Journal of Econometrics 189 (2015) 1–23

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Robust inference on average treatment effects with possibly more covariates than observations[☆]

Max H. Farrell[*]

University of Chicago Booth School of Business, 5807 South Woodlawn Avenue, Chicago, IL 60637, United States

ARTICLE INFO

Article history:
Received 30 October 2013
Received in revised form
3 April 2015
Accepted 7 June 2015
Available online 25 June 2015

JEL classification:
C21
C31
C52

Keywords:
High-dimensional sparse model
Heterogeneous treatment effects
Uniform inference
Model selection
Doubly-robust estimator
Unconfoundedness
Group lasso

ABSTRACT

This paper concerns robust inference on average treatment effects following model selection. Under selection on observables, we construct confidence intervals using a doubly-robust estimator that are robust to model selection errors and prove their uniform validity over a large class of models that allows for multivalued treatments with heterogeneous effects and selection amongst (possibly) more covariates than observations. The semiparametric efficiency bound is attained under appropriate conditions. Precise conditions are given for any model selector to yield these results, and we specifically propose the group lasso, which is apt for treatment effects, and derive new results for high-dimensional, sparse multinomial logistic regression. Both a simulation study and revisiting the National Supported Work demonstration show our estimator performs well in finite samples.

1. Introduction

Model selection has always had a place in empirical economics, whether or not it is formally acknowledged. A key problem in modern empirical work is that researchers face datasets with large numbers of variables, sometimes more than observations. A complementary problem is that economic theory and prior knowledge may mandate controlling for certain variables, but are generally silent regarding functional form. These two problems force researchers to search for a model that is simultaneously parsimonious and adequately flexible. Many formal methods are computationally infeasible with a large number of variables. A typical response to this challenge is to iteratively search over a small set of alternative specifications, guided only by the researcher's taste and intuition. But no matter the approach used, subsequent inference almost never takes accounts for this "specification search" and the resulting confidence intervals are not robust to model selection mistakes, and hence are unreliable in empirical work.

This problem is particularly important in estimating average treatment effects under selection on observables, because in this framework using the right covariates is crucial for identification and correct inference. In this context, we provide an easy-to-implement and objective method for covariate selection and post-selection inference on average treatment effects.[1] We establish four main results for multivalued treatments effects with arbitrary

[☆] An online supplement contains additional proofs and simulations. I am deeply grateful to Matias Cattaneo for advice and support. I am indebted to Xuming He, Lutz Kilian, and Jeffrey Smith for thoughtful feedback and discussions. I thank Victor Chernozhukov for pointing to the relevant latest results, obtained in joint work Alexandre Belloni and Christian Hansen, and the latter two authors for conversations in the early stages of this project. I benefited from discussions with Rosa Matzkin, Blaise Melly, and Jack Porter. I also thank the co-editor, Han Hong, and two reviewers for their detailed comments and suggestions that improved the paper.

[*] Tel.: +1 773 834 0161.
E-mail address: max.farrell@chicagobooth.edu.
URL: http://faculty.chicagobooth.edu/max.farrell/.

[1] Treatment effects, missing data, measurement error, and data combination models are equivalent under selection on observables. Thus, all our results immediately apply to those contexts. For reviews of these literatures, see Tsiatis (2006), Heckman and Vytlacil (2007), Imbens and Wooldridge (2009), and Wooldridge (2010).

http://dx.doi.org/10.1016/j.jeconom.2015.06.017
0304-4076/© 2015 Elsevier B.V. All rights reserved.

heterogeneity in observables and heteroskedasticity. First, we show that a doubly-robust estimator is robust to model selection errors. These estimators were initially developed for robustness to parametric misspecification, but are now known to be robust to selection.[2] By taking explicit account of the model selection stage and its inherent selection errors, we derive precise conditions required for any model selector to deliver confidence intervals for average treatment effects that are uniformly valid over a large class of data-generating processes. Second, we show that a simple refitting procedure allows researchers to augment variables chosen according economic theory with data-driven selection to deliver flexible inference that remains uniformly valid. Third, we prove that our estimator is asymptotically linear, and standard conditions imposed in the program evaluation literature, semiparametrically efficient bound. Fourth, we derive new results for multinomial (and binary) logistic regression, the most widely used model for treatment assignment.

Inference following model selection is notoriously difficult. In a sequence of papers, Leeb and Pötscher (2005, 2008a,b), Pötscher and Leeb (2009) have shown that inference relying too heavily on model selection cannot be made uniformly valid. Loosely speaking, uniform validity of a confidence interval captures the idea that the interval should have the same quality (coverage) for many data-generating processes. This theoretical property is practically important because it implies greater reliability in applications. Our proposed methods for post model selection inference build upon the path-breaking recent work of Belloni et al. (2014).

The crucial insight that leads to uniform inference is to change the goal of model selection away from perfect *covariate* selection (the oracle property) and to high-quality approximation of the underlying *functions*. This fundamental shift in focus allows us to circumvent, without contradicting, the impossibility results of Leeb and Pötscher. Valid post-selection inference has attracted considerable attention during the preparation of this paper: in contexts and with methods quite different from ours, contributions have been made by Belloni et al. (2013), Berk et al. (2013), Zhang and Zhang (2014), Efron (2014), van de Geer et al. (2014), and Belloni et al. (2014), among others.

Our approach, based on the doubly-robust estimator, has several key features. The name "doubly-robust" reflects that it is robust to misspecification of either the treatment equation (propensity score) or the outcome equation, a property obtained by combining inverse probability weighting and regression imputation. First, we show that this robustness extends to model selection, enabling us to allow for selection errors in both equations without impacting inference. Second, we capture arbitrary treatment effect heterogeneity (dependence of the effect on an individual's observed characteristics), which is crucial in empirical work. With such heterogeneity, the average treatment effect and the treatment on the treated differ, and hence we present results for both. Third, the doubly-robust estimator also stems from the semiparametric efficient moment conditions, and hence we obtain the semiparametric efficiency bound, even under heteroskedasticity, under standard additional conditions. Thus, Pötscher's (2009) result that sparse estimators have large confidence sets is also circumvented. Taking all these features together enables us to obtain uniform inference over such a large class of treatment effects models.

In recent independent work, Belloni et al. (2014), propose a similar approach. Their main focus is inference on the linear part of a partially linear model, which motivates an estimator quite different from ours, but it will recover the average treatment effect in the special case of a binary treatment where the effect is constant across observables. However, their Section 5, developed independently from our work, considers heterogeneous effects and proposes an estimator based on the efficient influence function, similar to ours. There are two broad differences. First, we allow for multivalued treatments, which offers a larger set of estimands and can thus enhance the understanding of program impacts.[3] In this context we propose a group lasso based approach that naturally exploits the already-present structure of treatment effects data to improve model selection by pooling information across treatment levels. This is particularly natural in the multivalued case, but even in the binary case there is still a grouped structure in the outcome regressions, though not in treatment assignment (i.e., in propensity score estimation). Second, although in both cases the doubly-robust estimator is used for average treatment effects[4] (following a quite different model selection step), we show that this estimator has two benefits: (i) it may require weaker conditions on the first stage (see Assumption 3); and (ii) it does not require using variables selected for the treatment equation in the outcome model, and vice versa ("post double selection"), and indeed, doing may require additional assumptions (see Assumption 5).

Our analysis is conducted under selection on observables, which has a long tradition and remains quite popular in empirical economics.[5] Covariates play three crucial roles in this framework. First, using more observed covariates as proxies, and more flexibly, may help account for unobserved confounding and hence increase the plausibility of unconfoundedness. Second, some observed variables may not be part of the causal mechanism under study, and should be excluded. Third, the efficient conditioning set are those variables that drive the outcome, not necessarily those important for treatment assignment. This reasoning mandates contradicting goals for practitioners: a large, rich set of controls on the one hand, and parsimony on the other. Our approach is a formal, theory-driven attempt to reconcile this contradiction.

A special feature of our analysis is that we match the empirical realities of large datasets by considering selection from amongst (possibly) more covariates than observations, so-called *high-dimensional* data. The goal of variable selection is to find a small model that is nonetheless sufficiently flexible to capture unknown features of the data-generating process required for inference. If a small model can perfectly capture the unknown feature it is said to be *exactly sparse*. More realistic is *approximate sparsity*, when the bias from using a small model is well-controlled, but nonzero. Sparsity is a natural framework for thinking about model selection. Indeed, any time only a few of the available variables are used, a sparsity assumption has effectively been made. It is common empirical practice to report results from several small models, but for these results to be valid one must assume these specifications give high-quality, sparse representations of the unknown features. The alternative we provide involves selecting a sparse, yet flexible, model from among a large set of variables. Results may then be compared with more traditional methods.

With the aim of mimicking common empirical practice we estimate the propensity score with multinomial logistic regression, coupled with group lasso selection (Yuan and Lin, 2006). Our

[2] Doubly-robust estimation and its role in program evaluation is discussed by Robins and Rotnitzky (1995), van der Laan and Robins (2003), Kang and Schafer (2007, with discussion), Tan (2010), and references therein.

[3] Discussion and applications may be found in, for example Imbens (2000), Lechner (2001), Imai and van Dyk (2004), Abadie (2005), Cattaneo (2010), and Cattaneo and Farrell (2011).

[4] They use different asymptotic variance estimators, and for treatment effects on the treated they do not exploit the simplification discussed in Remark 1.

[5] For other approaches and reviews of the literature, see, e.g., Holland (1986), Hahn (1998), Horowitz and Manski (2000), Chen et al. (2004, 2008), Bang and Robins (2005), Abadie and Imbens (2006), Wooldridge (2007), and references therein.

Table 1
Analysis of NSW demonstration: treatment effects on the treated and confidence intervals for various specifications.

Specifications:	Number of variables		Sample sizes[c]		ATT	95% CI
	Before selection[a]	After selection[b]	Control	Treated		
Experimental Benchmark	–	–	260	185	1794	[110, 3479]
Doubly-Robust Estimates						
Specification 1 (No Selection)	N/A	11	1211	185	1664	[−276, 3604]
DW02 (Informal Selection)	??	15	1058	185	2528	[149, 4908]
Refitting after Group Lasso Selection	171	20/6	1735	185	1737	[33, 3441]

Notes: All analyses use the DW99 subsample and PSID comparison group. Specifications vary, but all estimates and standard errors of from the method defined in Section 5 with the exception of the partially linear model.

[a] Not counting the intercept. The total set of variables considered by DW02 is not known.

[b] For the group lasso estimators, the two numbers given are for those used in the outcome regressions and propensity score, respectively. For other doubly-robust estimators all variables are used in the propensity score and outcome models.

[c] The full sample begins with 2490 comparisons and 185 treated units. Control observations outside the range of estimated propensity scores in the treated sample are discarded.

results are stated in the language of treatment effects, but apply to general data structures and are of independent interest in the high-dimensional literature.[6] Much of the literature has focused on linear models (see Buhlmann and van de Geer (2011) for a survey), while prior studies of nonlinear models often assume exact sparsity or present limited results.[7] Furthermore, these studies often use high-level conditions that can be hard to verify. In contrast, we obtain sharp results for logistic regression under the same simple and intuitive conditions used for linear modeling by exploiting mathematical techniques of self-concordant functions put forth by Bach (2010). We also provide extensions to prior work on linear models needed to apply them in treatment effect estimation.

Finally, we offer numerical evidence on the finite sample performance of our procedure. In a small simulation study we find that our procedure delivers very accurate coverage of confidence intervals even for models where covariate selection is difficult, either because of a low signal-to-noise ratio or lack of sparsity, thus highlighting the uniform validity of inference. We also apply our method to the widely-used National Supported Work Demonstration data (LaLonde, 1986) and find very accurate estimates and tight confidence intervals (see Table 1).

The paper proceeds as follows. Section 2 gives short, self-contained overview. Section 2.3 collects notation. Section 3 describes the treatment effect models. Sparse models are discussed in Section 4, which shows how several commonly used models fit in this framework. Section 5 presents our estimation method and complete results on treatment effect inference. Theoretical results for the group lasso are in Section 6. Section 7 presents the numerical evidence and Section 8 concludes. The main proofs are presented in Appendix, while the remainder are available in a supplement (see Appendix D).

2. Overview of results and notation

Here we give an overview of the paper, including treatment effect inference (Section 2.1), our new results for the group lasso (Section 2.2), and notation used throughout (Section 2.3).

2.1. *Treatment effects and results on post-selection inference*

We consider a multivalued treatment, with status indicated by $D \in \{0, 1, \dots, \mathcal{T}\}$. Interest lies in mean effects of the treatment on a scalar outcome Y. Let $\{Y(t)\}_{t=0}^{\mathcal{T}}$ be the (latent) potential outcomes: $Y(t)$ is the outcome a unit would have under $D = t$ and is only observed for units with $D = t$; that is, $Y = \sum_{t=0}^{\mathcal{T}} \mathbb{1}\{D = t\}Y(t)$. Many interesting parameters combine means of potential outcomes, and having multivalued treatments allows for a wider range of estimands. Define the mean of one potential outcome as $\mu_t = \mathbb{E}[Y(t)]$. To fix ideas, $\mu_1 - \mu_0$ is the average treatment effect in the binary case ($D \in \{0, 1\}$). Sections 3 and 5 consider more general average effects, including effects on treated groups. For simplicity, in this section we focus on a single μ_t.

We use the selection on observables framework to identify μ_t. For a vector of covariates X, define the generalized propensity score and conditional outcome regressions as

$$p_t(x) = \mathbb{P}[D = t|X = x] \quad \text{and} \quad \mu_t(x) = \mathbb{E}[Y|D = t, X = x].$$

For identification it is sufficient to assume that $\mathbb{E}[Y(t)|D, X] = \mathbb{E}[Y(t)|X]$ (mean independence) and $p_t(X)$ is bounded away from zero (overlap) for all treatment levels. Broadly, these two assumptions imply that units from one treatment group are good proxies for other treatments and that there are always such proxies available (see Section 3).

For an i.i.d. sample $\{(y_i, d_i, x_i')\}_{i=1}^n$ and model-selection-based estimators $\hat{p}_t(x_i)$ and $\hat{\mu}_t(x_i)$, we estimate μ_t with

$$\hat{\mu}_t = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{1}\{d_i = t\}(y_i - \hat{\mu}_t(x_i))}{\hat{p}_t(x_i)} + \hat{\mu}_t(x_i)\right\}.$$

This doubly-robust estimator combines regression imputation and inverse probability weighting, and remains consistent if either the model $p_t(x)$ or $\mu_t(x)$ is misspecified. Following widespread empirical practice, we estimate $\hat{p}_t(x_i)$ with multinomial logistic regression and $\hat{\mu}_t(x_i)$ linearly (see Section 6). The choice of covariates in $\hat{p}_t(x_i)$ and $\hat{\mu}_t(x_i)$ impacts consistency, efficiency, and finite sample performance. Covariate selection based on ad hoc, iterative searches is common in empirical work, but is not formal, objective, or replicable. Balancing tests are also common, but have the additional drawback of assuming the same covariates are important for outcomes and treatment assignment, and more generally do not weight the covariates by their importance for bias.

On the other hand, our proposed procedure gives practitioners an easy to implement, fully objective tool to perform data-driven covariate selection and treatment effect inference, with replicable results.[8] Importantly, we do not preclude the addition of variables

[6] Our techniques build on prior studies, in particular Bickel et al. (2009), Lounici et al. (2011), Obozinski et al. (2011), Belloni and Chernozhukov (2011), Belloni et al. (2012), Belloni and Chernozhukov (2013), and Belloni et al. (2014).

[7] Examples include van de Geer (2008) and Negahban et al. (2012), whose bounds do not imply our results. Bach (2010) only gives an error bound on coefficients in exactly sparse logistic regression, which cannot yield our results; and does not consider prediction error or post-selection estimation. In independent work, Kwemou (2012) and Belloni et al. (2013) also apply Bach's (2010) tools, but are focused on different goals. Vincent and Hansen (2014) apply the group lasso to multinomial logistic regression, but do not derive any theoretical results.

[8] For the final step, the doubly-robust estimator is available in STATA and the package of Cattaneo et al. (2013). The covariate selection stage is easily implemented in R.

known to be important from economic theory or prior knowledge. Our procedure is intended to supplement these variables with a flexible set of controls, guarding against misspecification or overfitting.

The following theorem is an example of the more general results presented in Section 5.2, wherein we also define V_t and $\hat{V}_t$.

Theorem 1. *Consider a sequence $\{P_n\}$ of data-generating processes that obey, for each n, Assumptions 1 and 2. If the first stage obeys*

(i) $\sum_{i=1}^n (\hat{p}_t(x_i) - p_t(x_i))^2/n = o_{P_n}(1)$ *and* $\sum_{i=1}^n (\hat{\mu}_t(x_i) - \mu_t(x_i))^2/n = o_{P_n}(1)$ *and*

(ii) $\left[\sum_{i=1}^n \mathbb{1}\{d_i = t\}(\hat{p}_t(x_i) - p_t(x_i))^2/n\right]^{1/2}\left[\sum_{i=1}^n \mathbb{1}\{d_i = t\}(\hat{\mu}_t(x_i) - \mu_t(x_i))^2/n\right]^{1/2} = o_{P_n}(n^{-1/2})$,

then $\sqrt{n}(\hat{\mu}_t - \mu_t) \to_d N(0, V_t)$ and $\hat{V}_t/V_t \to_{P_n} 1$. For each n, let $\boldsymbol{P}_n$ be the set of data-generating processes obeying Assumptions 1 and 2 and (i) *and* (ii) *above. Then for $c_\alpha = \Phi^{-1}(1-\alpha/2)$*

$$\sup_{P \in \boldsymbol{P}_n} \left| \mathbb{P}_P \left[\mu_t \in \left\{ \hat{\mu}_t \pm c_\alpha \sqrt{\hat{V}_t/n} \right\} \right] - (1-\alpha) \right| \to 0.$$

This result establishes the uniform validity of an asymptotic confidence interval for μ_t, overcoming all the post model selection inference challenges: robustness to model selection errors, selecting a model that is small but flexible enough to capture the features of the underlying data generating process, and still retaining efficiency under standard conditions (see Section 5.3). Intuitively, this is similar to (but distinct from) overcoming pretesting bias in other contexts. Also, although our discussion is in terms of covariate selection in high-dimensional, sparse models, the inference result is generic for any first stage estimator.

The two conditions placed on the first stage are analogous to the commonly-used, high-level requirement in semiparametrics that first stage components converge faster than $n^{-1/4}$. However exploiting features of the doubly-robust estimator yields weaker conditions. The first is a mild consistency requirement. The second requires a rate on the product of errors and is thus easier to satisfy if one function is easier to estimate, e.g. more smooth or more sparse. In model selection, the rates for the first stage depend on the sample size, the number of covariates considered, and the sparsity level. Importantly, the rate will depend on the total number of covariates only logarithmically, allowing for a large number. We propose to use the group lasso and prove that these estimators satisfy (i) and (ii).

2.2. Model selection stage

We propose refitting following group lasso selection, and show that it meets all requirements on the model selector. The group lasso is well-suited to program evaluation applications because covariates are penalized according to their overall contribution in all treatment groups. This has two consequences. First, information from all treatments is pooled when doing selection, and hence a weaker signal may be extracted, which improves the selection properties. Second, the selected variables are common to all treatment levels. From a practical point of view this is desirable, as interest rarely lies in a single μ_t, but rather a collection, and substantial commonality is expected in the variables important for different treatment levels.

We consider high-dimensional, sparse models for $p_t(x)$ and $\mu_t(x)$. These are defined by a p-dimensional vector X^* based on the original variables X. The X^* may consist of any combination of the original variables, interactions, flexible parametric transformations, and/or nonparametric series terms (such as splines or polynomials). A model is approximately sparse if there are $s < n$ of these terms that yield a good approximation ($s \to \infty$ is allowed). To build intuition, suppose that $\mu_t(x)$ obeys a p-dimensional linear model. Then the sparsity assumption is that there is an s-dimensional submodel with sufficiently small specification bias. In the nonparametric case, sparsity is weaker than (but analogous to) the familiar assumption that a small set of basis functions can approximate the unknown objects well. In practice researchers employ a hybrid of these approaches, which is covered by our results. Section 4 gives more details and examples.

We form $\hat{p}_t(x)$ and $\hat{\mu}_t(x)$ in two steps (complete details in Section 6). First, the group lasso is applied separately to multinomial logistic and least squares regression to select covariates from X^*. We then estimate $p_t(x)$ and $\mu_t(x)$ by refitting unpenalized models using the selected variables, possibly augmented with controls suggested by prior work or economic theory. It is not desirable for a model selector to discard theory and prior work, and our procedure explicitly avoids this. We also allow for using logistic-selected variables in the linear model refitting and vice versa, but this is not necessary for uniformity nor efficiency.

Our main results give precise bounds for the number of covariates selected and the estimation error, both for the penalized and unpenalized estimates. Section 6 results gives nonasymptotic bounds, with exact constants. Such results are complex and so we give the following intuitive, asymptotic result. (The notation O_{P_n} is defined in Section 2.3.)

Corollary 1. *Suppose the biases from the best s_d- and s_y-term approximations to $p_t(x)$ and $\mu_t(x)$ are order $\sqrt{s_d/n}$ and $\sqrt{s_y/n}$, respectively. Then under the assumptions in Section 6, and $\delta > 0$ described therein, with high probability we have:*

1. $\sum_{i=1}^n (\hat{p}_t(x_i) - p_t(x_i))^2/n = O_{P_n}\left(n^{-1} s_d \log(p \vee n)^{3/2+\delta}\right)$ *and*
2. $\sum_{i=1}^n (\hat{\mu}_t(x_i) - \mu_t(x_i))^2/n = O_{P_n}\left(n^{-1} s_y \log(p \vee n)^{3/2+\delta}\right)$.

These two results for our proposed group lasso estimators can be directly used to verify the high-level conditions in Theorem 1. Specifically, if $s_d s_y \log(p)^{3+2\delta} = o(n)$, conditions (i) and (ii) of Theorem 1 are met (requiring $s^2 = o(n)$, up to log factors, as found in other results in the literature). Further, it is clear how the doubly-robust estimator can help: if one function is more smooth or more sparse, s_d or s_y will be lower, easing the restriction. Section 6.3 gives further results: showing that the number of variables selected is the same order as the sparsity level, and provides bounds on the logistic and linear coefficients directly. Both these results are important for certain steps in treatment effect estimation that are not reflected in the simple statement of Theorem 1. These results appear to be entirely new for the multinomial logistic regression, for any version of the lasso. From a practical point of view, these results provide formal justification for using multinomial logistic regression, coupled with group lasso selection and post-selection refitting.

2.3. Notation

We collect here notation to be used for the rest of the paper. The data generating process (DGP) is denoted by P_n and is defined by the joint law of the random variables $(Y, D, X')'$. For a given n, $\{(y_i, d_i, x_i')'\}_{i=1}^n$ constitute draws from P_n. The DGP may vary with n, along with features such as parameters, distributions, and so forth, as discussed in Section 4.2. This is generally suppressed for clarity. We adopt the following conventions.

Treatments. Define the treatment sets $\overline{\mathbb{N}}_{\mathcal{T}} = \{0, 1, 2, \ldots, \mathcal{T}\}$ and $\mathbb{N}_{\mathcal{T}} = \{1, 2, \ldots, \mathcal{T}\}$. No order is assumed in the treatments. For each unit i, d_i indicates treatment assignment, and define $d_i^t = \mathbb{1}\{d_i = t\}$. Let $n_t = \sum_{i=1}^n d_i^t$ be the number of individuals with treatment t and define $\underline{n} = \min_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} n_t$ and $\overline{n} = \max_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} n_t$. Further define $\overline{\mathcal{T}} = \mathcal{T} + 1$.

Vectors. Define $\mathbb{N}_p = \{1, 2, \ldots, p\}$. For a doubly-indexed collection of scalars $\{\delta_{t,j} : t \in \overline{\mathbb{N}}_{\mathcal{T}}, j \in \mathbb{N}_p\}$, define $\delta_{\cdot,j} \in \mathbb{R}^{\overline{\mathcal{T}}}$ as the vector that collects over all t for fixed j; $\delta_{t,\cdot} \in \mathbb{R}^p$ collects over $j \in \mathbb{N}_p$ for fixed t; and $\delta_{\cdot,\cdot} \in \mathbb{R}^{p\times\overline{\mathcal{T}}}$ the concatenation of all $\delta_{t,\cdot}$. For simplicity, we write δ_t for $\delta_{t,\cdot}$. When considering the multinomial logistic model, t will vary only over $\mathbb{N}_{\mathcal{T}}$ but the notation will be maintained. For a set $S \subset \mathbb{N}_p$, let $\delta_{t,S} \in \mathbb{R}^{\text{card}(S)}$ be the vector of $\{\delta_{t,j} : j \in S\}$ for fixed t and similarly let $\delta_{\cdot,S} \in \mathbb{R}^{|S|\times\overline{\mathcal{T}}} = \{\delta_{t,j} : t \in \overline{\mathbb{N}}_{\mathcal{T}}, j \in S\}$.

Norms. Single bars will be either absolute value or cardinality of a set, and will be clear from the context. For a vector v, let $\|v\|_1$ and $\|v\|_2$ denote the ℓ_1 and ℓ_2 norms, respectively. For the group lasso, define the mixed ℓ_2/ℓ_1 norm as $|||\delta_{\cdot,\cdot}|||_{2,1} = \sum_{j\in\mathbb{N}_p} \|\delta_{\cdot,j}\|_2$. It will always be the case that the ("outer") ℓ_1 norm is over the covariates and the ("inner") ℓ_2 norm is over the treatments (in our application). When discussing the multinomial logistic model, treatments will be restricted to $\mathbb{N}_{\mathcal{T}}$ with no change in notation.

Data-Generating Processes. The set of all P_n considered is $\boldsymbol{P}_n$. For sequences, $\{P_n\} = \{P_n : n \geq 1, P_n \in \boldsymbol{P}_n\}$. Expectations and probabilities are taken against P_n, though notationally suppressed. For asymptotic arguments dependence on n is explicit, so that $O_{P_n}(\cdot)$ and $o_{P_n}(\cdot)$ have their usual meaning with the understanding that the measure P_n is used for each n.

For a set of scalars $\{m_t\}_{t=1}^{\mathcal{T}}$, let $\hat{p}_t(\{m_t\}_{\mathbb{N}_{\mathcal{T}}}) = \exp(m_t)[1 + \sum_{t\in\mathbb{N}_{\mathcal{T}}} \exp(m_t)]^{-1}$ denote the multinomial logit function. Empirical expectation will be denoted $\mathbb{E}_n[w_i] = \sum_{i=1}^n w_i/n$ and $\mathbb{E}_{n,t}[w_i] = \sum_{i\in\mathbb{I}_t} w_i/n_t = \sum_{i=1}^n d_i^t w_i/n_t$.

3. Treatment effects model

In this section we formally define the treatment effects model and the parameters of interest. Recall that $D \in \{0, 1, \ldots, \mathcal{T}\}$ indicates treatment status, $\{Y(t)\}_{t\in\overline{\mathbb{N}}_{\mathcal{T}}}$ are the (latent) potential outcomes, and $Y(t)$ is only observed for units with $D = t$; that is, $Y = \sum_{t\in\overline{\mathbb{N}}_{\mathcal{T}}} Y(t)$. The building blocks of many general estimands are the averages

$$\mu_t = \mathbb{E}[Y(t)], \quad t \in \overline{\mathbb{N}}_{\mathcal{T}}, \quad \text{and} \quad \mu_{t,t'} = \mathbb{E}[Y(t)|D = t'],$$
$$t, t' \in \overline{\mathbb{N}}_{\mathcal{T}} \times \overline{\mathbb{N}}_{\mathcal{T}}. \tag{1}$$

In the binary case, the average treatment effect is $\mu_1 - \mu_0$ and the treatment on the treated is $\mu_{1,1} - \mu_{0,1}$. A multivalued treatment allows for a large range of interesting estimands. To fix ideas, we keep as running examples two leading cases. First, the so-called dose–response function: the $(\mathcal{T} + 1)$-vector $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_{\mathcal{T}})'$. Second, define $\boldsymbol{\tau}$ as the $\mathcal{T}$-vector with element t given by $\mu_{t,t} - \mu_{0,t}$. This gives the effect of each treatment relative to the baseline $t = 0$, only for those who received that treatment. These are by no means the only interesting estimands constructed from μ_t and $\mu_{t,t'}$; many others are given by Lechner (2001), Heckman and Vytlacil (2007), and others.

The following two conditions are sufficient to identify μ_t and $\mu_{t,t'}$.

Assumption 1 (*Identification*)**.** For all $t \in \overline{\mathbb{N}}_{\mathcal{T}}$ and almost surely X, P_n obeys:

(a) (Mean independence) $\mathbb{E}[Y(t)|D, X = x] = \mathbb{E}[Y(t)|X = x]$, and
(b) (Overlap) $\mathbb{P}[D = t|X = x] \geq p_{\min} > 0$ for all $t \in \overline{\mathbb{N}}_{\mathcal{T}}$.

This assumption is a form of "ignorability" coined by Rosenbaum and Rubin (1983). This model allows arbitrary treatment effect heterogeneity in observables, but not unobservables. This assumption is standard in the program evaluation literature, and its plausibility has been discussed at length, so we omit a general discussion (see, e.g., Imbens (2004), Wooldridge (2010, Chapter 21), and references therein). However, in the context of model selection, three remarks are warranted.

First, in place of Assumption 1(a), it is more common to instead assume full conditional independence: $Y \perp\!\!\!\perp D|X$. However, as observed by Heckman et al. (1997), the weaker mean independence is sufficient. For our purposes, the "gap" between the two assumptions is important. Suppose full independence holds only conditional on a set of variables strictly larger than the variables entering the mean functions (e.g. the excess variables affect higher moments). In this case, because mean independence is still sufficient, we need not aim to select the larger set. Full independence is important for the efficiency discussed in Section 5.3.

Second, the covariates may, in general, include instruments for treatment status, but they are not known as such. This is standard, but left implicit, in discussions of ignorability. If instruments are present, and selected for estimation, efficiency suffers but unbiasedness is not harmed. Efficiency bounds in this context typically (implicitly) assume there are no instruments in X. Assumption 1(b) rules out perfect predictors. Section 5.3 offers further discussion.

Finally, the main drawback of Assumption 1(a) is that it does not give identification of average effects on transformations of $Y(t)$. However, we are expressly interested in model selection on the mean function of the level of $Y(t)$, and hence Assumption 1(a) is more natural. To operationalize model selection, structure must be placed on $\mathbb{E}[Y(t)|X = x]$, and hence functional form conditions tied to mean independence are not limiting per se. If the parameter of interest is changed, say to $\mathbb{E}[\log(Y(t))]$, and a sparsity assumption is made for $\mathbb{E}[\log(Y(t))|X = x]$, then our method applies.

Assumption 1 yields identification of μ_t and $\mu_{t,t'}$ using either inverse weighting or regression, and double robustness follows from combining the two strategies. Recall the notation $p_t(x) = \mathbb{P}[D = t|X = x]$ and $\mu_t(x) = \mathbb{E}[Y|D = t, X = x]$. Applying Assumption 1 we find that

$$\mathbb{E}\big[\psi_t\big(Y, D, \mu_t(X), p_t(X), \mu_t\big)\big]$$
$$= \mathbb{E}\left[\frac{\mathbb{1}\{D = t\}Y}{p_t(X)} + \mu_t(X) - \frac{\mathbb{1}\{D = t\}\mu_t(X)}{p_t(X)} - \mu_t\right] = 0 \tag{2}$$

and

$$\mathbb{E}\big[\psi_{t,t'}\big(Y, D, \mu_t(X), p_t(X), p_{t'}(X), \mu_{t,t'}\big)\big]$$
$$= \mathbb{E}\left[\frac{\mathbb{1}\{D = t'\}\mu_t(X)}{p_{t'}} + \frac{p_{t'}(X)}{p_{t'}}\frac{\mathbb{1}\{D = t\}(Y - \mu_t(X))}{p_t(X)} - \mu_{t,t'}\right]$$
$$= 0, \tag{3}$$

where $p_t = \mathbb{P}[D = t]$. The moment condition (2) holds if either $p_t(x)$ or $\mu_t(x)$ is misspecified. For $\mu_{t,t'}$, if $\mu_t(x)$ is misspecified, both $p_t(X)$ and $p_{t'}(X)$ must be correctly specified, while if $\mu_t(x)$ is correct, both propensity scores may be misspecified. It is important to note that the forms of $\psi_t(\cdot)$ and $\psi_{t,t'}(\cdot)$ are fixed, so the function itself does not depend on the sample size even if its arguments do. Our estimator is a plug-in version of this moment condition.

Remark 1 (*Simplifications for $\mu_{t,t}$*)**.** Identification of $\mu_{t,t}$ does not require Assumption 1. $Y(t)$ is fully observed for the sub-population of interest and so a simple average will deliver $\mu_{t,t} = \mathbb{E}[\mathbb{1}\{D = t\}Y]/p_t$. Note that (3) reduces to this when $t = t'$. For $\boldsymbol{\tau}$ this means we must only estimate the function $\mu_t(x_i)$ for $t = 0$. Intuitively, we must use comparison group observations to proxy for treated units, but not the other way around. Thus, for certain parameters of interest, Assumption 1 can be weakened to hold only for the comparison group. However, we cover generic estimands, without necessarily specifying a comparison group, and so we maintain Assumption 1 for simplicity, rather than keeping track of hosts of special cases. ■

Remark 2 (*Efficient Influence Functions*)**.** The efficient influence functions in this model are exactly $\psi_t(\cdot)$ and $\psi_{t,t'}(\cdot)$, and so our estimators have the interpretation of being plug-in versions of these, and indeed, will be asymptotically linear with this influence function (see Section 5.3). ■

4. Approximately sparse models

We now formalize approximate sparsity. Let X_Y^* and X_D^* be p-dimensional transformations of the covariates X, with $p > n$ allowed. These transformations are specific to the outcome and treatment models, but may overlap. They do not vary with t, nor depend on the DGP. Some examples are given below in Section 4.1. For the multinomial logistic model it is convenient to work with the log-odds ratio. We take $p_0(x) = 1 - \sum_{t \in \mathbb{N}_{\mathcal{T}}} p_t(x)$ and write

$$\log\left(\frac{p_t(x)}{p_0(x)}\right) = x_D^{*\prime}\gamma_t^* + B_t^D, \quad t \in \mathbb{N}_{\mathcal{T}}. \tag{4}$$

Similarly, write the outcome regressions as

$$\mu_t(x) = x_Y^{*\prime}\beta_t^* + B_t^Y, \quad t \in \overline{\mathbb{N}}_{\mathcal{T}}. \tag{5}$$

The terms $B_t^D = B_t^D(x)$ and $B_t^Y = B_t^Y(x)$ are bias terms arising from the parametric specification. As discussed below, these encompass the usual nonparametric bias as well. Approximate sparsity requires that only a small number of the X^* are needed to make the bias small. Define $S_*^D = \bigcup_{\mathbb{N}_{\mathcal{T}}} \operatorname{supp}(\gamma_t^*)$ and $S_*^Y = \bigcup_{\overline{\mathbb{N}}_{\mathcal{T}}} \operatorname{supp}(\beta_t^*)$, so that these sets capture all variables important for treatment and outcomes, respectively. We assume that there are some $s_d < n$ and $s_y < n$, such that for $|S_*^D| = s_d$ and $|S_*^Y| = s_y$, the biases B_t^D and B_t^Y are sufficiently small. This is made precise by defining the bounds:

$$\begin{aligned}&\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) - p_t(x_i))^2]^{1/2} \le b_s^d \quad \text{and} \\ &\mathbb{E}_n[B_t^Y(x_i)^2]^{1/2} \vee \mathbb{E}_{n,t}[B_t^Y(x_i)^2]^{1/2} \le b_s^y.\end{aligned} \tag{6}$$

Note that the former bias bound is placed directly on the propensity score because it is the ultimate object of interest, rather than on the linearization of the log-odds.

While a great deal of overlap is expected, in practice it is likely that a few covariates will be more or less important for different treatments, and so we do not require that the supports of $\gamma_t^*, t \in \mathbb{N}_{\mathcal{T}}$ or $\beta_t^*, t \in \overline{\mathbb{N}}_{\mathcal{T}}$ are constant over t, nor that S_*^D overlaps with S_*^Y. Instead, it may be better to think of $\mathbb{N}_p \setminus S_*^D$ and $\mathbb{N}_p \setminus S_*^Y$ as the "common nonsupports" of the treatment and outcome equations. When it is clear from the context we will abbreviate both X_D^* and X_Y^* by X^* (and their realizations by x_i^*) and refer to them generically as "covariates", and further write s for either s_d or s_y. We assume $\mathbb{E}_n[(x_{i,j}^*)^2] = 1$ without loss of generality (see Remark 4).

4.1. Parametric and nonparametric examples

To concretize the sparse model idea, we now discuss how several models commonly used in practice fit into this framework. These include parametric and nonparametric models for $p_t(x)$ and $\mu_t(x)$, and hybrids of these. A common theme to all examples will be comparison to the *oracle* model: the model that knows the true support in advance. Our uniform inference results include all these examples as special cases because, loosely speaking, we obtain uniformity over DGPs where $p_t(x)$ and $\mu_t(x)$ have sparse representations. We aim for an accessible discussion of each model, and defer technicalities to the literature (Raskutti et al., 2010; Rudelson and Zhou, 2013; Belloni et al., 2014).

Example 1 (*Oracle Parametric Model*)**.** Assume models (4) and (5) hold with $B_t^D = B_t^Y = 0$ and $X_D^* = X_Y^* = X$. Let $p = s = \dim(X)$. All covariates are used in all modeling. If dimension is fixed this is the textbook parametric model, see for example Wooldridge (2010). Alternatively, the dimension can be diverging, but more slowly than n. We are not aware of any work which covers this case explicitly, though for the first stage, He and Shao (2000) cover linear and logistic regression, and their results easily extend to multinomial logistic models.

The vast majority of treatment effect studies adopt this model (with dimension fixed), taking the set of covariates as given. In our framework, this is equivalent to the researcher having prior knowledge of which covariates are important and which are not. Such knowledge no doubt plays an important role, but it cannot cover all situations or all variables. Furthermore, as more data become available, the researcher does not increase the complexity of their model. ■

Example 2 (*Exactly Sparse Parametric Model*)**.** Retain the exact parametric structure of the prior example, but let $\dim(X) = p$ be possibly larger than n, and assume that S_*^Y and S_*^D are unknown sets of cardinality less than n. Model selection must be performed. Often, researchers (implicitly) rely on the *oracle property*, that S_*^Y and S_*^D can be found with probability approaching one, and conduct inference conditioning on this event. This approach cannot be made uniformly valid and has poor finite sample properties, as shown by Leeb and Pötscher (2005, 2008a,b) and Pötscher and Leeb (2009). ■

Example 3 (*Approximately Sparse Parametric Model*)**.** Again suppose a purely parametric model, so that $X_D^* = X_Y^* = X$ and $\dim(X) = p$, possibly greater than n. Suppose that there exist coefficients $\gamma_{\cdot,\cdot}^0$ and $\beta_{\cdot,\cdot}^0$ such that $\log[p_t(x)/p_0(x)] = x_D^{*\prime}\gamma_t^0$ and $\mu_t(x) = x'\beta_t^0$ exactly, but instead of any coefficients being precisely zero, suppose they may be ordered such that $|\gamma_{t,j}^0| \propto j^{-\alpha_\gamma}$ and $|\beta_{t,j}^0| \propto j^{-\alpha_\beta}$, with α_γ and α_γ at least one. Then, there exist s_d and s_y that are $o(n)$ such that Eqs. (4) and (5), and other conditions needed, are satisfied for $\gamma_{t,j}^* = \gamma_{t,j}^0$ for $j \le s_d$ and $\beta_{t,j}^* = \beta_{t,j}^0$ for $j \le s_y$ and the rest truncated to zero. That is S_*^D and S_*^Y collect the largest coefficients and $B_t^D = \sum_{\mathbb{N}_p \setminus S_*^D} x_j \gamma_{t,j}^0$, and similarly for B_t^Y. ■

Example 4 (*Semiparametric Model*)**.** Assume $p_t(x)$ and $\mu_t(x)$ are unknown functions that can be well-approximated by a linear combination of s_d and s_y basis functions, respectively (e.g. are sufficiently smooth). In (4) and (5), $\gamma_{\cdot,\cdot}^*$ and $\beta_{\cdot,\cdot}^*$ are the coefficients of these approximations, while B_t^D and B_t^Y are the usual nonparametric biases. $X_D^* = R_D(X)$ and $X_Y^* = R_Y(X)$ are series terms used in the approximation. Standard semiparametric analyses, such as Hirano et al. (2003), Imbens et al. (2007), or Cattaneo (2010), can be viewed in this context as oracle models that know in advance which terms yield the best approximation, typically assumed to be the first terms. Instead, we only require that some s_d (or s_y) of a set of p series terms give good approximations. This allows for greater flexibility in applications, where there is no knowledge of which series terms to use, and the researcher may want to mix terms from different bases. ■

Example 5 (*Mixed Parametric and Semiparametric Model*)**.** Partition $X = (X_1, X_2)$. Suppose that the true log-odds function satisfies $\log[p_t(x)/p_0(x)] = x_1'\gamma_t^1 + h_t(x_2) + B_t^1(x)$, where $B_t^1(x)$ is a specification bias and $h_t(\cdot)$ is a smooth unknown function. For a set of basis functions $R_D(x_2)$, there will exist coefficients γ_t^2 such that $h_t(x_2) = R_D(x_2)'\gamma_t^2 + B_t^2(x_2)$ and so

$$\log\left(\frac{p_t(x)}{p_0(x)}\right) = x_D^{*\prime}\gamma_t^* + B_t^D, \quad x_D^* = (x_1', R_D(x_2)')',$$
$$\gamma_t^* = (\gamma_t^{1\prime}, \gamma_t^{2\prime})', \text{ and } B_t^D = B_t^1 + B_t^2.$$

We require that some collection of variables and series terms give a good, sparse approximation, without placing explicit conditions on how many of either. Implicitly, one will restrict the other. For example, if the dimension of the parametric part is large, then we require that $h_t(\cdot)$ can be more easily approximated. We treat $\mu_t(x)$ the same. This example is closest to actual practice, where some variables (e.g. dummies) enter in a known way and should not be considered part of a nonparametric object, while other covariates must be considered flexibly. ■

It is important to note that misspecification of the type guarded against by double robustness can arise in any type of model. In parametric cases, this is most often functional form misspecification. While this type of misspecification does not occur in nonparametrics, others are possible, such as shape restrictions or separability assumptions being incorrect, or omitting relevant variables. None of these errors disappear asymptotically, and all of them are guarded against by use of the doubly-robust estimator.

4.2. *Conceptual considerations in n-varying DGPs*

Much of the DGP, including parameters and distributions, is allowed to depend on n. Perhaps the most salient features that do not depend on n are the set of treatments and the functions ψ_t and $\psi_{t,t'}$. It is likely that our results can be extended to accommodate a growing number of treatments, but that is beyond the scope of our study. In the models (4) and (5), X^*, $\gamma^*_{\cdot,\cdot}$, and $\beta^*_{\cdot,\cdot}$ must depend on n by construction. Our results on estimation of these models are nonasymptotic. For treatment effect inference, we use triangular array asymptotics to retain the dependence on n of the DGP. The interpretation of the results does, and should, change depending on what is assumed about the DGP. To illustrate, let us return to Examples 2 and 4.

First, consider the simple parametric models of Example 2. In this case, $\mu_t = \mathbb{E}[\mathbb{E}[Y(t)|X]] = \mathbb{E}[X']\beta_t^*$, which depends on n by construction, as the dimension is diverging. It may seem unnatural that the parameter to be estimated depends on n, as we typically think of "true" parameters being features of a (large) fixed study population. However, with a diverging number of covariates, there is no fixed DGP. Indeed, if we estimate $\mu_t = \mu_t^{(n_1)}$ based upon n_1 observations, and then proceed to gather n_2 *more* observations, when we re-estimate our target is now $\mu_t^{(n_1+n_2)} \neq \mu_t^{(n_1)}$. One possible resolution is as follows. First, the parameter of interest is $\mu_t^{(\infty)} = \mathbb{E}[Y(t)]$, which is defined without reference to covariates. We can view each successive n-dependent μ_t as an approximation of $\mu_t^{(\infty)}$ based upon $p = p_n$ covariates. Note well that in our thought experiment, $p_{n_1} \neq p_{n_1+n_2}$, and so additional variables should have been collected for all $n_1 + n_2$ samples.

Contrast this with the semiparametric model in Example 4. It is common to assume the population DGP is fixed over n. The treatment effects may be constructed in terms of the underlying variables, e.g. $\mu_t^{(\infty)} = \mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y(t)|X]]$, with X^* serving only the purpose of aiding in approximating the regression functions. Model selection is performed on series terms, not underlying variables, to estimate the coefficients $\gamma^*_{\cdot,\cdot}$ and $\beta^*_{\cdot,\cdot}$. If $\mu_t = \mathbb{E}[X_Y^{*\prime}]\beta_t^* + \mathbb{E}[B_t^Y]$ does not depend on n, the bias term, by definition, exactly compensates for the n-dependence in $\mathbb{E}[X_Y^{*\prime}]\beta_t^*$. We emphasize that our inference results allow for general n-dependence in the DGP, and interpretation by the econometrician must take careful account of any conceptual assumptions.

5. Main results on treatment effect estimation and inference

In this section we present results on uniformly valid treatment effect inference. We first present the estimators and conditions required for a generic first stage to yield uniform inference. Although our focus is on model selection and sparsity, our results are more general, showcasing the benefits of doubly robust estimation for any model in Section 4 where Assumption 3 (which does not refer to selection or sparsity) can be satisfied.

5.1. *Estimation procedure with a generic model selector*

The moment functions $\psi_t(\cdot)$ and $\psi_{t,t'}(\cdot)$ of Eqs. (2) and (3) have fixed and known form, and so for estimators $\hat{p}_t(x)$ and $\hat{\mu}_t(x)$, we can define

$$\hat{\mu}_t = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{d_i^t(y_i - \hat{\mu}_t(x_i))}{\hat{p}_t(x_i)} + \hat{\mu}_t(x_i)\right\} \tag{7}$$

and

$$\hat{\mu}_{t,t'} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{d_i^{t'}\hat{\mu}_t(x_i)}{\hat{p}_{t'}} + \frac{\hat{p}_{t'}(x_i)}{\hat{p}_{t'}}\frac{d_i^t(y_i - \hat{\mu}_t(x_i))}{\hat{p}_t(x_i)}\right\}, \tag{8}$$

where $\hat{p}_t = n_t/n$. By combining these estimators appropriately we can construct estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\tau}}$ for the dose–response function $\boldsymbol{\mu}$ and the vector $\boldsymbol{\tau}$, respectively, and any other estimand. Notice that when $t = t'\hat{\mu}_{t,t}$ is an average over the appropriate subpopulation: $\hat{\mu}_{t,t} = \mathbb{E}_{n,t}[y_i]$.

Although in this section we allow for generic estimates $\hat{p}_t(x)$ and $\hat{\mu}_t(x)$, it is important to distinguish between estimates based upon selected sets that have no "additional randomness" and those that do. Model selection based estimation will naturally have two steps: first data-driven selection and then refitting to ameliorate the shrinkage bias and allow the researcher to augment the selected variables. Let $\tilde{S}^D$ and $\tilde{S}^Y$ be the selected sets and $\hat{S}^D$ and $\hat{S}^Y$ be the final sets of variables used in the refitting. We will say that these contain no "additional randomness" if the added variables (i.e. $\hat{S} \setminus \tilde{S}$, for Y or D) are nonrandomly selected, such as from economic theory or prior knowledge. On the other hand, the added variables may be selected from a random process beyond that included in $\tilde{S}$. The leading example would be using logistic-selected variables in the regressions or vice versa. Then the variables used in $\hat{\mu}_t(x_i)$ depend not only on the randomness of $\tilde{S}^Y$, but also on that of $\tilde{S}^D$, and hence on $\{d_i\}_{i=1}^n$. Additional conditions are required for the estimators with additional randomness.

The choice of method is in part dependent on the assumptions of the underlying model. To illustrate, first, return to Example 2, where we have a purely parametric model with $X = X_D^* = X_Y^*$. The researcher may want to set $\hat{S}^D \supset \tilde{S}^D \cup \tilde{S}^Y$, in order to have a better chance that $S_*^Y \subset \hat{S}^D$. The set $\hat{S}^D$ now contains additional randomness due to $\tilde{S}^Y$. Conversely, consider Example 4. It is natural to include "low-order" basis functions for each underlying covariate, say linear and quadratic polynomials. Thus, the researcher may want to include these in $\hat{S}$, whether or not selected by the group lasso. However, there is no reason that the series terms useful for approximating the functions $\mu_t(x)$ would be useful for $p_t(x)$, or vice versa, and no additional randomness is injected.

We now state the sufficient conditions used for treatment effect estimation and inference. For exposition, we present these in three groups: those concerning the underlying DGP, requirements of $\hat{p}_t(x)$ and $\hat{\mu}_t(x)$ in the "no additional randomness" case, and finally the additional conditions to allow for "additionally random" selected sets. Begin with conditions on the DGP. Let $U \equiv Y(t) - \mu_t(X)$ and impose the following conditions.

Assumption 2 (*Data Generating Process*)**.** P_n obeys the following, with bounds uniform in n.

(a) $\{(y_i, d_i, x_i')'\}_{i=1}^n$ is an i.i.d. sample from $(Y, D, X')'$.
(b) The covariates X^* have bounded support, with $\max_{j\in\mathbb{N}_p}|X_j^*| \le \mathcal{X} < \infty$. Transformations may depend on n but not the underlying data generating process.
(c) $\mathbb{E}[|U|^4 \mid X] \le \mathcal{U}^4$.
(d) $\min_{j\in\mathbb{N}_p,\ t\in\bar{\mathbb{N}}_{\mathcal{T}}} \mathbb{E}[X_j^{*2}U^2] \wedge \mathbb{E}[X_j^{*2}(\mathbb{1}\{D=t\} - p_t(X))^2]$ is bounded away from zero.
(e) For some $r > 0$: $\mathbb{E}[|\mu_t(x_i)\mu_{t'}(x_i)|^{1+r}]$ and $\mathbb{E}[|u_i|^{4+r}]$ are bounded.

These conditions are mild and intuitive, and not unique to high-dimensional models or model selection. Assumption 2(a) restricts attention to cross-sectional applications. The condition of bounded covariates is unlikely to be a limitation in practice. Any X^* that are underlying variables will naturally be bounded in applications. This condition is automatically satisfied for most common choices of basis functions employed in nonparametric estimation. The rest are moment conditions on the potential outcome models, including allowing the errors to be heteroskedastic and non-Gaussian. The uniform bounds in n are needed for array asymptotics.

We now give precise conditions on the model selector sufficient for uniformly valid inference.

Assumption 3 (*First Stage Restrictions*)**.** The estimators $\hat{p}_t(x)$ and $\hat{\mu}_t(x)$ obey the following for a sequence $\{P_n\}$, uniformly in $t \in \bar{\mathbb{N}}_{\mathcal{T}}$.

(a) $\mathbb{E}_n[(\hat{p}_t(x_i) - p_t(x_i))^2] = o_{P_n}(1)$ and $\mathbb{E}_n\left[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2\right] = o_{P_n}(1)$.
(b) $\mathbb{E}_n[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2]^{1/2}\mathbb{E}_n[(\hat{p}_t(x_i) - p_t(x_i))^2]^{1/2} = o_{P_n}(n^{-1/2})$.

These two collectively play the same role as the commonly-used, high-level requirement in semiparametrics that each first-step component separately converges at $n^{-1/4}$ at least.[9] Indeed, Belloni et al. (2014) employ just such a condition for each component. However, by making use of the doubly-robust property we have the weaker conditions shown.[10] The first is a mild consistency requirement. The second requires an explicit rate on the product of errors, and hence if one function is relatively easy to estimate Assumption 3(b) can be satisfied even if the other does not converge at $n^{-1/4}$. This formalizes the benefit of doubly-robust estimation in general. In high-dimensional, sparse modeling specifically the rates for the first stage depend on the sample size, the number of covariates considered, and the sparsity level. Thus, if one function requires fewer covariates to estimate, i.e. smaller p or s, then greater complexity can be allowed for in the other (capturing, in particular, their relative smoothness).

The so-called "additional-randomness" estimators are more specific to the (approximately) sparse model context, and so we now codify the sparsity requirements of Section 4 and then give the additional conditions required for these estimators.

Assumption 4 (*Sparsity*)**.** For each n, P_n obeys (4)–(6), with $|S_*^Y| = s_d$ and $|S_*^D| = s_y$.

Assumption 5 (*Regularity Conditions for Union Estimators*)**.** For a sequence $\{P_n\}$, $\log(p) = o(n^{1/3})$ and the estimators $p_t(x)$ and $\hat{\mu}_t(x)$ obey the following, uniformly $t \in \bar{\mathbb{N}}_{\mathcal{T}}$:

$$\left(\max_{i\in\mathbb{I}_t}|u_i|\right)\left|\mathbb{E}_n[(\hat{p}_t(x_i) - p_t(x_i))^2]\right| = o_{P_n}(n^{-1/2}) \quad \text{and}$$

$$\left\|\hat{\gamma}_t - \gamma_t^*\right\|_1 \vee \|\hat{\beta}_t - \beta_t^*\|_1 = o_{P_n}(\log(p \vee n)^{-1/2}).$$

These conditions are needed to apply bounds for self-normalized sums (de la Peña et al., 2009). Belloni et al. (2012) were the first to use these techniques in high-dimensional, sparse models. The first condition is high-level, but can be verified with conditions on the errors and a bound for estimation. For the former, Belloni et al. (2012) assume that $\max_{i\in\mathbb{N}_n}|u_i| = O_{P_n}(n^{1/q})$ for some $q > 2$. A larger q eases the restriction in Assumption 5 but at the expense of stronger conditions on the noise distribution. For example, if u_i are assumed Gaussian, q can be taken to be any (large) positive number.

Remark 3 (*Linear Probability Models*)**.** Our results cover use of a linear probability model for $p_t(x)$, instead of the multinomial logistic form. All we require is a sufficiently high-quality approximation of the unknown function, and hence if Assumptions 3 and 5 if appropriate,[11] are met then uniform inference is possible using a linear probability model. Our group lasso results (Theorems 7 and 8) can be used directly to verify these conditions. In the same vein, multinomial logistic regression can be used to estimate $\mu_t(x)$ if the outcome Y is discretely valued. ■

5.2. Theoretical results

We now come to our main results on inference on average treatment effects. Most of our discussion will concern μ_t and $\boldsymbol{\mu}$; similar points apply to results for $\mu_{t,t'}$ and $\boldsymbol{\tau}$. Our first result formalizes consistency of our estimates under misspecification.

Theorem 2 (*Double Robustness*)**.** *Consider a sequence $\{P_n\}$ of data-generating processes. Suppose that for some $p_t^0(x)$ and $\mu_t^0(x)$, $\mathbb{E}_n[(\hat{p}_t(x_i) - p_t^0(x_i))^2] = o_{P_n}(1)$ and $\mathbb{E}_n[(\hat{\mu}_t(x_i) - \mu_t^0(x_i))^2] = o_{P_n}(1)$. Let Assumptions 1 and 2 hold for each n, with the regularity conditions also holding for $p_t^0(x)$ and $\mu_t^0(x)$. If $p_t^0(x) = p_t(x)$ or $\mu_t^0(x) = \mu_t(x)$, then $\left|\hat{\mu}_t - \mu_t\right| = o_{P_n}(1)$.*

This theorem formalizes the double-robustness property of our estimators: the propensity score or regression may be misspecified if the limiting objects are well-behaved. Compare to Assumption 3(a). The nearly identical result for $\mu_{t,t'}$ is omitted to save space.

We now turn to our main inference results. First we demonstrate a Bahadur representation of a generic $\hat{\mu}_t$ or $\hat{\mu}_{t,t'}$. These are shown to be equivalent to a sample average of the moment functions $\psi_t(\cdot)$ and $\psi_{t,t'}(\cdot)$, respectively, after proper centering and scaling, evaluated at the true $p_t(x_i)$ and $\mu_t(x_i)$. Using these results, asymptotic normality can be obtained for general estimands. We state explicit results for the leading examples $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$.

An asymptotic variance formula is needed to state the results. Define the conditional variance of the potential outcomes $\sigma_t^2(x) =$

[9] See Newey and McFadden (1994) and Chen (2007), and references therein.

[10] Many studies in the semiparametric literature relax or do not rely on the $n^{1/4}$ condition, allowing the nonparametric portion to converge at a slower rate, at any rate, or in some cases be inconsistent; examples include Powell et al. (1989), Newey (1990), Robins et al. (2008), Cattaneo et al. (2014a, 2013, 2014b), among others.

[11] Assumption 5 can be slightly weakened in this case due to the linear link function.

$\mathbb{E}[U^2|D = t, X = x]$ and the $\mathcal{T}$-square matrix $V_{\boldsymbol{\mu}}$ with elements

$$\begin{aligned} V_{\boldsymbol{\mu}}[t, t'] = {} & \mathbb{1}\{t = t'\}\mathbb{E}\left[\frac{\sigma_t^2(X)}{p_t(X)}\right] \\ & + \mathbb{E}\left[(\mu_t(X) - \mu_t)(\mu_{t'}(X) - \mu_{t'})\right] \\ \equiv {} & V_{\boldsymbol{\mu}}^W(t) + V_{\boldsymbol{\mu}}^B(t, t'). \end{aligned}$$

Straightforward plug-in estimators for these two components are given by[12]

$$\hat{V}_{\boldsymbol{\mu}}^W(t) = \mathbb{E}_n\left[\frac{d_i^t(y_i - \hat{\mu}_t(x_i))^2}{\hat{p}_t(x_i)^2}\right] \quad \text{and}$$

$$\hat{V}_{\boldsymbol{\mu}}^B(t, t') = \mathbb{E}_n\left[(\hat{\mu}_t(x_i) - \hat{\mu}_t)(\hat{\mu}_{t'}(x_i) - \hat{\mu}_{t'})\right].$$

Our first result gives the asymptotic behavior of $\hat{\mu}_t$ and $\hat{\boldsymbol{\mu}}$ for a sequence of DGPs.

Theorem 3 (*Estimation of Average Treatment Effects*)**.** *Consider a sequence $\{P_n\}$ of data-generating processes that obey Assumptions 1–3 for each n. If $\hat{\mu}_t(x_i)$ and $\hat{p}_t(x_i)$ do not have additional randomness in the estimated supports, we have:*

1. $\sqrt{n}(\hat{\mu}_t - \mu_t) = \sum_{i=1}^n \psi_t(y_i, d_i^t, \mu_t(x_i), p_t(x_i), \mu_t)/\sqrt{n} + o_{P_n}(1)$;
2. $V_{\boldsymbol{\mu}}^{-1/2}\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \to_d \mathcal{N}(0, I_{\mathcal{T}})$*; and*
3. $\hat{V}_{\boldsymbol{\mu}}^W(t) - V_{\boldsymbol{\mu}}^W(t) = o_{P_n}(1)$ *and* $\hat{V}_{\boldsymbol{\mu}}^B(t, t') - V_{\boldsymbol{\mu}}^B(t, t') = o_{P_n}(1)$.

If, in addition, Assumptions 4 and 5 hold, then the same is true when the supports contain additional randomness.

Theorem 3 itself may appear standard, but what is nonstandard is that the model selection step of the estimation has been explicitly accounted for. This immediately gives the following uniform inference results.

Corollary 2 (*Uniformly Valid Inference*)**.** *Let $\boldsymbol{P}_n$ be the set of data-generating processes satisfying the conditions of Theorem 3 for a given n and $G : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}$ be a fixed, twice uniformly continuously differentiable function with gradient ∇_G such that $\liminf_{n\to\infty} \|\nabla_G(\boldsymbol{\mu})\|_2$ is bounded away from zero. Then for $c_\alpha = \Phi^{-1}(1 - \alpha/2)$, we have:*

$$\sup_{P \in \boldsymbol{P}_n}\left|\mathbb{P}_P\left[G(\boldsymbol{\mu}) \in \left\{G(\hat{\boldsymbol{\mu}}) \pm c_\alpha\sqrt{\nabla_G(\hat{\boldsymbol{\mu}})'\hat{V}_{\boldsymbol{\mu}}\nabla_G(\hat{\boldsymbol{\mu}})/n}\right\}\right] - (1 - \alpha)\right| \\ \to 0.$$

Corollary 2 shows that these procedures are uniformly valid over the class of DGPs we consider, and hence will be reliable in applications. The crucial insight that leads to uniform inference is to change the goal of model selection away from perfect *covariate* selection (the oracle property) and to high-quality approximation of the underlying *functions* (here $p_t(\cdot)$ and $\mu_t(\cdot)$). This fundamental shift in focus allows us to avoid the uniformity problems demonstrated by Leeb and Pötscher. Assumption 3 formalizes exactly the quality of approximation needed. Such an approximation can be found for any element in $\boldsymbol{P}_n$, and hence inference is uniformly valid over that class. This method of proving uniformity follows Belloni et al. (2014) and Romano (2004), and is distinct from the approach of Andrews and Guggenberger (2009).

Results for the treatment effects on the treated are similar. The variance formula for $\boldsymbol{\tau}$ is slightly more cumbersome. Define the $\mathcal{T}$-square matrix $V_{\boldsymbol{\tau}}$ with elements

$$\begin{aligned} V_{\boldsymbol{\tau}}[t, t'] = {} & \mathbb{1}\{t = t'\}\mathbb{E}\left[\frac{p_t(X)}{p_t^2}\left[\sigma_t^2(X) + (\mu_t(X) - \mu_0(X)\right.\right. \\ & \left.\left. - \ \mu_{t,t} + \mu_{0,t})^2\right]\right] + \mathbb{E}\left[\frac{p_t(X)p_{t'}(X)}{p_t p_{t'} p_0(X)}\sigma_0^2(X)\right] \\ \equiv {} & V_{\boldsymbol{\tau}}^W(t) + V_{\boldsymbol{\tau}}^B(t, t'). \end{aligned}$$

Straightforward plug-in estimators for these two components are given by

$$\hat{V}_{\boldsymbol{\tau}}^W(t) = \mathbb{E}_n\left[\frac{d_i^t}{\hat{p}_t^2}\left[\left(y_i - \hat{\mu}_0(x_i) - \hat{\mu}_{t,t} + \hat{\mu}_{0,t}\right)^2\right]\right] \quad \text{and}$$

$$\hat{V}_{\boldsymbol{\tau}}^B(t, t') = \mathbb{E}_n\left[\frac{\hat{p}_t(x_i)\hat{p}_{t'}(x_i)}{\hat{p}_t\hat{p}_{t'}\hat{p}_0(x_i)^2}d_i^0(y_i - \hat{\mu}_0(x_i))^2\right].$$

Note that we need not estimate $\mu_t(x)$ and $\sigma_t^2(x)$, due to the simplification in Remark 1. With this notation, we have the following results. Proofs are so similar to those for Theorem 3 and Corollary 2 that we omit them.

Theorem 4 (*Estimation of Treatment Effects on Treated Groups*)**.** *Consider a sequence $\{P_n\}$ of data-generating processes that obey Assumptions 1–3 for each n. Then under P_n, as $n \to \infty$, if $\hat{\mu}_t(x_i)$ and $\hat{p}_t(x_i)$ do not have additional randomness in the estimated supports:*

1. $\sqrt{n}(\hat{\mu}_{t,t'} - \mu_{t,t'}) = \sum_{i=1}^n \psi_{t,t'}(y_i, d_i^t, \mu_t(x_i), p_t(x_i), p_{t'}(x_i), \mu_{t,t'})/\sqrt{n} + o_{P_n}(1)$;
2. $V_{\boldsymbol{\tau}}^{-1/2}\sqrt{n}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \to_d \mathcal{N}(0, I_{\mathcal{T}})$*; and*
3. $\hat{V}_{\boldsymbol{\tau}}^W(t) - V_{\boldsymbol{\tau}}^W(t) = o_{P_n}(1)$ *and* $\hat{V}_{\boldsymbol{\tau}}^B(t, t') - V_{\boldsymbol{\tau}}^B(t, t') = o_{P_n}(1)$.

If, in addition, Assumptions 4 and 5 hold, then the same is true when the supports contain additional randomness.

Corollary 3 (*Uniformly Valid Inference*)**.** *Let $\boldsymbol{P}_n$ be the set of data-generating processes satisfying the conditions of Theorem 4 for a given n and $G : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}$ be a fixed, twice uniformly continuously differentiable function with gradient ∇_G such that $\liminf_{n\to\infty} \|\nabla_G(\boldsymbol{\tau})\|_2$ is bounded away from zero. Then for $c_\alpha = \Phi^{-1}(1 - \alpha/2)$, we have:*

$$\sup_{P \in \boldsymbol{P}_n}\left|\mathbb{P}_P\left[G(\boldsymbol{\tau}) \in \left\{G(\hat{\boldsymbol{\tau}}) \pm c_\alpha\sqrt{\nabla_G(\hat{\boldsymbol{\tau}})'\hat{V}_{\boldsymbol{\tau}}\nabla_G(\hat{\boldsymbol{\tau}})/n}\right\}\right] - (1 - \alpha)\right| \\ \to 0.$$

5.3. *Efficiency considerations*

The prior theoretical results are aimed at delivering robust inference. In this section, we briefly discuss the efficiency of our estimator according to two criteria: semiparametric efficiency and oracle efficiency. To put each on sound conceptual footing we separate discussion and restrict to an appropriate set of models.

For semiparametric efficiency, $p_t(x)$ and $\mu_t(x)$ are nonparametric objects, as in Example 4, X are fixed-dimension variables and the DGP does not vary with n. If we "upgrade" the mean independence of Assumption 1(a) to full, namely $\{Y(t)\}_{\mathbb{N}_{\mathcal{T}}} \perp\!\!\!\perp D|X$, then Theorems 3 and 4 immediately yield asymptotic linearity and semiparametric efficiency, attaining Hahn's (1998) or Cattaneo's (2010) bounds. This requires there be no (known) instruments for treatment status in X, as implicitly assumed in those works, else the bound may change (Hahn, 2004).

Turning to oracle efficiency, an alternative to our robust approach is to prove that the true support can be found with probability approaching one (the oracle property), then conduct inference conditioning on this event. This approach cannot be made uniformly valid, but may be of interest in the exactly sparse models of Example 2 (there is no "true" support in approximately sparse models), because discovering the true support is equivalent to finding the variables in the causal mechanism (White and Lu, 2011), if one exists. This may be interesting in its own right, or for future applications by way of hypothesis generation. The post oracle selection estimator is made efficient by using only the variables important for $\mu_t(x_i) = \mathbb{E}[Y|D = t, x_i]$. This amounts to entirely removing the instrumental variables indexed by $S_*^D \setminus S_*^Y$, whose inclusion would, in general, reduce efficiency, though not

[12] Estimators can also be based on sample averages of outer products of influence functions, which would include the covariance term that vanishes in expectation.

increase bias. Further, $S_*^Y \setminus S_*^D$ are excluded from propensity score estimation.

Perfect selection requires two strong conditions: (i) an orthogonality condition on the Gram matrices that restricts the correlation between the variables in and out of the true support (Bach, 2008), and (ii) a *beta-min* condition bounding the nonzero coefficients away from zero. Intuitively, highly correlated variables cannot be distinguished, nor can coefficients sufficiently close to zero be found with certainty. Both bounds may depend on n, and in particular the lower bound on the coefficients may vanish at an appropriate rate. Under such conditions, it is straightforward to show that S_*^Y and S_*^D can be found with probability approaching one.

6. Group lasso selection and estimation

We now give details for group lasso model selection and estimation, and make the refitting precise. Section 6.1 discusses penalty choices and implementation. Restricted and sparse eigenvalues, key quantities in our bounds, are discussed in Section 6.2. Our main nonasymptotic results are stated in Section 6.3. These results are of interest more generally in the literature on high-dimensional sparse models. Finally, Section 6.4 gives asymptotic rates and verifies the conditions of Section 5.

We first select covariates by applying the group lasso penalty to the multinomial logistic loss (for the propensity scores) and to least squares loss (to estimate the outcome regression). The loss functions are defined as

$$\mathcal{M}(\gamma_{\cdot,\cdot}) = \sum_{t\in\mathbb{N}_{\mathcal{T}}} \mathbb{E}_n\left[-d_i^t \log\left(\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_{\mathcal{T}}})\right)\right] \quad \text{and}$$

$$\mathcal{E}(\beta_{\cdot,\cdot}) = \sum_{t\in\overline{\mathbb{N}}_{\mathcal{T}}} \mathbb{E}_{n,t}[(y_i - x_i^{*\prime}\beta_t)^2].$$

Then, the group lasso estimates for the propensity score coefficients, denoted $\tilde{\gamma}_{\cdot,\cdot}$, and the regression coefficients, $\tilde{\beta}_{\cdot,\cdot}$, respectively solve

$$\begin{aligned} \tilde{\gamma}_{\cdot,\cdot} &= \operatorname*{arg\,min}_{\gamma_{\cdot,\cdot}\in\mathbb{R}^{p\mathcal{T}}} \left\{\mathcal{M}(\gamma_{\cdot,\cdot}) + \lambda_D \left|\!\left|\!\left|\gamma_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}\right\} \quad \text{and} \\ \tilde{\beta}_{\cdot,\cdot} &= \operatorname*{arg\,min}_{\beta_{\cdot,\cdot}\in\mathbb{R}^{p\overline{\mathcal{T}}}} \left\{\mathcal{E}(\beta_{\cdot,\cdot}) + \lambda_Y \left|\!\left|\!\left|\beta_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}\right\}, \end{aligned} \tag{9}$$

where λ_D and λ_Y are the penalty parameters discussed below and $\left|\!\left|\!\left|\gamma_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}$ is the mixed ℓ_2/ℓ_1 norm.

To ameliorate the downward bias induced by the penalty and to allow for researcher-added variables, we refit unpenalized models.[13]Let $\tilde{S}^D = \{j : \|\tilde{\gamma}_{\cdot,j}\|_2 > 0\}$ and $\tilde{S}^Y = \{j : \|\tilde{\beta}_{\cdot,j}\|_2 > 0\}$ be the selected covariates and $\hat{S}^D$ and $\hat{S}^Y$ those used in refitting.[14] We require $\hat{S} \supset \tilde{S}$ and $|\hat{S}| \leq s$ for D and Y (we will prove that $|\tilde{S}| \leq s$ in both cases). The refitting estimators solve

$$\begin{aligned} \hat{\gamma}_{\cdot,\cdot} &= \operatorname*{arg\,min}_{\gamma_{\cdot,\cdot},\ \mathrm{supp}(\gamma_t)=\hat{S}^D} \left\{\mathcal{M}(\gamma_{\cdot,\cdot})\right\} \quad \text{and} \\ \hat{\beta}_{\cdot,\cdot} &= \operatorname*{arg\,min}_{\beta_{\cdot,\cdot},\ \mathrm{supp}(\beta_t)=\hat{S}^Y} \left\{\mathcal{E}(\beta_{\cdot,\cdot})\right\}. \end{aligned} \tag{10}$$

Remark 4 (*Weighted Penalties*)**.** The group lasso penalty can be weighted in two ways. First, one may weight the ℓ_2 portion, as in $\lambda_D \sum_{j\in\mathbb{N}_p} \|\boldsymbol{X}_j\gamma_{\cdot,j}\|_2$, where $\boldsymbol{X}_j$ is the design matrix for covariate j, across all the treatments. Other weight matrices are possible, but with this choice, the estimate is invariant to within group (treatment) reparameterizations, and is thus scale invariant for each covariate. We therefore assume $\mathbb{E}_n[(x_{i,j}^*)^2] = 1$ without loss of generality.

Second, the ℓ_1 norm can be weighted to give a penalty of the form $\lambda_D \sum_{j\in\mathbb{N}_p} w_j\|\gamma_{\cdot,j}\|_2$. Two common choices for w_j are the number of variables in group j or an adaptive penalty from a pilot estimate. Our groups are equally sized, and although adaptive procedures may improve oracle properties (Zou, 2006; Wei and Huang, 2010), our goal is not perfect selection. ■

6.1. Choice of penalty

We must specify choices of λ_D and λ_Y for programs (9). From a theoretical point of view, these must be chosen so that the penalty dominates the noise, which is captured by the magnitude of the score in the dual of the $|||\cdot|||_{2,1}$ norm, with high probability. To achieve this, we set

$$\begin{aligned} \lambda_D &= \frac{2\mathcal{X}\sqrt{\mathcal{T}}}{\sqrt{n}}\left(1 + \frac{\log(p \vee n)^{3/2+\delta_D}}{\sqrt{\mathcal{T}}}\right)^{1/2} \quad \text{and} \\ \lambda_Y &= \frac{4\mathcal{X}\mathcal{U}\sqrt{\overline{\mathcal{T}}}}{\sqrt{\underline{n}}}\left(1 + \frac{\log(p \vee \underline{n})^{3/2+\delta_Y}}{\sqrt{\overline{\mathcal{T}}}}\right)^{1/2}, \end{aligned} \tag{11}$$

for some $\delta_D > 0$ and $\delta_Y > 0$. With these choices, $\lambda_D > 2\max_{j\in\mathbb{N}_p} \|\mathbb{E}_n[(p_t(x_i) - d_i^t)x_{i,j}^*]\|_2$ and $\lambda_Y > 4\max_{j\in\mathbb{N}_p} \|\mathbb{E}_{n,t}[u_i x_{i,j}^*]\|_2$ with probability $1 - \mathcal{P}$ for a small (and shrinking) $\mathcal{P}$. In generic terms, λ is of the form $\Lambda(1 + r_n)$, where Λ is an upper bound on the true score and r_n is a rate that depends on n and p.[15] The specific rate chosen serves to balance the rate of convergence against the concentration effect: a smaller r_n would increase the rate of convergence, but at the expensive of lowering the concentration probability $1 - \mathcal{P}$. In Appendix we show that (for appropriate δ and n or $\underline{n}$) the concentration probability is given by

$$\mathcal{P} = \frac{4\sqrt{\log(2p)(1 + 64\log(12p)^2)}}{\log(p \vee n)^{3/2+\delta}}. \tag{12}$$

There are two practical methods to make these choices for feasible for implementation. When $\hat{p}_t(x)$ and $\hat{\mu}_t(x)$ are used to estimate average treatment effects, the decreased sensitivity of the final estimate to the first stage, thanks to the doubly-robust estimator, in turn results in less sensitivity to the choice of penalty (through the sparsity).[16] The first option is an iterative procedure to estimate the unknown $\mathcal{X}$ and $\mathcal{U}$ in λ_Y and λ_D, as employed by Belloni et al. (2012) (validity of this procedure may be established along the same lines as in that study). We use $\max_{i\leq n}\max_{j\in\mathbb{N}_p} |x_{i,j}^*|$ for $\mathcal{X}$ and estimate $\mathcal{U}$ by iteration: given an initial estimate $\hat{\mu}_t^{(0)}(x)$, set $\hat{\mathcal{U}}^{(k)} = \mathbb{E}_n[(y_i - \hat{\mu}_t^{(k-1)}(x_i))^4]^{1/4}$, where $\hat{\mu}_t^{(k)}(x_i)$, $k > 0$, is based on Eq. (10). In implementation we found 10 iterations more than sufficient, and based the initial estimate

[13] The bias is away from the pseudo-true coefficients of the sparse parametric representation, $\gamma_{\cdot,\cdot}^*$ and $\beta_{\cdot,\cdot}^*$. There is no relation to specification biases B_t^D and B_t^Y.

[14] When $\mathrm{supp}(\gamma_t^*)$ and $\mathrm{supp}(\beta_t^*)$ do not vary much over t, the group lasso is known to have better properties than the ordinary lasso in terms of selection and convergence. Obozinski et al. (2011) give a sharp bound on the overlap necessary to yield improvements, while Huang and Zhang (2010), Kolar et al. (2011), and Lounici et al. (2011) also demonstrate advantages of the group lasso approach. These works show, among other things, that the group lasso advantage increases with large $\mathcal{T}$, and with the group structure, may perform better with smaller samples. We defer to the works cited for a formal discussion.

[15] The slight differences in the two are as follows. The full sample has information on the logistic coefficients, so n appears instead of $\underline{n}$. No error bound appears in λ_D because the errors are bounded by one. The multiple 4 for λ_Y, instead of 2, can be traced to the quadratic loss. These forms are determined at heart by the maximal inequality of Lounici et al. (2011).

[16] To our knowledge, no formal results exist on "optimal" penalty parameter choices for inference in high-dimensional problems nor are any procedures free of user-specified choices.

on ridge regression (with penalty chosen by cross validation). A second option is to select λ_Y and λ_D directly by cross-validation. This has the appealing feature that the precise forms of Eq. (11) need not be characterized and estimated. If interest lies in the underlying functions $p_t(x)$ and $\mu_t(x)$, cross validation is appropriate as it minimizes a relevant loss function. Formal results establishing the validity of cross-validation are not available, but it performs well in practice.

6.2. Restricted eigenvalues

The local behavior of optimizations (9) and (10) is captured by their respective Hessians, which involve the second moment matrix of the covariates. The eigenvalues of such matrices will be explicit in our bounds. We are interested in finite sample bounds, and so we will only discuss the empirical Gram matrices (see Remark 5). Define

$$Q = \mathbb{E}_n[x_i^* x_i^{*\prime}] \quad \text{and} \quad Q_t = \mathbb{E}_{n,t}[x_i^* x_i^{*\prime}]. \tag{13}$$

In high-dimensional data, both are singular, and so we use restricted eigenvalues and sparse eigenvalues (Bickel et al., 2009).

For the multinomial logistic regression, the minimal restricted eigenvalue is defined by

$$\kappa_D^2 \le \min_\delta \left\{ \frac{\sum_{t\in\mathbb{N}_\mathcal{T}} \delta_t' Q \delta_t}{\|\delta_{\cdot,S_D^*}\|_2^2} : \delta \in \mathbb{R}^{p\mathcal{T}} \setminus \{0\}, \left|\left|\left|\delta_{\cdot,\{S_*^D\}^c}\right|\right|\right|_{2,1} \le 4 \left|\left|\left|\delta_{\cdot,S_*^D}\right|\right|\right|_{2,1} \right\}. \tag{14}$$

For least squares estimation we instead use

$$\kappa_Y^2 \le \min_\delta \left\{ \frac{\sum_{t\in\overline{\mathbb{N}}_\mathcal{T}} \delta_t' Q_t \delta_t}{\|\delta_{\cdot,S_Y^*}\|_2^2} : \delta \in \mathbb{R}^{p\overline{\mathcal{T}}} \setminus \{0\}, \left|\left|\left|\delta_{\cdot,\{S_*^Y\}^c}\right|\right|\right|_{2,1} \le 3 \left|\left|\left|\delta_{\cdot,S_*^Y}\right|\right|\right|_{2,1} \right\}. \tag{15}$$

Note that Q appears for κ_D, whereas the Q_t are used in κ_Y. The restricted set, or cone constraint, requires the magnitude of $\delta_{\cdot,\cdot}$ off the true support be small relative to the true support, measured in the group lasso norm.[17] We will show that $(\tilde{\gamma}_{\cdot,\cdot} - \gamma_{\cdot,\cdot}^*)$ and $(\tilde{\beta}_{\cdot,\cdot} - \beta_{\cdot,\cdot}^*)$ obey the respective constraints.

In contrast, the refitting errors $(\hat{\gamma}_{\cdot,\cdot} - \gamma_{\cdot,\cdot}^*)$ and $(\hat{\beta}_{\cdot,\cdot} - \beta_{\cdot,\cdot}^*)$ from (10) may not obey the cone constraint, but are sparse by construction. This motivates the use of sparse eigenvalues. For a set $S \subset \mathbb{N}_p$ and a $p \times p$ matrix $\tilde{Q}$, define

$$\underline{\phi}\{\tilde{Q}, S\}^2 = \min_{\delta\in\mathbb{R}^p,\, \mathrm{supp}(\delta)=S} \frac{\delta'\tilde{Q}\delta}{\|\delta\|_2^2} \quad \text{and}$$
$$\overline{\phi}\{\tilde{Q}, S\}^2 = \max_{\delta\in\mathbb{R}^p,\, \mathrm{supp}(\delta)=S} \frac{\delta'\tilde{Q}\delta}{\|\delta\|_2^2}. \tag{16}$$

Finally, it will be useful to define a bound on $\overline{\phi}\{\tilde{Q}, S\}$ over all subsets of a certain size. To this end, for any integer m, define $\overline{\overline{\phi}}(\tilde{Q}, m) = \max_{S\subset\mathbb{N}_p,\, |S|\le m} \overline{\phi}\{\tilde{Q}, S\}$.

We take these quantities to be primitive, and defer to the literature. For example, van de Geer and Buhlmann (2009), Huang and Zhang (2010), Raskutti et al. (2010), Rudelson and Zhou (2013), and Belloni et al. (2014). In particular, Huang and Zhang (2010) show that the group lasso may need fewer observations to satisfy conditions on $\underline{\phi}\{\tilde{Q}, S\}$.

Remark 5. Often, invertibility of Q and Q_t relies on their convergence to nonsingular population counterparts.[18] Some of the papers cited use this approach and our results can be restated in this way by conditioning on the event that Q and Q_t are close to their counterparts in the appropriate sense, and adjusting the probability with which the conclusions hold. We instead take bounds to be infinite if the minimum eigenvalues are zero. ■

6.3. Finite sample theoretical results

We now have the necessary notation and assumptions to state our theoretical results on group lasso estimation, beginning with multinomial logistic regression, followed by a terse treatment of linear models. Corollary 1 is a special case of the results in this section, see Section 6.4.

Our first result is a nonasymptotic bound on the group lasso estimates from (9).

Theorem 5 (*Group Lasso Estimation of Multinomial Logistic Models*)**.** *Suppose Assumptions* 1(b), 2(a), 2(b), 2(c), *and* 4 *hold. Define* $A_p = p_{\min}/(0 \vee (p_{\min} - b_s^d))$ *and*

$$R_\mathcal{M} = \left(A_p/p_{\min}\right)^{\overline{\mathcal{T}}} \mathcal{T} A_K \left(6\lambda_D\sqrt{|S_*|}\kappa_D^{-1} + 8b_s^d\sqrt{\mathcal{T}}\right),$$

for $A_K > 2\kappa_D^2\{\kappa_D^2 - (2/3)\mathcal{X}\sqrt{\mathcal{T}}(30\lambda_D|S_*| + 100\sqrt{|S_*|}\kappa_D b_s^d\sqrt{\mathcal{T}} + 80\kappa_D^2(b_s^d)^2\mathcal{T}\lambda_D^{-1})\}^{-1}$. *Then with probability* $1 - \mathcal{P}$, *we have*

1. $\max_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2]^{1/2} \le R_\mathcal{M} + b_s^d$,
2. $\max_{t\in\mathbb{N}_\mathcal{T}} \left\|\tilde{\gamma}_t - \gamma_t^*\right\|_1 \le R_\mathcal{M}\sqrt{|\tilde{S}^D \cup S_D^*|/\underline{\phi}\{Q, \tilde{S}^D \cup S_D^*\}}$,
3. *and* $|\tilde{S}^D| \le 8sL_n\left(\min\{\overline{\overline{\phi}}(Q, m) : m \in \mathbb{N}_Q^D\}\right)$,

where $\mathbb{N}_Q^D = \left\{m \in \{1, 2, \ldots, n\} : m > 8sL_n\overline{\overline{\phi}}(Q, m)\right\}$ *and* $L_n = \mathcal{T}\left((R_\mathcal{M} + b_s^d)/(\lambda_D\sqrt{s})\right)^2$.

This theorem is new to the literature, to the best of our knowledge. Much of the detail involves capturing the finite sample behavior of the Hessian and Gram matrices. We discuss the features of this result in the following remarks.

- The Hessian of $\mathcal{M}(\gamma_{\cdot,\cdot})$ is $\mathbb{E}_n[\mathcal{H}_i \otimes x_i^* x_i^{*\prime}]$ for a $\mathcal{T}$-square matrix $\mathcal{H}_i$ that depends on the coefficients and x_i^* through the estimated probabilities $\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})$. The error $R_\mathcal{M}$ depends on how well-controlled is this matrix. The factors $p_{\min}$, A_p, and A_K capture the behavior of $\mathcal{H}_i$ and κ_D^{-1} accounts for the rest. Under overlap, the true probabilities are bounded below by $p_{\min}$, and hence $p_{\min}^{-\overline{\mathcal{T}}}$ captures the nonsingularity of the population version of $\mathcal{H}_i$. To get to this point requires two steps. First, the sparse parametric representations $\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})$ must also be bounded away from zero, leading to the factor of A_p. This is essentially a bias condition, which in the asymptotic case holds trivially: A_p may be chosen arbitrarily close to one as $b_s^d \to 0$. Second, A_K controls the neighborhood in which $\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})$ is also bounded away from zero. Intuitively (and asymptotically), the estimate will be in a small (shrinking) neighborhood of the $\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})$. In asymptotics A_K may be chosen arbitrarily close to 2, which stems from the factor of 1/2 in a quadratic expansion of $\mathcal{M}(\cdot)$. A

[17] The multiplier of 4 in the constraint for κ_D is traceable to the nonlinear model.

[18] This is standard in fixed-dimension models, and has been used for diverging-dimensions parametric models (He and Shao, 2000) and nonparametrics (Newey, 1997; Huang, 2003; Cattaneo and Farrell, 2013; Belloni et al., 2015; Chen and Christensen, forthcoming). The eigenvalue assumptions employed in those works are conceptually the same as the restricted eigenvalues used here, only restricted to the $p < n$ case.

lower bound on A_K is required in finite samples to ensure that $\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})$ is positive, and hence the two-term expansion is valid. This is analogous to Belloni and Chernozhukov's (2011) "restricted nonlinear impact coefficient" approach, also used by Belloni et al. (2014) with a central difference that A_K is captured in our bound directly.

- The maximal sparse eigenvalues are crucial to the bound on $|\tilde{S}^D|$. In many prior results, the latter is bounded using the largest eigenvalue of Q itself, i.e. $\overline{\overline{\phi}}(Q, n)$. Adapting the technique of Belloni and Chernozhukov (2013) to the present case, we are able to find a tighter bound, which yields sparsity proportional to s under weaker conditions. This is crucial for refitting.
- For the linear model the constants in the group lasso bounds can offset the (logarithmic) suboptimality in rate (Huang and Zhang, 2010; Lounici et al., 2011), and this may be true here as well. This is application dependent however.

The error bounds for post-selection estimation are more complex and depend in part on the good properties of the initial group lasso fit. The following theorem gives our results.

Theorem 6 (*Post-Selection Multinomial Logistic Regression*)**.** *Suppose the conditions of Theorem 5 hold. To save notation, let $S_D = \hat{S}_D \cup S_D^*$ and $\underline{\phi} = \underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}$. Then for*

$$A_K > 2\left\{\frac{\underline{\phi}^2}{\underline{\phi}^2 - \mathcal{X}\sqrt{\mathcal{T}}(\lambda_D|S_D| + b_s^d\underline{\phi}\sqrt{\mathcal{T}}\sqrt{|S_D|})}\right\}$$
$$\vee\left\{\frac{\underline{\phi}}{\underline{\phi} - 2R_\mathcal{M}\mathcal{X}\sqrt{\mathcal{T}}\sqrt{|S_D|}}\right\}$$

define $R'_\mathcal{M} = (A_p/p_{\min})^{\overline{\mathcal{T}}}\,\mathcal{T}A_K\left(\lambda_D\sqrt{|S_D|}\underline{\phi}^{-1}/2 + b_s^d\sqrt{\mathcal{T}}\right)$ *and*

$$R''_\mathcal{M} = \{R_\mathcal{M}\} \vee \left\{R'_\mathcal{M} + \left[R'_\mathcal{M}R_\mathcal{M} + (A_p/p_{\min})^{\overline{\mathcal{T}}}\,\mathcal{T}A_K R_\mathcal{M}^2\right]^{1/2}\right\}.$$

Then with probability $1 - \mathcal{P}$, $\max_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2]^{1/2} \le R''_\mathcal{M} + b_s^d$, and $\max_{t\in\mathbb{N}_\mathcal{T}} \left\|\hat{\gamma}_t - \gamma_t^*\right\|_1 \le \left(|S^D|/\underline{\phi}\right)^{1/2} R''_\mathcal{M}$.

It is not readily discernible if these bounds improve upon the initial fit. This will depend on the DGP, the selection success of the initial fit, and any added variables. In this result, further lower bounds on A_K are required to handle the sparse eigenvalues, compared to the restricted version in Theorem 5. The role played by A_K is the same in both cases, as with the other factors.

It is worth noting that, despite the complexity of multinomial logistic regression, the conditions for Theorems 5 and 6 are simple and intuitive, and match those used for linear models.

We now give our results for group lasso estimation of the conditional outcome regressions. In computing $\mu_t(x_i)$ for $d_i^t \neq 1$ we are performing out of sample prediction, which slightly complicates the bounds. Our first result is on the initial group lasso fit.

Theorem 7 (*Group Lasso Estimation of Linear Models*)**.** *Suppose Assumptions 1(b), 2(a)–2(c), and 4 hold. To save notation, let $S_Y = \tilde{S}^Y \cup S_Y^*$. Define*

$$R_\mathcal{E} = \left(\frac{3\lambda_Y\sqrt{s}}{\kappa_Y} + 2b_s^y\right).$$

Then with probability $1 - \mathcal{P}$, we have

1. $\max_{t\in\overline{\mathbb{N}}_\mathcal{T}} \mathbb{E}_n[(x_i^{*\prime}\tilde{\beta}_t - \mu_t(x_i))^2]^{1/2} \le \left(\overline{\phi}\{Q, S_Y\}/\underline{\phi}\{Q_t, S_Y\}\right)^{1/2} R_\mathcal{E} + b_s^y$,
2. $\max_{t\in\overline{\mathbb{N}}_\mathcal{T}} \left\|\tilde{\beta}_t - \beta_t^*\right\|_1 \le \left(|S_Y|/\underline{\phi}\{Q, S_Y\}\right)^{1/2} (\overline{\phi}\{Q, S_Y\}/\underline{\phi}\{Q_t, S_Y\})^{1/2} R_\mathcal{E}$,
3. *and* $|\tilde{S}^Y| \le 32sL_n\left\{\min_{m\in\mathbb{N}_Q^Y} \sum_{t\in\overline{\mathbb{N}}_\mathcal{T}} \overline{\overline{\phi}}(Q_t, m)\right\}$,

where $\mathbb{N}_Q^Y = \left\{m \in \{1, 2, \ldots, \overline{n}\} : m > 32sL_n \sum_{t\in\overline{\mathbb{N}}_\mathcal{T}} \overline{\overline{\phi}}(Q_t, m)\right\}$ *and* $L_n = \left((R_\mathcal{E} + b_s^y)/(\lambda_Y\sqrt{s})\right)^2$.

This theorem generalizes Lounici et al. (2011) to the nonparametric, approximately sparse case, improves the sparsity bound, and gives out of sample prediction (imputation) results. The analogous generalization for within sample prediction loss (e.g. multitask learning), $\mathbb{E}_{n,t}[(x_i^{*\prime}\tilde{\beta}_t - \mu_t(x_i))^2]^{1/2}$, may be found in the Supplement (see Appendix D).

For refitting, we are predicting for the entire sample and so we utilize the general results given by Belloni et al. (2012) for post-selection estimation of least squares. The following result is a direct implication of their Lemma 7 and our Theorem 7.

Theorem 8 (*Post-Selection Linear Regression*)**.** *Suppose* $\log(p) = o(n^{1/3})$ *in addition to the conditions of Theorem 7. Then for constants A_1 –A_4 not depending on n nor the DGP:*

$$\mathbb{E}_n[(x_i'\hat{\beta}_t - \mu_t(x_i))^2]^{1/2} \le A_1\sqrt{\frac{s(\mathcal{T} \wedge \log(s\mathcal{T}))}{n\underline{\phi}\{Q, S_Y^*\}}}$$
$$+A_2\sqrt{\frac{|\hat{S}_Y \backslash S_Y^*|\log(p\mathcal{T})}{n\underline{\phi}\{Q, S_Y^{FP}\}}} + A_3\sqrt{\mathbb{E}_n[(x_i^{*\prime}\tilde{\beta}_t - \mu_t(x_i))^2]}$$

and $\max_{t\in\overline{\mathbb{N}}_\mathcal{T}} \|\hat{\beta}_t - \beta_t^*\|_1 \le A_4(|\hat{S}_Y \cup S_Y^*|\mathbb{E}_n[(x_i'\hat{\beta}_t - \mu_t(x_i))^2]/\underline{\phi}\{Q, \hat{S}_Y \cup S_Y^*\})^{1/2}$.

As above, the performance of the refitting procedure depends in part on the success of the initial group lasso fit. Indeed, the middle term is dropped if the true support union is found. The constants A_k, $k = 1, 2, 3, 4$ are not given explicitly but are known to be absolute bounds (de la Peña et al., 2009) under Assumption 2. This result is less precise than Theorems 5 and 6, but sufficient to verify Assumptions 3 and 5.

6.4. Asymptotic analysis and verification of high-level conditions

This section derives rates of convergence for the group lasso estimates and uses these results to verify Assumptions 3 and 5 in Section 5. For simplicity, we only state results for the post-selection estimators that we recommend in practice. In reducing the finite sample results of Theorems 6 and 8 to rates we retain the dependence on n, p, s, and the bias. Note that the number of treatments is fixed, and the overlap assumption ensures that all $n_t \propto n$. Further, the various (restricted and sparse) eigenvalues are commonly taken to be bounded (or bounded away from zero) in asymptotic analyses. This accounts for the remaining factors in the bounds. For multinomial logistic regression, we obtain the following result.

Corollary 4 (*Asymptotics for Multinomial Logistic Regression*)**.** *Suppose the conditions of Theorem 6 hold and further that* (i) $\lambda_D s_d = o(1)$, (ii) κ_D *is bounded away from zero, and* (iii) $\min_{S:|S|=O(s)} \underline{\phi}\{Q, S\}$ *is bounded away from zero and* $\overline{\overline{\phi}}(Q, \cdot)$ *is bounded, uniformly in* $\mathbb{N}_Q^D$*. Then*

1. $|\tilde{S}^D| = O_{P_n}(s_d)$,
2. $\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2] = O_{P_n}(n^{-1}s_d \log(p \vee n)^{3/2+\delta_D} + (b_s^d)^2)$, *and*
3. $\|\hat{\gamma}_t - \gamma_t^*\|_1 = O_{P_n}\left(\sqrt{n^{-1}s_d^2 \log(p \vee n)^{3/2+\delta_D}} + b_s^d\sqrt{s_d}\right)$.

Similarly, we have the following for the linear models.

Corollary 5 (*Asymptotics for Linear Regression*). *Suppose the conditions of* Theorem 8 *hold and further that* (i) $\lambda_Y\sqrt{s_y} = o(1)$, (ii) κ_Y *is bounded away from zero, and* (iii) *uniformly in* $\overline{\mathbb{N}}_\mathcal{T}$, $\min_{S:|S|=O(s)} \underline{\phi}\{Q_t, S\} \wedge \underline{\phi}\{Q, S\}$ *is bounded away from zero and* $\overline{\overline{\phi}}(Q, \cdot) \vee \overline{\overline{\phi}}(Q_t, \cdot)$ *is bounded uniformly in* $\mathbb{N}_Q^Y$. *Then*

1. $|\tilde{S}^Y| = O_{P_n}(s_y)$,
2. $\mathbb{E}_n[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2] = O_{P_n}\left(n^{-1}s_y \log(p \vee n)^{3/2+\delta_Y} + (b_s^y)^2\right)$, *and*
3. $\|\tilde{\beta}_t - \beta_t^*\|_1 = O_{P_n}\left(\sqrt{n^{-1}s_y^2 \log(p \vee n)^{3/2+\delta_Y}} + b_s^y\sqrt{s}\right)$.

It is now straightforward to verify the requirements of Section 5. Assumption 3(b) requires

$$(n^{-1}s_d \log(p \vee n)^{3/2+\delta_D} + (b_s^d)^2)(n^{-1}s_y \log(p \vee n)^{3/2+\delta_Y} + (b_s^y)^2) = o\left(n^{-1}\right).$$

Under the common assumption that $b_s = O(\sqrt{s/n})$, we require $s_d s_y \log(p \vee n)^{3+\delta_D+\delta_Y} = o(n)$. Both this, and the display above, clearly show how the sparsity and smoothness of the two functions interact due to the double robustness. Assumption 5 can be verified similarly.

These rates of convergence (i.e. part 2 of each corollary) are optimal up to factor $\log(p \vee n)^{1/2+\delta}$. At heart, this loss appears to stem from the maximal inequality used to establish the concentration probability of (12). In practice, this is unlikely to be a limitation. As mentioned above, the use of group lasso can yield improvements in the constants if the data obey a grouped sparsity pattern, as is expected for treatment effects data, and may even yield improvements in the detection of the sparse signal, further offsetting the suboptimal log factor (see for example Lounici et al. (2011) or Obozinski et al. (2011)). Alternative methods could, in principle, yield a rate improvement. Chief among these would be lasso-penalized linear probability models (see also Remark 3) or separate logistic regressions. The group lasso approach adopted here reflects common practice, and so it may be preferred. In any case, the log factors do not impact the treatment effect inference.

7. Numerical and empirical evidence

7.1. Simulation study

We conducted a Monte Carlo exercise to study how our estimator behaves as the propensity score and regression functions change, and the model selection problem becomes more or less difficult.[19] For simplicity we focus on the average effect of a binary treatment. We generated 1000 observations $(y_i, d_i, x_i')'$ from the models in Example 3, using both $p = 1000$ and $p = 1500$. The covariates include an intercept, with the remainder drawn from $N(0, \Sigma)$, with covariance $\Sigma[j_1, j_2] = 2^{-|j_1-j_2|}$, $2 \leq j_1, j_2 \leq p$. Errors are standard Normal. The crucial aspects of the DGP are the coefficient vectors β_0^0, β_1^0, and γ^0, which are defined to vary with the positive scalars $\rho_\beta, \rho_\gamma, \alpha_\beta$, and α_γ, as follows:

$$\beta_0^0 = \rho_\beta(-1, 1, -1, 2^{-\alpha_\beta}, -3^{-\alpha_\beta}, \ldots, j^{-\alpha_\beta}, \ldots, p^{-\alpha_\beta})',$$
$$\gamma^0 = \rho_\gamma(1, -1, 1, -2^{-\alpha_\gamma}, 3^{-\alpha_\gamma}, \ldots, j^{-\alpha_\gamma}, \ldots, -p^{-\alpha_\gamma})',$$

with $\beta_1^0 = -\beta_0^0$. The ρ multipliers affect the signal-to-noise ratio, but not the sparsity. For smaller values distinguishing the large and small coefficients is more difficult for a given sample. The exponents α control the sparsity, where a sparse representation is not possible for small values.

Fig. 1 shows the empirical coverage rates of 95% confidence intervals for $\mu_1 - \mu_0$ for different DGPs, for $p = 1000$ and 1500. Panels (a) and (c) show coverage as the multipliers ρ_β and ρ_γ range over 0.01 (weak signal) to 1 (strong), with $\alpha_\beta = \alpha_\gamma = 2$. Panels (b) and (d) vary the sparsity exponents α_β and α_γ over 1/8 (not sparse) to 4 (very sparse), with $\rho_\beta = \rho_\gamma = 1$. Of 1000 observations total, the (mean) size of the comparison group declines from roughly 500 to 300 as ρ_γ increases and 450 to 300 as α_γ increases, over their given ranges. Coverage is accurate over all signal strengths, and breaks down only when neither $\mu_t(x_i)$ nor $p_t(x_i)$ is sparse, which is exactly when Assumption 3(b) (or condition (ii) of Theorem 1) cannot be satisfied. Note that coverage accuracy is retained when only one function is sparse, showcasing the double-robustness property.

The penalty parameters λ_D and λ_Y are chosen using the iterative procedure described in Section 6.1, with $\delta_D = 4.5$ and $\delta_Y = 5$ throughout. Different DGPs exhibit different sensitivity to these values. Results using penalties chosen via 10-fold cross-validation appear in Fig. 2, which also exhibits excellent coverage across all sparse designs.[20]

7.2. Empirical application

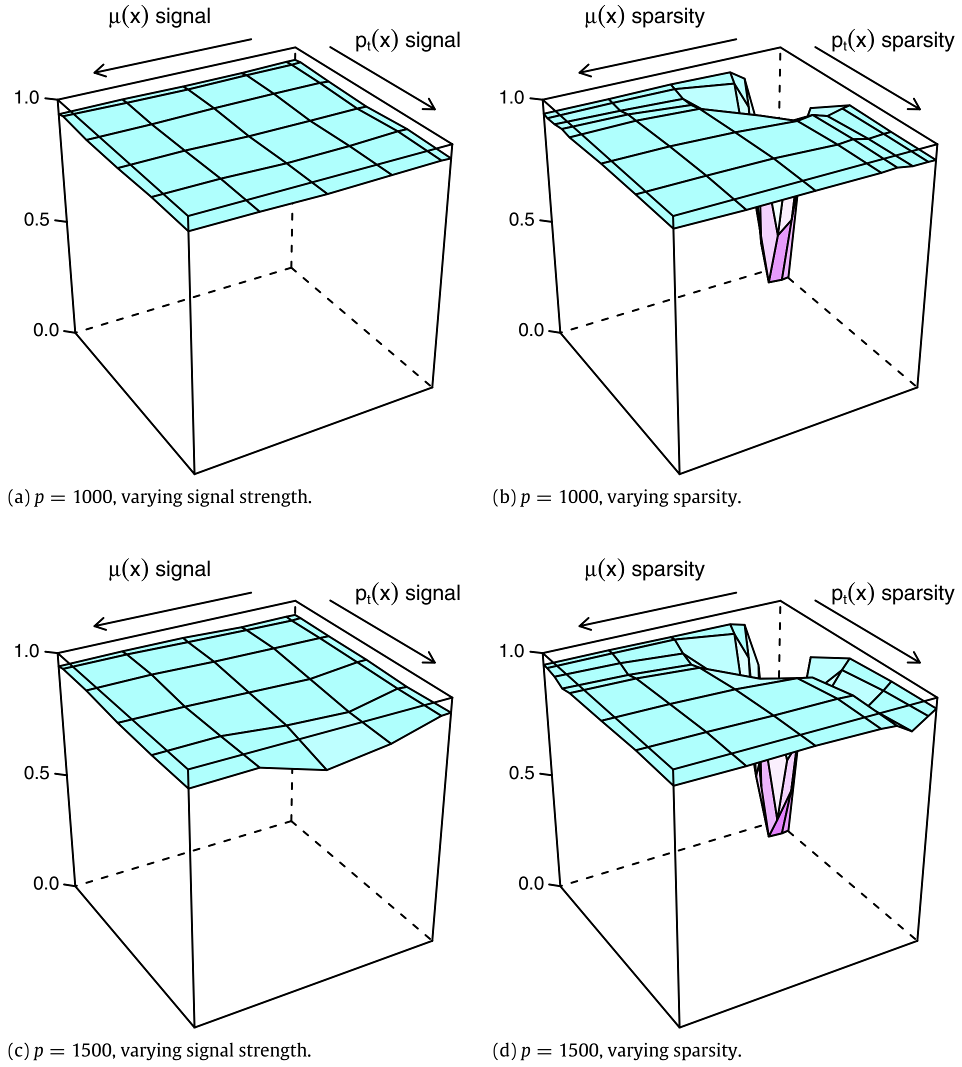
To illustrate the role that model selection can play in a real-world application, we revisit the National Supported Work (NSW) demonstration. The NSW has been analyzed numerous times since LaLonde (1986). Our aim is a simple study of model selection, not a comprehensive or conclusive evaluation of the NSW. We focus on the subsample used by Dehejia and Wahba (1999) and the Panel Study of Income Dynamics (PSID) comparison sample, taking as given their data definitions, sample selection, and trimming rules. Detailed discussion of these choices, and the NSW program may be found in Dehejia and Wahba (1999, 2002) (hereafter DW99 and DW02) and Smith and Todd (2005), and references therein. Briefly, the outcome of interest is earnings following a job training program. The dataset includes a treatment indicator, post-treatment earnings (1978), two years of pre-treatment earnings (1974[21] and 1975), as well as age, education, a marital status, and indicators for Black and Hispanic. Thus, X consists of seven variables. We will keep the estimator fixed: all estimates will be based on the doubly-robust estimator with standard errors from Section 5.2. We will compare the following specifications for X^*:

1. **No Selection**: X, $(\text{earn}1974)^2$, $(\text{earn}1975)^2$, $(\text{age})^2$, and $(\text{educ})^2$.
2. **Informally Selected:** The above, plus $\mathbb{1}\{\text{educ} < \text{HS}\}$, $\mathbb{1}\{\text{earn}1974 = 0\}$, $\mathbb{1}\{\text{earn}1975 = 0\}$, and $(\mathbb{1}\{\text{earn}1974 = 0\} \times \text{Hispanic})$. This specification was selected by DW02 using an informal balance test.

[19] The supplemental Appendix contains the additional results (see Appendix D).

[20] The R routines appear unstable for nonsparse designs, thus the analogues to panels (b) and (d) of Fig. 1 are omitted. See the supplement for limited versions. This will be explored for future software development.

[21] This naming follows DW99, but the variable may be measured outside 1974, see discussion in the works cited.



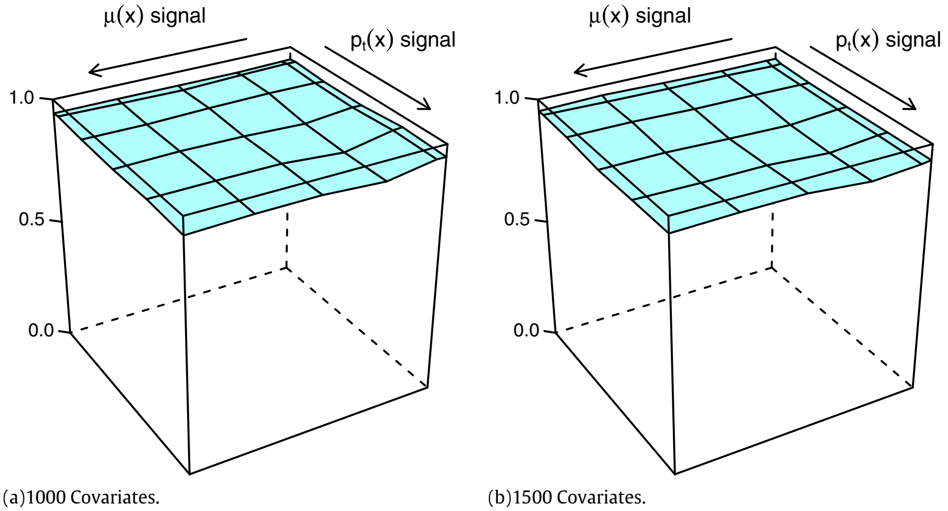
(a) $p = 1000$, varying signal strength.

(b) $p = 1000$, varying sparsity.

(c) $p = 1500$, varying signal strength.

(d) $p = 1500$, varying sparsity.

Fig. 1. Empirical coverage of 95% confidence intervals, varying signal strength and sparsity of $p_t(x)$ and $\mu_t(x)$.



(a)1000 Covariates.

(b)1500 Covariates.

Fig. 2. Empirical coverage of 95% confidence intervals, penalty chosen with cross-validation, varying signal strength of $p_t(x)$ and $\mu_t(x)$.

3. **Group Lasso Selection:** X, $\mathbb{1}\{\text{educ} < \text{HS}\}$, $\mathbb{1}\{\text{earn1974} = 0\}$, $\mathbb{1}\{\text{earn1975} = 0\}$, all possible first-order interactions, and all polynomials up to order five of the continuous covariates (age, educ, earn1974, earn1975).

For specifications 1 and 2, the same covariates are in the outcome and treatment models. All specifications include an intercept and we include education and pre-treatment income in the refitting step following model selection. We follow DW99 and DW02 and trim comparisons with estimated propensity score larger (smaller) than the maximum (minimum) in the treated sample.[22]

Table 1 presents results from these three specifications, and includes the experimental arm of the NSW. The group lasso based estimate performs very well: the point estimate is accurate and the interval is tight. Selecting from 171 possible covariates allows for a great deal of flexibility, but the sparsity of the estimate keeps the variance well-controlled. The no-selection point estimate is accurate, but fails to yield significance, while the specification of DW02 yields a significant, but overly high estimate and wide confidence interval. The benefits of explicit model selection are clear.

8. Discussion

This paper proposed a method that achieves uniformly valid inference on mean effects of a multivalued treatment even after model selection among possibly more covariates than observations. We demonstrated robustness to model selection errors, misspecification, and heterogeneous effects in observables. To accomplish this, a doubly-robust estimator was employed and shown to have excellent properties following model selection. We proved new results on group lasso estimation, which we argue is natural for treatment effects data. Multinomial logistic regression was studied in some detail. Numerical evidence shows that our method is quite promising for applications.

A key outstanding question in this work and in the high-dimensional, sparse modeling literature more generally, is penalty parameter choice. Very little work has been done in this area, which is a crucial gap in implementability of these techniques. We plan to develop a formal choice for the penalty parameter that is appropriately optimal. Tuning parameter selection in semi- and nonparametric analysis, and its impact on estimation and inference, is becoming better understood, and parallel developments must take place in model selection contexts.

Appendix A. Proofs for treatment effect inference

The proofs in this section are asymptotic. Order symbols hold for the sequence being considered, as a shorthand for the more formal versions given in e.g. Assumption 3. C will denote a generic positive constant, which may be a matrix. Define the set of indexes $\mathbb{I}_t = \{i : d_i = t\}$. The online supplement contains much greater detail. We make frequent use of the linearization

$$\frac{1}{a} = \frac{1}{b} + \frac{b-a}{ab} = \frac{1}{b} + \frac{b-a}{b^2} + \frac{(b-a)^2}{ab^2}. \tag{A.1}$$

Proof of Theorem 2. See supplemental Appendix. □

Proof of Theorem 3.1 without Additional Randomness. With $\psi_t(\cdot)$ defined in Eq. (2), we have $\sqrt{n}(\hat{\mu}_t - \mu_t) = \sqrt{n}\mathbb{E}_n[\psi_t(y_i, d_i^t, \mu_t(x_i), p_t(x_i), \mu_t)] + R_1 + R_2$, where

$$R_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} d_i^t (y_i - \mu_t(x_i)) \left(\frac{1}{\hat{p}_t(x_i)} - \frac{1}{p_t(x_i)} \right)$$

and

$$R_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{\mu}_t(x_i) - \mu_t(x_i)) \left(1 - \frac{d_i^t}{\hat{p}_t(x_i)} \right).$$

The proof proceeds by showing that both R_1 and R_2 are $o_{P_n}(1)$. Applying the first equality in Eq. (A.1), we rewrite R_1 as

$$R_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} d_i^t u_i \left(\frac{p_t(x_i) - \hat{p}_t(x_i)}{\hat{p}_t(x_i) p_t(x_i)} \right).$$

Applying Assumptions 1(b) and 2(c) and the first-stage consistency condition of Assumption 3(a):

$$\begin{aligned} \mathbb{E}\left[R_1^2 | \{x_i, d_i\}_{i=1}^n\right] &= \mathbb{E}_n \left[\frac{d_i^t \sigma_t^2(x_i)}{\hat{p}_t(x_i)^2 p_t(x_i)^2} \left(p_t(x_i) - \hat{p}_t(x_i)\right)^2 \right] \\ &\leq C\mathbb{E}_n[(p_t(x_i) - \hat{p}_t(x_i))^2] = o_{P_n}(1). \end{aligned}$$

Next, again using Eq. (A.1), we have $R_2 = R_{21} + R_{22}$, where

$$R_{21} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{\mu}_t(x_i) - \mu_t(x_i)) \left(\frac{p_t(x_i) - d_i^t}{p_t(x_i)} \right)$$

and

$$R_{22} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{\mu}_t(x_i) - \mu_t(x_i))(\hat{p}_t(x_i) - p_t(x_i)) \left(\frac{d_i^t}{\hat{p}_t(x_i) p_t(x_i)} \right).$$

For the first term $\mathbb{E}\left[R_{21}^2 | \{x_i\}_{i=1}^n\right] \leq C\mathbb{E}_n\left[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2\right] = o_{P_n}(1)$, by the first-stage consistency condition of Assumption 3(a). Next,

$$\begin{aligned} |R_{22}| \leq \sqrt{n} &\left(\max_{i \leq n} \frac{1}{\hat{p}_t(x_i) p_t(x_i)} \right) \\ &\times \sqrt{\mathbb{E}_n[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2] \mathbb{E}_n[(\hat{p}_t(x_i) - p_t(x_i))^2]} = o_{P_n}(1). \end{aligned}$$

by Hölder's inequality, Assumption 1(b) and the rate condition of Assumption 3(b). □

Proof of Theorem 3.1 with Additional Randomness. We must reconsider the remainders R_1 and R_2. For the former, applying Eq. (A.1), we find $R_1 = R_{11} + R_{12}$, where

$$R_{11} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d_i^t u_i}{p_t(x_i)^2} \left(p_t(x_i) - \hat{p}_t(x_i)\right) \quad \text{and}$$

$$R_{12} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d_i^t u_i}{p_t(x_i)^2 \hat{p}_t(x_i)} \left(\hat{p}_t(x_i) - p_t(x_i)\right)^2.$$

For R_{11}, we first add and subtract the parametric representation to get $R_{11} = R_{111} + R_{112}$, where,

$$R_{111} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d_i^t u_i}{p_t(x_i)^2} \left(\hat{p}_t(\{x_i^{*\prime} \gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - \hat{p}_t(\{x_i^{*\prime} \hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})\right) \quad \text{and}$$

$$R_{112} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d_i^t u_i}{p_t(x_i)^2} \left(p_t(x_i) - \hat{p}_t(\{x_i^{*\prime} \gamma_t^*\}_{\mathbb{N}_\mathcal{T}})\right).$$

By a two-term mean-value expansion $R_{111} = R_{111a} + R_{111b}$, with

$$\begin{aligned} R_{111a} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d_i^t u_i}{p_t(x_i)^2} \sum_{t \in \mathbb{N}_\mathcal{T}} &\left\{ \hat{p}_t(\{x_i^{*\prime} \gamma_t^*\}_{\mathbb{N}_\mathcal{T}})(1 - \hat{p}_t(\{x_i^{*\prime} \gamma_t^*\}_{\mathbb{N}_\mathcal{T}})) \right. \\ &\times \left. \left(x_i^{*\prime}(\hat{\gamma}_t - \gamma_t^*)\right) \right\} \end{aligned}$$

$$\text{and} \quad R_{111b} = \frac{1}{2\sqrt{n}} \sum_{i=1}^{n} \frac{d_i^t u_i}{p_t(x_i)^2} \upsilon_i' \bar{\mathcal{H}} \upsilon_i,$$

where $\upsilon_i = \{x_i^{*\prime}(\hat{\gamma}_t - \gamma_t^*)\}_{\mathbb{N}_\mathcal{T}}$ and $\overline{\mathcal{H}} = \mathcal{H}(\{x_i^{*\prime} \gamma_t^* + m_t x_i^{*\prime} \hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})$ for appropriate scalars m_t and the $\mathcal{T}$-square Hessian matrix $\mathcal{H}(\{x_i^{*\prime} \gamma_t\}_{\mathbb{N}_\mathcal{T}})$ (defined in Appendix B).

[22] A formal treatment of trimming is beyond the scope of the present study. The goal of our analysis is illustrative, and hence we take DW99's trimming as given. This issue is discussed by DW99, DW02, and Smith and Todd (2005).

For R_{111a}, consider each term in the sum over $\mathbb{N}_{\mathcal{T}}$ one at a time; let $R_{111a} = \sum_{t \in \mathbb{N}_{\mathcal{T}}} R_{111a}(t)$. Let t' denote the original treatment under consideration. Define

$$\Sigma_{t,j} = \mathbb{E}\left[(x^*_{i,j})^2 \sigma^2_{t'}(x_i) \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})^2 \right.$$
$$\left. \times (1 - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}))^2 / p_{t'}(x_i)^3 \right].$$

Then proceed as follows:

$$\begin{aligned}
&R_{111a}(t) \\
&= \sum_{j \in \hat{S}_D} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(x^*_{i,j} \frac{d_i^{t'} u_i \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})(1 - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}))}{p_{t'}(x_i)^2 \Sigma_{t,j}^{1/2}} \right) \right\} \\
&\quad \times \Sigma_{t,j}^{1/2} (\hat{\gamma}_{t,j} - \gamma^*_{t,j}) \\
&\leq \left(\max_{j \in \mathbb{N}_p} \Sigma_{t,j}^{1/2} \right) \\
&\quad \times \left(\max_{j \in \mathbb{N}_p} \frac{1}{\sqrt{n}} \sum_{i=1}^n x^*_{i,j} \frac{d_i^{t'} u_i \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})(1 - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}))}{p_{t'}(x_i)^2 \Sigma_{t,j}^{1/2}} \right) \\
&\quad \times \left\| \hat{\gamma}_t - \gamma_t^* \right\|_1 \\
&= O(1) O_{P_n}(\log(p)) \left\| \hat{\gamma}_t - \gamma_t^* \right\|_1 = o_{P_n}(1).
\end{aligned}$$

Convergence follows under Assumption 5. For the penultimate equality, it follows from Assumptions 1(b), 2(b) and 2(c) that $\max_{j \in \mathbb{N}_p} \Sigma_{t,j} = O(1)$. Finally, the center factor is shown to be $O_{P_n}(\log(p))$ by applying the moderate deviation theory for self-normalized sums of de la Peña et al. (2009, Theorem 7.4) and in particular Belloni et al. (2012, Lemma 5). To apply this lemma, first note that the summand of the center factor has bounded third moment and second moment bounded away from zero, from Assumptions 1(b), 2(b), 2(c), and the requirements of Assumptions 3 and 5. $\Sigma_{t,j}$ normalizes the second moment, and the lemma applies under Assumption 4 and the first restriction of Assumption 5.

For R_{111b}, the results of Tanabe and Sagae (1992) coupled with Assumption 3 give $v_i' \bar{\mathcal{H}} v_i \leq C \|v_i\|_2^2$. Thus, using Assumption 1(b) to bound $\max_{i \leq n} p_t(x_i)^{-2} < C$, we find R_{111b} may be bounded as follows:

$$\begin{aligned}
|R_{111b}| &\leq C \sum_{t \in \mathbb{N}_{\mathcal{T}}} \sqrt{n} (\max_{i \in \mathbb{I}_t} |u_i|) \mathbb{E}_n \left[|x_i^{*\prime}(\hat{\gamma}_t - \gamma_t^*)|^2 \right] \\
&\leq C\mathcal{T} \max_{t \in \mathbb{N}_{\mathcal{T}}} \left| \sqrt{n} (\max_{i \in \mathbb{I}_t} |u_i|) \mathbb{E}_n \right. \\
&\qquad \times \left. \left[|\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})|^2 \right] \right| = o_{P_n}(1),
\end{aligned}$$

by the union bound and Assumption 5, using Assumptions 1(b) and 3(a) to apply Eq. (B.15) with the inequality reversed.

A variance bound may be applied to R_{112} as in the previous proof, and we have $|R_{112}| = O_{P_n}(b_s) = o_{P_n}(1)$ by Markov's inequality.

Next, R_{12} is simply bounded by

$$\begin{aligned}
|R_{12}| &\leq \sqrt{n} (\max_{i \in \mathbb{I}_t} |u_i|) \left(\max_{i \in \mathbb{I}_t} \frac{1}{p_t(x_i)^2 \hat{p}_t(x_i)} \right) \mathbb{E}_n \left[\left(\hat{p}_t(x_i) - p_t(x_i) \right)^2 \right] \\
&\leq O_{P_n}(1) \sqrt{n} (\max_{i \in \mathbb{I}_t} |u_i|) \mathbb{E}_n \left[\left(\hat{p}_t(x_i) - p_t(x_i) \right)^2 \right] = o_{P_n}(1),
\end{aligned}$$

where the rate follows from Assumptions 1(b), 2 and 3, and this tends to zero by Assumption 5.

As in the prior proof, write $R_2 = R_{21} + R_{22}$. The same bound is used for R_{22}. However, for R_{21}, add and subtract the pseudotrue values to get $R_{21} = R_{211} + R_{212}$, where

$$R_{211} = \frac{1}{\sqrt{n}} \sum_{i=1}^n (x_i^{*\prime} \hat{\beta}_t - x_i^* \beta_t^*) \left(\frac{p_t(x_i) - d_i^t}{p_t(x_i)} \right) \quad \text{and}$$

$$R_{212} = \frac{1}{\sqrt{n}} \sum_{i=1}^n (x_i^* \beta_t^* - \mu_t(x_i)) \left(\frac{p_t(x_i) - d_i^t}{p_t(x_i)} \right).$$

For the first term, define $\tilde{\Sigma}_{t,j} = \mathbb{E}\left[(x^*_{i,j})^2 (d_i^t - p_t(x_i))^2 / p_t(x_i)^2 \right]$ and then proceed as follows:

$$\begin{aligned}
R_{211} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{p_t(x_i) - d_i^t}{p_t(x_i)} \right) \sum_{j \in \hat{S}_Y} x^*_{i,j} (\hat{\beta}_{t,j} - \beta^*_{t,j}) \\
&= \sum_{j \in \hat{S}_Y} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{x^*_{i,j}(p_t(x_i) - d_i^t)/p_t(x_i)}{\tilde{\Sigma}_{t,j}^{1/2}} \right\} \tilde{\Sigma}_{t,j}^{1/2} (\hat{\beta}_{t,j} - \beta^*_{t,j}) \\
&\leq \left(\max_{j \in \mathbb{N}_p} \tilde{\Sigma}_{t,j}^{1/2} \right) \left(\max_{j \in \mathbb{N}_p} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{x^*_{i,j}(p_t(x_i) - d_i^t)/p_t(x_i)}{\tilde{\Sigma}_{t,j}^{1/2}} \right) \\
&\quad \times \left\| \hat{\beta}_t - \beta_t^* \right\|_1 \\
&= O(1) O_{P_n}(\log(p)) \left\| \hat{\beta}_t - \beta_t^* \right\|_1 = o_{P_n}(1),
\end{aligned}$$

where the final line follows exactly as above. A variance bound may be applied to R_{212} as in the previous proof, and we have $|R_{212}| = O_{P_n}(b_s) = o_{P_n}(1)$ by Markov's inequality. □

Proof of Theorem 3.2. This follows from the prior result and Assumption 2(e). □

Proof of Theorem 3.3. We begin with $\hat{V}_W(t)$. Expanding the square and using Eq. (A.1), rewrite $\hat{V}^W_{\boldsymbol{\mu}}(t) = \mathbb{E}_n[d_i^t u_i^2 p_t(x_i)^{-2}] + R_{W,1} + R_{W,2} + R_{W,3}$ where

$$R_{W,1} = \mathbb{E}_n \left[\frac{d_i^t u_i^2}{\hat{p}_t(x_i)^2 p_t(x_i)^2} \left(\hat{p}_t(x_i) - p_t(x_i) \right) \left(\hat{p}_t(x_i) + p_t(x_i) \right) \right],$$

$$R_{W,2} = \mathbb{E}_n \left[\frac{d_i^t (\mu_t(x_i) - \hat{\mu}_t(x_i))^2}{\hat{p}_t(x_i)^2} \right], \quad \text{and}$$

$$R_{W,3} = 2\mathbb{E}_n \left[\frac{d_i^t u_i (\mu_t(x_i) - \hat{\mu}_t(x_i))}{\hat{p}_t(x_i)^2} \right].$$

Using Hölder's inequality, Assumptions 1(b), 2(e) and 3(a), we have the following:

$$\begin{aligned}
R_{W,1} &\leq \left(\max_{i \in \mathbb{I}_t} \frac{\hat{p}_t(x_i) + p_t(x_i)}{\hat{p}_t(x_i)^2 p_t(x_i)^2} \right) \mathbb{E}_n[d_i^t |u_i|^4]^{1/2} \\
&\quad \times \mathbb{E}_n[d_i^t (\hat{p}_t(x_i) - p_t(x_i))^2]^{1/2} = o_{P_n}(1), \\
R_{W,2} &\leq \left(\max_{i \in \mathbb{I}_t} \frac{1}{\hat{p}_t(x_i)^2} \right) \mathbb{E}_n[d_i^t (\hat{\mu}_t(x_i) - \mu_t(x_i))^2] = o_{P_n}(1), \\
\text{and,} \quad R_{W,3} &\leq 2 \left(\max_{i \in \mathbb{I}_t} \frac{1}{\hat{p}_t(x_i)^2} \right) \mathbb{E}_n[d_i^t |u_i|^2]^{1/2} \\
&\quad \times \mathbb{E}_n[d_i^t (\hat{\mu}_t(x_i) - \mu_t(x_i))^2]^{1/2} = o_{P_n}(1),
\end{aligned}$$

where $\mathbb{E}_n[|u_i|^4] = O_{P_n}(1)$ from the inequality of von Bahr and Esseen (1965). From the same inequality it follows that $\mathbb{E}_n[d_i^t u_i^2 p_t(x_i)^{-2}] - V^W_{\boldsymbol{\mu}}(t)| = o_{P_n}(1)$, under Assumptions 1(b) and 2(c).

Next consider the "between" variance estimator, $\hat{V}^B_{\boldsymbol{\mu}}$. For any $t\overline{\mathbb{N}}_{\mathcal{T}}$ and $t' \in \overline{\mathbb{N}}_{\mathcal{T}}$, define

$$R_{B,1}(t, t') = \mathbb{E}_n \left[(\hat{\mu}_t(x_i) - \mu_t(x_i))(\hat{\mu}_{t'}(x_i) - \mu_{t'}(x_i)) \right],$$

$$R_{B,2}(t, t') = \hat{\mu}_t \mathbb{E}_n \left[\hat{\mu}_{t'}(x_i) - \mu_{t'}(x_i) \right], \quad \text{and}$$

$$R_{B,3}(t, t') = \mathbb{E}_n \left[\mu_t(x_i)(\hat{\mu}_{t'}(x_i) - \mu_{t'}(x_i)) \right].$$

From Hölder's inequality, Assumption 3(a), Theorem 3.2, the von Bahr and Esseen inequality, and Assumptions 2(c) and 2(e) it follows that $R_{B,k}(t,t') = o_{P_n}(1)$ for $k \in \mathbb{N}_3$ and all pairs $(t,t') \in \mathbb{N}_t^2$. With this in mind, we decompose

$$\begin{aligned}
&\hat{V}^B_{\boldsymbol{\mu}}(t,t') \\
&\quad = \mathbb{E}_n\left[\mu_t(x_i)\mu_{t'}(x_i)\right] - \hat{\mu}_t\mathbb{E}_n\left[\mu_{t'}(x_i)\right] - \hat{\mu}_{t'}\mathbb{E}_n\left[\mu_t(x_i)\right] + \hat{\mu}_t\hat{\mu}_{t'} \\
&\qquad + R_{B,1}(t,t') + R_{B,2}(t,t') \\
&\qquad + R_{B,2}(t',t) + R_{B,3}(t,t') + R_{B,3}(t',t).
\end{aligned}$$

Consistency of $\hat{V}^B_{\boldsymbol{\mu}}(t,t')$ now follows from the von Bahr and Esseen inequality and Theorem 3.2. □

Proof of Corollary 2. Suppose the result did not hold. Then, there would exist a subsequence $P_m \in \boldsymbol{P}_m$, for each m, such that

$$\lim_{m\to\infty}\left|\mathbb{P}_{P_m}\left[G(\boldsymbol{\mu}) \in \left\{G(\hat{\boldsymbol{\mu}}) \pm c_\alpha\sqrt{\nabla_G(\hat{\boldsymbol{\mu}})\hat{V}\nabla_G'(\hat{\boldsymbol{\mu}})/n}\right\}\right] - (1-\alpha)\right| > 0.$$

But this contradicts Theorem 3, under which $(\nabla_G(\hat{\boldsymbol{\mu}})\hat{V}\nabla_G'(\hat{\boldsymbol{\mu}})/n)^{-1/2}(G(\hat{\boldsymbol{\mu}}) - G(\boldsymbol{\mu}))$ is asymptotically standard normal under the sequence P_m. □

Appendix B. Proofs for group Lasso selection and estimation of multinomial logistic models

This section is nonasymptotic. We use generic notation X^*, δ, etc. The online supplement has greater detail (see Appendix D).

B.1. Lemmas

The following three lemmas are needed for the proofs of Theorems 5 and 6. Due to space considerations, only a short sketch of the proofs will be given, highlight the main ideas in each. Full details are available in the online supplement (see Appendix D).

Lemma B.1 (*Score Bound*). *For* λ_D *and* $\mathcal{P}$ *defined in Eqs.* (11) *and* (12) *we have*

$$\mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[(p_t(x_i) - d_i^t)x^*_{i,j}]\|_2 \geq \frac{\lambda_D}{2}\right] \leq \mathcal{P}.$$

Proof. The residuals $v_{t,i} = p_t(x_i) - d_i^t$ are conditionally mean-zero by definition and satisfy $\mathbb{E}[v^2_{t,i}|x_i] \leq 1$. Using this, Assumption 2(a), and the definition of $\mathcal{X}$, we find that $\mathbb{E}\left[\|\mathbb{E}_n[v_{t,i}x^*_{i,j}]\|_2^2\right] \leq \mathcal{X}^2\mathcal{T}/n$, uniformly in $j \in \mathbb{N}_p$. Define the mean-zero random variables $\xi_{t,j} = (\mathbb{E}_n[v_{t,i}x^*_{i,j}])^2 - \frac{1}{n}\mathbb{E}[V_t^2X_j^{*2}]$ and set $r_n = \mathcal{T}^{-1/2}\log(p\vee n)^{3/2+\delta}$. Then

$$\begin{aligned}
&\mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[(p_t(x_i) - d_i^t)x^*_{i,j}]\|_2 \geq \frac{\lambda_D}{2}\right] \\
&\quad \leq \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\sum_{t\in\mathbb{N}_\mathcal{T}}\xi_{t,j} \geq \frac{\mathcal{X}^2\mathcal{T}r_n}{n}\right] \leq \mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\sum_{t\in\mathbb{N}_\mathcal{T}}\xi_{t,j}\right|\right]\frac{n}{\mathcal{X}^2\mathcal{T}r_n}
\end{aligned}$$

where final line follows from Markov's inequality. Next, applying Lemma 9.1 of Lounici et al. (2011), Jensen's inequality, and Assumption 2(c), we find that

$$\begin{aligned}
&\mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\sum_{t\in\mathbb{N}_\mathcal{T}}\xi_{t,j}\right|\right] \\
&\quad \leq 4\log(2p)^{1/2}\left(\sum_{t\in\mathbb{N}_\mathcal{T}}\frac{\mathcal{X}^4}{n^2} + \sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\mathbb{E}_n[v_{t,i}x^*_{i,j}]\right|^4\right]\right)^{1/2}.
\end{aligned}$$

Again using Lemma 9.1 of Lounici et al. (2011), and Assumptions 2(a) and 2(b), we bound the expectation in the second term above as follows:

$$\mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\mathbb{E}_n[v_{t,i}x^*_{i,j}]\right|^4\right] \leq \frac{64\log(12p)^2\mathcal{X}^4}{n^2}.$$

Collecting these results proves the Lemma. □

Lemma B.2 (*Estimate Sparsity*). *With probability at least* $1 - \mathcal{P}$

$$|\tilde{S}^D| \leq \frac{4}{\lambda_D^2}\overline{\phi}\{Q,\tilde{S}^D\}\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2\right].$$

Proof. From the Karush–Kuhn–Tucker conditions for (9), for all $t \in \mathbb{N}_\mathcal{T}$, if $\tilde{\gamma}_{\cdot,j} \neq 0$ it must satisfy

$$\mathbb{E}_n[x^*_{i,j}(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)] = \lambda_D\frac{\tilde{\gamma}_{t,j}}{\|\tilde{\gamma}_{\cdot,j}\|_2}.$$

Taking the ℓ_2-norm over $t \in \mathbb{N}_\mathcal{T}$ for fixed $j \in \tilde{S}^D$, adding and subtracting the true propensity score, using the triangle inequality, the score bound (B.1), collecting terms, squaring both sides, and summing over $j \in \tilde{S}^D$ (i.e. applying $\|\cdot\|_2^2$ over $j \in \tilde{S}^D$ to both sides) yields

$$\begin{aligned}
\sum_{j\in\tilde{S}^D}\lambda_D^2 &\leq 4\sum_{j\in\tilde{S}^D}\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n[x^*_{i,j}(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))]^2 \\
&\leq 4\overline{\phi}\{Q,\tilde{S}^D\}\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2\right].
\end{aligned}$$

The result now follows, as the left-hand side is equal to $|\tilde{S}^D|\lambda_D^2$. □

Lemma B.3 (*Bounds in* ℓ_2/ℓ_1 *norm*). *With probability* $1 - \mathcal{P}$ *the vector* $\tilde{\delta}_{\cdot,\cdot} = \tilde{\gamma}_{\cdot,\cdot} - \gamma^*_{\cdot,\cdot}$ *satisfies* $\left\|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right\|_{2,1} \leq 5a_n$ *and* $\left\|\left|\tilde{\delta}_{\cdot,S_*}\right|\right\|_{2,1} \leq a_n$ *where* $a_n := \max\left\{\kappa_D^{-1}\sqrt{|S_*|}, 2\lambda_D^{-1}b_s^d\sqrt{\mathcal{T}}\right\}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}$.

Proof. By the Cauchy–Schwarz inequality and Lemma B.1,

$$\begin{aligned}
&\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(p_t(x_i) - d_i^t)x_i^{*\prime}\tilde{\delta}_t\right] \\
&\quad \leq \sum_{j\in\mathbb{N}_p}\sqrt{\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(p_t(x_i) - d_i^t)x^*_{i,j}\right]^2}\sqrt{\sum_{t\in\mathbb{N}_\mathcal{T}}\tilde{\delta}^2_{t,j}} \\
&\quad \leq \max_{j\in\mathbb{N}_p}\left\{\left\|\mathbb{E}_n\left[(p_t(x_i) - d_i^t)x^*_{i,j}\right]\right\|_2\right\}\sum_{j\in\mathbb{N}_p}\left\|\tilde{\delta}_{\cdot,j}\right\|_2 \leq \frac{\lambda_D}{2}\left\|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right\|_{2,1},
\end{aligned} \tag{B.1}$$

with probability at least $1 - \mathcal{P}$. Applying the Cauchy–Schwarz inequality, the bias condition of Assumption 4, and Cauchy–Schwarz again yields

$$\begin{aligned}
&\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))x_i^{*\prime}\tilde{\delta}_t\right] \\
&\quad \leq \sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2\right]^{1/2}\mathbb{E}_n\left[(x_i^{*\prime}\tilde{\delta}_t)^2\right]^{1/2} \\
&\quad \leq b_s^d\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(x_i^{*\prime}\tilde{\delta}_t)^2\right]^{1/2} \\
&\quad \leq b_s^d\sqrt{\mathcal{T}}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}.
\end{aligned} \tag{B.2}$$

Combining Eqs. (B.1) and (B.2), we have, probability at least $1-\mathscr{P}$,

$$\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})-d_i^t)x_i^{*\prime}\tilde{\delta}_t\right]=\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n\left[(p_t(x_i)-d_i^t)x_i^{*\prime}\tilde{\delta}_t\right]$$
$$+\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})-p_t(x_i))x_i^{*\prime}\tilde{\delta}_t\right]$$
$$\leq\frac{\lambda_D}{2}\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}+b_s^d\sqrt{\mathcal{T}}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}. \tag{B.3}$$

By the optimality of $\tilde{\delta}_{\cdot,\cdot}$, $\mathcal{M}(\gamma_{\cdot,\cdot}^*+\tilde{\delta}_{\cdot,\cdot})+\lambda_D\left|\left|\left|\gamma_{\cdot,\cdot}^*+\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\leq\mathcal{M}(\gamma_{\cdot,\cdot}^*)+\lambda_D\left|\left|\left|\gamma_{\cdot,\cdot}^*\right|\right|\right|_{2,1}$, and so

$$\lambda_D\left\{\left|\left|\left|\gamma_{\cdot,\cdot}^*\right|\right|\right|_{2,1}-\left|\left|\left|\gamma_{\cdot,\cdot}^*+\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\right\}\geq\mathcal{M}(\gamma_{\cdot,\cdot}^*+\tilde{\delta}_{\cdot,\cdot})-\mathcal{M}(\gamma_{\cdot,\cdot}^*)$$
$$\geq\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})-d_i^t)x_i^{*\prime}\tilde{\delta}_t\right],$$

applying the convexity of $\mathcal{M}$. Using the bound in Eq. (B.3) and rearranging we find that

$$0\leq\lambda_D\left\{\left|\left|\left|\gamma_{\cdot,\cdot}^*\right|\right|\right|_{2,1}-\left|\left|\left|\gamma_{\cdot,\cdot}^*+\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\right\}+\frac{\lambda_D}{2}\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}$$
$$+b_s^d\sqrt{\mathcal{T}}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}.$$

Dividing through λ_D and decomposing the supports, we find that

$$0\leq\frac{1}{2}\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}+\left\{\left|\left|\left|\gamma_{\cdot,\cdot}^*\right|\right|\right|_{2,1}-\left|\left|\left|\gamma_{\cdot,\cdot}^*+\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\right\}$$
$$+\frac{b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}$$
$$=\frac{1}{2}\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}+\frac{1}{2}\left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}+\left|\left|\left|\gamma_{\cdot,S_*}^*\right|\right|\right|_{2,1}$$
$$-\left|\left|\left|\gamma_{\cdot,S_*}^*+\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}-\left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}$$
$$+\frac{b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2},$$

because $\gamma_{\cdot,S_*^c}^*=0$. Collecting terms and applying the triangle inequality yields

$$\frac{1}{2}\left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}\leq\frac{1}{2}\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}+\left|\left|\left|\left|\gamma_{\cdot,S_*}^*\right|\right|\right|_{2,1}-\left|\left|\left|\gamma_{\cdot,S_*}^*+\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}\right|$$
$$+\frac{b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}$$
$$\leq\frac{1}{2}\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}+\left|\left|\left|\gamma_{\cdot,S_*}^*-\left(\gamma_{\cdot,S_*}^*+\tilde{\delta}_{\cdot,S_*}\right)\right|\right|\right|_{2,1}$$
$$+\frac{b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}$$
$$=\frac{1}{2}\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}+\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}+\frac{b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}.$$

Therefore with probability at least $1-\mathscr{P}$

$$\left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}\leq 3\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}+\frac{2b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}. \tag{B.4}$$

Consider two cases based on the upper bound in (B.4). First, suppose that $\tilde{\delta}_{\cdot,\cdot}$ obeys the cone constraint of Eq. (14) in the definition of κ_D^2. This implies

$$\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\leq 5\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}$$
$$\leq 5\sqrt{|S_*|}\left\|\tilde{\delta}_{\cdot,S_*}\right\|_2\leq\frac{5\sqrt{|S_*|}}{\kappa_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}, \tag{B.5}$$

by the Cauchy–Schwarz inequality, the restricted eigenvalue definition of Eq. (14), and noting that $\sum_{t\in\mathbb{N}_{\mathcal{T}}}\tilde{\delta}_t'Q\tilde{\delta}_t=\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]$. Collecting across the second and third inequalities yields

$$\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}\leq\frac{\sqrt{|S_*|}}{\kappa_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}. \tag{B.6}$$

On the other hand, if the cone constraint fails, then $\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}<\frac{1}{4}\left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}$. Using this for the first and third inequalities, and Eq. (B.4) for the second, we have

$$\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\leq\frac{5}{4}\left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}$$
$$\leq\frac{15}{4}\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}+\frac{5}{2}\frac{b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}$$
$$\leq\frac{15}{16}\left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}+\frac{5}{2}\frac{b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}.$$

Combining the right hand side of the first line with third lines yields $\left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}\leq 8b_s^d\sqrt{\mathcal{T}}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}/\lambda_D$. Plugging this back into the last line we obtain the bound

$$\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\leq 10\frac{b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}, \tag{B.7}$$

while instead, plugging it into the failure of the cone constraint yields

$$\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}\leq 2\frac{b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}. \tag{B.8}$$

Combining Eqs. (B.5) and (B.7) gives the first claim of the lemma and Eqs. (B.6) and (B.8) give the second. □

B.2. Proof of Theorem 5

Define $\tilde{\delta}_{\cdot,\cdot}=\tilde{\gamma}_{\cdot,\cdot}-\gamma_{\cdot,\cdot}^*$. By the optimality of $\tilde{\delta}_{\cdot,\cdot}$, we have

$$\mathcal{M}(\gamma_{\cdot,\cdot}^*+\tilde{\delta}_{\cdot,\cdot})+\lambda_D\left|\left|\left|\gamma_{\cdot,\cdot}^*+\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\leq\mathcal{M}(\gamma_{\cdot,\cdot}^*)+\lambda_D\left|\left|\left|\gamma_{\cdot,\cdot}^*\right|\right|\right|_{2,1}.$$

Rearranging and subtracting the score, we have

$$\mathcal{M}(\gamma_{\cdot,\cdot}^*+\tilde{\delta}_{\cdot,\cdot})-\mathcal{M}(\gamma_{\cdot,\cdot}^*)-\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})-d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t$$
$$\leq\lambda_D\left\{\left|\left|\left|\gamma_{\cdot,\cdot}^*\right|\right|\right|_{2,1}-\left|\left|\left|\gamma_{\cdot,\cdot}^*+\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\right\}$$
$$-\sum_{t\in\mathbb{N}_{\mathcal{T}}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}})-d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t. \tag{B.9}$$

The proof proceeds by deriving a further upper bound to the right and a quadratic lower bound of the left. The combination of these will yield a bound on $\mathbb{E}_n[(x_i^{*\prime}\tilde{\delta}_t)^2]^{1/2}$.

Begin with the right side of Eq. (B.9). For the penalized difference of coefficients we have $\left|\left|\left|\gamma_{\cdot,S_*^c}^*\right|\right|\right|_{2,1}-\left|\left|\left|\gamma_{\cdot,S_*^c}^*+\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}=$

$\lvert\lvert\lvert\tilde{\delta}_{\cdot,S_*^c}\rvert\rvert\rvert_{2,1}$, because $\gamma^*_{\cdot,S_*^c}=0$. Therefore,

$$\begin{aligned}
&\lvert\lvert\lvert\gamma^*_{\cdot,\cdot}\rvert\rvert\rvert_{2,1}-\lvert\lvert\lvert\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot}\rvert\rvert\rvert_{2,1}\\
&\quad=\lvert\lvert\lvert\gamma^*_{\cdot,S_*}\rvert\rvert\rvert_{2,1}-\lvert\lvert\lvert\gamma^*_{\cdot,S_*}+\tilde{\delta}_{\cdot,S_*}\rvert\rvert\rvert_{2,1}-\lvert\lvert\lvert\tilde{\delta}_{\cdot,S_*^c}\rvert\rvert\rvert_{2,1}\\
&\quad\leq\lvert\lvert\lvert\gamma^*_{\cdot,S_*}\rvert\rvert\rvert_{2,1}-\lvert\lvert\lvert\gamma^*_{\cdot,S_*}+\tilde{\delta}_{\cdot,S_*}\rvert\rvert\rvert_{2,1}\\
&\quad\leq\left|\lvert\lvert\lvert\gamma^*_{\cdot,S_*}\rvert\rvert\rvert_{2,1}-\lvert\lvert\lvert\gamma^*_{\cdot,S_*}+\tilde{\delta}_{\cdot,S_*}\rvert\rvert\rvert_{2,1}\right|\\
&\quad\leq\lvert\lvert\lvert\gamma^*_{\cdot,S_*}-\left(\gamma^*_{\cdot,S_*}+\tilde{\delta}_{\cdot,S_*}\right)\rvert\rvert\rvert_{2,1}=\lvert\lvert\lvert\tilde{\delta}_{\cdot,S_*}\rvert\rvert\rvert_{2,1},
\end{aligned}$$

where the first inequality reflects dropping the nonpositive final term (the norm is nonnegative) and the third inequality follows from the triangle inequality. Using this result for the first term and the bound (B.3) for the second, the right side of Eq. (B.9) is bounded by

$$\begin{aligned}
&\lambda_D\lvert\lvert\lvert\tilde{\delta}_{\cdot,S_*}\rvert\rvert\rvert_{2,1}+\frac{\lambda_D}{2}\lvert\lvert\lvert\tilde{\delta}_{\cdot,\cdot}\rvert\rvert\rvert_{2,1}+b_s^d\sqrt{\mathcal{T}}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}\\
&\quad\leq\left(\lambda_D\left\{\frac{\sqrt{|S_*|}}{\kappa_D}\vee\frac{2b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\right\}+\frac{\lambda_D}{2}\left\{\frac{5\sqrt{|S_*|}}{\kappa_D}\vee\frac{10b_s^d\sqrt{\mathcal{T}}}{\lambda_D}\right\}\right.\\
&\qquad\left.+\;b_s^d\sqrt{\mathcal{T}}\right)\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}\\
&\quad\leq\left(6\frac{\lambda_D\sqrt{|S_*|}}{\kappa_D}+8b_s^d\sqrt{\mathcal{T}}\right)\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2},
\end{aligned}\tag{B.10}$$

where the second inequality applies Lemma B.3 and the third bounds the maximum by the sum.

Now turn to the left side of Eq. (B.9). Our goal is to show that this is bounded below by a quadratic function. We apply the bounds for Bach's (2010) modified self-concordant functions. To show that $\mathcal{M}(\cdot)$ belongs to this class, we must bound the third derivative in terms of the Hessian. Recall that $\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})=\exp\{x_i^{*\prime}\gamma_t\}/\left(1+\sum_{\mathbb{N}_\mathcal{T}}\exp\{x_i^{*\prime}\gamma_t\}\right)$ and the $\mathcal{T}$-square matrix $\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})$ has $(t,t')\in\mathbb{N}_\mathcal{T}^2$ entry given by

$$\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})_{[t,t']}=\begin{cases}\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})(1-\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})) & \text{if } t=t'\\ -\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})\hat{p}_{t'}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}}) & \text{if } t\neq t'.\end{cases}$$

First, note that $\mathcal{M}(\gamma_{\cdot,\cdot})$ can be written as

$$\mathcal{M}(\gamma_{\cdot,\cdot})=\mathbb{E}_n\left[\log\left(1+\sum_{t\in\mathbb{N}_\mathcal{T}}\exp\{x_i^{*\prime}\gamma_t\}\right)-\sum_{t\in\mathbb{N}_\mathcal{T}}d_i^t(x_i^{*\prime}\gamma_t)\right].$$

Define $F:\mathbb{R}^\mathcal{T}\to\mathbb{R}$ as $F(w)=\log\left(1+\sum_{t\in\mathbb{N}_\mathcal{T}}\exp(w_t)\right)$, so that $\mathcal{M}(\gamma_{\cdot,\cdot})=\mathbb{E}_n\left[F(w_i)-\sum_{t\in\mathbb{N}_\mathcal{T}}d_i^t w_{i,t}\right]$, where $w_{i,t}=x_i^{*\prime}\gamma_t$ and $w_i=\{w_{i,t}\}_{\mathbb{N}_\mathcal{T}}$. Then for any $w\in\mathbb{R}^\mathcal{T}$, $v\in\mathbb{R}^\mathcal{T}$, and scalar α, define $g(\alpha)=F(w+\alpha v):\mathbb{R}\to\mathbb{R}$. We verify the conditions of Bach (2010, Lemma 1) for this $g(\alpha)$ and $F(w)$. This involves finding the third derivative of $g(\alpha)$, and bounding it in terms of the second (i.e. the Hessian). To this end, note that the multinomial function has the property that $\partial\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})/\partial\gamma_t=\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})(1-\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}}))x_i^*$ and $\partial\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})/\partial\gamma_{t',\cdot}=-\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})\hat{p}_{t'}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})x_i^*$. From these, we find

$$g'(\alpha)=v'F'(w+\alpha v)=\sum_{t\in\mathbb{N}_\mathcal{T}}v_t\hat{p}_t(w+\alpha v)\quad\text{and}$$

$$g''(\alpha)=v'F''(w+\alpha v)v=v'\mathcal{H}(w+\alpha v)v.$$

To bound $g'''(\alpha)$, we again use the derivatives of $\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})$ to find the derivatives of elements $\mathcal{H}(w)$. Routine calculations give, for any $r\neq s\neq t$:

$$\begin{aligned}
\partial\mathcal{H}(w)_{t,t}/\partial w_t&=\hat{p}_t(w)(1-\hat{p}_t(w))(1-2\hat{p}_t(w))\\
&=\mathcal{H}(w)_{t,t}(1-2\hat{p}_t(w))\\
\partial\mathcal{H}(w)_{t,t}/\partial w_r&=-\hat{p}_t(w)\hat{p}_r(w)(1-\hat{p}_t(w))+\hat{p}_t(w)^2\hat{p}_r(w)\\
&=\mathcal{H}(w)_{t,t}(\hat{p}_t(w)\hat{p}_r(w)(1-\hat{p}_t(w))^{-1}-\hat{p}_r(w))\\
\partial\mathcal{H}(w)_{t,s}/\partial w_t&=-\hat{p}_t(w)\hat{p}_s(w)(1-2\hat{p}_t(w))\\
&=\mathcal{H}(w)_{t,s}(1-2\hat{p}_t(w))\\
\partial\mathcal{H}(w)_{t,s}/\partial w_r&=-\hat{p}_t(w)\hat{p}_s(w)(-2\hat{p}_r(w))\\
&=\mathcal{H}(w)_{t,s}(-2\hat{p}_r(w)).
\end{aligned}$$

Each derivative returns the same Hessian element multiplied by term bounded by 2 in absolute value. Let a_r represent this factor. Then we bound

$$\begin{aligned}
g'''(\alpha)&=\left|\sum_{r\in\mathbb{N}_\mathcal{T}}v_r\left.\frac{\partial v'\mathcal{H}(\tilde{w})v}{\partial w_r}\right|_{\tilde{w}=w+\alpha v}\right|\\
&=\left|\sum_{r\in\mathbb{N}_\mathcal{T}}v_r v'\mathcal{H}(w+\alpha v)va_r\right|\\
&\leq\sum_{r\in\mathbb{N}_\mathcal{T}}v'\mathcal{H}(w+\alpha v)v|v_r||a_r|\leq 2v'\mathcal{H}(w+\alpha v)v\sum_{r\in\mathbb{N}_\mathcal{T}}|v_r|\\
&=2\|v\|_1 g''(\alpha)\leq 2\sqrt{\mathcal{T}}\|v\|_2 g''(\alpha).
\end{aligned}$$

Applying Bach's (2010) Lemma 1 to each observation, as in Belloni et al. (2013), with $w_i=\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}$ and $v_i=\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}$ we get the lower bound

$$\begin{aligned}
&M(\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot})-\mathcal{M}(\gamma^*_{\cdot,\cdot})-\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t\\
&\quad\geq\mathbb{E}_n\left[\frac{v_i'\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})v_i}{4\mathcal{T}\|v_i\|_2^2}\left(e^{-2\|v_i\|_2}+2\|v_i\|_2-1\right)\right]\\
&\quad\geq\mathbb{E}_n\left[\frac{v_i'\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})v_i}{4\mathcal{T}\|v_i\|_2^2}\left(2\|v_i\|_2^2-\frac{4}{3}\|v_i\|_2^3\right)\right],
\end{aligned}\tag{B.11}$$

where the second inequality follows from Belloni et al. (2013, Lemma 9).

Tanabe and Sagae (1992, Theorem 1) give $\mathcal{H}(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})\geq\phi_{\min}\{\mathcal{H}(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})\}\mathcal{I}_\mathcal{T}$, in the positive definite sense, where $\phi_{\min}(A)$ denotes the smallest eigenvalue of A and $\mathcal{I}_T$ is the $\mathcal{T}\times\mathcal{T}$ identity matrix. Then

$$\begin{aligned}
\phi_{\min}\{\mathcal{H}(\{x_i^*\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})\}&\geq\det\{\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})\}\\
&=\prod_{t\in\overline{\mathbb{N}}_\mathcal{T}}\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})\geq\left(p_{\min}/A_p\right)^{\overline{\mathcal{T}}},
\end{aligned}$$

where $p_0(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})=1-\sum_{t\in\mathbb{N}_\mathcal{T}}\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})$ and the first inequality is also due to Tanabe and Sagae (1992). These results imply that $v_i'\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})v_i\geq(p_{\min}/A_p)^{\overline{\mathcal{T}}}v_i'\mathcal{I}_\mathcal{T}v_i=(p_{\min}/A_p)^{\overline{\mathcal{T}}}\|v_i\|_2^2$ and therefore

$$\begin{aligned}
&\mathbb{E}_n\left[\frac{v_i'\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})v_i}{4\mathcal{T}\|v_i\|_2^2}\left(2\|v_i\|_2^2-\frac{4}{3}\|v_i\|_2^3\right)\right]\\
&\quad\geq\left(p_{\min}/A_p\right)^{\overline{\mathcal{T}}}\frac{1}{4\mathcal{T}}\mathbb{E}_n\left[2\|v_i\|_2^2-\frac{4}{3}\|v_i\|_2^3\right]\\
&\quad=\left(p_{\min}/A_p\right)^{\overline{\mathcal{T}}}\frac{1}{\mathcal{T}}\frac{\mathbb{E}_n[\|v_i\|_2^2]}{2}\left(1-\frac{2}{3}\frac{\mathbb{E}_n[\|v_i\|_2^3]}{\mathbb{E}_n[\|v_i\|_2^2]}\right).
\end{aligned}\tag{B.12}$$

Recall that $v_i=\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}$. To prove a quadratic lower bound, consider two cases, depending on whether

$$\frac{1}{2}\left(1-\frac{2}{3}\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^3]}{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}\right)$$

is above or below $1/A_K$. In the first case, combining equations (B.11) and (B.12) gives

$$\mathcal{M}(\gamma^*_{\cdot,\cdot} + \tilde{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) - \sum_{t \in \mathbb{N}_{\mathcal{T}}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) - d_i^t) x_i^{*\prime} \right] \tilde{\delta}_t$$
$$\geq (p_{\min}/A_p)^{\overline{\mathcal{T}}} \frac{1}{\mathcal{T}} \frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]}{A_K}. \tag{B.13}$$

Now consider the second case, where this bound does not hold. By Assumption 2(b), the Cauchy–Schwarz inequality, and the conclusion of Lemma B.3

$$\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_1 = \sum_{t \in \mathbb{N}_{\mathcal{T}}} \sum_{j \in \mathbb{N}_p} \left| x^*_{i,j} \tilde{\delta}_{t,j} \right| \leq \mathcal{X} \left\| \tilde{\delta}_{\cdot,\cdot} \right\|_1 \leq \sqrt{\mathcal{T}} \mathcal{X} \left\| \left| \tilde{\delta}_{\cdot,\cdot} \right| \right\|_{2,1}$$
$$\leq \sqrt{\mathcal{T}} \mathcal{X} \left\{ \frac{5\sqrt{|S_*|}}{\kappa_D} \vee \frac{10 b_s^d \sqrt{\mathcal{T}}}{\lambda_D} \right\} \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}.$$

Hence, by subadditivity (to bound the ℓ_2 norm by the ℓ_1 norm),

$$\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^3] \leq \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2 \|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_1]$$
$$\leq \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{3/2} \sqrt{\mathcal{T}} \mathcal{X} \left\{ \frac{5\sqrt{|S_*|}}{\kappa_D} \vee \frac{10 b_s^d \sqrt{\mathcal{T}}}{\lambda_D} \right\}.$$

Thus

$$\frac{1}{A_K} > \frac{1}{2} \left(1 - \frac{2}{3} \frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^3]}{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]} \right)$$
$$\geq \frac{1}{2} \left(1 - \frac{2}{3} \frac{\mathcal{X}\sqrt{\mathcal{T}}}{\kappa_D \lambda_D} \left(5\lambda_D \sqrt{|S_*|} + 10 \kappa_D b_s^d \sqrt{\mathcal{T}} \right) \right.$$
$$\left. \times \; \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2} \right),$$

which is equivalent to

$$\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}$$
$$> \left(1 - \frac{2}{A_K} \right) \frac{3}{2} \frac{\kappa_D \lambda_D}{\mathcal{X}\sqrt{\mathcal{T}}} \left(5\lambda_D \sqrt{|S_*|} + 10 \kappa_D b_s^d \sqrt{\mathcal{T}} \right)^{-1} := r_n.$$

Because $\mathcal{M}(\gamma^*_{\cdot,\cdot} + \delta_{\cdot,\cdot}) - \mathcal{M}(\gamma_{\cdot,\cdot}) - \sum_{t \in \mathbb{N}_{\mathcal{T}}} \mathbb{E}_n\Big[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) - d_i^t)$ $x_i^{*\prime}\Big]\delta_t$ is convex in $\delta_{\cdot,\cdot}$, and hence any line segment lies above the function, we know that $\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2} > r_n$, so we have

$$\mathcal{M}(\gamma^*_{\cdot,\cdot} + \tilde{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma_{\cdot,\cdot}) - \sum_{t \in \mathbb{N}_{\mathcal{T}}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) - d_i^t) x_i^{*\prime} \right] \tilde{\delta}_t$$
$$\geq r_n^2 \geq r_n^2 \frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}}{r_n}$$
$$= r_n \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}.$$

Combining this result with Eqs. (B.9) and (B.10), we have

$$\left(1 - \frac{2}{A_K} \right) \frac{3}{2} \frac{\kappa_D \lambda_D}{\mathcal{X}\sqrt{\mathcal{T}}} \left(5\lambda_D \sqrt{|S_*|} + 10 \kappa_D b_s^d \sqrt{\mathcal{T}} \right)^{-1}$$
$$\times \mathbb{E}_n[\|\{x_i^{*\prime}\delta_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}$$
$$\leq \left(6 \frac{\lambda_D \sqrt{|S_*|}}{\kappa_D} + 8 b_s^d \sqrt{\mathcal{T}} \right) \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2},$$

which is impossible under the restriction on A_K. Therefore, Eq. (B.13) must hold.[23] Combining this with Eqs. (B.9) and (B.10), we find that

$$(p_{\min}/A_p)^{\overline{\mathcal{T}}} \frac{1}{\mathcal{T}} \frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]}{A_K}$$
$$\leq \left(6 \frac{\lambda_D \sqrt{|S_*|}}{\kappa_D} + 8 b_s^d \sqrt{\mathcal{T}} \right) \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}.$$

Thus, dividing through and applying the union bound we find that

$$\max_{t \in \mathbb{N}_{\mathcal{T}}} \mathbb{E}_n[(x_i^{*\prime}\tilde{\delta}_t)^2]^{1/2} \leq \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_{\mathcal{T}}}\|_2^2]^{1/2}$$
$$\leq (A_p/p_{\min})^{\overline{\mathcal{T}}} \mathcal{T} A_K \left(6 \frac{\lambda_D \sqrt{|S_*|}}{\kappa_D} + 8 b_s^d \sqrt{\mathcal{T}} \right). \tag{B.14}$$

To bound the propensity score error, we apply the mean value theorem and the form of $\partial \hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_{\mathcal{T}}})/\partial \gamma_t$. We must linearize with respect to t only (recall that $\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}})$ depends on all of $\tilde{\gamma}_{\cdot,\cdot}$). To this end, define M_t as the $\mathcal{T}$-vector with entry t given by $x_i^{*\prime}\gamma_t^* + \tilde{m}_t x_i^{*\prime}\tilde{\gamma}_t$ for a scalar $\tilde{m}_t \in [0, 1]$ and entries $t' \in \mathbb{N}_{\mathcal{T}} \setminus \{t\}$ equal to $x_i^{*\prime}\gamma_{t'}$. Then we have

$$\left| \hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) \right| = \left| \hat{p}_t(M_t)[1 - \hat{p}_t(M_t)] x_i^{*\prime}\tilde{\delta}_t \right|$$
$$\leq \left| x_i^{*\prime}\tilde{\delta}_t \right|. \tag{B.15}$$

Using this result coupled with the triangle inequality, the bias condition, and Eq. (B.14), we find

$$\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - p_t(x_i))^2]^{1/2}$$
$$\leq \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}))^2]^{1/2}$$
$$+ \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_{\mathcal{T}}}) - p_t(x_i))^2]^{1/2}$$
$$\leq \mathbb{E}_n\left[(x_i^{*\prime}\tilde{\delta}_t)^2 \right]^{1/2} + b_s^d$$
$$\leq (A_p/p_{\min})^{\overline{\mathcal{T}}} \mathcal{T} A_K \left(6 \frac{\lambda_D \sqrt{|S_*|}}{\kappa_D} + 8 b_s^d \sqrt{\mathcal{T}} \right) + b_s^d.$$

The ℓ_1 bound follows from Eq. (B.14), the Cauchy–Schwarz inequality, and Eq. (16):

$$\left\| \tilde{\gamma}_t - \gamma_t^* \right\|_1 \leq \sqrt{|\tilde{S}^D \cup S_D^*|} \left\| \tilde{\gamma}_t - \gamma_t^* \right\|_{2,p}$$
$$\leq \left(\frac{|\tilde{S}^D \cup S_D^*|}{\underline{\phi}\{Q, \tilde{S}^D \cup S_D^*\}} \right)^{1/2} \mathbb{E}_n[(x_i^{*\prime}(\tilde{\gamma}_t - \gamma_t^*))^2]^{1/2}.$$

Finally, we bound the size of the selected set of coefficients. First, note that optimality of $\tilde{\gamma}_{\cdot,\cdot}$ ensures that $|\tilde{S}^D| \leq n$. Then, restating the conclusion Lemma B.2 using the notation of the Theorem and the rate result (B.14), then bounding $\overline{\phi}$ by $\overline{\overline{\phi}}$ we find that

$$|\tilde{S}^D| \leq |S_D^*| 4 L_n \overline{\overline{\phi}}\{Q, |\tilde{S}^D|\}.$$

The argument now parallels that used by Belloni and Chernozhukov (2013), relying on their result on the sublinearity of sparse eigenvalues. Let $\lceil m \rceil$ be the ceiling function and note that $\lceil m \rceil \leq 2m$. For any $m \in \mathbb{N}_Q^D$, suppose that $|\tilde{S}^D| > m$. Then,

$$|\tilde{S}^D| \leq |S_D^*| 4 L_n \overline{\overline{\phi}}\{Q, m(|\tilde{S}^D|/m)\}$$
$$\leq \left\lceil |\tilde{S}^D|/m \right\rceil |S_D^*| 4 L_n \overline{\overline{\phi}}\{Q, m\}$$
$$\leq (|\tilde{S}^D|/m) |S_D^*| 8 L_n \overline{\overline{\phi}}\{Q, m\}.$$

[23] This analysis is conceptually similar to using Belloni and Chernozhukov's (2011) restricted nonlinearity impact coefficient, but our characterization is different.

Rearranging gives $m \leq |S_D^*| 8 L_n \overline{\overline{\phi}}\{Q, m\}$ whence $m \notin \mathbb{N}_Q^D$. Minimizing over $\mathbb{N}_Q^D$ gives the result. □

B.3. Proof of Theorem 6

Define $\hat{\delta}_{\cdot,\cdot} = \hat{\gamma}_{\cdot,\cdot} - \gamma^*_{\cdot,\cdot}$. Many of the arguments parallel those for Theorem 5. The key differences are that a quadratic lower bound for $\mathcal{M}(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) - \sum_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t) x_i^{*\prime} \right] \hat{\delta}_t$ may occur, but is not necessary, and $\hat{\delta}_{\cdot,\cdot}$ may not belong to the cone of the restricted eigenvalues, but obeys the sparse eigenvalue constraints.

We first give a suitable upper bound for $\mathcal{M}(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) - \sum_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t) x_i^{*\prime} \right] \hat{\delta}_t$. By the Cauchy–Schwarz inequality and the definition of the sparse eigenvalues of Eq. (16),

$$
\begin{aligned}
\left|\left|\left| \hat{\delta}_{\cdot,\cdot} \right|\right|\right|_{2,1} &\leq \sqrt{\left|\hat{S}_D \cup S_D^*\right|} \sqrt{\sum_{t \in \mathbb{N}_\mathcal{T}} \sum_{j \in \hat{S}_D \cup S_D^*} \hat{\delta}_{t,j}^2} \\
&\leq \sqrt{\left|\hat{S}_D \cup S_D^*\right|} \sqrt{\sum_{t \in \mathbb{N}_\mathcal{T}} \underline{\phi}\left\{Q, \hat{S}_D \cup S_D^*\right\}^{-2} \hat{\delta}_t' Q \hat{\delta}_t} \\
&= \sqrt{\left|\hat{S}_D \cup S_D^*\right|} \underline{\phi}\left\{Q, \hat{S}_D \cup S_D^*\right\}^{-1} \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}.
\end{aligned}
\tag{B.16}
$$

Following identical steps to Eqs. (B.1)–(B.3), but with $\hat{\delta}_{\cdot,\cdot}$ in place of $\tilde{\delta}_{\cdot,\cdot}$, and then using the above bound, we have

$$
\begin{aligned}
&\left| \sum_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t) x_i^{*\prime} \right] \hat{\delta}_t \right| \\
&\quad \leq \frac{\lambda_D}{2} \left|\left|\left| \hat{\delta}_{\cdot,\cdot} \right|\right|\right|_{2,1} + b_s^d \sqrt{\mathcal{T}} \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \\
&\quad \leq \left(\frac{\lambda_D}{2} \frac{\sqrt{|\hat{S}_D \cup S_D^*|}}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}} + b_s^d \sqrt{\mathcal{T}} \right) \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}.
\end{aligned}
\tag{B.17}
$$

Next we turn to $\mathcal{M}(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot})$. By optimality of the post selection estimator $\mathcal{M}(\hat{\gamma}_{\cdot,\cdot}) \leq \mathcal{M}(\tilde{\gamma}_{\cdot,\cdot})$, as $\tilde{S}^D \subset \hat{S}_D$ by construction, and hence $\mathcal{M}(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) \leq \mathcal{M}(\tilde{\gamma}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot})$. By the mean value theorem, for scalars $\{m_t \in [0, 1]\}_{\mathbb{N}_\mathcal{T}}$ we have

$$
\begin{aligned}
&\mathcal{M}(\gamma^*_{\cdot,\cdot} + \tilde{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) \\
&\quad = \sum_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n \left[(d_i^t - \hat{p}_t(\{x_i^{*\prime}\gamma_t^* + m_t x_i^{*\prime}\tilde{\delta}_t\})) x_i^{*\prime}\tilde{\delta}_t \right] \\
&\quad = \sum_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n \left[(d_i^t - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})) x_i^{*\prime}\tilde{\delta}_t \right] \\
&\qquad + \sum_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - \hat{p}_t(\{x_i^{*\prime}\gamma_t^* + m_t x_i^{*\prime}\tilde{\delta}_t\})) x_i^{*\prime}\tilde{\delta}_t \right], \\
&\quad \leq \frac{\lambda_D}{2} \left|\left|\left| \tilde{\delta}_{\cdot,\cdot} \right|\right|\right|_{2,1} + b_s^d \sqrt{\mathcal{T}} \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \\
&\qquad + \sum_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n \left[m_t (x_i^{*\prime}\tilde{\delta}_t)^2 \right]. \\
&\quad \leq \left(\frac{\lambda_D}{2} \frac{\sqrt{|\hat{S}_D \cup S_D^*|}}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}} + b_s^d \sqrt{\mathcal{T}} \right) \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \\
&\qquad + \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2],
\end{aligned}
\tag{B.18}
$$

where the first inequality follows from Eq. (B.3) and the same steps as in (B.15) while the second applies (B.16) with $\tilde{\delta}_{\cdot,\cdot}$ and $m_t \leq 1$.[24]

Collecting the bounds of (B.17) and (B.18), and the definition of $R_\mathcal{M}$ gives

$$
\begin{aligned}
&\mathcal{M}(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) - \sum_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t) x_i^{*\prime} \right] \hat{\delta}_t \\
&\quad \leq \left(\frac{\lambda_D}{2} \frac{\sqrt{|\hat{S}_D \cup S_D^*|}}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}} + b_s^d \sqrt{\mathcal{T}} \right) \\
&\qquad \times \left(\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} + R_\mathcal{M} \right) + R_\mathcal{M}^2.
\end{aligned}
\tag{B.19}
$$

Next, we turn to a lower bound. Consider the same two cases as in the proof of Theorem 5. In the first case, we have the quadratic lower bound:

$$
\begin{aligned}
&M(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma^*_{\cdot,\cdot}) - \sum_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t) x_i^{*\prime} \right] \hat{\delta}_t \\
&\quad \geq \left(p_{\min} / A_p \right)^{\mathcal{T}} \frac{\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}{\mathcal{T} A_K}.
\end{aligned}
\tag{B.20}
$$

In the other case, this bound may not hold. Arguing as in the proof of Theorem 5, but applying Eq. (B.16), we get

$$
\begin{aligned}
&\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_1 \\
&\quad \leq \sqrt{\mathcal{T}} \mathcal{X} \sqrt{|\hat{S}_D \cup S_D^*|} \underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}^{-1} \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}.
\end{aligned}
$$

Therefore, as above, we find

$$
\begin{aligned}
&\mathcal{M}(\gamma^*_{\cdot,\cdot} + \hat{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma_{\cdot,\cdot}) - \sum_{t \in \mathbb{N}_\mathcal{T}} \mathbb{E}_n \left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t) x_i^{*\prime} \right] \hat{\delta}_t \\
&\quad \geq r_n \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2},
\end{aligned}
\tag{B.21}
$$

$$
\text{with } r_n = \frac{3}{2} \left(1 - \frac{2}{A_K} \right) \frac{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}}{\mathcal{X} \sqrt{\mathcal{T}} \sqrt{|\hat{S}_D \cup S_D^*|}}.
$$

Collecting the upper bound of (B.19) and the lower bounds (B.20) and (B.21) we have

$$
\begin{aligned}
&\left\{ \left(p_{\min} / A_p \right)^{\mathcal{T}} \frac{1}{\mathcal{T}} \frac{\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}{A_K} \right\} \wedge \left\{ r_n \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \right\} \\
&\quad \leq \left(\frac{\lambda_D}{2} \frac{\sqrt{|\hat{S}_D \cup S_D^*|}}{\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}} + b_s^d \sqrt{\mathcal{T}} \right) \\
&\qquad \times \left(\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} + R_\mathcal{M} \right) + R_\mathcal{M}^2.
\end{aligned}
\tag{B.22}
$$

Suppose the linear term is the minimum. The restrictions on A_K imply, algebraically, that Eq. (B.22) yields

$$
\begin{aligned}
r_n \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} &\leq (r_n/3) \left(\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} + R_\mathcal{M} \right) + R_\mathcal{M}^2 \\
&\leq (r_n/3) \left(\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} + 2R_\mathcal{M} \right).
\end{aligned}
$$

Canceling the r_n and solving yields $\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \leq R_\mathcal{M}$. On the other hand, if the quadratic term is the minimum, define

$$
\begin{aligned}
R_\mathcal{M}' &= \left(A_p / p_{\min} \right)^{\mathcal{T}} \mathcal{T} A_K \\
&\quad \times \left(2^{-1} \lambda_D \sqrt{|\hat{S}_D \cup S_D^*|} / \underline{\phi}\{Q, \hat{S}_D \cup S_D^*\} + b_s^d \sqrt{\mathcal{T}} \right).
\end{aligned}
$$

[24] Applying the steps of Eq. (B.16) to $\tilde{\delta}_{\cdot,\cdot}$ is preferred to using the results of Lemma B.3 because it leads to the tidier expression involving $\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}$, but the latter method could be substituted.

With this notation and the quadratic term being the minimum, Eq. (B.22) becomes

$$\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2] \\ \leq R'_\mathcal{M}\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^1/2 + R'_\mathcal{M}R_\mathcal{M} + \left(A_p/p_{\min}\right)^{\overline{\mathcal{T}}}\mathcal{T}A_K R_\mathcal{M}^2.$$

Then, because $a^2 \leq ab + c$ implies that $a \leq b + \sqrt{c}$, we have

$$\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \leq R'_\mathcal{M} + \left(R'_\mathcal{M}R_\mathcal{M} + \left(A_p/p_{\min}\right)^{\overline{\mathcal{T}}}\mathcal{T}A_K R_\mathcal{M}^2\right)^{1/2}.$$

Combining the bounds on $\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}$ from the two cases gives

$$\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \\ \leq \{R_\mathcal{M}\} \vee \left\{R'_\mathcal{M} + \left(R'_\mathcal{M}R_\mathcal{M} + \left(A_p/p_{\min}\right)^{\overline{\mathcal{T}}}\mathcal{T}A_K R_\mathcal{M}^2\right)^{1/2}\right\}.$$

From this bound on the log-odds, we bound the propensity score and the ℓ_1 rate:

$$\max_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2]^{1/2} \\ \leq \{R_\mathcal{M}\} \vee \left\{R'_\mathcal{M} + \left(R'_\mathcal{M}R_\mathcal{M} + \left(A_p/p_{\min}\right)^{\overline{\mathcal{T}}}\mathcal{T}A_K R_\mathcal{M}^2\right)^{1/2}\right\} + b_s^d;$$

$$\max_{t\in\mathbb{N}_\mathcal{T}} \left\|\hat{\gamma}_t - \gamma_t^*\right\|_1 \leq \left(\frac{|\tilde{S}^D \cup S_D^*|}{\underline{\phi}\{Q, \tilde{S}^D \cup S_D^*\}}\right)^{1/2} \{R_\mathcal{M}\} \\ \vee \left\{R'_\mathcal{M} + \left(R'_\mathcal{M}R_\mathcal{M} + \left(A_p/p_{\min}\right)^{\overline{\mathcal{T}}}\mathcal{T}A_K R_\mathcal{M}^2\right)^{1/2}\right\},$$

by arguments parallel to those used in the proof of Theorem 5. □

Appendix C. Proofs for group lasso selection and estimation of linear models

See supplemental Appendix.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2015.06.017.

References

Abadie, A., 2005. Semiparametric difference-in-differences estimators. Rev. Econom. Stud. 72, 1–19.

Abadie, A., Imbens, G.W., 2006. Large sample properties of matching estimators for average treatment effects. Econometrica 74, 235–267.

Andrews, D.W. K., Guggenberger, P., 2009. Incorrect asymptotic size of subsampling procedures based on post-consistent model selection estimators. J. Econometrics 152, 19–27.

Bach, F.R., 2008. Consistency of the group lasso and multiple kernel learning. J. Mach. Learn. Res. 9, 1179–1225.

Bach, F.R., 2010. Self-concordant analysis for logistic regression. Electron. J. Stat. 4, 384–414.

Bang, H., Robins, J.M., 2005. Doubly robust estimation in missing data and causal inference models. Biometrics 61, 962–972.

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429.

Belloni, A., Chernozhukov, V., 2011. ℓ_1-Penalized quantile regression in high-dimensional sparse models. Ann. Statist. 39, 82–130.

Belloni, A., Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. Bernoulli 19, 521–547.

Belloni, A., Chernozhukov, V., Chetverikov, D., Kato, K., 2015. Some new asymptotic theory for least squares series: Pointwise and uniform results. J. Econometrics 186, 345–366.

Belloni, A., Chernozhukov, V., Fernandez-Val, I., Hansen, C., 2014, Program Evaluation with High-Dimensional Data. Arxiv preprint arXiv:1311:2645.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection amongst high-dimensional controls. Rev. Econom. Stud. 81, 608–650.

Belloni, A., Chernozhukov, V., Wei, Y., 2013, Honest Confidence Regions for Logistic Regression with a Large Number of Controls. arXiv:1304.3969.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., 2013. Valid post-selection inference. Ann. Statist. 4, 802–837.

Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of LASSO and dantzig selector. Ann. Statist. 37, 1705–1732.

Buhlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data. In: Springer Series in Statistics, Springer-Verlag, Berlin.

Cattaneo, M.D., 2010. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. J. Econometrics 155, 138–154.

Cattaneo, M.D., Crump, R.K., Jansson, M., 2013. Generalized Jackknife estimators of weighted average derivatives. J. Amer. Statist. Assoc. 108, 1243–1256.

Cattaneo, M.D., Drukker, D.M., Holland, A.D., 2013. Estimation of multivalued treatment effects under conditional independence. The Stata J. 13, 407–450.

Cattaneo, M.D., Farrell, M.H., 2011. Efficient estimation of the dose response function under ignorability using subclassification on the covariates. In: Drukker, D. (Ed.), Advances in Econometrics: Missing Data Methods, vol. 27A. Emerald Group Publishing Limited, pp. 93–127.

Cattaneo, M.D., Farrell, M.H., 2013. Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators. J. Econometrics 174, 127–143.

Cattaneo, M.D., Jansson, M., Newey, W.K., 2014a, Alternative Asymptotics and the Partially Linear Model with Many Regressors. Working Paper.

Cattaneo, M.D., Jansson, M., Newey, W.K., 2014b. Small bandwidth asymptotics for density-weighted average derivatives. Econometric Theory 30, 176–200.

Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier (Chapter 76).

Chen, X., Christensen, T.M., 2015. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. J. Econometrics (forthcoming).

Chen, X., Hong, H., Tarozzi, A., 2004, Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects. Cowles Foundation Discussion Paper No. 1644.

Chen, X., Hong, H., Tarozzi, A., 2008. Semiparametric efficiency in GMM models with auxiliary data. Ann. Statist. 36, 808–843.

de la Peña, V.H., Lai, T.L., Shao, Q.-M., 2009. Self-Normalized Processes: Limit Theory and Statistical Applications, Probability and Its Applications, Springer.

Dehejia, R.H., Wahba, S., 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. J. Amer. Statist. Assoc. 94, 1053–1062.

Dehejia, R.H., Wahba, S., 2002. Propensity score-matching methods for nonexperimental causal studies. Rev. Econ. Stat. 84, 151–161.

Efron, B., 2014. Estimation and accuracy after model selection. J. Amer. Statist. Assoc. 109, 991–1007.

Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66, 315–331.

Hahn, J., 2004. Functional restriction and efficiency in causal inference. Rev. Econ. Stat. 84, 73–76.

He, X., Shao, Q.-M., 2000. On parameters of increasing dimensions. J. Multivariate Anal. 73, 1201–1235.

Heckman, J., Ichimura, H., Todd, P., 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Rev. Econom. Stud. 64, 605–654.

Heckman, J., Vytlacil, E.J., 2007. Econometric evaluation of social programs, Part I. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. VIB. Elsevier Science B.V., pp. 4780–4874.

Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161–1189.

Holland, P.W., 1986. Statistics and causal inference. J. Amer. Statist. Assoc. 81, 945–960.

Horowitz, J.L., Manski, C.F., 2000. Nonparametric analysis of randomized experiments with missing covariate and outcome data. J. Amer. Statist. Assoc. 95, 77–84.

Huang, J.Z., 2003. Local asymptotics for polynomial spline regression. Ann. Statist. 31, 1600–1635.

Huang, J., Zhang, T., 2010. The benefit of group sparsity. Ann. Statist. 38, 1978–2004.

Imai, K., van Dyk, D.A., 2004. Causal inference with general treatment regimes: generalizing the propensity score. J. Amer. Statist. Assoc. 99, 854–866.

Imbens, G.W., 2000. The role of the propensity score in estimating dose–response functions. Biometrika 87, 706–710.

Imbens, G.W., 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. Rev. Econ. Stat. 86, 4–29.

Imbens, G.W., Newey, W.K., Ridder, G., 2007, Mean-Squared-Error Calculations for Average Treatment Effects. Working Paper.

Imbens, G.W., Wooldridge, J.M., 2009. Recent developments in the econometrics of program evaluation. J. Econ. Lit. 47, 5–86.

Kang, J.D. Y., Schafer, J.L., 2007. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statist. Sci. 22, 523–539.

Kolar, M., Lafferty, J., Wasserman, L., 2011. Union support recovery in multi-task learning. J. Mach. Learn. Res. 12, 2415–2435.

Kwemou, M., 2012, Non-asymptotic Oracle Inequalities for the Lasso and Group Lasso in High Dimensional Logistic Model. Arxiv preprint arXiv:1206.0710.

LaLonde, R.J., 1986. Evaluating the econometric evaluations of training programs with experimental data. Am. Econ. Rev. 76, 604–620.

Lechner, M., 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Lechner, M., Pfeiffer, E. (Eds.), Econometric Evaluations of Active Labor Market Policies. Physica. Heidelberg, pp. 43–58.

Leeb, H., Pötscher, B.M., 2005. Model selection and inference: facts and fiction. Econometric Theory 21, 21–59.

Leeb, H., Pötscher, B.M., 2008a. Can one estimate the unconditional distribution of post-model-selection estimators? Econometric Theory 24, 338–376.

Leeb, H., Pötscher, B.M., 2008b. Sparse estimators and the oracle property, or the return of Hodges' estimator. J. Econometrics 142, 201–211.

Lounici, K., Pontil, M., van de Geer, S., Tsybakov, A.B., 2011. Oracle inequalities and optimal inference under group sparsity. Ann. Statist. 39, 2164–2204.

Negahban, S.N., Ravikumar, P., Wainwright, M.J., Yu, B., 2012. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statist. Sci. 27, 538–557.

Newey, W.K., 1990. Efficient instrumental variables estimation of nonlinear models. Econometrica 58, 809–837.

Newey, W.K., 1997. Convergence rates and asymptotic normality for series estimators. J. Econometrics 79, 147–168.

Newey, W.K., McFadden, D.L., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics. In: Handbook of Econometrics, vol. 4. Elsevier, pp. 2111–2245 (Chapter 36).

Obozinski, G., Wainwright, M.J., Jordan, M.I., 2011. Support union recovery in high-dimensional multivariate regression. Ann. Statist. 39, 1–47.

Pötscher, B.M., 2009. Confidence sets based on sparse estimators are necessarily large. Sankhyā 71-A, 1–18.

Pötscher, B.M., Leeb, H., 2009. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. J. Multivariate Anal. 100, 2065–2085.

Powell, J.L., Stock, J.H., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1403–1430.

Raskutti, G., Wainwright, M.J., Yu, B., 2010. Restricted eigenvalue properties for correlated Gaussian designs. J. Mach. Learn. Res. 11, 2241–2259.

Robins, J., Li, L., Tchetgen, E., van der Vaart, A., 2008. Higher order influence functions and minimax estimation of nonlinear functionals. In: Nolan, D., Speed, T. (Eds.), Probability and Statistics: Essays in Honor of David A. Freedman, vol. 2. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Robins, J.M., Rotnitzky, A., 1995. Semiparametric efficiency in multivariate regression models with missing data. J. Amer. Statist. Assoc. 90, 122–129.

Romano, J.P., 2004. On non-parametric testing, the uniform behaviour of the t-test, and related problems. Scand. J. Stat. 31, 567–584.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rudelson, M., Zhou, S., 2013. Reconstruction from anisotropic random measurements. IEEE Trans. Inform. Theory 59, 3434–3447.

Smith, J.A., Todd, P.E., 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? J. Econometrics 125, 305–353.

Tan, Z., 2010. Bounded, efficient and doubly robust estimation with inverse weighting. Biometrika 97, 661–682.

Tanabe, K., Sagae, M., 1992. An exact Cholesky decomposition and the generalized inverse of the variance–covariance matrix of the multinomial distribution, with applications. J. R. Stat. Soc. Ser. B Stat. Methodol. 54, 211–219.

Tsiatis, A.A., 2006. Semiparametric Theory and Missing Data. Springer, New York.

van de Geer, S., 2008. High-dimensional generalized linear models and the Lasso. Ann. Statist. 36, 614–645.

van de Geer, S., Buhlmann, P., 2009. On the conditions used to prove oracle results for the Lasso. Electron. J. Stat. 3, 1360–1392.

van de Geer, S., Buhlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42, 1166–1202.

van der Laan, M., Robins, J.M., 2003. Unified Methods for Censored Longitudinal Data and Causality. Springer-Verlag.

Vincent, M., Hansen, N.R., 2014. Sparse group lasso and high dimensional multinomial classification. Comput. Statist. Data Anal. 71, 771–786.

von Bahr, B., Esseen, C.-G., 1965. Inequalities for the rth absolute moment of a sum of random variables, $1 \leqq r \leqq 2$. Ann. Math. Stat. 36, 299–303.

Wei, F., Huang, J., 2010. Consistent group selection in high-dimensional linear regression. Bernoulli 16, 1369–1384.

White, H., Lu, X., 2011. Causal diagrams for treatment effect estimation with application to efficient covariate selection. Rev. Econ. Stat. 93, 1453–1459.

Wooldridge, J.M., 2007. Inverse probability weighted estimation for general missing data problems. J. Econometrics 141, 1281–1301.

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data, second ed.. MIT Press, Cambridge.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68, 46–67.

Zhang, C.-H., Zhang, S.S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc. Ser. B 76, 217–242.

Zou, H., 2006. The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.