to derive model-assisted confidence intervals, that is, satisfying either property

(G2) or (G3) described in Section 3.1. For concreteness, we focus on a PS based, OR assisted approach to obtain confidence intervals with property (G2), that is, being valid if the propensity score model used is correctly specified but the outcome regression model may be misspecified. See Section 3.5 for further discussion of related issues.

Consider a logistic propensity score model (4) and a generalized linear outcome model with a canonical link,

$$E(Y|T=1,X) = m^1(X;\alpha^1) = \psi\{\alpha^{1\mathrm{T}} f(X)\}, \tag{51}$$

that is, model (1) with the vector of covariate functions $g^1(X)$ taken to be the same as $f(X)$ in model (4). This choice of covariate functions can be more justified than in the setting of Section 3.3, because OR model (51) plays an assisting role when confidence intervals for μ^1 are concerned. Our point estimator of μ^1 is $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ as defined in (10), where $\hat{\pi}^1_{\mathrm{RCAL}}(X) = \pi(X; \hat{\gamma}^1_{\mathrm{RCAL}})$ and $\hat{m}^1_{\mathrm{RWL}}(X) = m^1(X; \hat{\alpha}^1_{\mathrm{RWL}})$. The estimator $\hat{\gamma}^1_{\mathrm{RCAL}}$ is a regularized calibrated estimator of γ from Tan (2017) as in Section 3.3. But $\hat{\alpha}^1_{\mathrm{RWL}}$ is a regularized weighted likelihood estimator of α^1, defined as a minimizer of

$$\ell_{\mathrm{RWL}}(\alpha^1; \hat{\gamma}^1_{\mathrm{RCAL}}) = \ell_{\mathrm{WL}}(\alpha^1; \hat{\gamma}^1_{\mathrm{RCAL}}) + \lambda \|\alpha^1_{1:p}\|_1, \tag{52}$$

where $\lambda \ge 0$ is a tuning parameter and $\ell_{\mathrm{WL}}(\alpha^1; \hat{\gamma}^1_{\mathrm{RCAL}})$ is the weighted likelihood loss as follows, with $w(X;\gamma) = \{1-\pi(X;\gamma)\}/\pi(X;\gamma) = \mathrm{e}^{-\gamma^{\mathrm{T}} f(X)}$,

$$\ell_{\mathrm{WL}}(\alpha^1; \hat{\gamma}^1_{\mathrm{RCAL}}) = \tilde{E}\big(T w(X; \hat{\gamma}^1_{\mathrm{RCAL}})[-Y\alpha^{1\mathrm{T}} f(X) + \Psi\{\alpha^{1\mathrm{T}} f(X)\}]\big). \tag{53}$$

The regularized weighted least-squares estimator $\hat{\alpha}^1_{\mathrm{RWL}}$ for a linear outcome model in Section 3.3 is recovered in the special case of the identity link, $\psi(u) = u$ and $\Psi(u) = u^2/2$. In addition, the Kuhn–Tucker–Karush condition for minimizing (52) remains the same as (17)–(18), and hence the estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ can be put in the prediction form (19), which ensures the boundedness property that $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ always falls within the range of the observed outcomes $\{Y_i : T_i = 1, i = 1, \dots, n\}$ and the predicted values $\{\hat{m}^1_{\mathrm{RWL}}(X_i) : T_i = 0, i = 1, \dots, n\}$.

With possible model misspecification, the target value $\bar{\alpha}^1_{\mathrm{WL}}$ is defined as a minimizer of the expected loss $E\{\ell_{\mathrm{WL}}(\alpha^1; \bar{\gamma}^1_{\mathrm{CAL}})\}$. From a functional perspective, we write $\ell_{\mathrm{WL}}(\alpha^1;\gamma) = \kappa_{\mathrm{WL}}(\alpha^{1\mathrm{T}} f; \gamma)$, where for a function $h(x)$ which may not be in the form $\alpha^{1\mathrm{T}} f$,

$$\kappa_{\mathrm{WL}}(h;\gamma) = \tilde{E}\big(T w(X;\gamma)[-Yh(X) + \Psi\{h(X)\}]\big).$$

As $\kappa_{\mathrm{WL}}(h;\gamma)$ is convex in h by the convexity of $\Psi(\cdot)$, the Bregman divergence associated with $\kappa_{\mathrm{WL}}(h;\gamma)$ is defined as

$$D_{\mathrm{WL}}(h', h; \gamma) = \kappa_{\mathrm{WL}}(h';\gamma) - \kappa_{\mathrm{WL}}(h;\gamma) - \langle \nabla\kappa_{\mathrm{WL}}(h;\gamma), h' - h\rangle,$$

where $\nabla\kappa_{\mathrm{WL}}(h;\gamma)$ denotes the gradient of $\kappa_{\mathrm{WL}}(h;\gamma)$ with respect to $(h_1, \dots, h_n)$ with $h_i = h(X_i)$. The symmetrized Bregman divergence is

$$\begin{aligned} D^{\dagger}_{\mathrm{WL}}(h', h; \gamma) &= D_{\mathrm{WL}}(h', h; \gamma) + D_{\mathrm{WL}}(h, h'; \gamma) \\ &= \tilde{E}\big(T w(X;\gamma)[\psi\{h'(X)\} - \psi\{h(X)\}]\{h'(X) - h(X)\}\big). \end{aligned} \tag{54}$$

The following result establishes the convergence of $\hat{\alpha}^1_{\mathrm{RWL}}$ to $\bar{\alpha}^1_{\mathrm{WL}}$ in the L_1 norm $\|\hat{\alpha}^1_{\mathrm{RWL}} - \bar{\alpha}^1_{\mathrm{WL}}\|_1$ and the symmetrized Bregman divergence $D^{\dagger}_{\mathrm{WL}}(\hat{h}^1_{\mathrm{RWL}}, \bar{h}^1_{\mathrm{WL}}; \bar{\gamma}^1_{\mathrm{CAL}})$, where $\hat{h}^1_{\mathrm{RWL}}(X) = \hat{\alpha}^{1\mathrm{T}}_{\mathrm{RWL}} f(X)$ and $\bar{h}^1_{\mathrm{WL}}(X) = \bar{\alpha}^{1\mathrm{T}}_{\mathrm{WL}} f(X)$. In the case of the identity link, $\psi(u) = u$, the symmetrized Bregman divergence $D^{\dagger}_{\mathrm{WL}}(\hat{h}^1_{\mathrm{RWL}}, \bar{h}^1_{\mathrm{WL}}; \bar{\gamma}^1_{\mathrm{CAL}})$ becomes $Q_{\mathrm{WL}}(\hat{m}^1_{\mathrm{RWL}}, \bar{m}^1_{\mathrm{WL}}; \bar{\gamma}^1_{\mathrm{CAL}})$ in (36). Inequality (55) also reduces to (37) in Theorem 2 with the choices $C_2 = 1$ and $C_3 = \eta_2 = \eta_3 = 0$.

ASSUMPTION 3. Assume that $\psi(\cdot)$ is differentiable and denote $\psi_2(u) = \mathrm{d}\psi(u)/\mathrm{d}u$. Suppose that the following conditions are satisfied:

(i) $\psi_2\{\bar{h}^1_{\mathrm{WL}}(X)\} \le C_1$ almost surely for a constant $C_1 > 0$;

(ii) $\psi_2\{\bar{h}^1_{\mathrm{WL}}(X)\} \ge C_2$ almost surely for a constant $C_2 > 0$;

(iii) $\psi_2(u) \le \psi_2(u')\mathrm{e}^{C_3|u-u'|}$ for any (u, u'), where $C_3 \ge 0$ is a constant.

(iv) $C_0 C_3 (A_1 - 1)^{-1} \xi_3^2 \nu_2^{-2} C_2^{-1} |S_{\alpha^1}| \lambda_1 \le \eta_2$ for a constant $0 \le \eta_2 < 1$ and $C_0 C_3 \mathrm{e}^{3\eta_{01}} (A_1 - 1)^{-1} \xi_2^{-2} C_2^{-1} (M_{01} |S_\gamma| \lambda_0) \le \eta_3$ for a constant $0 \le \eta_3 < 1$, where $(\eta_{01}, \nu_2, \xi_2, \xi_3, M_{01})$ are as in Theorem 2.

THEOREM 5. *Suppose that Assumptions* 1, 2 *and* 3(*ii*)–(*iv*) *hold. If* $\log\{(1+p)/\epsilon\}/n \le 1$, *then for* $A_0 > (\xi_0+1)/(\xi_0-1)$ *and* $A_1 > (\xi_1+1)/(\xi_1-1)$, *we have with probability at least* $1 - 8\epsilon$,

$$
\begin{aligned}
&D^\dagger_{\mathrm{WL}}(\hat{m}^1_{\mathrm{RWL}}, \bar{m}^1_{\mathrm{WL}}) + \mathrm{e}^{\eta_{01}}(A_1 - 1)\lambda_1 \|\hat{\alpha}^1_{\mathrm{RWL}} - \bar{\alpha}^1_{\mathrm{WL}}\|_1 \\
&\quad \le \mathrm{e}^{4\eta_{01}} \xi_4^{-2} (M_{01} |S_\gamma| \lambda_0^2) + \mathrm{e}^{2\eta_{01}} \xi_3^2 (\nu_3^{-2} |S_{\alpha^1}| \lambda_1^2),
\end{aligned}
\tag{55}
$$

where $\xi_4 = \xi_2(1-\eta_3)^{1/2} C_2^{1/2}$, $\nu_3 = \nu_2^{1/2}(1-\eta_2)^{1/2} C_2^{1/2}$, *and* $(\eta_{01}, \nu_2, \xi_2, \xi_3, M_{01})$ *are as in Theorem* 2.

REMARK 12. We discuss the conditions involved in Theorem 5. Assumption 3(i) is not needed, but will be used in later results. Assumption 3(iii), adapted from Huang and Zhang (2012), is used along with Assumption 1(i) to bound the curvature of $D^\dagger_{\mathrm{WL}}(h', h; \bar{\gamma}^1_{\mathrm{CAL}})$ and then with Assumption 3(iv) to achieve a localized analysis when handling a non-quadratic loss function. Assumption 3(ii) is used for two distinct purposes. First, it is combined with Assumptions 2(ii)–(iii) to yield a compatibility condition for $\tilde{\Sigma}_\alpha = \tilde{E}[T w(X; \bar{\gamma}^1_{\mathrm{CAL}}) \psi_2\{\bar{h}^1_{\mathrm{WL}}(X)\} f(X) f^{\mathrm{T}}(X)]$, which is the sample version of the Hessian of the expected loss $E\{\ell_{\mathrm{WL}}(\alpha^1; \bar{\gamma}^1_{\mathrm{CAL}})\}$ at $\alpha^1 = \bar{\alpha}^1_{\mathrm{WL}}$, that is, $\Sigma_\alpha = E[T w(X; \bar{\gamma}^1_{\mathrm{CAL}}) \psi_2\{\bar{h}^1_{\mathrm{WL}}(X)\} \times f(X) f^{\mathrm{T}}(X)]$. Second, Assumption 3(ii) is also used in deriving a quadratic inequality to be inverted in our strategy to deal with the dependency of $\hat{\alpha}^1_{\mathrm{RWL}}$ on $\hat{\gamma}^1_{\mathrm{RCAL}}$ as mentioned in Remark 7. As seen from the proofs in Supplementary Material, similar results as in Theorem 5 can be obtained with Assumption 3(ii) replaced by the weaker condition that for some constant $\tau_1 > 0$,

$$
b^{\mathrm{T}} \Sigma_\gamma b \le (b^{\mathrm{T}} \Sigma_\alpha b)/\tau_1 \quad \forall b \in \mathbb{R}^{1+p},
$$

provided that the condition on A_1 and Assumption 3(iv) are modified accordingly, depending on τ_1. This extension is not pursued here for simplicity.

Now we study the proposed estimator $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ for μ^1, with the regularized estimators $\hat{\gamma}^1_{\mathrm{RCAL}}$ and $\hat{\alpha}^1_{\mathrm{RWL}}$ obtained using logistic propensity score model (4) and generalized linear outcome model (51). Theorem 6 gives an error bound for $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$, allowing that both models (4) and (51) may be misspecified, but depending on additional terms in the presence of misspecification of model (4). Denote $h(X; \alpha^1) = \alpha^{1\mathrm{T}} f(X)$ and for $r \ge 0$,

$$
\Lambda_0(r) = \sup_{j=0,1,\ldots,p,\, \|\alpha^1 - \bar{\alpha}^1_{\mathrm{WL}}\|_1 \le r} \left| E\left[\psi_2\{h(X; \alpha^1)\} f_j(X) \left\{ \frac{T}{\bar{\pi}^1_{\mathrm{CAL}}(X)} - 1 \right\} \right] \right|.
$$

As a special case, the quantity $\Lambda_0(0)$ is defined as

$$
\Lambda_1 = \sup_{j=0,1,\ldots,p} \left| E\left[\psi_2\{\bar{h}^1_{\mathrm{WL}}(X)\} f_j(X) \left\{ \frac{T}{\bar{\pi}^1_{\mathrm{CAL}}(X)} - 1 \right\} \right] \right|.
$$

By the definition of $\bar{\gamma}^1_{\text{CAL}}$, it holds that $E[\{T/\bar{\pi}^1_{\text{CAL}}(X) - 1\} f_j(X)] = 0$ for $j = 0, 1, \dots, p$ whether or not model (4) is correctly specified. But $\Lambda_0(r)$ is in general either zero or positive respectively if model (4) is correctly specified or misspecified, except in the case of linear outcome model (9) where $\Lambda_0(r)$ is automatically zero because $\psi_2(\cdot)$ is constant.

THEOREM 6. *Suppose that Assumptions 1, 2 and 3 hold. If* $\log\{(1+p)/\epsilon\}/n \le 1$, *then for* $A_0 > (\xi_0 + 1)/(\xi_0 - 1)$ *and* $A_1 > (\xi_1 + 1)/(\xi_1 - 1)$, *we have with probability at least* $1 - 12\epsilon$,

$$\begin{aligned} &|\hat{\mu}^1(\hat{m}^1_{\text{RWL}}, \hat{\pi}^1_{\text{RCAL}}) - \hat{\mu}^1(\bar{m}^1_{\text{WL}}, \bar{\pi}^1_{\text{CAL}})| \\ &\quad \le M_{21}|S_\gamma|\lambda_0^2 + M_{22}|S_\gamma|\lambda_0\lambda_1 + M_{23}|S_{\alpha^1}|\lambda_0\lambda_1 + \eta_{11}\Lambda_0(\eta_{11}), \end{aligned} \tag{56}$$

where M_{21}, M_{22} *and* M_{23} *are positive constants, depending only on* $(A_0, C_0, B_0, \xi_0, \nu_0, \eta_0)$, $(A_1, D_0, D_1, \xi_1, \nu_1, \eta_1)$, *and* $(C_1, C_2, C_3, \eta_2, \eta_3)$, $\eta_{11} = (A_1 - 1)^{-1} M_2(|S_\gamma|\lambda_0 + |S_{\alpha^1}|\lambda_1)$, *and* M_2 *is a constant such that the right-hand side of* (55) *is upper bounded by* $\mathrm{e}^{\eta_{01}} M_2(|S_\gamma|\lambda_0\lambda_1 + |S_{\alpha^1}|\lambda_1^2)$. *If, in addition, condition* (35) *holds, then we have with probability at least* $1 - 14\epsilon$,

$$\begin{aligned} &|\hat{\mu}^1(\hat{m}^1_{\text{RWL}}, \hat{\pi}^1_{\text{RCAL}}) - \hat{\mu}^1(\bar{m}^1_{\text{WL}}, \bar{\pi}^1_{\text{CAL}})| \\ &\quad \le M_{24}|S_\gamma|\lambda_0^2 + M_{25}|S_\gamma|\lambda_0\lambda_1 + M_{26}|S_{\alpha^1}|\lambda_0\lambda_1 + \eta_{11}\Lambda_1, \end{aligned} \tag{57}$$

where M_{24}, M_{25} *and* M_{26} *are positive constants, similar to* M_{21}, M_{22} *and* M_{23}, *but, in addition, depending on* τ_0 *from* (35).

REMARK 13. Two different error bounds are obtained in Theorem 6. Because $\Lambda_0(\eta_{11}) \ge \Lambda_1$, the error bound (57) is tighter than (56), but with the additional condition (35), which requires that the generalized eigenvalues of Σ_γ relative to the gram matrix $E\{f(X)f^{\text{T}}(X)\}$ is bounded away from 0. In either case, the result shows that $\hat{\mu}^1(\hat{m}^1_{\text{RWL}}, \hat{\pi}^1_{\text{RCAL}})$ is doubly robust for μ^1 provided $(|S_\gamma| + |S_{\alpha^1}|)\lambda_1 = o(1)$, that is, $(|S_\gamma| + |S_{\alpha^1}|)(\log p)^{1/2} = o(n^{1/2})$. In addition, the error bounds imply that $\hat{\mu}^1(\hat{m}^1_{\text{RWL}}, \hat{\pi}^1_{\text{RCAL}})$ admits the $n^{-1/2}$ asymptotic expansion (20) provided $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$, when PS model (4) is correctly specified but OR model (51) may be misspecified, because the term involving $\Lambda_0(\eta_{11})$ or Λ_1 vanishes as discussed above. Unfortunately, expansion (20) may fail when PS model (4) is misspecified.

Similarly as Theorem 4, the following result establishes the convergence of $\hat{V}$ to V as defined in Proposition 1, allowing that both models (4) and (51) may be misspecified.

THEOREM 7. *Under the conditions of Theorem 6, if* $\log\{(1+p)/\epsilon\}/n \le 1$, *then we have with probability at least* $1 - 12\epsilon$,

$$\begin{aligned} |\tilde{E}(\hat{\varphi}_c^2 - \bar{\varphi}_c^2)| &\le 2M_{27}\{\tilde{E}(\bar{\varphi}_c^2)\}^{1/2}\{1 + \Lambda_0(\eta_{11})\}(|S_\gamma|\lambda_0 + |S_{\alpha^1}|\lambda_1) \\ &\quad + M_{27}\{1 + \Lambda_0^2(\eta_{11})\}(|S_\gamma|\lambda_0 + |S_{\alpha^1}|\lambda_1)^2, \end{aligned} \tag{58}$$

where M_{27} *is a positive constant depending only on* $(A_0, C_0, B_0, \xi_0, \nu_0, \eta_0)$, $(A_1, D_0, D_1, \xi_1, \nu_1, \eta_1)$ *and* $(C_1, C_2, C_3, \eta_2, \eta_3)$. *If, in addition, condition* (35) *holds, then we have with probability at least* $1 - 14\epsilon$,

$$\begin{aligned} &|\tilde{E}(\hat{\varphi}_c^2 - \bar{\varphi}_c^2)| \\ &\quad \le 2M_{28}\{\tilde{E}(\bar{\varphi}_c^2)\}^{1/2}\{(|S_\gamma|\lambda_0\lambda_1 + |S_{\alpha^1}|\lambda_1^2)^{1/2} + \Lambda_1(|S_\gamma|\lambda_0 + |S_{\alpha^1}|\lambda_1)\} \\ &\qquad + M_{28}\{(|S_\gamma|\lambda_0\lambda_1 + |S_{\alpha^1}|\lambda_1^2) + \Lambda_1^2(|S_\gamma|\lambda_0 + |S_{\alpha^1}|\lambda_1)^2\}, \end{aligned} \tag{59}$$

where M_{28} *is a positive constant, similar to* M_{27} *but, in addition, depending on* τ_0 *from* (35).

REMARK 14. Two different rates of convergence are obtained for $\hat{V}$ in Theorem 7. Similarly, as discussed in Remark 8, if $(|S_\gamma| + |S_{\alpha^1}|)(\log p)^{1/2} = o(n^{1/2})$, then inequality (58) implies the consistency of $\hat{V}$ for V, which is sufficient for applying the Slutsky theorem to establish confidence intervals for μ^1. With an additional condition (35), inequality (59) gives a faster rate of convergence of $\hat{V}$ to V, which is of order $n^{-1/2}$ provided $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$.

Combining Theorems 6–7 leads to the following result.

PROPOSITION 3. *Suppose that Assumptions 1, 2 and 3 hold, and $(|S_\gamma| + |S_{\alpha^1}|)\log(p) = o(n^{1/2})$. For sufficiently large constants A_0 and A_1, if logistic PS model (4) is correctly specified but OR model (51) may be misspecified, then $\bar{\pi}^1_{\mathrm{CAL}}(x) \equiv \pi^*(x)$, and (i)–(iii) in Proposition 1 hold. That is, a PS based, OR assisted confidence interval for μ^1 is obtained.*

REMARK 15. The conclusion of Proposition 3 remains valid if PS model (4) is misspecified but only locally such that $\Lambda_0(\eta_{11}) = O(\{\log(p)/n\}^{1/2})$ or $\Lambda_1 = O(\{\log(p)/n\}^{1/2})$, in the case of the error bound (56) or (57). Therefore, $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) \pm z_{c/2}\sqrt{\hat{V}/n}$ can be interpreted as an asymptotic $(1-c)$ confidence interval for the target value $\bar{\mu}^1 = E(\bar{\varphi})$ if model (4) is at most locally misspecified but model (51) may be arbitrarily misspecified. It is an interesting open problem to find broadly valid confidence intervals in the presence of model misspecification similarly as discussed in Remark 9 when a linear outcome model is used.

3.5. *Further discussion.*

Estimation of ATE. Our theory and methods are presented mainly on estimation of μ^1, but they can be directly extended for estimating μ^0 and hence ATE, that is, $\mu^1 - \mu^0$. Consider a logistic propensity score model (4) and a generalized linear outcome model,

$$E(Y|T=0, X) = m_0(X; \alpha^0) = \psi\{\alpha^{0\mathrm{T}} f(X)\}, \tag{60}$$

where $f(X)$ is the same vector of covariate functions as in the model (4) and α^0 is a vector of unknown parameters. Our point estimator of ATE is $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}}) - \hat{\mu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}})$, and that of μ^0 is

$$\hat{\mu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}}) = \tilde{E}\{\varphi(Y, 1-T, X; \hat{m}^0_{\mathrm{RWL}}, 1 - \hat{\pi}^0_{\mathrm{RCAL}})\},$$

where $\varphi(\cdot)$ is defined in (7), $\hat{\pi}^0_{\mathrm{RCAL}}(X) = \pi(X; \hat{\gamma}^0_{\mathrm{RCAL}})$, $\hat{m}^0_{\mathrm{RWL}}(X) = m_0(X; \hat{\alpha}^0_{\mathrm{RWL}})$, and $\hat{\gamma}^0_{\mathrm{RCAL}}$ and $\hat{\alpha}^0_{\mathrm{RWL}}$ are defined as follows. The estimator $\hat{\gamma}^0_{\mathrm{RCAL}}$ is defined similarly as $\hat{\gamma}^1_{\mathrm{RCAL}}$, but with the loss function $\ell_{\mathrm{CAL}}(\gamma)$ in (12) replaced by

$$\ell^0_{\mathrm{CAL}}(\gamma) = \tilde{E}\{(1-T)\mathrm{e}^{\gamma^{\mathrm{T}} f(X)} - T\gamma^{\mathrm{T}} f(X)\},$$

that is, T and γ in $\ell_{\mathrm{CAL}}(\gamma)$ are replaced by $1-T$ and $-\gamma$. The estimator $\hat{\alpha}^0_{\mathrm{RWL}}$ is defined similarly as $\hat{\alpha}^1_{\mathrm{RWL}}$, but with the loss function $\ell_{\mathrm{WL}}(\cdot; \hat{\gamma}^1_{\mathrm{RCAL}})$ in (53) replaced by

$$\ell^0_{\mathrm{WL}}(\alpha^0; \hat{\gamma}^0_{\mathrm{RCAL}}) = \tilde{E}((1-T)w^0(X; \hat{\gamma}^0_{\mathrm{RCAL}})[-Y\alpha^{0\mathrm{T}} g^0(X) + \Psi\{\alpha^{0\mathrm{T}} g^0(X)\}]),$$

where $w^0(X;\gamma) = \pi(X;\gamma)/\{1 - \pi(X;\gamma)\} = \mathrm{e}^{\gamma^{\mathrm{T}} f(X)}$. Under similar conditions as in Propositions 1 and 3, the estimator $\hat{\mu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}})$ admits the asymptotic expansion

$$\hat{\mu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}}) = \tilde{E}\{\varphi(Y, 1-T, X; \bar{m}^0_{\mathrm{WL}}, 1 - \bar{\pi}^0_{\mathrm{CAL}})\} + o_p(n^{-1/2}), \tag{61}$$

where $\bar{\pi}^0_{\mathrm{RCAL}}(X) = \pi(X; \bar{\gamma}^0_{\mathrm{RCAL}})$, $\bar{m}^0_{\mathrm{RWL}}(X) = m_0(X; \bar{\alpha}^0_{\mathrm{RWL}})$ and $\bar{\gamma}^0_{\mathrm{RCAL}}$ and $\bar{\alpha}^0_{\mathrm{RWL}}$ are the target values defined similarly as $\bar{\gamma}^1_{\mathrm{RCAL}}$ and $\bar{\alpha}^1_{\mathrm{RWL}}$. Then Wald confidence intervals for μ^0 and ATE can be derived from (20) and (61) similarly as in Propositions 1 and 3 and shown to be either doubly robust in the case of linear outcome models, or valid if PS model (4) is correctly specified but OR models (51) and (60) may be misspecified for nonlinear outcome models.

An interesting feature of our approach is that two different estimators of the propensity score are used when estimating μ^0 and μ^1. Similar ideas have been involved in several related methods (e.g., Vermeulen and Vansteelandt (2015); Chan, Yam and Zhang (2016)). On one hand, the estimators $\hat{\gamma}^0_{\mathrm{RCAL}}$ and $\hat{\gamma}^1_{\mathrm{RCAL}}$ are both consistent, and hence there is no self-contradiction at least asymptotically, when PS model (4) is correctly specified. On the other hand, if model (4) is misspecified, the two estimators may in general have different asymptotic limits, which can be an advantage from the following perspective. By definition, the augmented IPW estimators of μ^1 and μ^0 are obtained, depending on fitted propensity scores within the treated group and untreated groups separately, that is, $\{\pi(X_i; \gamma^1) : T_i = 1\}$ and $\{\pi(X_i; \gamma^0) : T_i = 0\}$. In the presence of model misspecification, allowing different γ^1 and γ^0 can be helpful in finding suitable approximations of the two sets of propensity scores, without being constrained by the then false assumption that they are determined by the same coefficient vector $\gamma^1 = \gamma^0$.

Estimation of ATT. There is a simple extension of our approach to estimation of ATT, that is, $\nu^1 - \nu^0$ as defined in Section 2. The parameter $\nu^1 = E(Y^1|T = 1)$ can be directly estimated by $\tilde{E}(TY)/\tilde{E}(T)$. For $\nu^0 = E(Y^0|T = 1)$, our point estimator is

$$\hat{\nu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}}) = \tilde{E}\{\varphi_{\nu^0}(Y, T, X; \hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}})\}/\tilde{E}(T),$$

where $\hat{\pi}^0_{\mathrm{RCAL}}(X)$ and $\hat{m}^0_{\mathrm{RWL}}(X)$ are the same fitted values as used in the estimator $\hat{\mu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}})$ for μ^0, and $\varphi_{\nu^0}(\cdot; \hat{m}_0, \hat{\pi})$ is defined as

$$\varphi_{\nu^0}(Y, T, X; \hat{m}_0, \hat{\pi}) = \frac{(1-T)\hat{\pi}(X)}{1-\hat{\pi}(X)} Y - \left\{\frac{1-T}{1-\hat{\pi}(X)} - 1\right\}\hat{m}_0(X).$$

The function $\varphi_{\nu^0}(\cdot; \hat{m}_0, \hat{\pi})$ can be derived, by substituting fitted values $(\hat{m}_0, \hat{\pi})$ for the true values (m^*_0, π^*) in the efficient influence function of ν^0 under a nonparametric model (Hahn (1998); Shu and Tan (2018)). The estimator $\tilde{E}\{\varphi_{\nu^0}(Y, T, X; \hat{m}^0, \hat{\pi})\}$ is doubly robust: it remains consistent for $E(TY^0)$ if either $\hat{m}^0 = m^*_0$ or $\hat{\pi} = \pi^*$. In addition, by straightforward calculation, the function $\varphi_{\nu^0}()$ is related to $\varphi()$ in (7) through the simple identity:

$$\varphi_{\nu^0}(Y, T, X; \hat{m}_0, \hat{\pi}) = \varphi(Y, 1-T, X; \hat{m}_0, 1-\hat{\pi}) - (1-T)Y. \tag{62}$$

As a result, $\hat{\nu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}})$ can be equivalently obtained as

$$\begin{aligned}\hat{\nu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}}) &= [\hat{\mu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}}) - \tilde{E}\{(1-T)Y\}]/\tilde{E}(T) \\ &= \tilde{E}\{T\hat{m}^0_{\mathrm{RWL}}(X)\}/\tilde{E}(T),\end{aligned}$$

where the second step follows from a similar equation for $\hat{\mu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}})$ as (19). Moreover, it can be shown using equation (62) that under similar conditions as in Propositions 1 and 3, the estimator $\hat{\nu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}})$ admits the asymptotic expansion

$$\begin{aligned}&\hat{\nu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}}) - \nu^0 \\ &\quad = \tilde{E}\{\varphi_{\nu^0}(Y, T, X; \bar{m}^0_{\mathrm{WL}}, \bar{\pi}^0_{\mathrm{CAL}}) - T\nu^0\}/\tilde{E}(T) + o_p(n^{-1/2}),\end{aligned}$$

similarly as (61) for $\hat{\mu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}})$. From this expansion, Wald confidence intervals for ν^0 and ATT can be derived and shown to be either doubly robust with linear OR model (60) or valid at least when PS model (4) is correctly specified.

Coupled estimating functions. We provide additional comments about the technical complication and alternative approaches mentioned in Section 3.4 when using nonlinear outcome models.

For a possibly nonlinear outcome model (51), the orthogonality conditions (23)–(24) lead to the following estimating functions:

$$\frac{\partial \tilde{E}\{\varphi(Y, T, X; \alpha^1, \gamma)\}}{\partial \alpha^1} = \tilde{E}\left[\left\{1 - \frac{T}{\pi(X;\gamma)}\right\}\psi_2\{\alpha^{1\mathrm{T}} f(X)\} f(X)\right], \tag{63}$$

$$\frac{\partial \tilde{E}\{\varphi(Y, T, X; \alpha^1, \gamma)\}}{\partial \gamma} = -\tilde{E}\left[T \frac{1-\pi(X;\gamma)}{\pi(X;\gamma)}\{Y - m^1(X;\alpha^1)\} f(X)\right], \tag{64}$$

where $\psi_2(\cdot)$ denotes the derivative of $\psi(\cdot)$. The two vectors of functions, (63) and (64), are intrinsically coupled in (α^1, γ), each depending on both γ and α^1, unless outcome model (51) is linear, and hence the dependency of (63) on α^1 vanishes. Such joint dependency presents both computational and statistical obstacles. In low-dimensional settings, estimating equations can be defined by setting (63)–(64) to zero (Kim and Haziza (2014); Vermeulen and Vansteelandt (2015)). But the pair of equations need to be solved simultaneously in (α^1, γ) instead of sequentially. For nonlinear estimating equations, there may be various difficulties in computation and asymptotic theory, for example, related to multiple solutions (Small, Wang and Yang (2000)). In high-dimensional settings, with (α^1, γ) in both (63) and (64), it seems impossible to sequentially define loss functions and regularized M-estimators in γ and then α^1 (or vice versa) as in the case of linear outcome models. It is interesting to investigate other regularization methods.

Another worthwhile strategy is to modify one of estimating functions (63)–(64) and derive model-assisted (not doubly robust) confidence intervals. The development in Section 3.4 involves replacing (63) by (31) but retaining (64), which lead to the estimators $\hat{\gamma}^1_{\mathrm{RCAL}}$ and $\hat{\alpha}^1_{\mathrm{RWL}}$ based on the loss functions $\ell_{\mathrm{CAL}}(\gamma)$ and $\ell_{\mathrm{WL}}(\alpha^1;\gamma)$. The resulting confidence intervals are PS based, OR assisted, that is, being valid if PS model (4) is correctly specified but OR model (51) may be misspecified. Alternatively, it is possible to develop an OR based, PS assisted approach which retains (63), but replaces (64) by score functions in OR model (51). This approach leads to the regularized maximum likelihood estimator $\hat{\alpha}^1_{\mathrm{RML}}$ in conjunction with a regularized estimator of γ based on a weighted calibration loss,

$$\ell_{\mathrm{WL}}(\gamma; \hat{\alpha}^1_{\mathrm{RML}}) = \tilde{E}\left[\psi_2\{\hat{\alpha}^{1\mathrm{T}}_{\mathrm{RML}} f(X)\}\{T \mathrm{e}^{-\gamma^{\mathrm{T}} f(X)} + (1-T)\gamma^{\mathrm{T}} f(X)\}\right]. \tag{65}$$

The gradient of (65) in γ is (63), with $\alpha^1 = \hat{\alpha}^1_{\mathrm{RML}}$. Similar results can be established as in Section 3.4, to provide valid confidence intervals for μ^1 if OR model (51) is correctly specified but PS model (4) may be misspecified. This work can be pursued elsewhere.

4. Simulation studies.

We conducted two simulation studies to compare inferences using $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}^1_{\mathrm{RML}})$, without or with post-Lasso refitting, and $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$. The design of the first study is modified and extended from Kang and Schafer (2007) to high-dimensional, sparse settings. Both continuous and binary outcomes are considered. The results are presented in the preprint Tan (2018), and similar conclusions can be drawn as discussed below. The second study presented here involves a simpler design for covariates and model misspecification with continuous outcomes, but provides results with post-Lasso refitting and larger ratios p/n.

4.1. *Implementation details.* Both the regularized calibrated and maximum likelihood methods are implemented in the R package `RCAL` (Tan (2019)). The penalized loss function (3) or (6) for computing $\hat{\alpha}^1_{\mathrm{RML}}$ or $\hat{\gamma}^1_{\mathrm{RML}}$ or (11), (13) or (52) for computing $\hat{\alpha}^1_{\mathrm{RWL}}$ or $\hat{\gamma}^1_{\mathrm{RCAL}}$ is minimized for a fixed tuning parameter λ, using algorithms similar to those in Friedman, Hastie and Tibshirani (2010), but with the coordinate descent method replaced by an active set method as in Osborne, Presnell and Turlach (2000) for solving each Lasso penalized least squares problem. In addition, the penalized loss (11) for computing $\hat{\gamma}^1_{\mathrm{RCAL}}$ is minimized using the algorithm in Tan (2017), where a nontrivial Fisher scoring step is involved for quadratic approximation.

The tuning parameter λ is determined using 5-fold cross validation based on the corresponding loss function as follows. For $k = 1, \ldots, 5$, let $\mathcal{I}_k$ be a random subsample of size $n/5$ from $\{1, 2, \ldots, n\}$. For a loss function $\ell(\gamma)$, either $\ell_{\mathrm{ML}}(\gamma)$ in (5) or $\ell_{\mathrm{CAL}}(\gamma)$ in (12), denote by $\ell(\gamma; \mathcal{I})$ the loss function obtained when the sample average $\tilde{E}(\cdot)$ is computed over only the subsample $\mathcal{I}$. The 5-fold cross-validation criterion is defined as $\mathrm{CV}_5(\lambda) = (1/5)\sum_{k=1}^5 \ell(\hat{\gamma}^{(k)}_\lambda; \mathcal{I}_k)$,, where $\hat{\gamma}^{(k)}_\lambda$ is a minimizer of the penalized loss $\ell(\gamma; \mathcal{I}^c_k) + \lambda\|\gamma_{1:p}\|_1$ over the subsample $\mathcal{I}^c_k$ of size $4n/5$, that is, the complement to $\mathcal{I}_k$. Then λ is selected by minimizing $\mathrm{CV}_5(\lambda)$ over the discrete set $\{\lambda^*/2^j : j = 0, 1, \ldots, 10\}$, where for $\hat{\pi}_0 = \tilde{E}(T)$, the value λ^* is computed as either $\lambda^* = \max_{j=1,\ldots,p} |\tilde{E}\{(T - \hat{\pi}_0) f_j(X)\}|$ when the likelihood loss (5) is used, or $\lambda^* = \max_{j=1,\ldots,p} |\tilde{E}\{(T/\hat{\pi}_0 - 1) f_j(X)\}|$ when the calibration loss (12) is used. It can be shown that in either case, the penalized loss $\ell(\gamma) + \lambda\|\gamma_{1:p}\|_1$ over the original sample has a minimum at $\gamma_{1:p} = 0$ for all $\lambda \ge \lambda^*$.

For computing $\hat{\alpha}^1_{\mathrm{RML}}$ or $\hat{\alpha}^1_{\mathrm{RWL}}$, cross validation is conducted similarly as above using the loss function $\ell_{\mathrm{ML}}(\alpha^1)$ in (2) or $\ell_{\mathrm{WL}}(\alpha^1; \hat{\gamma}^1_{\mathrm{RCAL}})$ in (53). In the latter case, $\hat{\gamma}^1_{\mathrm{RCAL}}$ is determined separately and then fixed during cross validation for computing $\hat{\alpha}^1_{\mathrm{RWL}}$.

4.2. *Simulation setup and results.* Let $X = (X_1, \ldots, X_p)$ be multivariate normal with means 0 and covariances $\mathrm{cov}(X_j, X_k) = 2^{-|j-k|}$ for $1 \le j, k \le p$. In addition, let $x^\dagger_j$ be $X_j + \{(X_j + 1)_+\}^2$ standardized with mean 0 and variance 1 for $j = 1, \ldots, 4$. Consider the following data-generating configurations:

(C1) Generate T given X from a Bernoulli distribution with $P(T = 1|X) = \{1 + \exp(-1 - X_1 - 0.5X_2 - 0.25X_3 - 0.125X_4)\}^{-1}$ and, independently, generate Y^1 given X from a Normal distribution with variance 1 and mean $E(Y^1|X) = X_1 + 0.5X_2 + 0.25X_3 + 0.125X_4$.

(C2) Generate T give X as in (C1) but, independently, generate Y^1 given X from a Normal distribution with variance 1 and mean $E(Y^1|X) = X^\dagger_1 + 0.5X^\dagger_2 + 0.25X^\dagger_3 + 0.125X^\dagger_4$.

(C3) Generate Y^1 given X as in (C1) but, independently, generate T given X from a Bernoulli distribution with $P(T = 1|X) = \{1 + \exp(-1 - X^\dagger_1 - 0.5X^\dagger_2 - 0.25X^\dagger_3 - 0.125X^\dagger_4)\}^{-1}$.

As in Section 2, the observed data consist of independent and identically distributed observations $\{(T_iY_i, T_i, X_i) : i = 1, \ldots, n\}$. Consider logistic propensity score model (4) and linear outcome model (9), both with $f_j(X) = X_j$ for $j = 1, \ldots, p$. Then the two models can be classified as follows, depending on the data configuration above:

(C1) PS and OR models both correctly specified;
(C2) PS model correctly specified, but OR model misspecified;
(C3) PS model misspecified, but OR model correctly specified.

See the Supplement for boxplots of X_j within $\{T = 1\}$ and $\{T = 0\}$ and scatterplots of Y against X_j within $\{T = 1\}$ for $j = 1, \ldots, 4$. Partly because $X^\dagger_j$ is a monotone function of X_j,

TABLE 1
Summary of results with linear outcome models

	(C1) cor PS, cor OR			(C2) cor PS, mis OR			(C3) mis PS, cor OR		
	RML	RML2	RCAL	RML	RML2	RCAL	RML	RML2	RCAL
				$n = 800$ and $p = 200$					
Bias	0.019	0.002	0.026	−0.038	−0.009	0.012	0.010	0.002	0.016
$\sqrt{\text{Var}}$	0.070	0.079	0.070	0.072	0.092	0.071	0.070	0.076	0.070
$\sqrt{\text{EVar}}$	0.068	0.076	0.068	0.072	0.094	0.069	0.069	0.075	0.068
Cov90	0.864	0.880	0.854	0.859	0.906	0.889	0.875	0.886	0.871
Cov95	0.922	0.947	0.914	0.914	0.950	0.938	0.941	0.950	0.942
				$n = 800$ and $p = 1000$					
Bias	0.031	−0.002	0.033	−0.038	−0.001	0.015	0.026	0.009	0.028
$\sqrt{\text{Var}}$	0.068	0.109	0.070	0.070	0.394	0.071	0.069	0.092	0.070
$\sqrt{\text{EVar}}$	0.067	0.112	0.067	0.070	0.388	0.068	0.068	0.089	0.068
Cov90	0.851	0.911	0.836	0.842	0.896	0.877	0.867	0.892	0.854
Cov95	0.922	0.949	0.915	0.912	0.944	0.929	0.924	0.933	0.923

Note: RML denotes $\hat{\mu}^1(\hat{m}^1_{\text{RML}}, \hat{\pi}^1_{\text{RML}})$, and RML2 denotes the variant with $\hat{m}^1_{\text{RML}}$ and $\hat{\pi}^1_{\text{RML}}$ replaced by the fitted values obtained by refitting OR and PS models only including the variables selected from the corresponding Lasso estimation. RCAL denotes $\hat{\mu}^1(\hat{m}^1_{\text{RWL}}, \hat{\pi}^1_{\text{RCAL}})$. Bias and Var are the Monte Carlo bias and variance of the points estimates. EVar is the mean of the variance estimates, and hence $\sqrt{\text{EVar}}$ also measures the L_2-average of lengths of confidence intervals. Cov90 or Cov95 is the coverage proportion of the 90% or 95% confidence intervals.

the misspecified OR model in (C1) or PS model in (C2) appears difficult to detect by standard model diagnosis.

For $n = 800$ and $p = 200$ or 1000, Table 1 summarizes the results for estimation of μ^1, based on 1000 repeated simulations. The methods RML and RCAL perform similarly to each other in terms of bias, variance and coverage in the cases (C1) and (C3). But RCAL leads to noticeably smaller absolute biases and better coverage than RML in the case (C2), correct PS and misspecified OR models. The post-Lasso refitting method RML2 yields coverage proportions closer to the nominal probabilities than RCAL, but consistently and, in the case (C2), substantially higher variances and wider confidence intervals. These properties can also be seen from the QQ plots of the estimates and t-statistics in the Supplement. See the Supplement for additional results and discussion including the oracle estimators (using submodels with only the covariates $X_1, \dots, X_4$).

5. Empirical application.

We provide an application to a medical study in Connors et al. (1996) on the effects of right heart catheterization (RHC). The study included $n = 5735$ critically ill patients admitted to the intensive care units of 5 medical centers. For each patient, the data consist of treatment status T ($= 1$ if RHC was used within 24 hours of admission and 0 otherwise), health outcome Y (survival time up to 30 days) and a list of 75 covariates X specified by medical specialists in critical care. For previous analyses, propensity score and outcome regression models were employed either with main effects only (Hirano and Imbens (2002); Vermeulen and Vansteelandt (2015)) or with interaction terms manually added (Tan (2006)).

To explore dependency beyond main effects, we consider a logistic propensity score model (4) and a logistic outcome model (51) for 30-day survival status $1\{Y > 30\}$, with the vector $f(X)$ including all main effects and two-way interactions of X except those with the fractions of nonzero values less than 46 (i.e., 0.8% of the sample size 5735). The dimension of $f(X)$ is $p = 1855$, excluding the constant. All variables in $f(X)$ are standardized with sample

TABLE 2
Estimates of 30*-day survival probabilities and ATE*

	IPW			Augmented IPW		
	RML	RML2	RCAL	RML	RML2	RCAL
μ^1	0.636 ± 0.026	0.660 ± 0.049	0.634 ± 0.023	0.636 ± 0.021	0.646 ± 0.035	0.635 ± 0.021
μ^0	0.690 ± 0.017	0.691 ± 0.019	0.687 ± 0.017	0.691 ± 0.016	0.693 ± 0.018	0.688 ± 0.016
ATE	-0.054 ± 0.031	-0.031 ± 0.053	-0.053 ± 0.029	-0.055 ± 0.025	-0.047 ± 0.039	-0.053 ± 0.025

Note: Estimate $\pm 2 \times$ standard error, including nominal standard errors for IPW.

means 0 and variances 1. We apply the estimators $\hat{\mu}^1(\hat{m}^1_{\mathrm{RWL}}, \hat{\pi}^1_{\mathrm{RCAL}})$ and $\hat{\mu}^0(\hat{m}^0_{\mathrm{RWL}}, \hat{\pi}^0_{\mathrm{RCAL}})$ using regularized calibrated (RCAL) estimation and the corresponding estimators such as $\hat{\mu}^1(\hat{m}^1_{\mathrm{RML}}, \hat{\pi}_{\mathrm{RML}})$ using regularized maximum likelihood (RML) estimation, similarly as in the simulation study. The Lasso tuning parameter λ is selected by 5-fold cross validation over a discrete set $\{\lambda^*/2^{j/4} : j = 0, 1, \dots, 24\}$, where λ^* is the value leading to a zero solution $\gamma_1 = \cdots = \gamma_p = 0$. We also compute the (ratio) IPW estimators, such as $\hat{\mu}^1_{\mathrm{rIPW}}$, along with nominal standard errors obtained by ignoring data-dependency of the fitted propensity scores.

Table 2 shows various estimates of survival probabilities and ATE. The IPW estimates from RCAL estimation of propensity scores have noticeably smaller nominal standard errors than RML estimation, for example, with the relative efficiency $(0.026/0.023)^2 = 1.28$ for estimation of μ^1. This improvement can also be seen from Figure S7 in the Supplementary Material, where the RCAL inverse probability weights are much less variable than RML weights. See Tan (2017) for additional results on covariate balance and parameter sparsity from RML and RCAL estimation of propensity scores.

The augmented IPW estimates and confidence intervals are similar to each other from RCAL and RML estimation. The estimate of μ^1 from RML with post-Lasso refitting appears problematic with a large standard error. However, the validity of RML confidence intervals depends on both PS and OR models being correctly specified, whereas that of RCAL confidence intervals holds even when the OR model is misspecified. While assessment of this difference is difficult with real data, Figure S7 shows that the sample influence functions for ATE using RCAL estimation appears to be more normally distributed especially in the tails than RML estimation.

Finally, the augmented IPW estimates here are smaller in absolute values, and also with smaller standard errors, than previous estimates based on main-effect models, about $-0.060 \pm 2 \times 0.015$ (Vermeulen and Vansteelandt (2015)). The reduction in standard errors might be explained by the well-known property that an augmented IPW estimator has a smaller asymptotic variance when obtained using a larger (correct) propensity score model.

Acknowledgments. The author thanks Cun-Hui Zhang and Zijian Guo for helpful discussions and two referees for constructive comments which led to various improvements in the article.

This work was supported in part by PCORI Grant ME-1511-32740.

SUPPLEMENTARY MATERIAL

Supplement to "Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data" (DOI: 10.1214/19-AOS1824SUPP; .pdf). We provide technical proofs and additional numerical results in Tan (2020).

REFERENCES

ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 597–623. MR3849336 https://doi.org/10.1111/rssb.12268

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and WEI, Y. (2018). Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *Ann. Statist.* **46** 3643–3675. MR3852664 https://doi.org/10.1214/17-AOS1671

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

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 https://doi.org/10.1111/rssb.12129

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CONNORS, A. F., SPEROFF, T., DAWSON, N. V. et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *J. Am. Med. Assoc.* **276** 889–897.

FARRELL, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *J. Econometrics* **189** 1–23 [Corrigendum available at http://faculty.chicagobooth.edu/max.farrell/]. MR3397349 https://doi.org/10.1016/j.jeconom.2015.06.017

FOLSOM, R. E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. In *Proceedings of the American Statistical Association*, *Social Statistics Section* 197–202.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

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 https://doi.org/10.1093/restud/rdr047

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242 https://doi.org/10.2307/2998560

HAINMUELLER, J. (2012). Entropy balancing for causal effects: Multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46.

HIRANO, K. and IMBENS, G. W. (2002). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Serv. Outcomes Res. Methodol.* **2** 259–278.

HUANG, J. and ZHANG, C.-H. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *J. Mach. Learn. Res.* **13** 1839–1864. MR2956344

IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263. MR3153941 https://doi.org/10.1111/rssb.12027

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

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 https://doi.org/10.1214/07-STS227

KIM, J. K. and HAZIZA, D. (2014). Doubly robust inference with missing data in survey sampling. *Statist. Sinica* **24** 375–394. MR3183689

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. *Monographs on Statistics and Applied Probability*. CRC Press, London. Second edition [of MR0727836]. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133 https://doi.org/10.1214/12-STS400

NEWEY, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147–168. MR1457700 https://doi.org/10.1016/S0304-4076(97)00011-0

OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20** 389–403. MR1773265 https://doi.org/10.1093/imanum/20.3.389

ROBINS, J. M. and ROTNITZKY, A. (2001). Comment on "Inference for semiparametric models: Some questions and an answer" by Bickel and Kwon. *Statist. Sinica* **11** 920–936.

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

ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist.* **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515

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 https://doi.org/10.1093/biomet/70.1.41

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.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. *Springer Series in Statistics*. Springer, New York. MR1140409 https://doi.org/10.1007/978-1-4612-4378-6

SHU, H. and TAN, Z. (2018). Improved estimation of average treatment effects on the treated: Local efficiency, double robustness, and beyond. arXiv:1808.01408.

SMALL, C. G., WANG, J. and YANG, Z. (2000). Eliminating multiple root problems in estimation. *Statist. Sci.* **15** 313–341. MR1819708 https://doi.org/10.1214/ss/1009213000

SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. MR1092986

TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101** 1619–1637. MR2279484 https://doi.org/10.1198/016214506000000023

TAN, Z. (2007). Comment: Understanding OR, PS and DR [MR2420458]. *Statist. Sci.* **22** 560–568. MR2420461 https://doi.org/10.1214/07-STS227A

TAN, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97** 661–682. MR2672490 https://doi.org/10.1093/biomet/asq035

TAN, Z. (2017). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. arXiv:1710.08074.

TAN, Z. (2018). Model-assisted inference for treatment effects using regularized calibrated estimationwith high-dimensional data. arXiv:1801.09817.

TAN, Z. (2019). RCAL: Regularized calibrated estimation. R package version 1.0.

TAN, Z. (2020). Supplement to "Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data." https://doi.org/10.1214/19-AOS1824SUPP.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

VERMEULEN, K. and VANSTEELANDT, S. (2015). Bias-reduced doubly robust estimation. *J. Amer. Statist. Assoc.* **110** 1024–1036. MR3420681 https://doi.org/10.1080/01621459.2014.958155

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838. MR0575027 https://doi.org/10.2307/1912934

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163 https://doi.org/10.2307/1912526

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026