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

BY ZHIQIANG TAN

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 et al. 1994).

A conventional approach for propensity score estimation (Rosenbaum & Rubin 1984) involves fitting a propensity score model (often logistic regres-
sion) 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 adjusted for 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 (that is, confidence intervals or hypothesis testing) about treatment effects.

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 estimates 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 1992) 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 (Sarndal et al. 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 $\hat{\gamma}_{\text{RCAL}}^1$, for the coefficients in the propensity score model and a regularized weighted likelihood estimator, denoted by $\hat{\alpha}_{\text{RWL}}^1$, 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 $\hat{\gamma}_{RCAL}^1$ 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 $\hat{\gamma}_{RCAL}^1$ can be used in conjunction with $\hat{\alpha}_{RWL}^1$ 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 $\hat{\alpha}_{RWL}^1$ and the resulting augmented IPW estimator with possible model misspecification, while building on related results on convergence of $\hat{\gamma}_{RCAL}^1$ to a target value $\bar{\gamma}_{CAL}^1$ in Tan (2017). In particular, a new strategy is developed to tackle the technical issue that the weighted likelihood loss for $\hat{\alpha}_{RWL}^1$ is defined depending on the estimator $\hat{\gamma}_{RCAL}^1$. As a result, we obtain the convergence of $\hat{\alpha}_{RWL}^1$ to a target value $\bar{\alpha}_{WL}^1$ 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., Buhlmann & van de Geer 2011), where $|S_\gamma|$ or $|S_{\alpha^1}|$ denotes the size of nonzero elements in $\bar{\gamma}_{CAL}^1$ or respectively $\bar{\alpha}_{WL}^1$. Furthermore, we establish an asymptotic expansion of the augmented IPW estimator based on $\hat{\gamma}_{RCAL}^1$ and $\hat{\alpha}_{RWL}^1$, 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 et al. 2014; van de Geer et al. 2014).

Related works. We compare and connect our work with related works in several areas. Non-penalized 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 et al. 2012; Hainmueller 2012; Imai & Ratovic 2014; Kim & Haziza 2014; Vermeulen & Vansteelandt 2015; Chan et al. 2016). The non-penalized version of the estimator $\hat{\alpha}_{RWL}^1$ for outcome regression models have
also been proposed in Kim & Haziza (2014) and Vermeulen & 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 et al. (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 et al. (2014) and Farrell (2015) also advocated post-Lasso refitting, to potentially improve finite-sample performance.

Another related work is Athey et al. (2016), where valid confidence intervals are obtained for the sample treatment effect \(n^{-1} \sum_{i} \{m^*_1(X_i) - m^*_0(X_i)\}\) (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 & Zhang 2014; van de Geer et al. 2014; Javanmard & Montanari 2014). Model-assisted inference does not seem to be addressed in these works, but can potentially be developed.

2. Setup. Suppose that \(\{(Y_i, T_i, X_i) : i = 1, \ldots, n\}\) 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 (Neyman 1923; 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 $\mu^1$ until Section 3.5 to discuss ATE and ATT.

Estimation of ATE is fundamentally a missing-data problem: only one potential outcome, $Y^0_i$ or $Y^1_i$, is observed and the other one is missing for each subject $i$. For identification of $(\mu^0, \mu^1)$ 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 & Rubin 1983).

Under these assumptions, $(\mu^0, \mu^1)$ and ATE are often estimated by imposing additional modeling assumptions on the outcome regression function $m_1^*(X) = E(Y|T = t,X)$ or the propensity score $\pi^*(X) = P(T = 1|X)$.

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

$$E(Y|T = 1, X) = m_1(X; \alpha^1) = \psi\{\alpha^{1T}g_1^1(X)\}, \quad (1)$$

where $\psi()$ is an (increasing) inverse link function, $g_1^1(x) = \{1, g_1^1(x), \ldots, g_q^1(x)\}^T$ is a vector of known functions such as $g_1^1(x) = (1, x^T)^T$, and $\alpha^1 = (\alpha_0^1, \alpha_1^1, \ldots, \alpha_q^1)^T$ is a vector of unknown parameters. Throughout, superscript $^T$ denotes a transpose, not the treatment variable $T$. Model (1) can be deduced from a generalized linear model with a canonical link (McCullagh & Nelder 1989).

Then the average negative log-(quasi-)likelihood function can be written (after dropping any dispersion parameter) as

$$\ell_{\text{ML}}(\alpha^1) = \tilde{E}\{T \left[ -Y \alpha^{1T}g_1^1(X) + \Psi\{\alpha^{1T}g_1^1(X)\} \right] \}, \quad (2)$$

where $\Psi(u) = \int_0^u \psi(u') \, du'$, which is convex in $u$. Throughout, $\tilde{E}()$ denotes the sample average. With high-dimensional data, a regularized maximum likelihood estimator, $\hat{\alpha}_{\text{RML}}^1$, can be defined by minimizing the loss $\ell_{\text{ML}}(\alpha^1)$ with the Lasso penalty (Tibshirani 1992),

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

where $\| \cdot \|_1$ denotes the $L_1$ norm, $\alpha_{1, q}^1 = (\alpha_1^1, \ldots, \alpha_q^1)^T$ excluding $\alpha_0^1$, and $\lambda \geq 0$ is a tuning parameter. The resulting estimator of $\mu^1$ is then

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

where $\hat{m}_{\text{RML}}^1(X) = m_1^1(X; \hat{\alpha}_{\text{RML}}^1)$, the fitted outcome regression function. Various theoretical results have been obtained on Lasso estimation in sparse,
high-dimensional regression (e.g., Buhlmann & van de Geer 2011). If model (1) is correctly specified, then it can be shown under suitable conditions that
\|\hat{\alpha}_{RML} - \alpha^*\|_1 = O_p(1)\|\alpha^*\|_0\{\log(q)/n\}^{1/2} and \hat{\mu}_{OR} = \mu_1 + O_p(1)\{\|\alpha^*\|_0 \log(q)/n\}^{1/2}, where \alpha^* is the true value for model (1) such that \(m_1^\star(X) = m_1^\star(X; \alpha^*)\) and \(\|\alpha^*\|_0\) is the size of nonzero elements in \(\alpha^*\).

Alternatively, consider a propensity score (PS) model

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

where \(\Pi()\) is an inverse link function, \(f(x) = \{1, f_1(x), \ldots, f_p(x)\}^T\) is a vector of known functions such as \(g_1(x) = (1, x^T)^T\), and \(\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_p)^T\) is a vector of unknown parameters. For concreteness, assume that model (4) is logistic regression with \(\pi(X; \gamma) = [1 + \exp\{-\gamma^T f(X)\}]^{-1}\), and hence the average negative log-likelihood function is

\(\ell_{\text{ML}}(\gamma) = \bar{E}\left[\log\{1 + e^{\gamma^T f(X)}\} - T \gamma^T f(X)\right].\)

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

\(\ell_{\text{RML}}(\gamma) = \ell_{\text{ML}}(\gamma) + \lambda\|\gamma_{1:p}\|_1,\)

where \(\gamma_{1:p} = (\gamma_1, \ldots, \gamma_p)^T\) excluding \(\gamma_0\), and \(\lambda \geq 0\) is a tuning parameter. The fitted propensity score is then \(\hat{\pi}_{\text{RML}}(X) = \pi(X; \hat{\gamma}_{\text{RML}})\). A (ratio) inverse probability weighted (IPW) estimator for \(\mu_1\) is

\(\hat{\mu}_{\text{IPW}}(\hat{\pi}_{\text{RML}}) = \bar{E}\left\{TY \right\} / \bar{E}\left\{T \hat{\pi}_{\text{RML}}(X)\right\}.\)

From previous works (e.g., Buhlmann & van de Geer 2011), if model (4) is correctly specified, then it can be shown under suitable conditions that \(\|\hat{\gamma}_{\text{RML}} - \gamma^*\|_1 = O_p(1)\|\gamma^*\|_0\{\log(p)/n\}^{1/2} and \hat{\mu}_{\text{IPW}}(\hat{\pi}_{\text{RML}}) = \mu_1 + O_p(1)\{\|\gamma^*\|_0 \log(p)/n\}^{1/2}, where \gamma^* is the true value for model (4) such that \(\pi^\star(X) = \pi(X; \gamma^*)\) and \(\|\gamma^*\|_0\) is the size of nonzero elements in \(\gamma^*\).

To attain consistency for \(\mu_1\), the estimator \(\hat{\mu}_{OR}\) or \(\hat{\mu}_{\text{IPW}}(\hat{\pi}_{\text{RML}})\) 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 et al. 1994)

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

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

\(\varphi(Y, T, X; \hat{m}, \hat{\pi}) = TY / \hat{\pi}(X) - \left\{T / \hat{\pi}(X) - 1\right\} \hat{m}(X).\)
See Kang & Schafer (2007) and Tan (2010) for reviews in low-dimensional settings. Recently, in high-dimensional settings, Belloni et al. (2014) and Farrell (2015) studied the estimator $\hat{\mu}(\hat{m}_{RML}, \hat{\pi}_{RML})$, using the fitted values $\hat{m}_{RML}(X)$ and $\hat{\pi}_{RML}(X)$ from Lasso penalized estimation or similar methods. Their results are mainly of two types. The first type shows double robustness: $\hat{\mu}(\hat{m}_{RML}, \hat{\pi}_{RML})$ remains consistent if either OR model (1) or PS model (4) is correctly specified. The second type establishes valid confidence intervals: $\hat{\mu}(\hat{m}_{RML}, \hat{\pi}_{RML})$ admits the usual influence function, $\hat{\mu}(\hat{m}_{RML}, \hat{\pi}_{RML}) = \tilde{E}\{\phi(Y,T,X; m^*, \pi^*)\} + o_p(n^{-1/2})$, 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 $\{\|\alpha^*\|_0 + \|\gamma^*\|_0\} \log(p) = o(n^{1/2})$ in Belloni et al. (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 $\hat{\mu}(\hat{m}_{RML}, \hat{\pi}_{RML})$ is obtained only under the assumption that both OR model (1) and PS model (4) are correctly specified, even though the point estimator $\hat{\mu}(\hat{m}_{RML}, \hat{\pi}_{RML})$ 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 $\mu^1$, 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,

$$E(Y|T = 1, X) = m^1(X; \alpha^1) = \alpha^{T_1} f(X),$$

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 $\mu^1$ is

$$\hat{\mu}^1(m_{RWL}, \hat{\pi}_{RCAL}) = \tilde{E}\{\phi(Y,T,X; \hat{m}_{RWL}, \hat{\pi}_{RCAL})\},$$

where $\phi()$ is defined in (7), $\hat{\pi}_{RCAL}(X) = \pi(X; \hat{\gamma}_{RCAL})$, $\hat{m}_{RWL}(X) = m^1(X; \hat{\alpha}_{RWL})$, and $\hat{\gamma}_{RCAL}$ and $\hat{\alpha}_{RWL}$ are defined as follows. The estimator $\hat{\gamma}_{RCAL}$ is a regularized calibrated estimator of $\gamma$ from Tan (2017), defined as a minimizer of
the Lasso penalized objective function,
\begin{equation}
\ell_{\text{RCAL}}(\gamma) = \ell_{\text{CAL}}(\gamma) + \lambda \| \gamma_1 \|_1,
\end{equation}
where $\lambda \geq 0$ is a tuning parameter and $\ell_{\text{CAL}}(\gamma)$ is the calibration loss,
\begin{equation}
\ell_{\text{CAL}}(\gamma) = \tilde{E} \left\{ T e^{-\gamma^T f(X)} + (1 - T) \gamma^T f(X) \right\}.
\end{equation}
The estimator $\hat{\alpha}_1^{\text{RWL}}$ is a regularized weighted least-squares estimator of $\alpha_1$, defined as a minimizer of
\begin{equation}
\ell_{\text{RWL}}(\alpha_1; \hat{\gamma}_1^{\text{RCAL}}) = \ell_{\text{WL}}(\alpha_1; \hat{\gamma}_1^{\text{RCAL}}) + \lambda \| \alpha_{1;p} \|_1,
\end{equation}
where $\ell_{\text{WL}}(\alpha_1; \hat{\gamma}_1^{\text{RCAL}})$ is the weighted least-squares loss,
\begin{equation}
\ell_{\text{WL}}(\alpha_1; \hat{\gamma}_1^{\text{RCAL}}) = \tilde{E} \left[ \frac{T_i - \hat{\pi}_1^{\text{RCAL}}(X_i)}{\hat{\pi}_1^{\text{RCAL}}(X_i)} \{ Y_i - \alpha_1^T f(X_i) \}^2 \right] / 2,
\end{equation}
and $\lambda \geq 0$ is a tuning parameter. That is, the observations in the treated group are weighted by $\left\{ 1 - \hat{\pi}_1^{\text{RCAL}}(X_i) \right\} / \hat{\pi}_1^{\text{RCAL}}(X_i)$, which differs slightly from the commonly used inverse propensity score weight $1/\hat{\pi}_1^{\text{RCAL}}(X_i)$.

There are simple and interesting interpretations of the preceding estimators. By the Karush–Kuhn–Tucker condition for minimizing (11), the fitted propensity score $\hat{\pi}_1^{\text{RCAL}}(X)$ satisfies
\begin{equation}
\frac{1}{n} \sum_{i=1}^{n} \frac{T_i}{\hat{\pi}_1^{\text{RCAL}}(X_i)} = 1,
\end{equation}
\begin{equation}
\frac{1}{n} \left| \sum_{i=1}^{n} \frac{T_i f_j(X_i)}{\hat{\pi}_1^{\text{RCAL}}(X_i)} - \sum_{i=1}^{n} f_j(X_i) \right| \leq \lambda, \quad j = 1, \ldots, p,
\end{equation}
where equality holds in (16) for any $j$ such that the $j$th estimate $\left( \hat{\gamma}_1^{\text{RCAL}} \right)_j$ is nonzero. Eq. (15) shows that the inverse probability weights, $1/\hat{\pi}_1^{\text{RCAL}}(X_i)$, with $T_i = 1$, sum to the sample size $n$, whereas Eq. (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 $\lambda$. The Lasso penalty is used to induce the box constraints on the gradient of $\ell_{\text{CAL}}(\gamma)$ instead of setting the gradient to 0.

By the Karush–Kuhn–Tucker condition for minimizing (13), the fitted outcome regression function $\hat{m}_1^{\text{RWL}}(X)$ satisfies
\begin{equation}
\frac{1}{n} \sum_{i=1}^{n} \frac{T_i (1 - f_j(X_i))}{\hat{\pi}_1^{\text{RCAL}}(X_i)} \{ Y_i - \hat{m}_1^{\text{RWL}}(X_i) \} = 0,
\end{equation}
\begin{equation}
\frac{1}{n} \left| \sum_{i=1}^{n} \frac{T_i (1 - f_j(X_i))}{\hat{\pi}_1^{\text{RCAL}}(X_i)} \{ Y_i - \hat{m}_1^{\text{RWL}}(X_i) \} f_j(X_i) \right| \leq \lambda, \quad j = 1, \ldots, p,
where equality holds in (18) for any $j$ such that the $j$th estimate $(\hat{\alpha}_1^{\text{RWL}})_j$ is nonzero. Eq. (17) implies that by simple calculation, the estimator $\hat{\mu}^1(\hat{m}_1^{\text{RWL}}, \hat{\pi}_1^{\text{CAL}})$ can be recast as

$$
\hat{\mu}^1(\hat{m}_1^{\text{RWL}}, \hat{\pi}_1^{\text{CAL}}) = \hat{E}\left[\hat{m}_1^{\text{RWL}}(X) + \frac{T}{\hat{\pi}_1^{\text{CAL}}(X)} \{Y - \hat{m}_1^{\text{RWL}}(X)\}\right] \\
= \hat{E}\left\{TY + (1 - T)\hat{m}_1^{\text{RWL}}(X)\right\},
$$

which takes the form of linear prediction estimators known in the survey literature (e.g., Sarndal et al. 1992): $\hat{E}\{TY + (1 - T)\hat{m}_1^{\text{RWL}}(X)\}$ for some fitted outcome regression function $\hat{m}_1^{\text{RWL}}(X)$. As a consequence, $\hat{\mu}^1(\hat{m}_1^{\text{RWL}}, \hat{\pi}_1^{\text{CAL}})$ always falls within the range of the observed outcomes $\{Y_i : T_i = 1, i = 1, \ldots, n\}$ and the predicted values $\{\hat{m}_1^{\text{RWL}}(X_i) : T_i = 0, i = 1, \ldots, n\}$. This boundedness property is not satisfied by the estimator $\hat{\mu}^1(\hat{m}_1^{\text{RML}}, \hat{\pi}_1^{\text{RML}})$.

We provide a high-dimensional analysis of the estimator $\hat{\mu}^1(\hat{m}_1^{\text{RWL}}, \hat{\pi}_1^{\text{CAL}})$ in Section 3.3, allowing for possible model misspecification. Our main result shows that under suitable conditions, the estimator $\hat{\mu}^1(\hat{m}_1^{\text{RWL}}, \hat{\pi}_1^{\text{CAL}})$ admits the asymptotic expansion

$$
\hat{\mu}^1(\hat{m}_1^{\text{RWL}}, \hat{\pi}_1^{\text{CAL}}) = \hat{E}\left\{\varphi(Y, T, X; \hat{m}_1^{\text{RWL}}, \hat{\pi}_1^{\text{CAL}})\right\} + o_p(n^{-1/2}),
$$

where $\hat{\pi}_1^{\text{CAL}}(X) = \hat{\pi}(X; \hat{\gamma}_1^{\text{CAL}})$, $\hat{m}_1^{\text{RWL}}(X) = \hat{m}^1(X; \hat{\alpha}_1^{\text{RML}})$ and $\hat{\gamma}_1^{\text{CAL}}$ and $\hat{\alpha}_1^{\text{RML}}$ are defined as follows. With possible model misspecification, the target value $\hat{\gamma}_1^{\text{CAL}}$ is defined as a minimizer of the expected calibration loss

$$
E\left\{\ell_1^{\text{CAL}}(\gamma)\right\} = E\left\{Te^{-\gamma^Tf(X)} + (1 - T)\gamma^Tf(X)\right\}.
$$

If model (4) is correctly specified, then $\hat{\pi}_1^{1,\text{CAL}}(X) = \pi^*(X)$. Otherwise, $\hat{\pi}_1^{1,\text{CAL}}(X)$ may differ from $\pi^*(X)$. The target value $\hat{\alpha}_1^{\text{RML}}$ is defined as a minimizer of the expected loss

$$
E\left\{\ell_1^{\text{RML}}(\alpha_1; \hat{\gamma}_1^{\text{CAL}})\right\} = E\left\{\frac{T\{1 - \hat{\pi}_1^{\text{CAL}}(X)/\hat{\pi}_1^{\text{CAL}}(X)\} \{Y - \alpha^T f(X)\}^2}{2}\right\}.
$$

If model (9) is correctly specified, then $\hat{m}_1^{\text{RML}}(X) = m^1_1(X)$. But $\hat{m}_1^{\text{RML}}(X)$ may in general differ from $m^1_1(X)$. The following result can be deduced from Theorems 3 and 4. Suppose that the Lasso tuning parameter is specified as $\lambda = A^1_0\{\log(p)/n\}^{1/2}$ for $\hat{\gamma}_1^{\text{RML}}$ and $\lambda = A^1_1\{\log(p)/n\}^{1/2}$ for $\hat{\alpha}^{\text{RML}}$, with some constants $A^1_0$ and $A^1_1$. Denote $S_\alpha = \{0\} \cup \{j : \hat{\alpha}_\text{RML,j} = 0, j = 1, \ldots, p\}$ and $S_{\alpha^1} = \{0\} \cup \{j : \hat{\alpha}_{\text{RML},j} \neq 0, j = 1, \ldots, p\}$.
Proposition 1. Suppose that Assumptions 1 and 2 hold as in Section 3.3, and \(|S_\alpha| + |S_\alpha^1| (log(p) = o(n^{1/2})\). Then for \(\hat{\gamma}_{RCAL}^1\) and \(\hat{\alpha}_{RWL}^1\) with sufficiently large constants \(A_0^\dagger\) and \(A_1^\dagger\), asymptotic expansion (20) is valid. Moreover, if either logistic PS model (4) or linear OR model (9) is correctly specified, then \(\hat{\pi}_{CAL}^1(x) \equiv \pi^*(x)\) or, respectively, \(\hat{m}_{WL}^1(x) \equiv m^1(x)\), and the following results hold:

(i) \(n^{1/2}\{\hat{\mu}^1(\hat{m}_{RWL}^1, \hat{\pi}_{RCAL}^1) - \mu_1^1\} \to_D N(0, V)\), where \(V = \text{var}\{\varphi(Y, T, X; m_{WL}^1, \pi_{CAL}^1)\}\);

(ii) a consistent estimator of \(V\) is

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

(iii) an asymptotic \((1 - c)\) confidence interval for \(\mu_1\) is \(\hat{\mu}^1(\hat{m}_{RWL}^1, \hat{\pi}_{RCAL}^1) \pm z_{c/2}\sqrt{\hat{V}/n}\), where \(z_{c/2}\) is the \((1 - c/2)\) quantile of \(N(0, 1)\).

That is, a doubly robust confidence interval for \(\mu_1\) is obtained.

Remark 1. We discuss two implications of Proposition 1. First, the estimator \(\hat{\mu}^1(\hat{m}_{RWL}^1, \hat{\pi}_{RCAL}^1)\) is also locally efficient, i.e., achieves the semiparametric efficiency bound, \(n^{-1}\text{var}\{\varphi(Y, T, X; m^1, \pi^*)\}\), for estimation of \(\mu_1\) (Hahn 1998), when both models (4) and (9) are correctly specified, because in this case \(\hat{\pi}_{CAL}^1(x) \equiv \pi^*(x)\) and \(\hat{m}_{WL}^1(x) \equiv m^1(x)\). 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 (Buhlmann & van de Geer 2011). Therefore, our method provides uniformly valid inference similarly as in Belloni et al. (2014, Corollary 1).

3.2. On construction of estimators. We point out basic ideas underlying the construction of the estimators \(\hat{\gamma}_{RCAL}^1\) and \(\hat{\alpha}_{RWL}^1\) such that the estimator \(\hat{\mu}^1(\hat{m}_{RWL}^1, \hat{\pi}_{RCAL}^1)\) 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 \(\alpha^1\) in model (1), which is assumed to converge in probability to a limit \(\tilde{\alpha}^1\). Denote \(\hat{m}^1(X) = m^1(X; \hat{\alpha}^1)\) and \(\hat{m}^1(X) = m^1(X; \tilde{\alpha}^1)\). Similarly, let \(\hat{\gamma}\) be some estimator of \(\gamma\) in model (4), which is assumed to converge in probability to a limit \(\tilde{\gamma}\). Denote \(\hat{\pi}(X) = \pi(X; \hat{\gamma})\) and \(\hat{\pi}(X) = \pi(X; \tilde{\gamma})\).

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

\[
\hat{\mu}^1(\hat{m}^1, \hat{\pi}) = \hat{\mu}^1(m^1, \pi) + \Delta_1 + \Delta_2 + o_p(n^{-1/2}),
\]
with
\[
\Delta_1 = (\hat{\alpha}^1 - \bar{\alpha}^1)^T \times \frac{\partial}{\partial \alpha^1} \tilde{E}\{\varphi(Y, T, X; \alpha^1, \gamma)\} \bigg|_{(\alpha^1, \gamma) = (\hat{\alpha}^1, \hat{\gamma})},
\]
\[
\Delta_2 = (\hat{\gamma} - \bar{\gamma})^T \times \frac{\partial}{\partial \gamma} \tilde{E}\{\varphi(Y, T, X; \alpha^1, \gamma)\} \bigg|_{(\alpha^1, \gamma) = (\hat{\alpha}^1, \hat{\gamma})},
\]
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 \(\Delta_1\) or \(\Delta_2\) 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, \(\Delta_1\) and \(\Delta_2\), 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 \(\Delta_1\) and \(\Delta_2\) 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
\[
(22) \quad \hat{\mu}^1(\hat{m}^1, \hat{\pi}) = \hat{\mu}^1(\bar{m}^1, \bar{\pi}) + o_p(n^{-1/2}),
\]
it seems necessary that the second terms in \(\Delta_1\) and \(\Delta_2\) should be \(o_p(1)\) (i.e., vanish in probability), and their population versions should satisfy
\[
(23) \quad \frac{\partial}{\partial \alpha^1} E\{\varphi(Y, T, X; \alpha^1, \gamma)\} \bigg|_{(\alpha^1, \gamma) = (\hat{\alpha}^1, \hat{\gamma})} = 0,
\]
\[
(24) \quad \frac{\partial}{\partial \gamma} E\{\varphi(Y, T, X; \alpha^1, \gamma)\} \bigg|_{(\alpha^1, \gamma) = (\hat{\alpha}^1, \hat{\gamma})} = 0.
\]
These conditions (23) and (24) are known to be sufficient for (22) to hold, under additional regularity conditions in low-dimensional settings (Kim & Haziza 2014; Vermeulen & Vansteelandt 2015). As discussed below, such orthogonality conditions are also key to the construction of our regularized estimators \(\hat{\gamma}^1_{\text{RCAL}}\) and \(\hat{\alpha}^1_{\text{RWL}}\) in high-dimensional settings.

We now examine implications of (23)–(24) in the concrete situation of Section 3.1. For linear OR model (9), \(\Delta_1\) reduces to
\[
(25) \quad \Delta_1 = (\hat{\alpha}^1 - \bar{\alpha}^1)^T \times \tilde{E} \left[ \left\{ 1 - \frac{T}{\bar{\pi}(X)} \right\} f(X) \right].
\]
For logistic PS model (4), \(\Delta_2\) reduces to
\[
(26) \quad \Delta_2 = - (\hat{\gamma} - \bar{\gamma})^T \times \tilde{E} \left[ T \frac{1 - \bar{\pi}(X)}{\bar{\pi}(X)} \{Y - \bar{m}^1(X)\} f(X) \right].
\]
Then conditions (23) and (24) become

\[
E \left\{ \left. \left( 1 - \frac{T}{\bar{\pi}(X)} \right) f(X) \right\} \right|_{\bar{\pi}(X)} = 0,
\]

(27)

\[
E \left\{ \left. T \left( \frac{1 - \bar{\pi}(X)}{\bar{\pi}(X)} \right) \{ Y - \bar{m}(X) \} f(X) \right\} \right|_{\bar{\pi}(X)} = 0.
\]

(28)

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

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

(29)

\[
\frac{\partial}{\partial \alpha} E\{\varphi(Y, T, X; \alpha, \gamma)\}\bigg|_{(\alpha, \gamma) = (\alpha^*, \gamma^*)} = 0,
\]

(30)

\[
\frac{\partial}{\partial \gamma} E\{\varphi(Y, T, X; \alpha, \gamma)\}\bigg|_{(\alpha, \gamma) = (\alpha^*, \gamma^*)} = 0,
\]

where \(\alpha^*\) and \(\gamma^*\) are the true values such that \(m(X; \alpha^*) = m^*(X)\) and \(\pi(X; \gamma^*) = \pi^*(X)\). While conditions (23)–(24) in general depend on the estimators (\(\hat{\alpha}, \hat{\gamma}\)) through their limits (\(\bar{\alpha}, \bar{\gamma}\)), such dependency is suppressed in (29)–(30) because (\(\bar{\alpha}, \bar{\gamma}\)) is assumed to coincide with (\(\alpha^*, \gamma^*\)) under correctly specified models. The identities (29) and (30) are known to hold for the AIPW estimating function \(\varphi()\) as a consequence of its double robustness (Robins & Rotnitzky 2001). Moreover, the relationship (29)–(30) is called an orthogonality property and exploited by Belloni et al. (2014) to establish asymptotic expansion (8) for the estimator \(\hat{\mu}(\hat{m}_{\text{RML}}, \hat{\pi}_{\text{RML}})\), 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 \((\hat{\alpha}_{\text{RML}}, \hat{\gamma}_{\text{RML}})\). In this context, the estimators \(\hat{\gamma}_{\text{CAL}}\) and \(\hat{\alpha}_{\text{WL}}\) are constructed with loss functions \(\ell_{\text{CAL}}(\gamma)\) and \(\ell_{\text{WL}}(\alpha; \gamma)\) in (12) and (14) such that

(31)

\[
\frac{\partial}{\partial \gamma} \ell_{\text{CAL}}(\gamma) = \frac{\partial}{\partial \alpha} \hat{E}\{\varphi(Y, T, X; \alpha, \gamma)\}
\]

\[
= \hat{E} \left\{ \left. \left( 1 - \frac{T}{\pi(X; \gamma)} \right) f(X) \right\} \right|_{\pi(X; \gamma)}.
\]
(32) \[
\frac{\partial}{\partial \alpha^1} \ell_{WL}(\alpha^1; \gamma) = \frac{\partial}{\partial \gamma} \tilde{E}\{\varphi(Y, T, X; \alpha^1, \gamma)\}
\]
\[= -\tilde{E}\left[T - \frac{\pi(X; \gamma)}{\pi(X; \gamma)} \{Y - \alpha^T f(X)\} f(X)\right].\]

As a result, conditions (27) and (28) are satisfied by \(\hat{\pi}_{\text{CAL}}^1(X) = \pi(X; \hat{\gamma}_{\text{CAL}}^1)\) and \(\hat{m}_{\text{WL}}^1(X) = m^1(X; \hat{\alpha}_{\text{WL}}^1)\), where \(\hat{\gamma}_{\text{CAL}}^1\) and \(\hat{\alpha}_{\text{WL}}^1\) are, respectively, the probability limits of \(\hat{\gamma}_{\text{RCAL}}^1\) and \(\hat{\alpha}_{\text{RWL}}^1\), defined as minimizers of the expected loss \(E\{\ell_{\text{CAL}}(\gamma)\}\) and \(E\{\ell_{\text{WL}}(\alpha^1; \gamma^1_{\text{CAL}})\}\). 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 \(\|\hat{\gamma}_{\text{RCAL}}^1 - \hat{\gamma}_{\text{CAL}}^1\|_1 = O_p(1)|S_\gamma|/\{\log(p)/n\}^{1/2}\) and \(\|\hat{\alpha}_{\text{RWL}} - \hat{\alpha}_{\text{WL}}\|_1 = O_p(1)(|S_\gamma| + |S_{\alpha^1}|)\{\log(p)/n\}^{1/2}\), as demonstrated in Theorems 1 and 2 later. Consequently, the products \(\Delta_1\) and \(\Delta_2\) 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 \(\hat{\gamma}_{\text{RCAL}}^1\) is called a regularized calibrated estimator of \(\gamma\) (Tan 2017), because in the extreme case of \(\lambda = 0\), Eqs. (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 \(\hat{\gamma}_{\text{RCAL}}^1\) can also be derived to reduce the variation associated with estimation of \(\alpha^1\) from linear OR model (9) for the estimator \(\hat{\mu}^1(m^1, \hat{\pi})\), when PS model (4) may be misspecified. Similarly, \(\hat{\alpha}_{\text{RWL}}^1\) is constructed to reduce the variation associated with estimation of \(\gamma\) from logistic PS model (4) for the estimator \(\hat{\mu}^1(m^1, \hat{\pi})\), when OR model (9) may be misspecified. By extending the meaning of calibrated estimation, we call \(\hat{\alpha}_{\text{RWL}}^1\) a regularized calibrated estimator of \(\alpha^1\) against model (4), as well as \(\hat{\gamma}_{\text{RCAL}}^1\) a regularized calibrated estimator of \(\gamma\) against model (9), when used to define \(\hat{\mu}^1(m^1, \hat{\pi})\).

**Remark 2.** Conditions (23)–(24) were previously used by Kim & Haziza (2014) and Vermeulen & Vansteelandt (2015) to construct an augmented IPW estimator \(\hat{\mu}^1(m^1, \hat{\pi})\) for \(\mu^1\) in low-dimensional settings, where \((\hat{\alpha}^1, \hat{\gamma})\) are non-penalized, 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 \((\hat{\alpha}^1, \hat{\gamma})\). 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 \(\hat{\alpha}^1 - \hat{\alpha}\) and \(\hat{\gamma} - \hat{\gamma}\) 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 \((\hat{\alpha}_{\text{RWL}}^1, \hat{\gamma}_{\text{RCAL}}^1)\) and achieve asymp-
to asymptotic expansion (22) for $\hat{\mu}^1(\hat{m}^{1\text{RWL}}, \hat{\pi}^{1\text{RCAL}})$, so that valid confidence intervals for $\mu^1$ can be obtained in high-dimensional settings.

Remark 3. Our approach based on (23)–(24) and that of Belloni et al. (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 $\varphi()$, but in different manners. Conditions (23)–(24) amount to requiring orthogonality to hold at whatever limits $(\hat{\alpha}^1, \hat{\gamma})$, 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 $(\hat{\alpha}^1, \hat{\gamma})$ 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 $\hat{\mu}^1(\hat{m}^{1\text{RWL}}, \hat{\pi}^{1\text{RCAL}})$ 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 $\hat{\alpha}^{1\text{RWL}}$ on $\hat{\pi}^{1\text{RCAL}}$, and what condition is required on the sparsity sizes of $\bar{\gamma}^{1\text{CAL}}$ and $\bar{\alpha}^{1\text{WL}}$.

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

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

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 $\bar{\gamma}^{1\text{CAL}}$ is defined as a minimizer of the expected calibration loss $E\{\ell_\text{CAL}(\gamma)\}$ as in Section 3.1. From a functional perspective, we write $\ell_\text{CAL}(\gamma) = \kappa_\text{CAL}(\gamma^Tf)$, where for a function $h(x)$,

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

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

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

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

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

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

$$\sum_{j \notin S} |b_j| \leq \xi_0 \sum_{j \in S} |b_j|. \quad (33)$$

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

**Assumption 1.** Suppose that the following conditions are satisfied:

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

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

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

(iv) $|S_\gamma| \lambda_0 \leq \eta_0$ for a sufficiently small constant $\eta_0 > 0$, depending only on $(A_0, C_0, \xi_0, \nu_0)$.

**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 $1 - 4e$,

$$D_{\text{CAL}}(\hat{h}_{\text{RCAL}}^1, \bar{h}_{\text{CAL}}^1) + (A_0 - 1)\lambda_0 \|\hat{\gamma}_{\text{RCAL}}^1 - \bar{\gamma}_{\text{CAL}}^1\|_1 \leq M_0 |S_\gamma| \lambda_0^2, \quad (34)$$

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 $E\{f(X)f^T(X)\}$ in conjunction with additional conditions such as for some constant $\tau_0 > 0$,

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

For example, (35) holds provided that $\pi^*(X)$ is bounded from below by a positive constant and $\pi(X; \hat{\gamma}_{CAL}^1)$ 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, $\pi(X; \hat{\gamma}_{CAL}^1)$ 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 $f_1(X), \ldots, f_p(X)$ are jointly (not just marginally) sub-Gaussian (Huang & 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 $\mu^1$ from existing works (Belloni et al. 2014) and our later results.

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

Now we turn to the regularized weighted least-squares estimator $\hat{\alpha}_{RWL}^1$. We develop a new strategy of inverting a quadratic inequality to address the dependency of $\hat{\alpha}_{RWL}^1$ on $\hat{\gamma}_{RCAL}^1$ and establish convergence of $\hat{\alpha}_{RWL}^1$ 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 $\lambda$ in (13) for defining $\hat{\alpha}_{RWL}^1$ is specified as $\lambda = A_1 \lambda_1$, with a constant $A_1 > 1$ and

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

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 $\tilde{\alpha}_{WL}^1$ is defined as a minimizer of
the expected loss $E \{ \ell_{\text{WL}}(\alpha^1; \gamma^1_{\text{CAL}}) \}$ as in Section 3.1. The following result gives the convergence of $\hat{\alpha}^1_{\text{RWL}}$ to $\hat{\alpha}^1_{\text{WL}}$ in the $L_1$ norm $\|\hat{\alpha}^1_{\text{RWL}} - \hat{\alpha}^1_{\text{WL}}\|_1$ and the weighted (in-sample) prediction error defined as

$$Q_{\text{WL}}(\hat{m}^1_{\text{RWL}}; \hat{m}^1_{\text{WL}}; \gamma^1_{\text{CAL}}) = \hat{E} \left[ T w(X; \gamma^1_{\text{CAL}})(\hat{m}^1_{\text{RWL}}(X) - \hat{m}^1_{\text{WL}}(X))^2 \right],$$

where $\hat{m}^1_{\text{RWL}}(X) = \hat{\alpha}^1_{\text{RWL}} f(X)$ and $\hat{m}^1_{\text{WL}}(X) = \hat{\alpha}^1_{\text{WL}} f(X)$. In fact, $Q_{\text{WL}}(\hat{m}^1_{\text{RWL}}; \hat{m}^1_{\text{WL}}; \gamma^1_{\text{CAL}})$ is the symmetrized Bregman divergence between $\hat{m}^1_{\text{RWL}}(X)$ and $\hat{m}^1_{\text{WL}}(X)$ associated with the loss $\kappa_{\text{WL}}(h; \gamma^1_{\text{CAL}}) = \hat{E}[T w(X; \gamma^1_{\text{CAL}})(Y - h(X))^2]/2$.

See Section 3.4 for further discussion.

**Assumption 2.** Suppose that the following conditions are satisfied:

(i) $Y^1 - \hat{m}^1_{\text{WL}}(X)$ is uniformly sub-Gaussian given $X$: $D^2_{\text{WL}}E(\exp\{|Y^1 - \hat{m}^1_{\text{WL}}(X)|^2/D^2_{\text{WL}}|X\} \leq D^2_{\text{WL}}$ for some positive constants $(D_0, D_1)$;

(ii) the compatibility condition holds for $\Sigma_\gamma$ with the subset $S_{\alpha^1} = \{0\} \cup \{j : \hat{\alpha}^1_{\text{WL},j} \neq 0, j = 1, \ldots, p\}$ and some constants $\nu_1 > 0$ and $\xi_1 > 1$;

(iii) $(1 + \xi_1)^2 \nu_1^{-2} |S_{\alpha^1}| \lambda_1 \leq \eta_1$ 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 $\log\{(1 + p)/\epsilon\}/n \leq 1$, then we have with probability at least $1 - 8\epsilon$,

$$Q_{\text{WL}}(\hat{m}^1_{\text{RWL}}; \hat{m}^1_{\text{WL}}; \gamma^1_{\text{CAL}}) + e^{\nu_01}(A_1 - 1)\lambda_1 ||\hat{\alpha}^1_{\text{RWL}} - \hat{\alpha}^1_{\text{WL}}||_1 \leq e^{4\nu_01} \xi_2^{-2} (M_{01} |S_{\gamma}| \lambda_0^2 + e^{2\nu_01} \xi_3^2 (\nu_2^{-2} |S_{\alpha^1}| \lambda_2^2),$$

where $\xi_2 = 1 - 2A_1/\{(\xi_1 + 1)(A_1 - 1)\} \in (0, 1]$, $\xi_3 = (A_1 + 1)(A_1 - 1)$, and $\nu_1 = (A_1 - 1)/\nu_1$, depending only on $(A_1, \xi_1, \nu_1, \eta_1)$, and $M_{01} = (D^2_0 + D^2_1)(e^{\nu_01} M_0 + \eta_0) + (D^2_0 + D_0 D_1)\eta_0$, $\eta_0 = (A_0 - 1)^{-1} M_{00} \eta_0 C_0$, and $\eta_0 = (A_0 - 1)^{-2} M_{00}^2 \eta_0$, 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 $\Sigma_\gamma$ as in Assumption 1(iii), but with the sparsity subset $S_{\alpha^1}$ from $\hat{\alpha}^1_{\text{WL}}$ instead of $S_S$ from $\gamma^1_{\text{CAL}}$. The matrix $\Sigma_\gamma$ is also the Hessian of the expected loss $E\{\ell_{\text{WL}}(\alpha^1; \gamma^1_{\text{CAL}})\}$ at $\alpha^1 = \hat{\alpha}^1_{\text{WL}}$, for reasons mentioned in Remark 5. Assumptions 2(ii)–(iii) are combined to derive a compatibility condition for the sample matrix $\tilde{\Sigma}_\gamma = \hat{E}[T w(X; \gamma^1_{\text{CAL}}) f(X)] f^T(X)$. Assumption 2(iii) can be relaxed such that $|S_{\alpha^1}| \lambda^2_2$ is sufficiently small under further side conditions, but it is already weaker than the sparsity condition, $|S_{\alpha^1}| \log(p) = o(\eta_0^{1/2})$, later needed for valid confidence intervals for $\mu^1$. Essentially, the conditions in Assumption 2 are comparable to those for high-dimensional analysis of standard Lasso estimators (e.g., Buhlmann & van de Geer 2011).
Remark 7. One of the key steps in our proof is to upper-bound the product
\[ (\hat{\alpha}^{1}_{\text{RWL}} - \hat{\alpha}^{1}_{\text{WL}})^{T} \hat{E} \left[ Tw(X; \hat{\gamma}^{1}_{\text{RCAL}}) \{ Y - \hat{m}^{1}_{\text{WL}}(X) \} f(X) \right], \]
which involves the estimated weight \( w(X; \hat{\gamma}^{1}_{\text{RCAL}}) \). If \( \hat{\gamma}^{1}_{\text{RCAL}} \) were replaced by \( \hat{\gamma}^{1}_{\text{CAL}} \), then it is standard to use the following bound,
\[ (\hat{\alpha}^{1}_{\text{RWL}} - \hat{\alpha}^{1}_{\text{WL}})^{T} \hat{E} \left[ Tw(X; \hat{\gamma}^{1}_{\text{CAL}}) \{ Y - \hat{m}^{1}_{\text{WL}}(X) \} f(X) \right] \leq \| \hat{\alpha}^{1}_{\text{RWL}} - \hat{\alpha}^{1}_{\text{WL}} \|_{1} \times \| \hat{E} \left[ Tw(X; \hat{\gamma}^{1}_{\text{CAL}}) \{ Y - \hat{m}^{1}_{\text{WL}}(X) \} f(X) \right] \|_{\infty}. \]
To handle the dependency on \( \hat{\gamma}^{1}_{\text{RCAL}} \), our strategy is to derive an upper bound of the difference between (38) and (39), depending on \( Q_{\text{WL}}(\hat{m}^{1}_{\text{RWL}}, \hat{m}^{1}_{\text{WL}}; \hat{\gamma}^{1}_{\text{CAL}}) \), which we seek to control. Carrying this bound leads to a quadratic inequality in \( Q_{\text{WL}}(\hat{m}^{1}_{\text{RWL}}, \hat{m}^{1}_{\text{WL}}; \hat{\gamma}^{1}_{\text{CAL}}) \), which can be inverted to obtain an explicit bound on \( Q_{\text{WL}}(\hat{m}^{1}_{\text{RWL}}, \hat{m}^{1}_{\text{WL}}; \hat{\gamma}^{1}_{\text{CAL}}) \). The resulting error bound (37) is of order \( \left( |S_{\gamma}| + |S_{\alpha}| \right) \log(p)/n \), much sharper than could be obtained using other approaches, for example, directly bounding \( \| \hat{E}[ Tw(X; \hat{\gamma}^{1}_{\text{RCAL}}) \{ Y - \hat{m}^{1}_{\text{WL}}(X) \} f(X) \|_{\infty} \). 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 \( \hat{\mu}^{1}(\hat{m}^{1}_{\text{RWL}}, \hat{\pi}^{1}_{\text{RCAL}}) \) for \( \mu^{1} \), depending on the regularized estimators \( \hat{\gamma}^{1}_{\text{RCAL}} \) and \( \hat{\alpha}^{1}_{\text{RWL}} \) from logistic propensity score model (4) and linear outcome regression model (9). The following result gives an error bound for \( \hat{\mu}^{1}(\hat{m}^{1}_{\text{RWL}}, \hat{\pi}^{1}_{\text{RCAL}}) \), allowing that both models (4) and (9) may be misspecified.

Theorem 3. Under the conditions of Theorem 2, if \( \log \{(1 + p)/\epsilon \}/n \leq 1 \), then we have with probability at least \( 1 - 10\epsilon \),
\[ \left| \hat{\mu}^{1}(\hat{m}^{1}_{\text{RWL}}, \hat{\pi}^{1}_{\text{RCAL}}) - \mu^{1}(\hat{m}^{1}_{\text{WL}}, \hat{\pi}^{1}_{\text{CAL}}) \right| \leq M_{11}|S_{\gamma}|\lambda_{0}^{2} + M_{12}|S_{\gamma}|\lambda_{0}\lambda_{1} + M_{13}|S_{\alpha}||\lambda_{0}\lambda_{1}, \]
where \( M_{11} = M_{13} + \sqrt{D_{0}^{2} + D_{1}^{2}}e^{\eta_{0}1}(e^{\eta_{1}1}M_{0}/2 + \eta_{02}), M_{12} = (A_{0} - 1)^{-1}M_{0}, M_{13} = A_{0}(A_{1} - 1)^{-1}M_{1}, \) and \( M_{1} \) is a constant such that the right hand side of (37) in Theorem 2 is upper-bounded by \( e^{\eta_{01}1}(|S_{\gamma}|\lambda_{0}\lambda_{1} + |S_{\alpha}||\lambda_{0}^{2}). \)

Theorem 3 shows that \( \hat{\mu}^{1}(\hat{m}^{1}_{\text{RWL}}, \hat{\pi}^{1}_{\text{RCAL}}) \) is doubly robust for \( \mu^{1} \) provided \( (|S_{\gamma}| + |S_{\alpha}|)|\lambda_{1}^{2} = o(1) \), that is, \( (|S_{\gamma}| + |S_{\alpha}||)\log(p) = o(n) \). In addition, Theorem 3 gives the \( n^{-1/2} \) asymptotic expansion (20) provided \( n^{1/2}(|S_{\gamma}| + |S_{\alpha}|)|\lambda_{1}^{2} = o(1) \), that is, \( (|S_{\gamma}| + |S_{\alpha}||)\log(p) = o(n^{1/2}) \). To obtain valid confidence intervals for \( \mu^{1} \) 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 $\hat{\varphi} = \varphi(T, Y, X; \hat{m}_{\text{WL}}, \hat{\pi}_{\text{CAL}})$ and $\hat{\varphi}_c = \hat{\varphi} - \hat{\mu}^1(\hat{m}_{\text{WL}}, \hat{\pi}_{\text{CAL}})$ such that $\hat{V} = \hat{E}(\hat{\varphi}_c^2)$. Similarly, denote $\bar{\varphi} = \varphi(T, Y, X; \bar{m}_{\text{WL}}, \bar{\pi}_{\text{CAL}})$ and $\bar{\varphi}_c = \bar{\varphi} - \bar{\mu}^1(\bar{m}_{\text{WL}}, \bar{\pi}_{\text{CAL}})$ such that $V = E(\bar{\varphi}_c^2)$.

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

$$
\left| \hat{E}
\left(
\hat{\varphi}_c^2 - \bar{\varphi}_c^2
\right)
\right| \leq 2M_{14}\{\hat{E}(\hat{\varphi}_c^2)\}^{1/2}(|S_\gamma|\lambda_0 + |S_{\alpha_1}|\lambda_1) + M_{14}(|S_\gamma|\lambda_0 + |S_{\alpha_1}|\lambda_1)^2,
$$

(41)

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 $1 - 12\epsilon$,

$$
\left| \hat{E}
\left(
\hat{\varphi}_c^2 - \bar{\varphi}_c^2
\right)
\right| \leq 2M_{15}\{\hat{E}(\hat{\varphi}_c^2)\}^{1/2}(|S_\gamma|\lambda_0\lambda_1 + |S_{\alpha_1}|\lambda_1^2)^{1/2} + M_{15}(|S_\gamma|\lambda_0\lambda_1 + |S_{\alpha_1}|\lambda_1^2),
$$

(42)

where $M_{15}$ is a positive constant, depending on $\tau_0$ 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 $\mu^1$ 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 $\mu^1$. 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 $\hat{\mu}^1(\hat{m}_{\text{WL}}, \hat{\pi}_{\text{CAL}}) \pm z_{c/2}\sqrt{\hat{V}/n}$ is an asymptotic $(1 - c)$ confidence interval for the target value $\mu^1 = E(\bar{\varphi})$, which in general differs from the true value $\mu^1$. The only effect of models (4) and (9) being correctly
specified is to ensure that \( \hat{\mu} \) coincides with \( \mu \). Such robustness 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 \( \hat{\mu}(\hat{m}_{\text{RML}}, \hat{\tau}_{\text{RML}}) \) can also be shown to converge to a target value \( \mu_{\text{ML}} \), which, by double robustness, coincides with \( \mu \) if either model (4) or (9) is correctly specified, but otherwise differs from \( \mu \). As discussed in Section 3.2, asymptotic expansion (8) for \( \hat{\mu}(\hat{m}_{\text{RML}}, \hat{\tau}_{\text{RML}}) \) is in general invalid, and so are the associated confidence intervals for \( \mu_{\text{ML}} \), 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

\[
\begin{align*}
   b_\gamma &= \hat{E}^{1/2} \left[ \{1/\hat{\pi}_{\text{CAL}}(X) - 1/\pi^*(X)\}^2 \right], \\
   b_{\alpha_1} &= \hat{E}^{1/2} \left[ \{\hat{m}_{\text{WL}}(X) - m^{1*}(X)\}^2 \right],
\end{align*}
\]

These bias terms can be linked to those in Belloni et al. (2014) and Farrell (2015) by the following bounds (shown in the Supplement):

\[
\begin{align}
   E \left( b_\gamma^2 \right) &\leq O(1) \inf_\gamma E \left[ \{1/\pi(X; \gamma) - 1/\pi^*(X)\}^2 \right], \\
   E \left( b_{\alpha_1}^2 \right) &\leq O(1) \inf_{\alpha} E \left[ \{m^1(X; \alpha^1) - m^{1*}(X)\}^2 \right],
\end{align}
\]

where \( O(1) \) in each inequality depends only on constants \((B_0, B_1)\) and \((B_0^*, B_1^*)\) such that \( \log\{\hat{\pi}_{\text{CAL}}(X)/(1 - \hat{\pi}_{\text{CAL}}(X))\} \in [B_0, B_1] \) 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 \( \pi^*(X) \geq e^{B_0^*}/(1 + e^{B_0^*}) \) almost surely for a constant \( B_0^* \in \mathbb{R} \). Then for sufficiently large constants \( A_0 \) and \( A_1 \),

\[
\begin{align}
   &|\hat{\mu}^1(\hat{m}_{\text{RML}}, \hat{\tau}_{\text{RML}}) - \hat{\mu}^1(\hat{m}_{\text{WL}}, \pi^*)| \\
   &\leq \left( |S_\gamma| + |S_{\alpha_1}| \right) O_p(\log(p)/\sqrt{n}) + b_\gamma O_p(n^{-1/2}) + b_\gamma b_{\alpha_1}, \\
   &|\hat{\mu}^1(\hat{m}_{\text{RML}}, \hat{\tau}_{\text{RML}}) - \hat{\mu}^1(\hat{m}_{\text{RML}}, \hat{\tau}_{\text{RML}})| \\
   &\leq \left( |S_\gamma| + |S_{\alpha_1}| \right) O_p(\log(p)/\sqrt{n}) + b_{\alpha_1} O_p(n^{-1/2}) + b_\gamma b_{\alpha_1}.
\end{align}
\]

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 \( \hat{\pi}_{\text{CAL}} \) replaced by \( \pi^* \) or \( \hat{m}_{\text{WL}} \) replaced by \( m^{1*} \) (or both) in the definition of \( V \).
Remark 10. For \( \hat{\mu}^1(\hat{m}^1_{\text{RML}}, \hat{\pi}^1_{\text{RML}}) \) to admit asymptotic expansion (8), the rate condition in Farrell (2015, Corrigendum) amounts to

\[
\left\{ |S_\gamma| \log^{\frac{3}{2}+\delta}(p)/n + b^2 \right\}^{\frac{1}{2}} \left\{ |S_{\alpha^1}| \log^{\frac{3}{2}+\delta}(p)/n + b_{\alpha^1}^2 \right\}^{\frac{1}{2}} = o_p(n^{-\frac{1}{2}})
\]

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 \( b^2 \) 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 \(|S_\gamma||S_{\alpha^1}| \log^{5+2\delta}(p) = o(n)\), which, ignoring \( \log(p) \) factors, is weaker than \(|\langle S_\gamma + |S_{\alpha^1}\rangle\log(p) = o(n^{1/2})\): 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 \( b^2 \) or \( b_{\alpha^1} \) is relatively large such that \(|S_{\alpha^1}| b^2 \) 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 \(|\langle S_\gamma + |S_{\alpha^1}\rangle\log(p) = o(n^{1/2})\). 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 \( \hat{\mu}^1(\hat{m}^1_{\text{RML}}, \hat{\pi}^1_{\text{RML}}) \) 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 \( (\hat{m}^1_{\text{RCAL}}, \hat{\pi}_{\text{RCAL}}) \) the corresponding series estimators. Assume that \( m^{1*}(x) \) and \( \pi^{*}(x) \) belong to Hölder classes with \( r_{m1} \) and \( r_{\pi} \)-times continuous derivatives. By bias and variance formulas in series estimation (Newey 1997), it might be expected from Proposition 2 that

\[
|\hat{\mu}^1(\hat{m}^1_{\text{RML}}, \hat{\pi}_{\text{RCAL}}) - \hat{\mu}^1(m^{1*}, \pi^{*})| \leq \left\{ p/n + (p^{-r_{\pi}/d} + p^{-r_{m1}/d})n^{-1/2} + p^{-(r_{\pi}+r_{m1})/d} \right\} O_p(1).
\]

By choosing \( p \propto n^{d/(r_{\pi}+r_{m1}+d)} \) to balance the 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

\[
\frac{r_{\pi} + r_{m1}}{r_{\pi} + r_{m1} + d} > \frac{1}{2}, \quad \text{i.e.,} \quad r_{\pi} + r_{m1} > d.
\]
As the mean squared errors of $\hat{\pi}_{RML}$ and $\hat{m}_{1RML}$ 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 $\hat{\mu}_{1}(\hat{m}_{1RML}, \hat{\pi}_{RML})$ satisfies asymptotic expansion (8) if
\[
\frac{r_\pi}{2r_\pi + d} + \frac{r_{m1}}{2r_{m1} + d} > \frac{1}{2},
\]
Condition (49) is weaker than (50), because $(r_\pi + r_{m1})/(r_\pi + r_{m1} + d) \geq r_\pi/(2r_\pi + d) + r_{m1}/(2r_{m1} + d)$ by Cauchy–Schwartz inequality, $2/(r_\pi + r_{m1} + d) \leq 1/(2r_\pi + d) + 1/(2r_{m1} + d)$, 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 $\mu^1$ 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 $\mu^1$ 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 non-linear 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 $\gamma$ and $\alpha^1$ 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,
\[
E(Y|T = 1, X) = m^1(X; \alpha^1) = \psi\{\alpha^{1T} f(X)\},
\]
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 $\mu^1$ are concerned. Our point estimator of $\mu^1$ is $\hat{\mu}^1(\hat{m}_{1RML}, \hat{\pi}_{RML})$ as defined in (10), where
\[ \hat{\pi}^{1}_{\text{RCAL}}(X) = \pi(X; \hat{\gamma}^{1}_{\text{RCAL}}) \] and \[ \hat{m}_{\text{RWL}}^{1}(X) = m^{1}(X; \hat{\alpha}^{1}_{\text{RWL}}). \] The estimator \( \hat{\gamma}^{1}_{\text{RCAL}} \) is a regularized calibrated estimator of \( \gamma \) from Tan (2017) as in Section 3.3. But \( \hat{\alpha}^{1}_{\text{RWL}} \) is a regularized weighted likelihood estimator of \( \alpha^{1} \), defined as a minimizer of

\[
\ell_{\text{RWL}}(\alpha^{1}; \hat{\gamma}^{1}_{\text{RCAL}}) = \ell_{\text{WL}}(\alpha^{1}; \hat{\gamma}^{1}_{\text{RCAL}}) + \lambda \|\alpha_{1:p}\|_{1},
\]

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

\[
(53) \quad \ell_{\text{WL}}(\alpha^{1}; \hat{\gamma}^{1}_{\text{RCAL}}) = \hat{E} \left(Tw(X; \hat{\gamma}^{1}_{\text{RCAL}}) [ -Y\alpha^{T}f(X) + \Psi\{\alpha^{T}f(X)\}] \right).
\]

The regularized weighted least-squares estimator \( \hat{\alpha}^{1}_{\text{RWL}} \) 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 \( \hat{\mu}^{1}(\hat{m}_{\text{RWL}}^{1}, \hat{\gamma}^{1}_{\text{RCAL}}) \) can be put in the prediction form (19), which ensures the boundedness property that \( \hat{\mu}^{1}(\hat{m}_{\text{RWL}}^{1}, \hat{\gamma}^{1}_{\text{RCAL}}) \) always falls within the range of the observed outcomes \( \{Y_i : T_i = 1, i = 1, \ldots, n\} \) and the predicted values \( \{\hat{m}_{\text{RWL}}^{1}(X_i) : T_i = 0, i = 1, \ldots, n\}. \)

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

\[
\kappa_{\text{WL}}(h; \gamma) = \hat{E} \left(Tw(X; \gamma) [-Yh(X) + \Psi\{h(X)\}] \right).
\]

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

\[
D_{\text{WL}}(h', h; \gamma) = \kappa_{\text{WL}}(h'; \gamma) - \kappa_{\text{WL}}(h; \gamma) - \langle \nabla \kappa_{\text{WL}}(h; \gamma), h' - h \rangle,
\]

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

\[
D_{\text{WL}}^{\dagger}(h', h; \gamma) = D_{\text{WL}}(h', h; \gamma) + D_{\text{WL}}(h, h'; \gamma)
\]

\[
= \hat{E} \left(Tw(X; \gamma) \left[ \psi\{h'(X)\} - \psi\{h(X)\} \right] \left\{h'(X) - h(X)\right\} \right).
\]

The following result establishes the convergence of \( \hat{\alpha}^{1}_{\text{RWL}} \) to \( \bar{\alpha}^{1}_{\text{WL}} \) in the \( L_1 \) norm \( \|\hat{\alpha}_{\text{RWL}}^{1} - \bar{\alpha}_{\text{WL}}^{1}\|_{1} \) and the symmetrized Bregman divergence \( D_{\text{WL}}^{\dagger}(\hat{\alpha}_{\text{RWL}}^{1}, \bar{\alpha}_{\text{WL}}^{1}; \hat{\gamma}_{\text{CAL}}^{1}) \), where \( \hat{\alpha}_{\text{RWL}}^{1}(X) = \hat{\alpha}_{\text{RWL}}^{1}f(X) \) and \( \bar{\alpha}_{\text{WL}}^{1}(X) = \bar{\alpha}_{\text{WL}}^{1}f(X) \). In the case of the identity link, \( \psi(u) = u \), the symmetrized Bregman divergence \( D_{\text{WL}}^{\dagger}(\hat{\alpha}_{\text{RWL}}^{1}, \bar{\alpha}_{\text{WL}}^{1}; \hat{\gamma}_{\text{CAL}}^{1}) \) becomes \( Q_{\text{WL}}(\hat{\alpha}_{\text{RWL}}^{1}, \bar{\alpha}_{\text{WL}}^{1}; \hat{\gamma}_{\text{CAL}}^{1}) \) 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 $\psi()$ is differentiable and denote $\psi_2(u) = d\psi(u)/du$. Suppose that the following conditions are satisfied:

(i) $\psi_2(\tilde{h}_{WL}^1(X)) \leq C_1$ almost surely for a constant $C_1 > 0$;
(ii) $\psi_2(\tilde{h}_{WL}^1(X)) \geq C_2$ almost surely for a constant $C_2 > 0$;
(iii) $\psi_2(u) \leq \psi_2(u')e^{C_3|u-u'|}$ for any $(u,u')$, where $C_3 \geq 0$ is a constant.
(iv) $C_0C_3(A_1 - 1)^{-1}\xi_3^2\nu_2^2C_2^{-1}|S_\alpha|\lambda_1 \leq \eta_2$ for a constant $0 \leq \eta_2 < 1$ and $C_0C_3e^{3\eta_1} (A_1 - 1)^{-1}\xi_3^2C_2^{-1}(M_0|S_\gamma|\lambda_0) \leq \eta_3$ for a constant $0 \leq \eta_3 < 1$, where $(\eta_0, \nu_2, \xi_3, M_0)$ are as in Theorem 2.

THEOREM 5. Suppose that Assumptions 1, 2, and 3(ii)–(iv) hold. If $\log\{(1 + p)/\epsilon\}/n \leq 1$, 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 $1 - 8\epsilon$,

$$D_{WL}^1(\hat{m}_{RWL}^1, \tilde{m}_{WL}^1) + e^{\eta_1}(A_1 - 1)\lambda_1||\hat{\alpha}_{RWL}^1 - \tilde{\alpha}_{WL}^1||1 \leq e^{4\eta_1}\xi_4^{-2} (M_0|S_\gamma|\lambda_0^2) + e^{2\eta_1}\xi_3^2 \left(\nu_2^2|S_\alpha|\lambda_2^2\right),$$

(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 $(\eta_0, \nu_2, \xi_3, M_0)$ 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 & Zhang (2012), is used along with Assumption 1(i) to bound the curvature of $D_{WL}^1(h', h; \tilde{\gamma}_{CAL}^1)$ 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 $\tilde{\Sigma}_\alpha = \tilde{E}[Tw(X; \tilde{\gamma}_{CAL}^1)\psi_2(\tilde{h}_{WL}^1(X))f(X)f^T(X)]$, which is the sample version of the Hessian of the expected loss $E[\ell_{WL}(\alpha; \tilde{\gamma}_{CAL}^1)]$ at $\alpha = \hat{\alpha}_{WL}^1$, that is, $\Sigma_\alpha = E[Tw(X; \tilde{\gamma}_{CAL}^1)\psi_2(\tilde{h}_{WL}^1(X))f(X)f^T(X)]$. Second, Assumption 3(ii) is also used in deriving a quadratic inequality to be inverted in our strategy to deal with the dependency of $\hat{\alpha}_{RWL}^1$ on $\tilde{\gamma}_{RCAL}^1$ 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 $\tau_1 > 0$,

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

provided that the condition on $A_1$ and Assumption 3(iv) are modified accordingly, depending on $\tau_1$. This extension is not pursued here for simplicity.

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

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

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

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

By the definition of $\hat{\gamma}^1_{\text{CAL}}$, it holds that $E[\{T/\hat{\pi}^1_{\text{CAL}}(X) - 1\} f_j(X)] = 0$ for $j = 0, 1, \ldots, p$ 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 $\log\{(1 + p)/\epsilon\}/n \leq 1$, 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 $1 - 12\epsilon$,

$$| \hat{\mu}^1(\hat{m}^{1}_{\text{RWL}}, \hat{\pi}^1_{\text{RCAL}}) - \mu^1(\hat{m}^{1}_{\text{RWL}}, \hat{\pi}^1_{\text{CAL}}) | \leq M_{21}|S_1|\lambda_0^2 + M_{22}|S_1|\lambda_0\lambda_1 + M_{23}|S_{\alpha^1}|\lambda_0\lambda_1 + \eta_{11}\Lambda_0(\eta_{11}),$$

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 $(C_1, C_2, C_3, \nu_2, \nu_3, \eta_3, \eta_{11}) = (A_1 - 1)^{-1} M_2 (|S_1|\lambda_0 + |S_{\alpha^1}|\lambda_1)$, and $M_2$ is a constant such that the right hand side of (55) is upper-bounded by $e^{\gamma_0} M_2(|S_1|\lambda_0\lambda_1 + |S_{\alpha^1}|\lambda_1^2)$. If, in addition, condition (35) holds, then we have with probability at least $1 - 14\epsilon$,

$$| \hat{\mu}^1(\hat{m}^{1}_{\text{RWL}}, \hat{\pi}^1_{\text{RCAL}}) - \mu^1(\hat{m}^{1}_{\text{RWL}}, \hat{\pi}^1_{\text{CAL}}) | \leq M_{24}|S_1|\lambda_0^2 + M_{25}|S_1|\lambda_0\lambda_1 + M_{26}|S_{\alpha^1}|\lambda_0\lambda_1 + \eta_{11}\Lambda_1,$$

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 $\gamma_0$ from (35).

**Remark 13.** Two different error bounds are obtained in Theorem 6. Because $\Lambda_0(\eta_{11}) \geq \Lambda_1$, the error bound (57) is tighter than (56), but with the additional condition (35), which requires that the generalized eigenvalues of $\Sigma_\gamma$ relative to the gram matrix $E\{f(X)f^T(X)\}$ is bounded away from 0. In
Theorem 7. Similarly as discussed in Remark 8, if \(|S_γ| + |S_{α1}| \lambda_1 = o(1)|\), that is, \((|S_γ| + |S_{α1}|)(\log p)^{1/2} = o(n^{1/2})\) in addition, the error bounds imply that \(\hat{\mu}^1(\hat{m}_{\text{RWL}}^1, \hat{\pi}_{\text{RCAL}}^1)\) admits the \(n^{-1/2}\) asymptotic expansion (20) provided \((|S_γ| + |S_{α1}|)\log(p) = o(n^{1/2})\), when PS model (4) is correctly specified but OR model (51) may be misspecified, because the term involving \(A_0(\eta_1)\) or \(A_1\) 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 \(\hat{\nu}\) to \(\nu\) as defined in Proposition 1, allowing that both models (4) and (51) may be misspecified.

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

\[
\left| \hat{E}\left(\hat{\varphi}_c^2 - \varphi_c^2\right) \right| \leq 2M_{27}\{\hat{E}(\hat{\varphi}_c^2)\}^{1/2}\{1 + \lambda_0(\eta_1)\}\{|S_γ|\lambda_0 + |S_{α1}|\lambda_1\}
+ M_{27}\{1 + \lambda_0^2(\eta_1)\}\{|S_γ|\lambda_0 + |S_{α1}|\lambda_1|^2\},
\]

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 \(1 - 14\epsilon\),

\[
\left| \hat{E}\left(\hat{\varphi}_c^2 - \varphi_c^2\right) \right| \leq 2M_{28}\{\hat{E}(\hat{\varphi}_c^2)\}^{1/2}\left\{\{|S_γ|\lambda_0\lambda_1 + |S_{α1}|\lambda_1^2\}^{1/2} + \lambda_1(\{|S_γ|\lambda_0 + |S_{α1}|\lambda_1\}) \right\}
+ M_{28}\left\{|S_γ|\lambda_0\lambda_1 + |S_{α1}|\lambda_1^2\} + \lambda_0(\{|S_γ|\lambda_0 + |S_{α1}|\lambda_1\})\right\},
\]

where \(M_{28}\) is a positive constant, similar to \(M_{27}\) but, in addition, depending on \(\tau_0\) from (35).

Remark 14. Two different rates of convergence are obtained for \(\hat{\nu}\) in Theorem 7. Similarly as discussed in Remark 8, if \((|S_γ| + |S_{α1}|)(\log p)^{1/2} = o(n^{1/2})\), then inequality (58) implies the consistency of \(\hat{\nu}\) for \(\nu\), which is sufficient for applying Slutsky Theorem to establish confidence intervals for \(\mu^1\). With an additional condition (35), inequality (59) gives a faster rate of convergence of \(\hat{\nu}\) to \(\nu\), which is of order \(n^{-1/2}\) provided \((|S_γ| + |S_{α1}|)\log(p) = o(n^{1/2})\).

Combining Theorems 6–7 leads to the following result.

Proposition 3. Suppose that Assumptions 1, 2, and 3 hold, and \((|S_γ| + |S_{α1}|)\log(p) = o(n^{1/2})\). 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 $\pi_{\text{CAL}}^{1}(x) \equiv \pi^*(x)$, and (i)-(iii) in Proposition 1 hold. That is, a PS based, OR assisted confidence interval for $\mu^1$ is obtained.

**Remark 15.** The conclusion of Proposition 3 remains valid if PS model (4) is misspecified but only locally such that $\Lambda_0(\eta_1) = O(\{\log(p)/n\}^{1/2})$ or $\Lambda_1 = O(\{\log(p)/n\}^{1/2})$, in the case of the error bound (56) or (57). Therefore, $\mu^1(\hat{m}_{\text{RWL}}^1, \hat{\pi}_{\text{RCAL}}^0) \pm z_{\alpha/2}\sqrt{\hat{V}/n}$ can be interpreted as an asymptotic (1 - $\alpha$) confidence interval for the target value $\hat{\mu} = E(\hat{\varphi})$ 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 $\mu^1$, but they can be directly extended for estimating $\mu^0$ and hence ATE, that is, $\mu^1 - \mu^0$. Consider a logistic propensity score model (4) and a generalized linear outcome model,

\[(60) \quad E(Y|T = 0, X) = m_0(X; \alpha^0) = \psi\{\alpha^{\text{et}} f(X)\},\]

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

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

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

\[\ell_{\text{CAL}}^0(\gamma) = \hat{E} \{(1 - T)\alpha^{\text{et}} f(X) - T\gamma^T f(X)\},\]

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

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

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

\[\hat{\mu}^0(\hat{m}_{\text{RWL}}^0, \hat{\pi}_{\text{RCAL}}^0) = \hat{E} \{\varphi(Y, 1 - T, X; \hat{m}_{\text{RWL}}^0, 1 - \hat{\pi}_{\text{CAL}}^0)\} + o_p(n^{-1/2}),\]
where $\tilde{\pi}_{\text{RCAL}}^0(X) = \pi(X; \tilde{\gamma}_{\text{RCAL}}^0)$, $\tilde{m}_{\text{RWL}}^0(X) = m_0(X; \tilde{\alpha}_{\text{RWL}}^0)$, and $\gamma^0_{\text{RCAL}}$ and $\alpha^0_{\text{RWL}}$ are the target values defined similarly as $\gamma^1_{\text{RCAL}}$ and $\alpha^1_{\text{RWL}}$. Then Wald confidence intervals for $\mu^0$ 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 $\mu^0$ and $\mu^1$. Similar ideas have been involved in several related methods (e.g., Vermueilen & Vansteelandt 2015; Chan et al. 2016). On one hand, the estimators $\tilde{\gamma}_{\text{RCAL}}^0$ and $\tilde{\gamma}_{\text{RCAL}}^1$ 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 $\mu^1$ and $\mu^0$ 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 $\gamma^1$ and $\gamma^0$ 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 $\hat{E}(TY)/\hat{E}(T)$. For $\nu^0 = E(Y^1|T = 1)$, our point estimator is

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

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

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

The function $\varphi_{\nu^0}(\cdot; \tilde{m}_0, \tilde{\pi})$ can be derived, by substituting fitted values ($\tilde{m}_0, \tilde{\pi}$) for the true values ($m_0^*, \pi^*$) in the efficient influence function of $\nu^0$ under a nonparametric model (Hahn 1998). The estimator $\hat{E}\{\varphi_{\nu^0}(Y, T, X; \tilde{m}_0, \tilde{\pi})\}$ is doubly robust: it remains consistent for $E(TY)$ if either $\tilde{m}_0 = m_0^*$ or $\tilde{\pi} = \pi^*$. In addition, by straightforward calculation, the function $\varphi_{\nu^0}(\cdot)$ is related to $\varphi(\cdot)$ in (7) through the simple identity:

$$\varphi_{\nu^0}(Y, T, X; \tilde{m}_0, \tilde{\pi}) = \varphi(Y, 1 - T, X; \tilde{m}_0, 1 - \tilde{\pi}) - (1 - T)Y.$$
As a result, $\hat{\nu}(\hat{m}_{\text{RWL}}^0, \hat{\pi}_{\text{RCAL}}^0)$ can be equivalently obtained as

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

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

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

similarly as (61) for $\hat{\mu}(\hat{m}_{\text{RWL}}^0, \hat{\pi}_{\text{RCAL}}^0)$. From this expansion, Wald confidence intervals for $\nu^0$ 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:

$$\begin{align*}
\frac{\partial \tilde{E}\{\varphi(Y, T, X; \alpha_1, \gamma)\}}{\partial \alpha_1} &= \tilde{E}\left[ \left\{ 1 - \frac{T}{\pi(X; \gamma)} \right\} \psi_2\{\alpha_1^T f(X)\} f(X) \right], \\
\frac{\partial \tilde{E}\{\varphi(Y, T, X; \alpha_1, \gamma)\}}{\partial \gamma} &= -\tilde{E}\left[ T \frac{1 - \pi(X; \gamma)}{\pi(X; \gamma)} \{Y - m_1(X; \alpha_1)\} f(X) \right],
\end{align*}$$

where $\psi_2()$ denotes the derivative of $\psi()$. The two vectors of functions, (63) and (64), are intrinsically coupled in $(\alpha_1, \gamma)$, each depending on both $\gamma$ and $\alpha_1$, unless outcome model (51) is linear and hence the dependency of (63) on $\alpha_1$ 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 & Haziza 2014; Vermeulen & Vansteelandt 2015). But the pair of equations need to be solved simultaneously in $(\alpha_1, \gamma)$ instead of sequentially. For nonlinear estimating equations, there may be various difficulties in computation and asymptotic theory, for example, related to multiple solutions (Small et al. 2000). In high-dimensional settings, with $(\alpha_1, \gamma)$ in both (63) and (64), it seems impossible to sequentially define loss functions and regularized M-estimators in $\gamma$ and then $\alpha_1$ (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 $\hat{\gamma}^{1}_{RCAL}$ and $\hat{\alpha}^{1}_{RWL}$ based on the loss functions $\ell_{CAL}(\gamma)$ and $\ell_{WL}(\alpha^{1}; \gamma)$. 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 $\hat{\alpha}^{1}_{RML}$ in conjunction with a regularized estimator of $\gamma$ based on a weighted calibration loss,

$$
\ell_{WL}(\gamma; \hat{\alpha}^{1}_{RML}) = \bar{E}\left[\psi_2\{\hat{\alpha}^{1T}_{RML}f(X)\}\left\{Te^{-\gamma^Tf(X)} + (1 - T)\gamma^Tf(X)\right\}\right].
$$

The gradient of (65) in $\gamma$ is (63), with $\alpha^{1} = \hat{\alpha}^{1}_{RML}$. Similar results can be established as in Section 3.4, to provide valid confidence intervals for $\mu^{1}$ 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 $\hat{\mu}^{1}(\hat{m}^{1}_{RML}, \hat{\pi}^{1}_{RML})$, without or with post-Lasso refitting, and $\hat{\mu}^{1}(\hat{m}^{1}_{RWL}, \hat{\pi}^{1}_{RCAL})$. The design of the first study is modified and extended from Kang & 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 $\hat{\alpha}^{1}_{RML}$ or $\hat{\gamma}^{1}_{RML}$ or (11), (13), or (52) for computing $\hat{\alpha}^{1}_{RWL}$ or $\hat{\gamma}^{1}_{RCAL}$ is minimized for a fixed tuning parameter $\lambda$, using algorithms similar to those in Friedman et al. (2010), but with the coordinate descent method replaced by an active set method as in Osborne et al. (2000) for solving each Lasso penalized least squares problem. In addition, the penalized loss (11) for computing $\hat{\gamma}^{1}_{RCAL}$ is minimized using the algorithm in Tan (2017), where a nontrivial Fisher scoring step is involved for quadratic approximation.

The tuning parameter $\lambda$ is determined using 5-fold cross validation based on the corresponding loss function as follows. For $k = 1, \ldots, 5$, let $T_k$ be
a random subsample of size \( n/5 \) from \( \{1, 2, \ldots, n\} \). For a loss function \( \ell(\gamma) \), either \( \ell_{\text{ML}}(\gamma) \) in (5) or \( \ell_{\text{CAL}}(\gamma) \) in (12), denote by \( \ell(\gamma; I) \) the loss function obtained when the sample average \( \tilde{E}() \) is computed over only the subsample \( I \). The 5-fold cross-validation criterion is defined as \( CV_5(\lambda) = (1/5) \sum_{k=1}^{5} \ell(\hat{\gamma}_k(\lambda); I_k) \), where \( \hat{\gamma}_k(\lambda) \) is a minimizer of the penalized loss \( \ell(\gamma; I_k) + \lambda \| \gamma \|_1 \) over the subsample \( I_k \) of size \( 4n/5 \), i.e., the complement to \( I_k \). Then \( \lambda \) is selected by minimizing \( CV_5(\lambda) \) over the discrete set \( \{ \lambda^*/2^j : j = 0, 1, \ldots, 10 \} \), where for \( \hat{\pi}_0 = \tilde{E}(T) \), the value \( \lambda^* \) 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 \( \ell(\gamma) + \lambda \| \gamma \|_1 \) over the original sample has a minimum at \( \gamma_{1:p} = 0 \) for all \( \lambda \geq \lambda^* \).

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

4.2. Simulation setup and results. Let \( X = (X_1, \ldots, X_p) \) be multivariate normal with means 0 and covariances \( \text{cov}(X_j, X_k) = 2^{-|j-k|} \) for \( 1 \leq j, k \leq p \). In addition, let \( X_j^\dagger = X_j + \{X_j + 1\}_+ \) for \( j = 1, \ldots, 4 \), where \( c_+ = \max(0, c) \). 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 \( T \) given \( X \) as in (C1), but, independently, generate \( Y^1 \) given \( X \) from a Normal distribution with variance 1 and mean \( E(Y^1|X) = X_1^\dagger + 0.5X_2^\dagger + 0.25X_3^\dagger + 0.125X_4^\dagger \).

(C3) Generate \( Y^1 \) given \( X \) as in (C1), but, independently, generate \( T \) given \( X \) from a Bernoulli distribution with \( P(T = 1|X) = 1 + \exp(-1 - X_1^\dagger - 0.5X_2^\dagger - 0.25X_3^\dagger - 0.125X_4^\dagger) \)^{-1}.

As in Section 2, the observed data consist of independent and identically distributed observations \( \{(T_i Y_i, T_i, X_i) : i = 1, \ldots, n\} \). 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.
Table 1

Summary of results with linear outcome models

<table>
<thead>
<tr>
<th></th>
<th>(C1) cor PS, cor OR</th>
<th>(C2) cor PS, mis OR</th>
<th>(C3) mis PS, cor OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>0.019</td>
<td>0.026</td>
<td>0.009</td>
</tr>
<tr>
<td>√Var</td>
<td>0.070</td>
<td>0.070</td>
<td>0.072</td>
</tr>
<tr>
<td>√EVar</td>
<td>0.068</td>
<td>0.068</td>
<td>0.072</td>
</tr>
<tr>
<td>Cov90</td>
<td>0.864</td>
<td>0.859</td>
<td>0.906</td>
</tr>
<tr>
<td>Cov95</td>
<td>0.922</td>
<td>0.914</td>
<td>0.914</td>
</tr>
</tbody>
</table>

n = 800 and p = 200

<table>
<thead>
<tr>
<th></th>
<th>RML</th>
<th>RML2</th>
<th>RCAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>0.010</td>
<td>0.012</td>
<td>0.010</td>
</tr>
<tr>
<td>√Var</td>
<td>0.076</td>
<td>0.071</td>
<td>0.071</td>
</tr>
<tr>
<td>√EVar</td>
<td>0.069</td>
<td>0.069</td>
<td>0.069</td>
</tr>
<tr>
<td>Cov90</td>
<td>0.875</td>
<td>0.906</td>
<td>0.889</td>
</tr>
<tr>
<td>Cov95</td>
<td>0.941</td>
<td>0.950</td>
<td>0.950</td>
</tr>
</tbody>
</table>

n = 800 and p = 1000

<table>
<thead>
<tr>
<th></th>
<th>RML</th>
<th>RML2</th>
<th>RCAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>0.016</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>√Var</td>
<td>0.076</td>
<td>0.071</td>
<td>0.071</td>
</tr>
<tr>
<td>√EVar</td>
<td>0.069</td>
<td>0.069</td>
<td>0.069</td>
</tr>
<tr>
<td>Cov90</td>
<td>0.886</td>
<td>0.906</td>
<td>0.889</td>
</tr>
<tr>
<td>Cov95</td>
<td>0.950</td>
<td>0.950</td>
<td>0.950</td>
</tr>
</tbody>
</table>

Note: RML denotes $\hat{\mu}_1(\hat{m}_1, \hat{\pi}_1)$, and RML2 denotes the variant with $\hat{m}_1$ and $\hat{\pi}_1$ replaced by the fitted values obtained by refitting OR and PS models only including the variables selected from the corresponding Lasso estimation. RCAL denotes $\hat{\mu}_1(\hat{m}_1, \hat{\pi}_1)$. Bias and Var are the Monte Carlo bias and variance of the point 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.

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 $X_j^*$ is a monotone function of $X_j$, 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 $\mu_1$, 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 sub-models with only the covariates $X_1, \ldots, X_4$).

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 & Imbens 2002; Vermeulen & 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 \( Y > 30 \), 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 means 0 and variances 1. We apply the estimators \( \hat{\mu}^1(\hat{\pi}^1_{\text{RML}}, \hat{\pi}^1_{\text{RCAL}}) \) and \( \hat{\mu}^0(\hat{\pi}^0_{\text{RML}}, \hat{\pi}^0_{\text{RCAL}}) \) using regularized calibrated (RCAL) estimation and the corresponding estimators such as \( \hat{\mu}^1(\hat{\pi}^1_{\text{RML}}, \hat{\pi}^1_{\text{RML}}) \) using regularized maximum likelihood (RML) estimation, similarly as in the simulation study. The Lasso tuning parameter \( \lambda \) is selected by 5-fold cross validation over a discrete set \( \{\lambda^* / 2^j / 4 : j = 0, 1, \ldots, 24\} \), where \( \lambda^* \) is the value leading to a zero solution \( \gamma_1 = \cdots = \gamma_p = 0 \). We also compute the (ratio) IPW estimators, such as \( \hat{\mu}^1_{\text{IPW}} \), 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 \( \mu^1 \). This improvement can also be seen from Figure S7 in the Supplement, 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 \( \mu^1 \) from RML with post-Lasso refitting appears problematic with a large standard error.

Table 2  
Estimates of 30-day survival probabilities and ATE

<table>
<thead>
<tr>
<th></th>
<th>RML</th>
<th>IPW</th>
<th>Augmented IPW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RML2</td>
<td>RCAL</td>
<td>RML2</td>
</tr>
<tr>
<td>( \mu^1 )</td>
<td>.630 ± .026</td>
<td>.660 ± .049</td>
<td>.634 ± .023</td>
</tr>
<tr>
<td>( \mu^0 )</td>
<td>.690 ± .017</td>
<td>.691 ± .019</td>
<td>.687 ± .017</td>
</tr>
<tr>
<td>ATE</td>
<td>-.054 ± .031</td>
<td>-.031 ± .053</td>
<td>-.053 ± .029</td>
</tr>
</tbody>
</table>

Note: Estimate ± 2×standard error, including nominal standard errors for IPW.
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 & 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.

**Acknowledgements.** 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. The research 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”:** (). We provide technical proofs and additional numerical results.

**REFERENCES**


DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY
110 FRELINGHUYSEN ROAD
PISCATAWAY, NJ 08854
E-MAIL: ztan@stat.rutgers.edu