Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations

Max H. Farrell*
University of Michigan

Job Market Paper

October 24, 2013

Abstract

This paper concerns robust inference on average treatment effects following model selection. In the selection on observables framework, we show how to construct confidence intervals based on a doubly-robust estimator that are robust to model selection errors and prove that they are valid uniformly over a large class of treatment effect models. The class allows for multivalued treatments with heterogeneous effects (in observables), general heteroskedasticity, and selection amongst (possibly) more covariates than observations. Our estimator attains the semiparametric efficiency bound under appropriate conditions. Precise conditions are given for any model selector to yield these results, and we show how to combine data-driven selection with economic theory. For implementation, we give a specific proposal for selection based on the group lasso and derive new technical results for high-dimensional, sparse multinomial logistic regression. A simulation study shows our estimator performs very well in finite samples over a wide range of models. Revisiting the National Supported Work demonstration data, our method yields accurate estimates and tight confidence intervals.

Keywords: High-dimensional sparse model, heterogeneous treatment effects, uniform inference, model selection, doubly-robust estimator, unconfoundedness.

*Email: maxhf@umich.edu. Web: http://www.umich.edu/~maxhf. I am deeply grateful to Matias Cattaneo for his continual guidance, encouragement, and especially his patience. I am also indebted to Xuming He, Lutz Kilian, and Jeffrey Smith for thoughtful feedback and helpful 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. Funding from the Rackham Graduate School and the Haber Fellowship is gratefully acknowledged.

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. No matter the approach used, inference almost never takes into account 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 heterogeneity in observables and heteroskedasticity. First and foremost, we show that a doubly-robust estimator is robust to model selection errors, a newly-discovered virtue of this class of estimators.[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 proposed estimator is asymptotically linear and attains the semiparametric efficiency bound, under standard conditions imposed in the program evaluation literature. Fourth, we derive new technical 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, 2008b, 2009, 2009) have shown that inference relying too heavily on model selection can not 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

[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).

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

reliability in applications. Our proposed methods for post model selection inference build upon the path-breaking recent work of Belloni, Chernozhukov, and Hansen (2013). We circumvent, without contradicting, the impossibility results of Leeb and Pötscher by not insisting on perfect selection, but rather explicitly accounting for inevitable model selection errors in the asymptotic approximations.[3]

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. 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, Chernozhukov, and Hansen (2013, draft dated July 19), propose a similar approach. Their main focus is a partially linear model, in which the coefficient of a treatment indicator will recover the average effect of a binary treatment only if the effect is constant across observables, but Section 5 of their most recent draft, developed independently from our work, considers heterogeneous effects. There are two broad differences in our approaches. First, we allow for multivalued treatments, which offers a larger set of estimands and can thus enhance the understanding of program impacts.[4] We show how to improve model selection in this context by pooling information across treatment levels. Second, although in both cases the doubly-robust estimator is used for average treatment effects following a (quite different) model selection step,[5] we exploit certain features of the estimator to produce two benefits: (i) our procedure requires demonstrably weaker conditions on the model selection stage (see Assumption 3); and (ii) none of our results require using variables selected for the treatment equation in the outcome model estimation, and vice versa (their "post double selection" method), and indeed, we show doing so requires stronger assumptions (see Assumption 4).

Our analysis is conducted under selection on observables, which has a long tradition and remains

[3] Efron (2013) and Berk, Brown, Buja, Zhang, and Zhao (2013) also propose methods for post-selection inference, both quite distinct from our method.

[4] 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).

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

quite popular in empirical economics.[6] Covariate selection has three crucial roles to play in this framework. First, using more observed covariates, and more flexibly, may help proxy for unobserved confounding and hence increase the plausibility of unconfoundedness. Second, it is natural that some variables are not part of the causal mechanism under study, and therefore should be excluded. Third, the efficient conditioning set must contain those variables that drive the outcome, which are 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 data sets 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*. A far more realistic scenario 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 used in practice.

With the aim of mimicking common empirical practice we estimate the propensity score with multinomial logistic regression. To handle this nonlinear model under approximate sparsity we employ the group lasso (Yuan and Lin 2006) coupled with a novel penalty that controls both the noise and bias simultaneously. In our view, the group lasso is particularly well-suited to multivalued treatments because it pools information across all treatment levels to aid selection. Our results are stated in the language of treatment effects, but apply to general data structures and are of independent interest.[7] To the best of our knowledge this is the first detailed study of an approximately sparse, nonlinear model in the high-dimensional literature. 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.[8] Furthermore,

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

[7] Our techniques build on prior studies, in particular Bickel, Ritov, and Tsybakov (2009), Lounici, Pontil, van de Geer, and Tsybakov (2009, 2011), Obozinski, Wainwright, and Jordan (2011), Belloni and Chernozhukov (2011b), Belloni, Chen, Chernozhukov, and Hansen (2012).

[8] Examples include van de Geer (2008), Belloni and Chernozhukov (2011a), or Negahban, Ravikumar, Wainwright, and Yu (2012). Bach (2010) only gives an error bound on coefficients in exactly sparse logistic regression, which can not yield our results; and does not consider prediction error or post-selection estimation. In independent work,

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 remainder of the paper proceeds as follows. In Section 2, we give a short, self-contained overview of the main results. For ease of reference, Section 2.3 collects all notation. Section 3 describes the treatment effect model. Sparse models are discussed in Section 4, which shows how several models that are commonly used in empirical work fit in this framework. Section 5 presents our estimation method and gives complete results on treatment effect inference. The proposed group lasso approach to sparse modeling is detailed in Section 6, including theoretical results on model selection and estimation. Section 7 presents the numerical evidence on finite sample behavior of our procedure. Section 8 concludes. The main proofs are presented in the Appendix, while the remainder are available in a supplement.

2 Overview of Results and Notation

Here we give an overview of the main contributions of the paper. We first discuss treatment effect inference with a general model selector. Then in Section 2.2 we discuss our new results for the group lasso, our proposed model selector. Section 2.3 collects notation to be used throughout.

2.1 Treatment Effects and Results on Post-Selection Inference

We consider a multivalued treatment, with status indicated by $D \in \{0, 1, \ldots, \mathcal{T}\}$. 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 $D = t$. $Y(t)$ 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

Belloni, Chernozhukov, and Wei (2013) study exactly sparse logistic regression, also using Bach's (2010) tools, but are focused on a different inference goal.

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 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

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

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).

Suppose we have an i.i.d. sample $\{(y_i, d_i, x_i')\}_{i=1}^n$ from (Y, D, X'). Then, for 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 $p_t(x)$ 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 $\hat{\mu}_t(x_i)$ is crucial, impacting consistency, efficiency, and finite sample performance. Covariate selection based on ad hoc, iterative searches is common in empirical evaluations, but is informal, not objective, and not replicable. Balancing tests are also commonly used in this context, 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.[9] Importantly, we do not preclude the addition of variables 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.

[9]For the final estimation step, the doubly-robust estimator is available in STATA 13 and the package of Cattaneo, Drukker, and Holland (forthcoming) (as the default). The covariate selection stage is easily implemented in `R`; code is available upon request. A self-contained STATA package is under development.

The following theorem is an example of the more general results presented in Section 5.2, wherein we also define V_t and $\hat{V}_t = \hat{V}^W_{\boldsymbol{\mu}}(t) + \hat{V}^B_{\boldsymbol{\mu}}(t,t)$.

Theorem 1. *Consider a sequence $\{P_n\}$ of data-generating processes that obey, for each n, Assumptions 1, and 2 below. We require two conditions on the model selector:*

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

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

Under these conditions, $\sqrt{n}(\hat{\mu}_t - \mu_t) \to_d N(0, V_t)$ *and* $\hat{V}_t/V_t \to_{P_n} 1$*. For each n, let $\boldsymbol{P}_n$ be the set of data-generating processes satisfying Assumption 1, and 2 and conditions (i) and (ii). Then*

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

where $c_\alpha = \Phi^{-1}(1-\alpha/2)$.

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 additional, standard conditions (see Section 5.3). Intuitively, this is similar to (but distinct from) overcoming pretesting bias in other contexts.

Two general conditions are placed on the model selector. The first is a mild consistency requirement. The second is analogous to the commonly-used, high-level requirement in semiparametrics that first-stage components converge faster than $n^{-1/4}$. However, because we use the doubly-robust moment condition we only have the product of the two estimation errors; this requirement can be easier to satisfy if one or the other function is easier to estimate (e.g. if one function is very smooth or very sparse). In high-dimensional models the rates for the first stage depend on the sample size, the number of covariates considered, and the sparsity level. We propose to use the group lasso and prove that these estimators satisfy (i) and (ii). Importantly, the rate will depend on the total number of covariates only logarithmically, allowing for a large number.

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. Formally, the group lasso gives the union of all supports. The group lasso is easy to implement, as discussed in Remark 4.

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, with $p > n$ allowed. 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 $p_t(x)$ and $\mu_t(x)$ follow p-dimensional parametric models. Then the sparsity assumption is that there is an s-dimensional model that has 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 detail 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 that are known to be important from 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 uniform inference nor efficiency (and requires stronger assumptions).

Our main results give precise bounds for the number of covariates selected and the estimation error, both for the penalized and unpenalized estimates. These bounds, given in Section 6.3, are nonasymptotic: exact constants are given for each bound and these bounds are valid for any given n, p, and s, provided our assumptions are met. 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 bounded by b_s^d and b_s^y, respectively. Then under the assumptions in Section 6.3, and $\delta > 0$ described therein, with high probability we have:*

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

These two results for our proposed group lasso estimators can be directly used to verify the high-level conditions in Theorem 1 above. The product of these rates makes explicit the advantage discussed above of using condition (ii) in Theorem 1, by showing exactly how the errors depend on

the total number of variables, the sparsity, and the bias. Section 6.3 also shows 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 aren't 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 population 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 n i.i.d. draws from P_n. In general, the DGP may vary with n, along with features such as parameters, distributions, an so forth, as discussed in Section 4.2. This is generally suppressed for notational clarity.

We further 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 (or, equivalently, normalize $\delta_{0,\cdot} = 0$). 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 DGP for a fixed n will be denoted by P_n. The set of all such P_n that we allow for will be $\boldsymbol{P}_n$. As shorthand for a sequence we will use $\{P_n\} = \{P_n : n \geq 1, P_n \in \boldsymbol{P}_n\}$. Expectations and probabilities will be understood to be taken against P_n,

though notationally suppressed: $\mathbb{E}[W] = \mathbb{E}_{P_n}[W]$ denotes the population expectation for a random variable W and $\mathbb{P}[A] = \mathbb{P}_{P_n}[A]$ the probability of event A. 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.

The empirical expectation will be denoted $\mathbb{E}_n[w_i] = \sum_{i=1}^n w_i/n$. Also, define $\mathbb{E}_{n,t}[w_i] = \sum_{i\in\mathbb{I}_t} w_i/n_t = \sum_{i=1}^n d_i^t w_i/n_t$ for observations with treatment 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

$$\mu_t = \mathbb{E}[Y(t)], \qquad t \in \overline{\mathbb{N}}_\mathcal{T}, \tag{1}$$

and

$$\mu_{t,t'} = \mathbb{E}[Y(t)|D = t'], \qquad t, t' \in \overline{\mathbb{N}}_\mathcal{T} \times \overline{\mathbb{N}}_\mathcal{T}, \tag{2}$$

In the binary case, the average treatment effect is given by $\mu_1 - \mu_0$, whereas the treatment on the treated is $\mu_{1,1} - \mu_{0,1}$. Having a multivalued treatment allows for a much larger range of interesting estimands. To fix ideas, we keep as running examples two leading cases from the literature. 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 vectors are by no means the only interesting estimands constructed from μ_t and $\mu_{t,t'}$; many others are discussed 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, two remarks on 1(a) 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, Ichimura, and Todd (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 of covariates. Our results of course hold under full independence, which is important for the efficiency discussed in Section 5.3 below.

Second, 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. Indeed, if the parameter of interest is changed, for example 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

$$\mathbb{E}\big[\psi_t\big(Y, D, \mu_t(X), p_t(X), \mu_t\big)\big] = \mathbb{E}\left[\frac{\mathbb{1}\{D = t\}Y}{p_t(X)} + \mu_t(X) - \frac{\mathbb{1}\{D = t\}\mu_t(X)}{p_t(X)} - \mu_t\right] = 0 \tag{3}$$

and

$$\begin{aligned}\mathbb{E}\big[\psi_{t,t'}\big(Y, D, \mu_t(X), p_t(X), p_{t'}(X), \mu_{t,t'}\big)\big] \\ = \mathbb{E}\left[\frac{\mathbb{1}\{D = t'\}\mu_t(X)}{p_{t'}} + \frac{p_{t'}(X)}{p_{t'}}\frac{\mathbb{1}\{D = t\}(Y - \mu_t(X))}{p_t(X)} - \mu_{t,t'}\right] = 0,\end{aligned} \tag{4}$$

where $p_t = \mathbb{P}[D = t]$. The moment condition (3) 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 will be a plug-in version of this moment condition.

Remark 1 (Simplifications for $\mu_{t,t}$)**.** Identification $\mu_{t,t}$ 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 (4) 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 control 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 control group, and so we maintain Assumption 1 for simplicity, rather than keeping track of hosts of special cases. ■

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

4 Sparse Models

We now formalize approximate sparsity. It is convenient to work with the linear log-odds ratio form of the multinomial model; the outcome model already being linear. 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, we take $p_0(x) = 1 - \sum_{t\in\mathbb{N}_\mathcal{T}} p_t(x)$ and write

$$\log\left(\frac{p_t(x)}{p_0(x)}\right) = x_D^{*\prime}\gamma_t^* + B_t^D, \qquad t \in \mathbb{N}_\mathcal{T}. \tag{5}$$

Similarly, write the outcome regressions as

$$\mu_t(x) = x_Y^{*\prime}\beta_t^* + B_t^Y, \qquad t \in \overline{\mathbb{N}}_\mathcal{T}, \tag{6}$$

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. When it is clear from the context we often abbreviate both X_D^* and X_Y^* by X^* (and their realizations by x_i^*) and refer to them generically as "covariates". Much discussion applies to both. We assume $\mathbb{E}_n[(x_i^*)^2] = 1$ without loss of generality (see Remark 5).

Approximate sparsity requires that only a small number of the X^* are needed to make the bias small. Define $S_*^D = \bigcup_{\overline{\mathbb{N}}_\mathcal{T}} \text{supp}(\gamma_t^*)$ and $S_*^Y = \bigcup_{\mathbb{N}_\mathcal{T}} \text{supp}(\beta_t^*)$, so that these sets capture all variables important for treatment and outcomes, respectively. We assume that there is some $s_d < n$ such that for $|S_*^D| = s_d$, and similarly $|S_*^Y| = s_y < n$, and B_t^D and B_t^Y are sufficiently small. 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 there is no confusion, we will write s for either s_d or s_y.

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, Wainwright, and Yu 2010, Rudelson and Zhou 2011, Belloni, Chernozhukov, and Hansen 2013).

Example 1 (Oracle parametric model)**.** Assume models (5) and (6) 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.[10]

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 access to prior knowledge of which covariates are important and which are not. Such knowledge no doubt plays an important role, but it can not cover all situations or all variables in a data set. 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 can never be made uniformly valid, and is known to have poor finite sample properties, as shown by Leeb and Pötscher. ■

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

[10] Even with diverging dimensions, the parametric multinomial logistic model relies on the independence of irrelevant alternatives.

of any coefficients being precisely zero, suppose they may be ordered such that $|\gamma^0_{t,j}| \propto j^{-\alpha_\gamma}$ and $|\beta^0_{t,j}| \propto j^{-\alpha_\beta}$, with α_γ and α_γ at least 2. With this rapid decay, there exist s_d and s_y that are $o(n)$ such that Equations (5) and (6), and other conditions needed, are satisfied for $\gamma^*_{t,j} = \gamma^0_{t,j}$ for $j \leq s_d$ and $\beta^*_{t,j} = \beta^0_{t,j}$ for $j \leq s_y$ and the rest truncated to zero. That is S^D_* and S^Y_* collect the largest coefficients and $B^D_t = \sum_{\mathbb{N}_p \setminus S^D_*} x_j \gamma^0_{t,j}$, and similarly for B^Y_t. ■

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 (5) and (6), $\gamma^*_{\cdot,\cdot}$ and $\beta^*_{\cdot,\cdot}$ are the coefficients of these approximations, while B^D_t and B^Y_t 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, Imbens, and Ridder (2003), Imbens, Newey, and Ridder (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 $\log[p_t(x)/p_0(x)] = x_1'\gamma^1_t + h_t(x_2) + B^1_t(x)$, where $B^1_t(x)$ 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^2_t such that $h_t(x_2) = R_D(x_2)'\gamma^2_t + B^2_t(x_2)$ and so

$$\log\left(\frac{p_t(x)}{p_0(x)}\right) = x^{*\prime}_D \gamma^*_t + B^D_t, \quad x^*_D = (x_1', R_D(x_2)')', \quad \gamma^*_t = (\gamma^{1\prime}_t, \gamma^{2\prime}_t)', \quad \text{and} \quad B^D_t = B^1_t + B^2_t.$$

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. ■

4.2 Conceptual considerations in n-varying DGPs

We close this section with a discussion of how the DGP may vary with sample size. 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 (5) and (6), X^*, $\gamma^*_{\cdot,\cdot}$, and $\beta^*_{\cdot,\cdot}$ must depend on n by construction. Our results on estimation of these models are nonasymptotic: exact constants

are provided that are defined for a fixed n. 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. We may now define $\mu_t = \mathbb{E}[\mathbb{E}[Y(t)|X]] = \mathbb{E}[X']\beta_t^*$, which depends on n by construction. That is, given an exact parametric specification for $\mathbb{E}[Y(t)|X]$ with a diverging number of covariates, the parameter to be estimated, μ_t, must depend on n. This may seem unnatural, as we typically think of the "true" parameters being features of a (large) fixed study population. However, with a diverging number of covariates, the idea of a fixed DGP is not clear. 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 model selector to yield uniform inference. We then give theoretical results, and close with a short discussion of efficiency.

5.1 Estimation Procedure with a Generic Model Selector

The moment functions $\psi_t(\cdot)$ and $\psi_{t,t'}(\cdot)$ of Equations (3) and (4) have fixed and known form, and so for (model selection based) 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 model selection based 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 $\{d_i\}_{i=1}^n$. Stronger 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 stronger 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 condition.

Assumption 2 (Data Generating Process)**.** *For each n, the following are true for the DGP P_n.*

(a) (y_i, d_i, x_i) is an i.i.d. sample from (Y, D, X), where the data generating process obeys Equations (5) *and* (6) *such that $|S_*^Y| = s_d$ and $|S_*^D| = s_y$.*

(b) The covariates X^* have bounded support, with $\max_{j\in\mathbb{N}_p} X_j^* \leq \mathcal{X} < \infty$, uniformly in n. Trans-

formations may depend on n but not the underlying data generating process.

(c) $\mathbb{E}[|U|^4 \mid X] \leq \mathcal{U}^4$, *uniformly in* n.

(d) $\min_{j \in \mathbb{N}_p,\ t \in \overline{\mathbb{N}}_\mathcal{T}} \mathbb{E}[X_j^{*2} U^2] \wedge \mathbb{E}[X_j^{*2}(\mathbb{1}\{D = t\} - p_t(X))^2]$ *is bounded away from zero, uniformly in* n.

(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, uniformly in* n.

The conditions of Assumption 2 are mild and intuitive. Assumption 2(a) restricts attention to cross-sectional applications and codifies the requirement that the underlying functions have sparse representations. 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. Finally, Assumptions 2(c), 2(d), and 2(e) are weak moment conditions on the potential outcome models, including allowing the errors to be heteroskedastic and non-Gaussian. Excepting the support requirements of 2(a) these conditions are not unique to high-dimensional models or model selection. Formalizing the requirement of uniform bounds in n is needed when doing array asymptotics.

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

Assumption 3 (Model Selector Restrictions)**.** *The model selection based estimators* $\hat{p}_t(x)$ *and* $\hat{\mu}_t(x)$ *obey the following for a sequence* $\{P_n\}$, *uniformly in* $t \in \overline{\mathbb{N}}_\mathcal{T}$.

(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) $\mathbb{E}_{n,t}[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2]^{1/2} \mathbb{E}_{n,t}[(\hat{p}_t(x_i) - p_t(x_i))^2]^{1/2} = o_{P_n}(n^{-1/2})$.

The first is a mild consistency requirement. The second is more interesting. It is analogous to the commonly-used, high-level requirement in semiparametrics that each first-step component converge at $n^{-1/4}$ at least.[11] Belloni, Chernozhukov, and Hansen (2013) use just such a condition. However, by making use of the doubly-robust property we have the weaker condition shown, involving the product. If one function is relatively easy to estimate Assumption 3(b) can be satisfied even if the other does not converge at $n^{-1/4}$. In high-dimensional, sparse models 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).

For our proposed group lasso selectors, recalling the results of Corollary 1, Assumption 3(a) will be satisfied if $(\sqrt{n^{-1} s_d \log(p \vee \underline{n})^{3/2+\delta_D}} + b_s^d \sqrt{s_d}) \to 0$ and $(\sqrt{n^{-1} s_y \log(p \vee \underline{n})^{3/2+\delta_Y}} + b_s^y) \to 0$.

[11] See, for example, Newey (1994), Newey and McFadden (1994), and Chen (2007), and references therein.

Further, 3(b) is satisfied if their product is $o(n^{-1/2})$, which clearly shows how the relative sparsities and smoothnesses may interact.

When considering the additional-randomness estimators, we need a stronger bound on the regression errors U and more conditions on the first stage.

Assumption 4 (Regularity conditions for union estimators)**.** *The model selection based estimators $p_t(x)$ and $\hat{\mu}_t(x)$ obey the following for a sequence $\{P_n\}$, uniformly $t \in \overline{\mathbb{N}}_{\mathcal{T}}$:*

$$\left(\max_{i \in \mathbb{I}_t} |u_i|\right) \left|\mathbb{E}_{n,t}[(\hat{p}_t(x_i) - p_t(x_i))^2]\right| = o_{P_n}(n^{-1/2}) \quad \textit{and} \quad \|\hat{\gamma}_t - \gamma_t^*\|_1 \vee \|\hat{\beta}_t - \beta_t^*\|_1 = o_{P_n}(\log(p)^{-1}).$$

These stronger conditions are needed because we must apply bounds for self-normalized sums (de la Peña, Lai, and Shao 2009). Belloni, Chen, Chernozhukov, and Hansen (2012) were the first to use these techniques in high-dimensional, sparse models. The first condition is a high-level condition that can be verified with conditions on the errors and a bound for estimation. For example, if we follow Belloni, Chen, Chernozhukov, and Hansen (2012) and assume that $\max_{i \in \mathbb{N}_n} |u_i| = O_{P_n}(n^{1/q})$ for some $q > 2$, then Assumption 4 is met under our group lasso results if both $\left(n^{1/2+1/q}\left[n^{-1} s_d \log(p \vee \underline{n})^{3/2+\delta_D} + (b_s^d)^2 s_d\right]\right)$ and $\left(\log(p)\left[\sqrt{n^{-1} s_d^2 \log(p \vee \underline{n})^{3/2+\delta_D}} + b_s^d s_d \vee b_s^y\right]\right)$ converge to zero. Note that as q increases, the stringency of the rate restriction decreases. For example, if the u_i are Gaussian, q can be taken to be any (large) positive number.

Remark 3 (Linear Probability Models)**.** Some authors advocate a linear probability model for the function $p_t(x)$, instead of the multinomial logistic form. Our results cover this case as well. Note that all we require are sufficiently high quality approximations to the underlying objects. If Assumptions 3, and 4 if appropriate,[12] 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 concerns 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 $|\hat{\mu}_t - \mu_t| = o_{P_n}(1)$.*

[12] Assumption 4 can be slightly weakened in this case due to the linear link function.

This theorem formalizes the double-robustness property of our estimators: the propensity score or regression may be misspecified if the limiting objects must be 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 $\boldsymbol{\tau}$.

Before giving the results for $\boldsymbol{\mu}$, we need an asymptotic variance formula. Let the conditional variance of the potential outcomes be $\sigma_t^2(x) = \mathbb{E}[U^2|D=t, X=x]$. Define the $\overline{\mathcal{T}}$-square matrix $V_{\boldsymbol{\mu}}$ with elements

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

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

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

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, 2, and 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. $V_{\boldsymbol{\mu}}^{-1/2}\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \to_d \mathcal{N}(0, I_{\overline{\mathcal{T}}})$*; 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, Assumption 4 holds, 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. For a fixed, twice uniformly continuously differentiable function $G : \mathbb{R}^{\overline{\mathcal{T}}} \to \mathbb{R}$ with gradient ∇_G such that $\liminf_{n\to\infty} \|\nabla_G(\boldsymbol{\mu})\|_2$ is bounded away from zero,*

we have:

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

where $c_\alpha=\Phi^{-1}(1-\alpha/2)$.

Corollary 2 shows that these procedures are uniformly valid over the class of DGPs we consider, and hence will be reliable in applications. This method of proving uniformity follows Belloni, Chernozhukov, and Hansen (2013) and Romano (2004), and is distinct from the approach of Andrews and Guggenberger (2009). By not relying on an oracle property, we avoid the uniformity problems demonstrated by Leeb and Pötscher, as discussed before.

Our results for the treatment effects on the treated, $\mu_{t,t'}$, are conceptually similar. The variance formula for $\boldsymbol{\tau}$ is slightly more cumbersome notationally. Define the $\mathcal{T}$-square matrix $V_{\boldsymbol{\tau}}$ with elements

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

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

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

Note that we needn't estimate $\mu_t(x)$ and $\sigma_t^2(x)$, again due to the simplification discussed in Remark 1. With this notation, we have the following results.

Theorem 4 (Estimation of Treatment Effects on Treated Groups)**.** *Consider a sequence* $\{P_n\}$ *of data-generating processes that obey Assumptions 1, 2, and 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, Assumption 4 holds, 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*. For a fixed, twice uniformly continuously differentiable*

function $G : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}$ *with gradient* ∇_G *such that* $\liminf_{n\to\infty} \|\nabla_G(\boldsymbol{\tau})\|_2$ *is bounded away from zero, we have:*

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

where $c_\alpha = \Phi^{-1}(1-\alpha/2)$.

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. We consider two efficiency criteria: semiparametric efficiency and oracle efficiency. The former deals with the variance of the final estimator, whereas the latter is directly about the efficacy of the model selection. To put each criterion on sound conceptual footing, we separate discussion and restrict each to the most appropriate set of models.

For semiparametric efficiency, assume that $p_t(x)$ and $\mu_t(x)$ are nonparametric objects, as in Example 4. Recall that 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 independence, namely $\{Y(t)\}_{\overline{\mathbb{N}}_{\mathcal{T}}} \perp\!\!\!\perp D|X$, then Theorems 3 and Theorem 4 immediately yield asymptotically linearity and semiparametric efficiency, attaining Hahn's (1998) and Cattaneo's (2010) bounds.

Let us turn to oracle efficiency. An alternative to our 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 causal setting when restricted to exactly sparse models (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). This may be interesting in its own right, or for future applications by way of hypothesis generation. Further, efficiency can be improved because only variables appearing in $\mu_t(x_i) = \mathbb{E}[Y|D = t, x_i]$ should be used, hence $S_*^D \setminus S_*^Y$ are not needed and $S_*^Y \setminus S_*^D$ can be ignored for propensity score estimation.

Perfect selection requires two strong conditions: (i) an orthogonality condition on the Gram matrixes that restricts the correlation between the variables in and out of the true support (Zhao and Yu 2006, Bach 2008), and (ii) a *beta-min* condition bounding the nonzero coefficients away from zero. Intuitively, highly correlated variables can not be distinguished, nor can coefficients sufficiently close to zero be found with certainty. Both bounds may depend on n. Under such conditions, it is possible 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. This section is quite technical. Our main theorems are given in Section 6.3. To set up these results, we first make precise how selection and refitting are implemented. Section 6.1 develops our (apparently) novel penalty choice for multinomial logistic regression. Restricted and sparse eigenvalues, key quantities in our bounds, are discussed in Section 6.2. Discussion will be model-specific so we use the general notation X^* and s.

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

$$\mathcal{M}(\gamma_{\cdot,\cdot}) = \sum_{t \in \mathbb{N}_{\mathcal{T}}} \mathbb{E}_n \left[d_i^t \log \left(\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_{\mathcal{T}}}) \right) \right] \qquad \text{and} \qquad \mathcal{E}(\beta_{\cdot,\cdot}) = \sum_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} \mathbb{E}_{n,t}[(y_i - x_i^{*\prime}\beta_t)^2],$$

where we denote the multinomial logit function as $\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_{\mathcal{T}}}) = \exp(x_i^{*\prime}\gamma_t) / \left(1 + \sum_{t \in \mathbb{N}_{\mathcal{T}}} \exp(x_i^{*\prime}\gamma_t)\right)$. Then, the group lasso estimates for the propensity score coefficients, denoted $\tilde{\gamma}_{\cdot,\cdot}$, solve

$$\tilde{\gamma}_{\cdot,\cdot} = \underset{\gamma_{\cdot,\cdot} \in \mathbb{R}^{p^{\mathcal{T}}}}{\arg\min} \left\{ \mathcal{M}(\gamma_{\cdot,\cdot}) + \lambda_D \left\| \gamma_{\cdot,\cdot} \right\|_{2,1} \right\}, \tag{9}$$

where λ_D is a penalty parameter discussed in detail below and $\left\| \gamma_{\cdot,\cdot} \right\|_{2,1}$ is the mixed ℓ_2/ℓ_1 norm defined above. Similarly, the regression estimates solve

$$\tilde{\beta}_{\cdot,\cdot} = \underset{\beta_{\cdot,\cdot} \in \mathbb{R}^{p^{\overline{\mathcal{T}}}}}{\arg\min} \left\{ \mathcal{E}(\beta_{\cdot,\cdot}) + \lambda_Y \left\| \beta_{\cdot,\cdot} \right\|_{2,1} \right\}. \tag{10}$$

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

$$\hat{\gamma}_{\cdot,\cdot} = \underset{\gamma_{\cdot,\cdot},\ \text{supp}(\gamma_t) = \hat{S}^D}{\arg\min} \{\mathcal{M}(\gamma_{\cdot,\cdot})\} \tag{11}$$

[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 $\text{supp}(\gamma_t^*)$ and $\text{supp}(\beta_t^*)$ will 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, Wainwright, and Jordan (2011) give a sharp bound on the overlap necessary to yield improvements, while Huang and Zhang (2010), Kolar, Lafferty, and Wasserman (2011), and Lounici, Pontil, van de Geer, and Tsybakov (2011) also demonstrate advantages of the group lasso approach. These works show, among other things, that the group lasso advantage increases as $\mathcal{T}$ increases, and with the group structure, may perform better with smaller samples. We defer to the works cited for a formal discussion.

and

$$\hat{\beta}_{.,.} = \underset{\beta_{.,.},\ \mathrm{supp}(\beta_t)=\hat{S}^Y}{\arg\min} \left\{ \mathcal{E}(\beta_{.,.}) \right\}. \tag{12}$$

6.1 Choice of Penalty

We now turn to choice of the penalty parameters λ_D and λ_Y. These must be chosen so that, with high probability, the penalty dominates the noise, which is captured by the magnitude of the score in the dual of the $|||\cdot|||_{2,1}$ norm.

For linear regression, we set

$$\lambda_Y = \frac{4\mathcal{X}\mathcal{U}\sqrt{\bar{\mathcal{T}}}}{\sqrt{\underline{n}}} \left(1 + \frac{\log(p \vee \underline{n})^{3/2+\delta_Y}}{\sqrt{\bar{\mathcal{T}}}} \right)^{1/2}, \tag{13}$$

for some $\delta_Y > 0$, so that

$$\lambda_Y > 4 \max_{j \in \mathbb{N}_p} \|\mathbb{E}_{n,t}[u_i x^*_{i,j}]\|_2,$$

with probability $1 - \mathcal{P}$ for small (and shrinking) $\mathcal{P}$, following Lounici, Pontil, van de Geer, and Tsybakov (2011). This penalty is of the form $\lambda_Y \propto \Lambda(1 + r_n)$, where Λ is an upper bound on the true score. The rate r_n balances the rate of convergence against the concentration effect: larger r_n slows the rate of convergence, but makes the probability of concentration of the group lasso estimate higher, by shrinking $\mathcal{P}$.

For the multinomial logistic regression, we instead find λ_D such that

$$\lambda_D > 2 \max_{j \in \mathbb{N}_p} \|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t) x^*_{i,j}]\|_2$$

with probability $1 - \mathcal{P}$. Note that $\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})$ appears instead of $p_t(x_i)$,[15] which implies that the bias and noise are simultaneously dominated. To achieve this, we set

$$\lambda_D = 2\mathcal{X}\sqrt{\mathcal{T}} \left[b_s^d + \frac{1}{\sqrt{\underline{n}}} \left(1 + \frac{\log(p \vee \underline{n})^{3/2+\delta_D}}{\sqrt{\mathcal{T}}} \right)^{1/2} \right], \tag{14}$$

for some $\delta_D > 0$. The form of λ_D is $\Gamma + \Lambda(1+r_n)^{1/2}$, where the added Γ bounds the bias contribution. To the best of our knowledge, choosing the penalty in this way to handle an approximately sparse, nonlinear model is new in the high-dimensional, sparse literature, and may be useful in future research.

[15] The multiple of 2 instead of 4 is a related technicality.

In the Appendix we show that, for $\delta = \delta_Y$ or δ_D, the concentration probability is given by

$$\mathcal{P} = \frac{4\sqrt{\log(2p)(1 + 64\log(12p)^2)}}{\log(p \vee \underline{n})^{3/2+\delta}}, \tag{15}$$

Remark 4. For given δ_Y and δ_D, the only unknown quantities in λ_Y and λ_D are $\mathcal{X}$, $\mathcal{U}$, and b_s^d. In practice, we set $b_s^d = 0$ for two reasons: first, bias estimation may be difficult; and second, the only consequence of a smaller penalty is (perhaps) a slight reduction in efficiency. We use $\max_{i \le n} \max_{j \in \mathbb{N}_p} |x_{i,j}^*|$ to estimate $\mathcal{X}$, after scaling (see Remark 5). We estimate $\mathcal{U}$ by iteration. Given an initial estimate $\hat{\mu}_t^{(0)}(x)$, $\hat{\mathcal{U}}^{(k)} = \mathbb{E}_n[(y_i - \hat{\mu}_t^{(k-1)}(x_i))^4]^{1/4}$, where $\hat{\mu}_t^{(k)}(x_i)$, $k > 0$, is based on Eqn. (12). The initial estimate can be least squares on a few variables, a regularized method tuned by cross validation, or other options. ■

Remark 5 (Weighted Penalties)**.** Two final remarks are in order regarding weighting the group lasso penalty. First, one may weight the ℓ_2 portion of the penalty, 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 matrixes are possible.) With this choice, the estimate is invariant to within group (treatment) reparameterizations, and is thus scale invariant for each covariate. We therefore assume throughout that $\mathbb{E}_n[(x_i^*)^2] = 1$ 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. ■

6.2 Restricted Eigenvalues

The local behavior of optimizations (9), (10), (11), and (12) is captured by their respective Hessians, which involve the second moment matrix of the covariates. The eigenvalues of such matrixes will be explicit in our bounds. We are interested in finite sample bounds, and so we will only discuss the empirical Gram matrixes (see Remark 6). Define

$$Q = \mathbb{E}_n[x_i^* x_i^{*\prime}] \qquad \text{and} \qquad Q_t = \mathbb{E}_{n,t}[x_i^* x_i^{*\prime}]. \tag{16}$$

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

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

$$\kappa_D^2 \leq \min_{\delta} \left\{ \frac{\sum_{t \in \mathbb{N}_{\mathcal{T}}} \delta_t' Q \delta_t}{\|\delta_{\cdot,S_D^*}\|_2^2} : \delta \in \mathbb{R}^{p\mathcal{T}} \setminus \{0\}, \left|\!\left|\!\left| \delta_{\cdot,\{S_*^D\}^c} \right|\!\right|\!\right|_{2,1} \leq 3 \left|\!\left|\!\left| \delta_{\cdot,S_*^D} \right|\!\right|\!\right|_{2,1} \right\}. \tag{17}$$

For least squares estimation we instead use

$$\kappa_Y^2 \leq \min_{\delta} \left\{ \frac{\sum_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} \delta_t' Q_t \delta_t}{\|\delta_{\cdot,S_Y^*}\|_2^2} : \delta \in \mathbb{R}^{p\overline{\mathcal{T}}} \setminus \{0\}, \left|\!\left|\!\left| \delta_{\cdot,\{S_*^Y\}^c} \right|\!\right|\!\right|_{2,1} \leq 3 \left|\!\left|\!\left| \delta_{\cdot,S_*^Y} \right|\!\right|\!\right|_{2,1} \right\}. \tag{18}$$

The only difference is that Q appears for κ_D, whereas 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. 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 (11) and (12)) may not obey the cone constraint, but are known to be sparse. This motivates the use of sparse eigenvalues. For a set $S \subset \mathbb{N}_p$ and a $p \times p$ matrix $\tilde{Q}$, define

$$\underline{\phi}\{\tilde{Q}, S\}^2 = \min_{\delta \in \mathbb{R}^p,\, \text{supp}(\delta) = S} \frac{\delta' \tilde{Q} \delta}{\|\delta\|_2^2} \qquad \text{and} \qquad \overline{\phi}\{\tilde{Q}, S\}^2 = \max_{\delta \in \mathbb{R}^p,\, \text{supp}(\delta) = S} \frac{\delta' \tilde{Q} \delta}{\|\delta\|_2^2}. \tag{19}$$

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 $\overline{\overline{\phi}}(\tilde{Q}, m) = \max_{S \subset \mathbb{N}_p,\, |S| \leq m} \overline{\phi}\{\tilde{Q}, S\}$.

We take these quantities to be primitive, and defer discussion to the literature. For example, see van de Geer and Buhlmann (2009), Huang and Zhang (2010), Raskutti, Wainwright, and Yu (2010), Rudelson and Zhou (2011), and Belloni, Chernozhukov, and Hansen (2013). In particular, Huang and Zhang (2010) show that the group lasso may need fewer observations to satisfy conditions on sparse eigenvalues.

Remark 6. Often, invertibility of Q and Q_t relies on their convergence to nonsingular population counterparts.[16] Some of the papers cited above verify conditions on the restricted and sparse eigenvalues by just this approach. Our theorems 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. ■

[16] 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, Belloni, Chen, Chernozhukov, and Kato 2012, Cattaneo and Farrell 2013).

6.3 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 Remarks 7 and 8.

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), and 2(c) hold and that* $\max_{i\leq n} b_{t,i}^d \leq b_s^d$. *Define* $A_p = 0 \vee (p_{\min}/(p_{\min} - b_s^d))$ *and*

$$R_{\mathcal{M}} = \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \frac{3\mathcal{T} A_K \lambda_D \sqrt{s}}{\kappa_D}, \qquad \textit{for} \qquad A_K > 2\frac{\kappa_D^2}{\kappa_D^2 - 8\mathcal{X}\sqrt{\mathcal{T}}\lambda_D s}.$$

Then with probability $1 - \mathcal{P}$

$$\max_{t\in\mathbb{N}_{\mathcal{T}}} \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_{\mathcal{T}}}) - p_t(x_i))^2]^{1/2} \leq R_{\mathcal{M}} + b_s^d,$$

$$\max_{t\in\mathbb{N}_{\mathcal{T}}} \|\tilde{\gamma}_t - \gamma_t^*\|_1 \leq \left(\frac{|\tilde{S}^D \cup S_D^*|}{\underline{\phi}\{Q, \tilde{S}^D \cup S_D^*\}}\right)^{1/2} R_{\mathcal{M}},$$

and

$$|\tilde{S}^D| \leq 8sL_n \left\{\min_{m\in\mathbb{N}_Q^D} \overline{\overline{\phi}}(Q,m)\right\},$$

where

$$\mathbb{N}_Q^D = \left\{m \in \{1,2,\dots n\} : m > 8sL_n\overline{\overline{\phi}}(Q,m)\right\}, \quad \textit{and} \quad L_n = \left(\frac{R_{\mathcal{M}}}{\lambda_D\sqrt{s}}\right)^2.$$

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 matrixes. 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 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 above $p_{\min}$, 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 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 (2011a) "restricted nonlinear impact coefficient" approach, 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 (2011b) 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, Pontil, van de Geer, and Tsybakov 2011), and this may be true here as well.

The error bounds for post-selection estimation are more complex and depends 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. For*

$$A_K > 2\left\{\frac{\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}^2}{\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}^2-\mathcal{X}\sqrt{\mathcal{T}}\lambda_D|\hat{S}_D\cup S_D^*|}\right\}\vee\left\{\frac{\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}}{\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}-2R_\mathcal{M}\mathcal{X}\sqrt{\mathcal{T}}\sqrt{|\hat{S}_D\cup S_D^*|}}\right\},$$

define

$$R'_\mathcal{M}=\left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}}\frac{\mathcal{T}A_K\lambda_D\sqrt{|\hat{S}_D\cup S_D^*|}}{2\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}}\quad\text{and}\quad R''_\mathcal{M}=\{R_\mathcal{M}\}\vee\left\{R'_\mathcal{M}+\left[R'_\mathcal{M}R_\mathcal{M}+\left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}}\mathcal{T}A_KR_\mathcal{M}^2\right]^{1/2}\right\}.$$

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

$$\max_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-p_t(x_i))^2]^{1/2}\leq R''_\mathcal{M}+b_s^d,$$

and

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

This result is the first study of post-selection estimation in approximately sparse logistic models, for any $\mathcal{T} \geq 1$. We explicitly capture the dependence on the loss function $\mathcal{M}(\gamma_{\cdot,\cdot})$ and the impact of the initial group lasso fit. It is not readily discernible if these bounds improve upon the group lasso estimates. This in part depends on the DGP and the selection success of the initial fit. It would be interesting to have an explicit characterization of the improvements offered by refitting. 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.

Remark 7 (Asymptotics for Multinomial Logistic Regression)**.** It is relatively straightforward to state asymptotic rates of convergence, as done in Corollary 1. The first conclusion there is immediate from Theorem 6 if in addition to the conditions required, we also impose that $\lambda_D\sqrt{s} = o(1)$, κ_D and $\min_{S:|S|=O(s)} \underline{\phi}\{Q, S\}$ are bounded away from zero, and $\overline{\overline{\phi}}(Q, \cdot)$ is bounded, uniformly in the set $\mathbb{N}_Q^D$. This also implies that $|\tilde{S}^D| = O_{P_n}(s)$ and $\|\gamma_t - \gamma_t^*\|_1 = O_{P_n}(\sqrt{n^{-1}s^2 \log(p \vee \underline{n})^{3/2+\delta}} + b_s^d s)$, which is important for verifying Assumption 4.

The rates of convergence for the propensity score estimates and the ℓ_1 error of the coefficients are minimax optimal up to a factor of $\log(p \vee \underline{n})^{1/2+\delta}$. A tighter bias condition (by $\sqrt{s}$) is required than in the linear model case, due to the bias in estimating the Hessian.[17] Inspection of the proof shows that this condition can be dropped in the binary case. ■

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 Assumption 2(a), 2(b), 2(c) hold and that* $\max_{i \leq n} b_{t,i}^y \leq b_s^y$*. Define*

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

Then with probability $1 - \mathcal{P}$

$$\max_{t \in \overline{\mathbb{N}}_{\mathcal{T}}} \mathbb{E}_n[(x_i^{*\prime}\tilde{\beta}_t - \mu_t(x_i))^2]^{1/2} \leq \left(\frac{\overline{\phi}\{Q, \tilde{S}^Y \cup S_Y^*\}}{\underline{\phi}\{Q_t, \tilde{S}^Y \cup S_Y^*\}} \right)^{1/2} R_{\mathcal{E}} + b_s^y,$$

[17] The more stringent requirement may be an artifact of the proof. However, it is worth noting that using a different proof method and considering only an oracle series estimator in a semiparametric model, Cattaneo (2010, Theorem B-1) found the same bias requirement.

$$\max_{t\in\overline{\mathbb{N}}_{\mathcal{T}}}\left\|\tilde{\beta}_t-\beta_t^*\right\|_1\leq\left(\frac{|\tilde{S}^Y\cup S_Y^*|}{\underline{\phi}\{Q,\tilde{S}^Y\cup S_Y^*\}}\right)^{1/2}\left(\frac{\overline{\phi}\{Q,\tilde{S}^Y\cup S_Y^*\}}{\underline{\phi}\{Q_t,\tilde{S}^Y\cup S_Y^*\}}\right)^{1/2}R_{\mathcal{E}},$$

and

$$|\tilde{S}^Y|\leq 32sL_n\left\{\min_{m\in\mathbb{N}_Q^Y}\sum_{t\in\overline{\mathbb{N}}_{\mathcal{T}}}\overline{\overline{\phi}}(Q_t,m)\right\},$$

where

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

This theorem generalizes Lounici, Pontil, van de Geer, and Tsybakov (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. multi-task learning), $\mathbb{E}_{n,t}[(x_i^{*\prime}\tilde{\beta}_t-\mu_t(x_i))^2]^{1/2}$, may be found in the Appendix.

For refitting, we are predicting for the entire sample and so we utilize the general results given by Belloni, Chen, Chernozhukov, and Hansen (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* $\log(p)=o(n^{1/3})$ *and* $\min_{j\in\mathbb{N}_p}\mathbb{E}[X_j^{*2}U^2]>0$ *hold in addition to the conditions of Theorem 7. Then*

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

and

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

where for absolute constants A_k*, k=1, 2, 3, 4 that do not depend on n nor the DGP.*

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, Lai, and Shao 2009). This result is less precise than Theorems 5 and 6, but sufficient to verify Assumptions 3 and 4.

Remark 8 (Asymptotics for Multinomial Logistic Regression)**.** As in Remark 7, we can now recover Corollary 1 by imposing that, uniformly in $\overline{\mathbb{N}}_{\mathcal{T}}$: κ_Y and $\min_{S:|S|=O(s)}\underline{\phi}\{Q_t,S\}\wedge\underline{\phi}\{Q,S\}$

are bounded away from zero, and $\overline{\overline{\phi}}(Q,\cdot) \vee \overline{\overline{\phi}}(Q_t,\cdot)$ is bounded, also uniformly in the set $\mathbb{N}_Q^Y$. This also yields $|\tilde{S}^Y| = O_{P_n}(s)$ and $\|\tilde{\beta}_t - \beta_t^*\|_1 = O_{P_n}(\sqrt{n^{-1}s^2\log(p \vee \underline{n})^{3/2+\delta}} + b_s^y\sqrt{s})$. ■

7 Numerical and Empirical Evidence

7.1 Simulation Study

To illustrate the uniform validity of our inference procedure we conducted a small-scale 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. For simplicity we focus on the average effect of a binary treatment with . We generate 1000 observations $(y_i, d_i, x_i')'$, with $p = 500$, from the models in Example 3. The covariates include an intercept, with the remainder drawn from $N(0,\Sigma)$, with covariance $\Sigma[j_1,j_2] = 2^{-|j_1-j_2|}, 2 \leq j_1, j_2 \leq 500$. Errors are standard Normal. The crucial aspect of the DGP are the coefficient vectors β_0^0, β_1^0, and γ^0. We consider a range of models, defined by positive scalars ρ_β, ρ_γ, α_β, and α_γ, as follows:

$$\beta_0^0 = \rho_\beta(-1, 1, -1, 2^{-\alpha_\beta}, -3^{-\alpha_\beta}, \ldots, j^{-\alpha_\beta}, \ldots, p^{-\alpha_\beta})', \qquad \beta_1^0 = -\beta_0^0, \qquad \text{and}$$
$$\gamma^0 = \rho_\gamma(1, -1, 1, -2^{-\alpha_\gamma}, 3^{-\alpha_\gamma}, \ldots, j^{-\alpha_\gamma}, \ldots, -p^{-\alpha_\gamma})'.$$

The multipliers ρ_β and ρ_γ affect the signal-to-noise ratio (the variance is fixed), but not the sparsity. For very small values distinguishing the large and small coefficients is difficult for a given sample size. The exponents α_β and α_γ control the sparsity, where for small values a sparse representation is not possible.

Figure 1 shows the empirical coverage rate of 95% confidence interval for $\mu_1 - \mu_0$ for different DGPs. Panel (a) shows the multipliers ρ_β and ρ_γ ranging over 0.01 (weak signal) to 1 (strong), with $\alpha_\beta = \alpha_\gamma = 2$. Panel (b) varies the sparsity exponents α_β and α_γ range 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 497 to 302 as ρ_γ increases and 444 to 303 as α_γ increases, over their given ranges. Coverage is exceedingly 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) can not be satisfied. Note that coverage accuracy is retained when only one function is sparse, showcasing the double-robustness property.

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

conclusion evaluation of the NSW. As such, 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[18] and 1975), as well as age, education, a marital status, and indicators for Black and Hispanic. Thus, X consists of seven variables.

Our goal is to highlight the role model selection has in inference, and hence we will be interested in comparing specifications. We will keep the estimator fixed: all estimates will be based on the doubly-robust estimator (8) (with standard errors from Section 5.2). We will consider the following three:

1. **No Selection**: X, (earn1974)^2, (earn1975)^2, (age)^2, and (educ)^2;
2. **Informally Selected:** The above, plus $\mathbb{1}\{\text{educ<HS}\}$, $\mathbb{1}\{\text{earn1974=0}\}$, $\mathbb{1}\{\text{earn1975=0}\}$, and ($\mathbb{1}\{\text{earn1974=0}\}\times$Hispanic). This specification was selected by DW02 using an informal balance test.
3. **Group Lasso Selection:** X, $\mathbb{1}\{\text{educ<HS}\}$, $\mathbb{1}\{\text{earn1974=0}\}$, $\mathbb{1}\{\text{earn1975=0}\}$, all possible interactions, and all polynomials up to order five of the continuous covariates (age, educ, earn1974, earn1975).

For specifications 1 and 2, we use the same covariates in the outcome and treatment models. In addition, 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 discard controls with estimated propensity score larger (smaller) than the maximum (minimum) in the treated sample.[19]

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 intervals. The benefits of explicit model selection are clear.

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

[19] 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).

8 Discussion

The main results of this paper established a method to achieve uniformly valid inference on average effects of multivalued treatments 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, we proved new results on group lasso estimation of multinomial logistic regression models. Numerical evidence shows that our method is quite promising for applications.

We handle very general treatment effects models, but restrict attention to studying impacts on the mean. A useful and natural extension would be to consider quantile treatment effects (Firpo 2007) or more generic moment condition based estimands (Cattaneo 2010). Under appropriate regularity conditions, it seems plausible that such an extension can be made. However, the first stage estimation is quite complex in our framework, and this extension would require additional nontrivial technical work. In additional, we plan to develop a formal choice for the penalty parameter that is optimal in some sense, beyond the simple discussion in Section 6.1.

Appendix A Proofs for Treatment Effect Inference

The proofs in this section are asymptotic in nature, compared to the nonasymptotic bounds of the next section. It shall be understood that asymptotic order symbols hold for the sequence being considered, as a shorthand for the more formal versions given in the assumptions (e.g. Assumption 3). C will denote a generic positive constant, which may be a matrix. Define the set of indexes $\mathbb{I}_t = \{i : d_i = t\}$.

A.1 Proofs for Average Treatment Effects

Proof of Theorem 2. SEE SUPPLEMENTAL APPENDIX. □

We first prove Theorem 3.1 assuming there is no additional randomness injected into the support estimates. Following this, we redo the proof to account for additional randomness. We then turn to the remaining portions of Theorem 3 and to Corollary 2, which require shorter arguments.

We make frequent use of the linearization

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

where the first inequality is readily verified, and the second re-applies the first.

Proof of Theorem 3.1 without Additional Randomness. With $\psi_t(\cdot)$ defined in Eqn. (3), 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 $o_{P_n}(1)$.

For R_1, applying the first equality in Eqn. (A.1), we rewrite R_1 as

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

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

$$\mathbb{E}\left[R_1^2 | \{x_i, d_i\}_{i=1}^n \right] = \mathbb{E}_n \left[\frac{d_i^t \sigma_t(x_i)}{p_t(x_i)^4} \left(p_t(x_i) - \hat{p}_t(x_i) \right)^2 \right] \leq C \mathbb{E}_n[(p_t(x_i) - \hat{p}_t(x_i))^2] = o_{P_n}(1).$$

Next, again using Eqn. (A.1) we have

$$1 - \frac{d_i^t}{\hat{p}_t(x_i)} = \frac{p_t(x_i) - d_i^t}{p_t(x_i)} + \frac{d_i^t(\hat{p}_t(x_i) - p_t(x_i))}{\hat{p}_t(x_i)p_t(x_i)}.$$

We use this to re-write $R_2 = R_{21} + R_{22}$, 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, as in R_1, we have

$$\mathbb{E}\left[R_{21}^2 | \{y_i, x_i\}_{i=1}^n\right] = \mathbb{E}_n\left[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2 \left(\frac{p_t(x_i)(1 - p_t(x_i))}{p_t(x_i)^2} \right) \right] \leq C\mathbb{E}_n\left[(\hat{\mu}_t(x_i) - \mu_t(x_i))^2 \right] = o_{P_n}(1),$$

by the first-stage consistency condition of Assumption 3(a). Next, by Hölder's inequality, Assumption 1(b) and the rate condition of Assumption 3(b)

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

This completes the proof, as $|R_1 + R_2| = o_{P_n}(1)$ by Markov's inequality and the triangle inequality. □

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

$$R_{11} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d_i^t u_i}{p_t(x_i)^2} (p_t(x_i) - \hat{p}_t(x_i)) \qquad \text{and} \qquad R_{12} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d_i^t u_i}{p_t(x_i)^2 \hat{p}_t(x_i)} (\hat{p}_t(x_i) - p_t(x_i))^2.$$

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

$$R_{111} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d_i^t u_i}{p_t(x_i)^2} \left(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - \hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) \right) \quad \text{and} \quad R_{112} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d_i^t u_i}{p_t(x_i)^2} \left(p_t(x_i) - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) \right).$$

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

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

and

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

where $v_i = \{x_i^{*\prime}(\hat{\gamma}_t - \gamma_t^*)\}_{\mathbb{N}_\mathcal{T}}$ 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.

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 $\Sigma_{t,j} = \mathbb{E}\left[(x_{i,j}^*)^2 \sigma_{t'}^2(x_i)\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})^2(1-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}\right.$ Then proceed as follows

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

Convergence follows under Assumption 4. 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, Lai, and Shao (2009, Theorem 7.4) and in particular Belloni, Chen, Chernozhukov, and Hansen (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 4. $\Sigma_{t,j}$ normalizes the second moment, and the lemma applies under the first restriction of Assumption 4.

Again by the results of Tanabe and Sagae (1992) and Assumption 3, $v_i'\bar{\mathcal{H}}v_i \leq C\|v_i\|_2^2$. 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:

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

by the union bound and Assumption 4, using the Assumptions 1(b) and 3(a) to apply Eqn. (B.12) 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

$$|R_{12}| \leq \sqrt{n}(\max_{i\in\mathbb{I}_t} |u_i|) \left(\max_{i\in\mathbb{I}_t} \frac{1}{p_t(x_i)^2 \hat{p}_t(x_i)} \right) \mathbb{E}_{n,t}\left[(\hat{p}_t(x_i) - p_t(x_i))^2\right]$$

$$\leq O_{P_n}(1)\sqrt{n}(\max_{i\in\mathbb{I}_t}|u_i|)\mathbb{E}_{n,t}\left[(\hat{p}_t(x_i)-p_t(x_i))^2\right] = o_{P_n}(1),$$

where the rate follows from Assumptions 1(b), 2, and 3, and this tends to zero by Assumption 4.

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

$$R_{211} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i^{*\prime}\hat{\beta}_t - x_i^*\beta_t^*)\left(\frac{p_t(x_i)-d_i^t}{p_t(x_i)}\right) \qquad \text{and} \qquad R_{212} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i^*\beta_t^* - \mu_t(x_i))\left(\frac{p_t(x_i)-d_i^t}{p_t(x_i)}\right)$$

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:

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

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 claim follows directly from the prior result under the moment conditions of Assumption 2(e). □

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

$$\begin{aligned}
R_{W,1} &= \mathbb{E}_n\left[\frac{d_i^t u_i^2}{\hat{p}_t(x_i)^2 p_t(x_i)^2}\left(\hat{p}_t(x_i)-p_t(x_i)\right)\left(\hat{p}_t(x_i)+p_t(x_i)\right)\right], \\
R_{W,2} &= \mathbb{E}_n\left[\frac{d_i^t(\mu_t(x_i)-\hat{\mu}_t(x_i))^2}{\hat{p}_t(x_i)^2}\right], \qquad \text{and} \qquad R_{W,3} = 2\mathbb{E}_n\left[\frac{d_i^t u_i(\mu_t(x_i)-\hat{\mu}_t(x_i))}{\hat{p}_t(x_i)^2}\right].
\end{aligned}$$

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

$$R_{W,1} \leq \left(\max_{i\in\mathbb{I}_t}\frac{\hat{p}_t(x_i)+p_t(x_i)}{\hat{p}_t(x_i)^2 p_t(x_i)^2}\right)\mathbb{E}_n[d_i^t|u_i|^4]^{1/2}\mathbb{E}_n[d_i^t(\hat{p}_t(x_i)-p_t(x_i))^2]^{1/2} = o_{P_n}(1),$$

$$R_{W,2} \leq \left(\max_{i\in\mathbb{I}_t}\frac{1}{\hat{p}_t(x_i)^2}\right)\mathbb{E}_n[d_i^t(\hat{\mu}_t(x_i)-\mu_t(x_i))^2] = o_{P_n}(1),$$

and,

$$R_{W,3} \leq 2\left(\max_{i\in\mathbb{I}_t}\frac{1}{\hat{p}_t(x_i)^2}\right)\mathbb{E}_n[d_i^t|u_i|^2]^{1/2}\mathbb{E}_n[d_i^t(\hat{\mu}_t(x_i)-\mu_t(x_i))^2]^{1/2} = o_{P_n}(1),$$

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 $\mathbb{E}_n[d_i^t u_i^2 p_t(x_i)^{-2}] - V_{\boldsymbol{\mu}}^W(t)| = o_{P_n}(1)$, under Assumptions 1(b) and 2(c).

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

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

$$R_{B,2}(t,t') = \hat{\mu}_t\mathbb{E}_n\left[\hat{\mu}_{t'}(x_i)-\mu_{t'}(x_i)\right], \quad \text{and} \quad R_{B,3}(t,t') = \mathbb{E}_n\left[\mu_t(x_i)(\hat{\mu}_{t'}(x_i)-\mu_{t'}(x_i))\right].$$

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

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

Consistency of $\hat{V}_{\boldsymbol{\mu}}^B(t,t')$ 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 $P_m \in \mathcal{P}_m$, 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. □

A.2 Proofs for Average Treatment Effects on Treated Groups

Proofs are similar to those for Theorem 3 and Corollary 2, and hence we omit them to save space.

Appendix B Proofs for Group Lasso Selection and Estimation

Unless otherwise noted, all bounds in this section are nonasymptotic. Further, as the proofs are segregated we will use the generic notation X^* and s for the covariates and sparsity level.

B.1 Proofs for Multinomial Logistic Models

B.1.1 Lemmas

Lemma B.1 (Score Bound)**.** *For λ_D and $\mathcal{P}$ defined respectively in Eqn.* (14) *and Eqn.* (15) *we have*

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

Proof. First, by the Cauchy-Schwarz inequality and Assumption 2(b) and the bias condition, we have

$$\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-p_t(x_i))x_{i,j}^*]\|_2 \leq \mathcal{X}\|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-p_t(x_i))^2]^{1/2}]\|_2 \leq \mathcal{X}b_s^d\sqrt{\mathcal{T}}.$$

Therefore, by the triangle inequality and the definition of λ_D, with $r_n = \mathcal{T}^{-1/2}\log(p\vee\underline{n})^{3/2+\delta}$,

$$\begin{aligned}
&\mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_{i,j}^*]\|_2 \geq \frac{\lambda_D}{2}\right] \\
&\quad\leq \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2 + \max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-p_t(x_i))x_{i,j}^*]\|_2 \geq \frac{\lambda_D}{2}\right] \\
&\quad= \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2 + \max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-p_t(x_i))x_{i,j}^*]\|_2 \geq \mathcal{X}\sqrt{\mathcal{T}}\left[b_s^d + \frac{1}{\sqrt{\underline{n}}}(1+r_n)^{1/2}\right]\right] \\
&\quad= \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2 \geq \frac{\mathcal{X}\sqrt{\mathcal{T}}}{\sqrt{\underline{n}}}(1+r_n)^{1/2}\right] \\
&\quad= \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2^2 \geq \frac{\mathcal{X}^2\mathcal{T}}{\underline{n}}(1+r_n)\right],
\end{aligned}$$

canceling the bias terms from each side the squaring.

The residuals $v_{t,i}$ are conditionally mean-zero by definition and satisfy $\mathbb{E}[v_{t,i}^2|x_i] \leq 1$. Using this, Assumption 2(a) and the definition of $\mathcal{X}$, we find that

$$\mathbb{E}\left[\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2^2\right] = \sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}\left[\mathbb{E}_n[v_{t,i}x_{i,j}^*]^2\right] = \sum_{t\in\mathbb{N}_\mathcal{T}}\frac{1}{n}\mathbb{E}[v_{t,i}^2(x_{i,j}^*)^2] \leq \frac{\mathcal{X}^2\mathcal{T}}{n}$$

uniformly in $j\in\mathbb{N}_p$. Define the mean-zero random variables $\xi_{t,j}$ as:

$$\xi_{t,j} = (\mathbb{E}_n[v_{t,i}x_{i,j}^*])^2 - \frac{1}{n}\mathbb{E}[V_t^2X_j^{*2}].$$

Thus, we further bound the probability as follows.

$$\mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2 \geq \frac{\mathcal{X}^2\mathcal{T}}{\underline{n}}(1+r_n)\right] = \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2^2 - \frac{\mathcal{X}^2\mathcal{T}}{n} \geq \frac{\mathcal{X}^2\mathcal{T}r_n}{n}\right]$$

$$\leq \mathbb{P}\left[\max_{j\in\mathbb{N}_p}\sum_{t\in\mathbb{N}_\mathcal{T}}\xi_{t,j} \geq \frac{\mathcal{X}^2\mathcal{T}r_n}{n}\right]$$
$$\leq \mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\sum_{t\in\mathbb{N}_\mathcal{T}}\xi_{t,j}\right|\right]\frac{n}{\mathcal{X}^2\mathcal{T}r_n}, \tag{B.1}$$

where final line follows from Markov's inequality.

Next, applying Lemma 9.1 of Lounici, Pontil, van de Geer, and Tsybakov (2011) (with their $m = 1$ and hence $c(m) = 2$) followed by Jensen's inequality and Assumption 2(c), we find that

$$\mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\sum_{t\in\mathbb{N}_\mathcal{T}}\xi_{t,j}\right|\right] \leq (8\log(2p))^{1/2}\mathbb{E}\left[\left(\sum_{t\in\mathbb{N}_\mathcal{T}}\max_{j\in\mathbb{N}_p}\xi_{t,j}^2\right)^{1/2}\right]$$
$$\leq (8\log(2p))^{1/2}\left(\mathbb{E}\left[\sum_{t\in\mathbb{N}_\mathcal{T}}\max_{j\in\mathbb{N}_p}\xi_{t,j}^2\right]\right)^{1/2}$$
$$\leq 4\log(2p)^{1/2}\left(\sum_{t\in\mathbb{N}_\mathcal{T}}\frac{\mathcal{X}^4}{n^2} + \sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\right|^4\right]\right)^{1/2}. \tag{B.2}$$

The leading 4 is $\sqrt{8}\sqrt{2}$, where $\sqrt{2}$ is a byproduct of applying the inequality $(a-b)^2 \leq 2(a^2+b^2)$ to $\xi_{t,j}^2$. Again using Lemma 9.1 of Lounici, Pontil, van de Geer, and Tsybakov (2011) (with their $m = 4$, and $c(m) = 12$ since $c(4) \geq (e^{4-1}-1)/2 + 2 \approx 11.54$), we bound the expectation in the second term above as follows:

$$\mathbb{E}\left[\max_{j\in\mathbb{N}_p}\left|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\right|^4\right] \leq [8\log(12p)]^{4/2}\mathbb{E}\left[\left(\sum_{i=1}^n\max_{j\in\mathbb{N}_p}\left|\frac{v_{t,i}x_{i,j}^*}{n}\right|^2\right)^{4/2}\right] \leq \frac{64\log(12p)^2\mathcal{X}^4}{n^2}, \tag{B.3}$$

using Assumptions 2(a) and 2(b).

Now, inserting the results of Eqns. (B.2) and (B.3) into Eqn. (B.1), we have

$$\mathbb{P}\left[\max_{j\in\mathbb{N}_p}\|\mathbb{E}_n[v_{t,i}x_{i,j}^*]\|_2 \geq \frac{\lambda_D}{4}\right] \leq \frac{4n\log(2p)^{1/2}}{\mathcal{T}\mathcal{X}^2 r_n}\left(\sum_{t\in\mathbb{N}_\mathcal{T}}\frac{\mathcal{X}^4}{n^2} + \sum_{t\in\mathbb{N}_\mathcal{T}}\frac{64\log(12p)^2\mathcal{X}^4}{n^2}\right)^{1/2}$$
$$\leq \frac{4\log(2p)^{1/2}}{r_n\sqrt{\mathcal{T}}}[1+64\log(12p)^2]^{1/2} = \mathcal{P},$$

using the choice $r_n = \mathcal{T}^{-1/2}\log(p\vee\underline{n})^{3/2+\delta}$. □

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

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

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

$$\mathbb{E}_n[x_{i,j}^*(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-d_i^t)] = \lambda_D\frac{\tilde{\gamma}_{t,j}}{\|\tilde{\gamma}_{\cdot,j}\|_2}. \tag{B.4}$$

Hence, 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, and the score bound (B.1), we find that

$$\begin{aligned}\lambda_D &= \left\|\mathbb{E}_n[x_{i,j}^*(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-d_i^t)]\right\|_2\\ &\leq \left\|\mathbb{E}_n[x_{i,j}^*(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)]\right\|_2 + \left\|\mathbb{E}_n[x_{i,j}^*(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))]\right\|_2\\ &\leq \lambda_D/2 + \left\|\mathbb{E}_n[x_{i,j}^*(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))]\right\|_2.\end{aligned}$$

Let $\boldsymbol{P}_t^*$ be the vector of $\{\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})\}_{i=1}^n$ and similarly for $\tilde{\boldsymbol{P}}_t$. Collecting terms, then squaring both sides and and summing over $j\in\tilde{S}^D$ (i.e. applying $\|\cdot\|_2^2$ over $j\in\tilde{S}^D$ to both sides) yields

$$\begin{aligned}\sum_{j\in\tilde{S}^D}\lambda_D^2 &\leq 4\sum_{j\in\tilde{S}^D}\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n[x_{i,j}^*(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))]^2\\ &= 4\sum_{t\in\mathbb{N}_\mathcal{T}}\frac{1}{n^2}\left\|\left[\boldsymbol{X}'(\tilde{\boldsymbol{P}}_t-\boldsymbol{P}_t^*)\right]_{j\in\tilde{S}^D}\right\|_2^2\\ &\leq 4\sum_{t\in\mathbb{N}_\mathcal{T}}\frac{\overline{\phi}\{Q,\tilde{S}^D\}}{n}\left\|\tilde{\boldsymbol{P}}_t-\boldsymbol{P}_t^*\right\|_{2,n}^2\\ &\leq 4\overline{\phi}\{Q,\tilde{S}^D\}\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}})-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))^2\right].\end{aligned}$$

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

Lemma B.3 (Cone Constraint)**.** *Define* $\tilde{\delta}_{\cdot,\cdot}=\tilde{\gamma}_{\cdot,\cdot}-\gamma_{\cdot,\cdot}^*$*. With probability* $1-\mathcal{P}$*,* $\tilde{\delta}_{\cdot,\cdot}$ *obeys the cone constraint required by the definition of* κ_D*.*

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

$$\begin{aligned}\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_i^{*\prime}\tilde{\delta}_t\right] &= \sum_{j\in\mathbb{N}_p}\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_{i,j}^*\right]\tilde{\delta}_{t,j}\\ &\leq \sum_{j\in\mathbb{N}_p}\sqrt{\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_{i,j}^*\right]^2}\sqrt{\sum_{t\in\mathbb{N}_\mathcal{T}}\tilde{\delta}_{t,j}^2}\\ &\leq \max_{j\in\mathbb{N}_p}\left\{\left\|\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_{i,j}^*\right]\right\|_2\right\}\sum_{j\in\mathbb{N}_p}\left\|\tilde{\delta}_{\cdot,j}\right\|_2\end{aligned}$$

$$\leq \frac{\lambda_D}{2}\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}, \tag{B.5}$$

with probability at least $1-\mathcal{P}$.

By the optimality of $\tilde{\delta}_{\cdot,\cdot}$, we have

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

implying

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

applying the convexity of $\mathcal{M}$. Using the bound in Eqn. (B.5) and rearranging we find that

$$0 \leq \lambda_D\left\{\left|\left|\left|\gamma^*_{\cdot,\cdot}\right|\right|\right|_{2,1}-\left|\left|\left|\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\right\}+\frac{\lambda_D}{2}\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}.$$

Canceling λ_D and decomposing the supports, we find that

$$\begin{aligned}0 &\leq \frac{1}{2}\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}+\left\{\left|\left|\left|\gamma^*_{\cdot,\cdot}\right|\right|\right|_{2,1}-\left|\left|\left|\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\right\} \\ &= \frac{1}{2}\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}+\frac{1}{2}\left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}+\left|\left|\left|\gamma^*_{\cdot,S_*}\right|\right|\right|_{2,1}-\left|\left|\left|\gamma^*_{\cdot,S_*}+\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}-\left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1},\end{aligned}$$

where the second line follows because $\gamma^*_{\cdot,S_*^c}=0$. Collecting terms and applying the triangle inequality yields

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

Hence $\tilde{\delta}_{\cdot,\cdot}$ belongs to the restricted set of (17). □

B.1.2 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

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

Rearranging and subtracting the score, we have

$$\begin{aligned}\mathcal{M}(\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot})-\mathcal{M}(\gamma^*_{\cdot,\cdot})-\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma^*_t\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t \\ \leq \lambda_D\left\{\left|\left|\left|\gamma^*_{\cdot,\cdot}\right|\right|\right|_{2,1}-\left|\left|\left|\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}\right\}-\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma^*_t\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t.\end{aligned} \tag{B.6}$$

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 $\mathbb{E}_n[(x_i^{*\prime}\tilde{\delta}_t)^2]^{1/2}$.

Let us begin with the right side of Eqn. (B.6). For the penalized difference of coefficients we have

$$\left|\left|\left|\gamma^*_{\cdot,S_*^c}\right|\right|\right|_{2,1}-\left|\left|\left|\gamma^*_{\cdot,S_*^c}+\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1}=\left|\left|\left|\tilde{\delta}_{\cdot,S_*^c}\right|\right|\right|_{2,1},$$

because $\gamma^*_{\cdot,S_*^c}=0$. Therefore,

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

where the second step follows from the triangle inequality and dropping the nonnegative norm, and the third by the triangle inequality again. Thus, using this result for the first term and the bound (B.5) for the second, we find that the right side of Eqn. (B.6) is bounded as follows, using the cone constraint, the Cauchy-Schwarz inequality, and the definition of κ_D from Eqn. (17),

$$\begin{aligned}\lambda_D\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}+\frac{\lambda_D}{2}\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1} &\leq \lambda_D\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}+\frac{\lambda_D}{2}\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1}+\frac{\lambda_D}{2}3\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1} \\ &= 3\lambda_D\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1} \\ &\leq 3\lambda_D\sqrt{|S_*|}\left\|\tilde{\delta}_{\cdot,S_*}\right\|_2 \\ &\leq \frac{3\lambda_D\sqrt{|S_*|}}{\kappa_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}.\end{aligned} \tag{B.7}$$

Note that $\sum_{t\in\mathbb{N}_\mathcal{T}}\tilde{\delta}_t'Q\tilde{\delta}_t=\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]$.

Now turn to the left side of Eqn. (B.6). 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 $\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})=\exp\{x_i^{*\prime}\gamma_t\}/\left(1+\sum_{\mathbb{N}_\mathcal{T}}\exp\{x_i^{*\prime}\gamma_t\}\right)$. Define the $\mathcal{T}$-square

matrix $\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})$ as having the $(t,t') \in \mathbb{N}_\mathcal{T}^2$ 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 $\mathcal{M}(\gamma_{.},.)$ can be written as

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

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 $\mathcal{M}(\gamma_{.},.) = \mathbb{E}_n\left[F(w_i) - \sum_{t\in\mathbb{N}_\mathcal{T}} d_i^t w_{i,t}\right]$, 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 he 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 $\partial\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})/\partial\gamma_{t'},. = -\hat{p}_t(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})\hat{p}_{t'}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})x_i^*$. From these, we find that

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

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 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

$$\begin{aligned} M(\gamma_{\cdot,\cdot}^* + \tilde{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma_{\cdot,\cdot}^*) - \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t \\ \geq \mathbb{E}_n\left[\frac{v_i'\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})v_i}{4\mathcal{T}\|v_i\|_2^2}\left(e^{-2\|v_i\|_2} + 2\|v_i\|_2 - 1\right)\right] \\ \geq \mathbb{E}_n\left[\frac{v_i'\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})v_i}{4\mathcal{T}\|v_i\|_2^2}\left(2\|v_i\|_2^2 - \frac{4}{3}\|v_i\|_2^3\right)\right], \end{aligned} \tag{B.8}$$

where the second inequality follows from Belloni, Chernozhukov, and Wei (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

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

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

$$\begin{aligned} \mathbb{E}_n\left[\frac{v_i'\mathcal{H}(\{x_i^{*\prime}\gamma_t\}_{\mathbb{N}_\mathcal{T}})v_i}{4\mathcal{T}\|v_i\|_2^2}\left(2\|v_i\|_2^2 - \frac{4}{3}\|v_i\|_2^3\right)\right] &\geq \left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}} \frac{1}{4\mathcal{T}}\mathbb{E}_n\left[2\|v_i\|_2^2 - \frac{4}{3}\|v_i\|_2^3\right] \\ &= \left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}} \frac{1}{\mathcal{T}}\frac{\mathbb{E}_n[\|v_i\|_2^2]}{2}\left(1 - \frac{2}{3}\frac{\mathbb{E}_n[\|v_i\|_2^3]}{\mathbb{E}_n[\|v_i\|_2^2]}\right). \end{aligned} \tag{B.9}$$

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.8) and (B.9) gives

$$M(\gamma_{\cdot,\cdot}^* + \tilde{\delta}_{\cdot,\cdot}) - \mathcal{M}(\gamma_{\cdot,\cdot}^*) - \sum_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)x_i^{*\prime}\right]\tilde{\delta}_t \geq \left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}} \frac{1}{\mathcal{T}}\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}{A_K}. \tag{B.10}$$

Now consider the second case, where this bound does not hold. By Lemma B.3, $\tilde{\delta}_{\cdot,\cdot}$ is in the cone defined by (17), and therefore

$$\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_1 = \sum_{t\in\mathbb{N}_\mathcal{T}}\sum_{j\in\mathbb{N}_p}\left|x_{i,j}^*\tilde{\delta}_{t,j}\right| \leq \mathcal{X}\left\|\tilde{\delta}_{\cdot,\cdot}\right\|_1 \leq \sqrt{\mathcal{T}}\mathcal{X}\left|\left|\left|\tilde{\delta}_{\cdot,\cdot}\right|\right|\right|_{2,1}$$

$$= \sqrt{\mathcal{T}}\mathcal{X}4\left|\left|\left|\tilde{\delta}_{\cdot,S_*}\right|\right|\right|_{2,1} \leq \sqrt{\mathcal{T}}\mathcal{X}4\sqrt{|S_*|}\left\|\tilde{\delta}_{\cdot,S_*}\right\|_2 \leq \sqrt{\mathcal{T}}\mathcal{X}4\sqrt{|S_*|}\kappa_D^{-1}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2},$$

using Assumption 2(b), the Cauchy-Schwarz inequality, decomposing the support of $\delta_{\cdot,\cdot}$, and then following the same steps as (B.7). Hence, by subadditivity,

$$\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^3] \leq \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_1] \leq \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{3/2}\sqrt{\mathcal{T}}\mathcal{X}4\sqrt{|S_*|}\kappa_D^{-1}.$$

Thus

$$\frac{1}{A_K} > \frac{1}{2}\left(1 - \frac{2}{3}\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^3]}{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}\right) \geq \frac{1}{2}\left(1 - \frac{\sqrt{\mathcal{T}}\mathcal{X}8\sqrt{|S_*|}}{3\kappa_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}\right),$$

which is equivalent to

$$\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} > 3\left(1 - \frac{2}{A_K}\right)\frac{\kappa_D}{8\mathcal{X}\sqrt{\mathcal{T}}\sqrt{|S_*|}} \equiv r_n.$$

Because $\mathcal{M}(\gamma^*_{\cdot,\cdot} + \delta_{\cdot,\cdot}) - \mathcal{M}(\gamma_{\cdot,\cdot}) - \sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - d_i^t)x_i^{*\prime}\right]\delta_t$ is convex in $\delta_{\cdot,\cdot}$, and hence any line segment lies above the function, we have know that $\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} > r_n$, 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 Equations (B.6) and (B.7), we have

$$3\left(1 - \frac{2}{A_K}\right)\frac{\kappa_D}{8\mathcal{X}\sqrt{\mathcal{T}}\sqrt{|S_*|}}\mathbb{E}_n[\|\{x_i^{*\prime}\delta_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \leq \frac{3\lambda_D\sqrt{|S_*|}}{\kappa_D}\mathbb{E}_n[\|\{x_i^{*\prime}\delta_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2},$$

which is impossible under the restriction on A_K.

Therefore, Eqn. (B.10) must hold.[20] Combining this with Equations (B.6) and (B.7), we find that

$$\left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}}\frac{1}{\mathcal{T}}\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}{A_K} \leq \frac{3\lambda_D\sqrt{|S_*|}}{\kappa_D}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}.$$

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

$$\max_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n[(x_i^{*\prime}\tilde{\delta}_t)^2]^{1/2} \leq \mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \leq \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}}\frac{3\mathcal{T}A_K\lambda_D\sqrt{|S_*|}}{\kappa_D}. \tag{B.11}$$

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

[20] Intuitively, the deviation $\tilde{\delta}_{\cdot,\cdot}$ is too large for the quadratic bound to hold, and so this analysis is conceptually similar to using Belloni and Chernozhukov's (2011a) restricted nonlinearity impact coefficient, but our characterization is different.

$\tilde{m}_t \in [0,1]$ and entries $t' \in \mathbb{N}_\mathcal{T} \setminus \{t\}$ equal to $x_i^{*\prime}\gamma_{t'}$. Then we have

$$\left|\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})\right| = \left|\hat{p}_t(M_t)[1-\hat{p}_t(M_t)]x_i^{*\prime}\tilde{\delta}_t\right| \leq \left|x_i^{*\prime}\tilde{\delta}_t\right|. \tag{B.12}$$

Using this result coupled with the triangle inequality, the bias condition, and Eqn. (B.11), we find

$$\begin{aligned}
\mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2]^{1/2} &\leq \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\tilde{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - \hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))^2]^{1/2} + \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2]^{1/2} \\
&\leq \mathbb{E}_n\left[(x_i^{*\prime}\tilde{\delta}_t)^2\right]^{1/2} + b_s^d \\
&\leq \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \frac{3\mathcal{T}A_K\lambda_D\sqrt{|S_*|}}{\kappa_D} + b_s^d.
\end{aligned}$$

The ℓ_1 bound follows from Eqn. (B.11) by the Cauchy-Schwarz inequality and the definition in Eqn. (19):

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

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.11), then bounding $\overline{\phi}$ by $\overline{\overline{\phi}}$ we find that

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

The argument now parallels that used by Belloni and Chernozhukov (2011b), 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,

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

Rearranging gives

$$m \leq |S_D^*|8L_n\overline{\overline{\phi}}\{Q, m\}$$

whence $m \notin \mathbb{N}_Q^D$. Minimizing over $\mathbb{N}_Q^D$ gives the result. □

B.1.3 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 Eqn. (19),

$$\begin{aligned}
\left|\!\left|\!\left|\hat{\delta}_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1} &= \sum_{j\in\hat{S}_D\cup S_D^*}\left\|\hat{\delta}_{\cdot,j}\right\|_2 \\
&\leq \sqrt{\left|\hat{S}_D\cup S_D^*\right|}\sqrt{\sum_{t\in\mathbb{N}_\mathcal{T}}\sum_{j\in\hat{S}_D\cup S_D^*}\hat{\delta}_{t,j}^2} \\
&= \sqrt{\left|\hat{S}_D\cup S_D^*\right|}\sqrt{\sum_{t\in\mathbb{N}_\mathcal{T}}\left\|\hat{\delta}_{\cdot,j}\right\|_2^2} \\
&\leq \sqrt{\left|\hat{S}_D\cup S_D^*\right|}\sqrt{\sum_{t\in\mathbb{N}_\mathcal{T}}\underline{\phi}\left\{Q,\hat{S}_D\cup S_D^*\right\}^{-2}\hat{\delta}_t'Q\hat{\delta}_t} \\
&= \sqrt{\left|\hat{S}_D\cup S_D^*\right|}\underline{\phi}\left\{Q,\hat{S}_D\cup S_D^*\right\}^{-1}\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}. \qquad \text{(B.13)}
\end{aligned}$$

Combining this bound with that of (B.5) yields

$$\left|\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_i^{*\prime}\right]\hat{\delta}_t\right| \leq \frac{\lambda_D}{2}\left|\!\left|\!\left|\hat{\delta}_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1} \leq \frac{\lambda_D}{2}\sqrt{\left|\hat{S}_D\cup S_D^*\right|}\underline{\phi}\left\{Q,\hat{S}_D\cup S_D^*\right\}^{-1}\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}. \qquad \text{(B.14)}$$

Next we $\mathcal{M}(\gamma^*_{\cdot,\cdot}+\hat{\delta}_{\cdot,\cdot})-\mathcal{M}(\gamma^*_{\cdot,\cdot})$. 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 the right side of the prior display is bounded by $\mathcal{M}(\tilde{\gamma}_{\cdot,\cdot})-\mathcal{M}(\gamma^*_{\cdot,\cdot})$. By the mean value theorem, for scalars $\{m_t\in[0,1]\}_{\mathbb{N}_\mathcal{T}}$, $\mathcal{M}(\tilde{\gamma}_{\cdot,\cdot})-\mathcal{M}(\gamma^*_{\cdot,\cdot})$, the bound of (B.5), the same steps in (B.12), and (B.13) with $\tilde{\delta}_{\cdot,\cdot}$:

$$\begin{aligned}
\mathcal{M}(\gamma^*_{\cdot,\cdot}+\tilde{\delta}_{\cdot,\cdot})-\mathcal{M}(\gamma^*_{\cdot,\cdot}) &= \sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(d_i^t-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*+m_tx_i^{*\prime}\tilde{\delta}_t\}))x_i^{*\prime}\tilde{\delta}_t\right] \\
&= \sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(d_i^t-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}}))x_i^{*\prime}\tilde{\delta}_t\right] \\
&\quad+\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-\hat{p}_t(\{x_i^{*\prime}\gamma_t^*+m_tx_i^{*\prime}\tilde{\delta}_t\}))x_i^{*\prime}\tilde{\delta}_t\right], \\
&\leq \frac{\lambda_D}{2}\left|\!\left|\!\left|\tilde{\delta}_{\cdot,\cdot}\right|\!\right|\!\right|_{2,1}+\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[m_t(x_i^{*\prime}\tilde{\delta}_t)^2\right]. \\
&\leq \frac{\lambda_D}{2}\frac{\sqrt{|\hat{S}_D\cup S_D^*|}}{\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}}\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}+\mathbb{E}_n[\|\{x_i^{*\prime}\tilde{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2], \qquad \text{(B.15)}
\end{aligned}$$

using that $m_t\leq 1$.

Collecting the bounds of (B.14) and (B.15), and the definition of $R_\mathcal{M}$ gives

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

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:

$$M(\gamma^*_{\cdot,\cdot}+\hat{\delta}_{\cdot,\cdot})-\mathcal{M}(\gamma^*_{\cdot,\cdot})-\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_i^{*\prime}\right]\hat{\delta}_t\geq\left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}}\frac{1}{\mathcal{T}}\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}{A_K}. \quad \text{(B.16)}$$

In the other case, this bound may not hold. Arguing as in the proof of Theorem 5, but applying Eqn. (B.13), we get

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

Therefore, as above, we find

$$\mathcal{M}(\gamma^*_{\cdot,\cdot}+\hat{\delta}_{\cdot,\cdot})-\mathcal{M}(\gamma_{\cdot,\cdot})-\sum_{t\in\mathbb{N}_\mathcal{T}}\mathbb{E}_n\left[(\hat{p}_t(\{x_i^{*\prime}\gamma_t^*\}_{\mathbb{N}_\mathcal{T}})-d_i^t)x_i^{*\prime}\right]\hat{\delta}_t\geq r_n\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}, \quad \text{(B.17)}$$

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 bounds of (B.14) and (B.15) and the lower bounds (B.16) and (B.17), and using the definition of $R_\mathcal{M}$, we have

$$\left\{\left(\frac{p_{\min}}{A_p}\right)^{\overline{\mathcal{T}}}\frac{1}{\mathcal{T}}\frac{\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]}{A_K}\right\}\wedge\left\{r_n\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}\right\}$$
$$\leq\frac{\lambda_D}{2}\frac{\sqrt{|\hat{S}_D\cup S_D^*|}}{\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}}\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2}+\frac{\lambda_D}{2}\frac{\sqrt{|\hat{S}_D\cup S_D^*|}}{\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}}R_\mathcal{M}+R_\mathcal{M}^2.$$

For some $A_1>1$, replace the restriction on A_K in the Theorem with the requirement that

$$A_K>2\left\{\frac{\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}^2}{\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}^2-(A_1/3)\mathcal{X}\sqrt{\mathcal{T}}\lambda_D|\hat{S}_D\cup S_D^*|}\right\}$$
$$\vee\left\{\frac{\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}}{\underline{\phi}\{Q,\hat{S}_D\cup S_D^*\}-(A_1/3)2R_\mathcal{M}\mathcal{X}\sqrt{\mathcal{T}}\sqrt{|\hat{S}_D\cup S_D^*|}}\right\}.$$

Suppose the linear term is the minimum. Then, with the restrictions on A_K (and hence r_n),

we have

$$r_n \mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \leq (r_n/A_1)\left(\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} + R_\mathcal{M}\right) + R_\mathcal{M}^2 \leq (r_n/A_1)\left(\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} + 2R_\mathcal{M}\right)$$

Therefore

$$\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \leq \frac{2R_\mathcal{M}}{A_1 - 1}.$$

On the other hand, if the quadratic term is the minimum, define

$$R'_\mathcal{M} = \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \frac{\mathcal{T} A_K \lambda_D \sqrt{|\hat{S}_D \cup S_D^*|}}{2\underline{\phi}\{Q, \hat{S}_D \cup S_D^*\}},$$

and we have

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

Then, because $a^2 \leq ab + c$ implies that $a \leq b + \sqrt{c}$, we have the final bound on the log-odds estimates:

$$\mathbb{E}_n[\|\{x_i^{*\prime}\hat{\delta}_t\}_{\mathbb{N}_\mathcal{T}}\|_2^2]^{1/2} \leq R'_\mathcal{M} + \left(R'_\mathcal{M}R_\mathcal{M} + \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \mathcal{T} A_K R_\mathcal{M}^2\right)^{1/2}. \tag{B.18}$$

From this bound on the log-odds estimates, we obtain the bound on the propensity score estimates and the ℓ_1 rate, given by,

$$\max_{t\in\mathbb{N}_\mathcal{T}} \mathbb{E}_n[(\hat{p}_t(\{x_i^{*\prime}\hat{\gamma}_t\}_{\mathbb{N}_\mathcal{T}}) - p_t(x_i))^2]^{1/2} \leq \left\{\frac{2R_\mathcal{M}}{A_1 - 1}\right\} \vee \left\{R'_\mathcal{M} + \left(R'_\mathcal{M}R_\mathcal{M} + \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \mathcal{T} A_K R_\mathcal{M}^2\right)^{1/2}\right\} + b_s^d,$$

and

$$\max_{t\in\mathbb{N}_\mathcal{T}} \|\hat{\gamma}_t - \gamma_t^*\|_1 \leq \left(\frac{|\tilde{S}^D \cup S_D^*|}{\underline{\phi}\{Q, \tilde{S}^D \cup S_D^*\}}\right)^{1/2} \left\{\frac{2R_\mathcal{M}}{A_1 - 1}\right\} \vee \left\{R'_\mathcal{M} + \left(R'_\mathcal{M}R_\mathcal{M} + \left(\frac{A_p}{p_{\min}}\right)^{\overline{\mathcal{T}}} \mathcal{T} A_K R_\mathcal{M}^2\right)^{1/2}\right\},$$

by arguments parallel to those used in the proof of Theorem 5. The results as stated now follow by setting $A_1 = 3$. □

B.2 Proofs for Linear Models

SEE SUPPLEMENTAL APPENDIX.

9 References

ABADIE, A. (2005): "Semiparametric difference-in-differences estimators," *Review of Economic Studies*, 72, 1–19.

ABADIE, A., AND G. W. IMBENS (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74(1), 235–267.

ANDREWS, D. W. K., AND P. GUGGENBERGER (2009): "Incorrect asymptotic size of subsampling procedures based on post-consistent model selection estimators," *Journal of Econometrics*, 152, 19–27.

BACH, F. R. (2008): "Consistency of the Group Lasso and Multiple Kernel Learning," *Journal of Machine Learning Research*, 9, 1179–1225.

——— (2010): "Self-concordant analysis for logistic regression," *Electronic Journal of Statistics*, 4, 384–414.

BANG, H., AND J. M. ROBINS (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–972.

BELLONI, A., D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN (2012): "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80(6), 2369–2429.

BELLONI, A., X. CHEN, V. CHERNOZHUKOV, AND K. KATO (2012): "On the Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Arxiv preprint arXiv:1212.0442*.

BELLONI, A., AND V. CHERNOZHUKOV (2011a): "ℓ_1-Penalized quantile regression in high-dimensional sparse models," *Annals of Statistics*, 39(1), 82–130.

——— (2011b): "Least Squares After Model Selection in High-dimensional Sparse Models," *Arxiv preprint arXiv:1001.0188v4*.

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2013): "Inference on Treatment Effects after Selection Amongst High-Dimensional Controls," *cemmap working paper CWP26/13*.

BELLONI, A., V. CHERNOZHUKOV, AND Y. WEI (2013): "Honest Confidence Regions for Logistic Regression with a Large Number of Controls," *arXiv:1304.3969v1*.

BERK, R., L. BROWN, A. BUJA, K. ZHANG, AND L. ZHAO (2013): "Valid Post-Selection Inference," *Annals of Statistics*, 4(2), 802–837.

BICKEL, P. J., Y. RITOV, AND A. B. TSYBAKOV (2009): "Simultaneous Analysis of LASSO and Dantzig Selector," *Annals of Statistics*, 37(4), 1705–1732.

BUHLMANN, P., AND S. VAN DE GEER (2011): *Statistics for High-Dimensional Data*, Springer Series in Statistics. Springer-Verlag, Berlin.

CATTANEO, M. D. (2010): "Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability," *Journal of Econometrics*, 155(2), 138–154.

CATTANEO, M. D., D. M. DRUKKER, AND A. D. HOLLAND (forthcoming): "Estimation of multivalued treatment effects under conditional independence," *The Stata Journal*.

CATTANEO, M. D., AND M. H. FARRELL (2011): "Efficient Estimation of the Dose Response Function under Ignorability using Subclassification on the Covariates," in *Advances in Econometrics: Missing Data Methods*, ed. by D. Drukker, vol. 27A, pp. 93–127. Emerald Group Publishing Limited.

——— (2013): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics*, 174, 127–143.

CHEN, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6B of *Handbook of Econometrics*, chap. 76. Elsevier.

CHEN, X., H. HONG, AND A. TAROZZI (2004): "Semiparametric Efficiency in GMM Models of Nonclassical Measurament Errors, Missing Data and Treatment Effects," Cowles Foundation Discussion Paper No. 1644.

——— (2008): "Semiparametric Efficiency in GMM Models With Auxiliary Data," *Annals of Statistics*, 36(2), 808–843.

DE LA PEÑA, V. H., T. L. LAI, AND Q.-M. SHAO (2009): *Self-Normalized Processes: Limit Theory and Statistical Applications*, Probability and Its Applications. Springer.

DEHEJIA, R. H., AND S. WAHBA (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94(448), 1053–1062.

——— (2002): "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *The Review of Economics and Statistics*, 84(1), 151–161.

EFRON, B. (2013): "Estimation and Accuracy after Model Selection," *Stanford University working paper*.

FIRPO, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259–276.

HAHN, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–331.

HE, X., AND Q.-M. SHAO (2000): "On Parameters of Increasing Dimensions," *Journal of Multivariate Analysis*, 73, 1201–35.

HECKMAN, J., H. ICHIMURA, AND P. TODD (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

HECKMAN, J., AND E. J. VYTLACIL (2007): "Econometric Evaluation of Social Programs, Part I," in *Handbook of Econometrics, vol. VIB*, ed. by J. Heckman, and E. Leamer, pp. 4780–4874. Elsevier Science B.V.

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

HOLLAND, P. W. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81(396), 945–960.

HOROWITZ, J. L., AND C. F. MANSKI (2000): "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95(449), 77–84.

HUANG, J., AND T. ZHANG (2010): "The Benefit of Group Sparsity," *Annals of Statistics*, 38(4), 1978–2004.

HUANG, J. Z. (2003): "Local asymptotics for polynomial spline regression," *Annals of Statistics*, 31(5), 1600–1635.

IMAI, K., AND D. A. VAN DYK (2004): "Causal Inference With General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical Association*, 99(467), 854–866.

IMBENS, G. W. (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87(3), 706–710.

——— (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, 86(1), 4–29.

IMBENS, G. W., W. K. NEWEY, AND G. RIDDER (2007): "Mean-Squared-Error Calculations for Average Treatment Effects," *working paper*.

IMBENS, G. W., AND J. M. WOOLDRIDGE (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

KANG, J. D. Y., AND J. L. SCHAFER (2007): "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22(4), 523–539.

KOLAR, M., J. LAFFERTY, AND L. WASSERMAN (2011): "Union Support Recovery in Multi-task Learning," *Journal of Machine Learning Research*, 12, 2415–2435.

LALONDE, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76(4), 604–620.

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

LEEB, H., AND B. M. PÖTSCHER (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

——— (2008a): "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?," *Econometric Theory*, 24, 338–376.

——— (2008b): "Sparse estimators and the oracle property, or the return of Hodges' estimator," *Journal of Econometrics*, 142, 201–211.

LOUNICI, K., M. PONTIL, S. VAN DE GEER, AND A. B. TSYBAKOV (2009): "Taking advantage of sparsity in multi-task learning," in *Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2009)*, pp. 73–82. Omnipress.

——— (2011): "Oracle Inequalities and Optimal Inference under Group Sparsity," *Annals of Statistics*, 39(4), 2164–2204.

NEGAHBAN, S. N., P. RAVIKUMAR, M. J. WAINWRIGHT, AND B. YU (2012): "A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers," *Statistical Science*, 27(4), 538–557.

NEWEY, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–1382.

——— (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147–168.

NEWEY, W. K., AND D. L. MCFADDEN (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. McFadden, vol. 4 of *Handbook of Econometrics*, chap. 36, pp. 2111–2245. Elsevier.

OBOZINSKI, G., M. J. WAINWRIGHT, AND M. I. JORDAN (2011): "Support Union Recovery in High-Dimensional Multivariate Regression," *Annals of Statistics*, 39(1), 1–47.

PÖTSCHER, B. M. (2009): "Confidence Sets Based on Sparse Estimators Are Necessarily Large," *Sankhyā*, 71-A, 1–18.

PÖTSCHER, B. M., AND H. LEEB (2009): "On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding," *Journal of Multivariate Analysis*, 100, 2065–2085.

RASKUTTI, G., M. J. WAINWRIGHT, AND B. YU (2010): "Restricted Eigenvalue Properties for Correlated Gaussian Designs," *Journal of Machine Learning Research*, 11, 2241–2259.

ROBINS, J. M., AND A. ROTNITZKY (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90(429), 122–129.

ROMANO, J. P. (2004): "On non-parametric testing, the uniform behaviour of the t-test, and related problems," *Scandinavian Journal of Statistics*, 31(4), 567–584.

ROSENBAUM, P. R., AND D. B. RUBIN (1983): "On the Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

RUDELSON, M., AND S. ZHOU (2011): "Reconstruction from anisotropic random measurements," *Arxiv preprint arXiv:1106.1151*.

SMITH, J. A., AND P. E. TODD (2005): "Does matching overcome LaLonde's critique of nonexperimental estimators?," *Journal of Econometrics*, 125, 305–353.

TAN, Z. (2010): "Bounded, efficient and doubly robust estimation with inverse weighting," *Biometrik*, 97, 661–682.

TANABE, K., AND M. SAGAE (1992): "An Exact Cholesky Decomposition and the Generalized Inverse of the Variance-Covariance Matrix of the Multinomial Distribution, with Applications," *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1), 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," *Annals of Statistics*, 36, 614–645.

VAN DE GEER, S., AND P. BUHLMANN (2009): "On the conditions used to prove oracle results for the Lasso," *Electronic Journal of Statistics*, 3, 1360–1392.

VAN DER LAAN, M., AND J. M. ROBINS (2003): *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag.

VON BAHR, B., AND C.-G. ESSEEN (1965): "Inequalities for the rth absolute moment of a sum of random variables, $1 \leqq r \leqq 2$," *Annals of Mathematical Statistics*, 36(1), 299–303.

WEI, F., AND J. HUANG (2010): "Consistent group selection in high-dimensional linear regression," *Bernoulli*, 16(4), 1369–1384.

WHITE, H., AND X. LU (2011): "Causal Diagrams for Treatment Effect Estimation with Application to Efficient Covariate Selection," *Review of Economics and Statistics*, 93(4), 1453–1459.

WOOLDRIDGE, J. M. (2007): "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 141, 1281–1301.

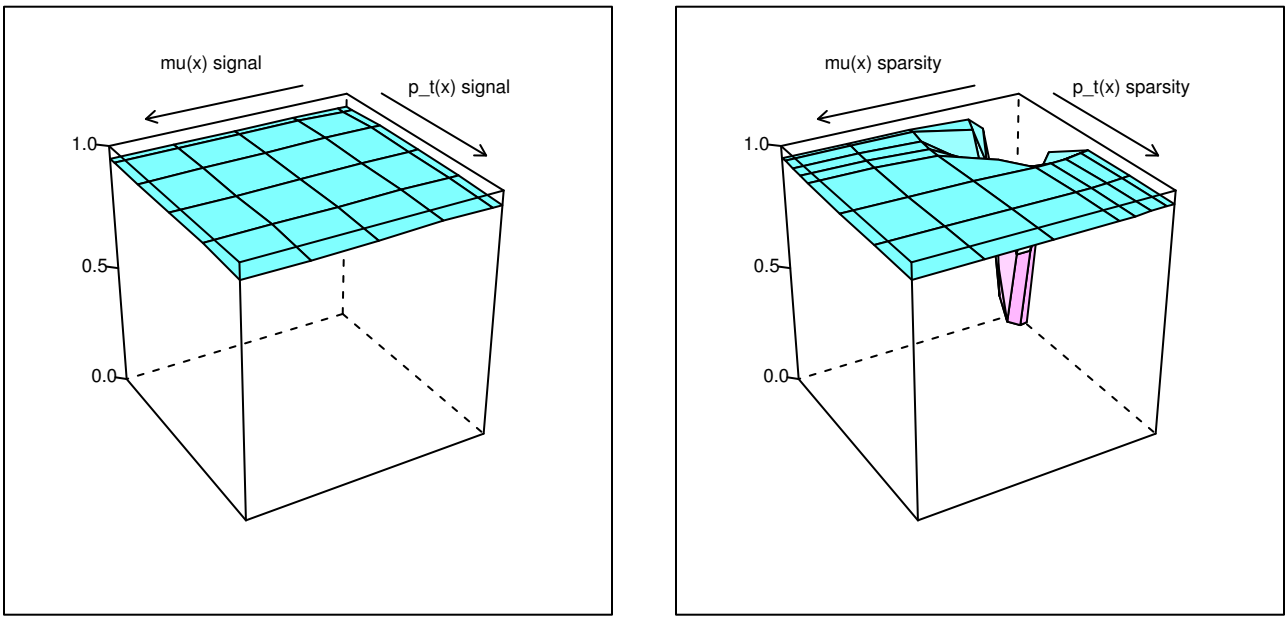
——— (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, 2 edn.

YUAN, M., AND Y. LIN (2006): "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society. Series B*, 68(1), 46–67.

ZHAO, P., AND B. YU (2006): "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

ZOU, H. (2006): "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101(476), 1418–1429.

Figure 1: Empirical Coverage of 95% Confidence Intervals, Varying Signal Strength and Sparsity of $p_t(x)$ and $\mu_t(x)$



(a) Varying Signal Strength **(b) Varying Sparsity**

Table 1: Analysis of NSW Demonstration: Treatment Effects on the Treated and Confidence Intervals for Various Specifications

	Number of Variables		Sample Sizes[c]			
Specifications:	Before selection[a]	After selection[b]	Control	Treated	ATT	95% CI
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 control 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 controls and 185 treated units. Control observations outside the range of estimated propensity scores in the treated sample are discarded.