Online Appendix for "Double/Debiased/Neyman Machine Learning of Treatment Effects"

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey

Contents

1. Practical Implementation and Empirical Examples

To illustrate the methods developed in this paper, we consider two empirical examples. In the first, we use the method outlined in the paper to estimate the effect of 401(k) eligibility on accumulated assets. In this example, the treatment variable is not randomly assigned and we aim to eliminate the potential biases due to the lack of random assignment by flexibly controlling for a rich set of variables. The second example reexamines the Pennsylvania Reemployment Bonus experiment which used a randomized control trial to investigate the incentive effect of unemployment insurance. Our goal in this supplement is to illustrate the use of our method and examine its empirical properties in two different settings: 1) an observational study where it is important to flexibly control for a large number of variables in order to overcome endogeneity, and 2) a randomized control trial where controlling for confounding factors is not needed for bias reduction but may produce more precise estimates.

Practical Details: Incorporating Uncertainty Induced by Sample Splitting. The results we report are based on repeated application of the method developed in the main paper as discussed in Section III. Specifically, we repeat the main estimation procedure 100 times repartitioning the

data in each replication. We then report the average of the ATE estimates from the 100 random splits as the "Mean ATE," and we report the median of the ATE estimates from the 100 splits as the "Median ATE." We then report the measures of uncertainty that account for sampling variability and variability across the splits, $\widehat{\sigma}^{\text{Mean ATE}}$ and $\widehat{\sigma}^{\text{Median ATE}}$, for the "Mean ATE" and the "Median ATE" respectively.

The effect of 401(k) Eligibility on Net Financial Assets. The key problem in determining the effect of 401(k) eligibility is that working for a firm that offers access to a 401(k) plan is not randomly assigned. To overcome the lack of random assignment, we follow the strategy developed in Poterba, Venti and Wise (1994*a*) and Poterba, Venti and Wise (1994*b*). In these papers, the authors use data from the 1991 Survey of Income and Program Participation and argue that eligibility for enrolling in a 401(k) plan in this data can be taken as exogenous after conditioning on a few observables of which the most important for their argument is income. The basic idea of their argument is that, at least around the time 401(k)'s initially became available, people were unlikely to be basing their employment decisions on whether an employer offered a 401(k) but would instead focus on income and other aspects of the job. Following this argument, whether one is eligible for a 401(k) may then be taken as exogenous after appropriately conditioning on income and other control variables related to job choice.

A key component of the argument underlying the exogeneity of 401(k) eligibility is that eligibility may only be taken as exogenous after conditioning on income and other variables related to job choice that may correlate with whether a firm offers a 401(k). Poterba, Venti and Wise (1994*a*) and Poterba, Venti and Wise (1994*b*) and many subsequent papers adopt this argument but control only linearly for a small number of terms. One might wonder whether such specifications are able to adequately control for income and other related confounds. At the same time, the power to learn about treatment effects decreases as one allows more flexible models. The principled use of flexible machine learning tools offers one resolution to this tension. The results presented below thus complement previous results which rely on the assumption that confounding effects can adequately be controlled for by a small number of variables chosen *ex ante* by the researcher.

In the example in this paper, we use the same data as in Chernozhukov and Hansen (2004). We use net financial assets - defined as the sum of IRA balances, 401(k) balances, checking accounts, U.S. saving bonds, other interest-earning accounts in banks and other financial institutions, other interest-earning assets (such as bonds held personally), stocks, and mutual funds less non-mortgage debt - as the outcome variable, Y, in our analysis. Our treatment variable, D, is an indicator for being eligible to enroll in a 401(k) plan. The vector of raw covariates, Z, consists of age, income, family size, years of education, a married indicator, a two-earner status indicator, a defined benefit pension status indicator, an IRA participation indicator, and a home ownership indicator.

In Table 1, we report estimates of the mean average treatment effect (Mean ATE) of 401(k) eligibility on net financial assets both in the partially linear model and allowing for heterogeneous

treatment effects using the interactive model outlined in Section discussed in the main text. To reduce the disproportionate impact of extreme propensity score weights in the interactive model we trim the propensity scores which are close to the bounds, with the cutoff points of 0.01 and 0.99. We present two sets of results based on sample-splitting using a 2-fold cross-fitting and 5-fold cross-fitting.

We report results based on five simple methods for estimating the nuisance functions used in forming the orthogonal estimating equations. We consider three tree-based methods, labeled "Random Forest", "Reg. Tree", and "Boosting", one ℓ_1-penalization based method, labeled "Lasso", and a neural network method, labeled "Neural Net". For "Reg. Tree," we fit a single CART tree to estimate each nuisance function with penalty parameter chosen by 10-fold cross-validation. The results in the "Random Forest" column are obtained by estimating each nuisance function with a random forest which averages over 1000 trees. The results in "Boosting" are obtained using boosted regression trees with regularization parameters chosen by 10-fold cross-validation. To estimate the nuisance functions using the neural networks, we use 8 hidden layers and a decay parameter of 0.01, and we set activation function as logistic for classification problems and as linear for regression problems.[1] "Lasso" estimates an ℓ_1-penalized linear regression model using the data-driven penalty parameter selection rule developed in Belloni et al. (2012). For "Lasso", we use a set of 275 potential control variables formed from the raw set of covariates and all second order terms, i.e. all squares and first-order interactions. For the remaining methods, we use the raw set of covariates as features.

We also consider two hybrid methods labeled "Ensemble" and "Best". "Ensemble" optimally combines four of the machine learning methods listed above by estimating the nuisance functions as weighted averages of estimates from "Lasso," "Boosting," "Random Forest," and "Neural Net". The weights are restricted to sum to one and are chosen so that the weighted average of these methods gives the lowest average mean squared out-of-sample prediction error estimated using 5-fold cross-validation. The final column in Table 1 ("Best") reports results that combines the methods in a different way. After obtaining estimates from the five simple methods and "Ensemble", we select the best methods for estimating each nuisance functions based on the average out-of-sample prediction performance for the target variable associated to each nuisance function obtained from each of the previously described approaches. As a result, the reported estimate in the last column uses different machine learning methods to estimate different nuisance functions. Note that if a single method outperformed all the others in terms of prediction accuracy for all nuisance functions, the estimate in the "Best" column would be identical to the estimate reported under that method.

Turning to the results, it is first worth noting that the estimated ATE of 401(k) eligibility on net financial assets is \$19,559 (not reported) with an estimated standard error of 1413 when no

[1]We also experimented with "Deep Learning" methods from which we obtained similar results for some tuning parameters. However, we ran into stability and computational issues and chose not to report these results in the empirical section.

control variables are used. Of course, this number is not a valid estimate of the causal effect of 401(k) eligibility on financial assets if there are neglected confounding variables as suggested by Poterba, Venti and Wise (1994*a*) and Poterba, Venti and Wise (1994*b*). When we turn to the estimates that flexibly account for confounding reported in Table 1, we see that they are substantially attenuated relative to this baseline that does not account for confounding, suggesting much smaller causal effects of 401(k) eligiblity on financial asset holdings. It is interesting and reassuring that the results obtained from the different flexible methods are broadly consistent with each other. This similarity is consistent with the theory that suggests that results obtained through the use of orthogonal estimating equations and any sensible method of estimating the necessary nuisance functions should be similar. Finally, it is interesting that these results are also broadly consistent with those reported in the original work of Poterba, Venti and Wise (1994*a*) and Poterba, Venti and Wise (1994*b*) which used a simple intuitively motivated functional form, suggesting that this intuitive choice was sufficiently flexible to capture much of the confounding variation in this example.

There are other interesting observations that can provide useful insights into understanding the finite sample properties of the "double ML" estimation method. First, the standard errors of the estimates obtained using 5-fold cross-fitting are considerably lower than those obtained from 2-fold cross-fitting for all methods. This fact suggests that having more observations in the auxiliary sample may be desirable. Specifically, 5-fold cross-fitting estimates uses more observations to learn the nuisance functions than 2-fold cross-fitting and thus likely learns them more precisely. This increase in precision in learning the nuisance functions may then translate into more precisely estimated parameters of interest. While intuitive, we note that this statement does not seem to be generalizable in that there does not appear to be a general relationship between the number of folds in cross-fitting and the precision of the estimate of the parameter of interest. Second, we also see that the standard errors of the Lasso estimates are noticeably larger than the standard errors coming from the other machine learning methods. We believe that this is due to the fact that the out-of-sample prediction errors from a linear model tend to be larger when there is a need to extrapolate. In our framework, if the main sample includes observations that are outside of the range of the observations in the auxiliary sample, the model has to extrapolate to those observations. The fact that the standard errors are lower in 5-fold cross-fitting than in 2-fold cross-fitting for the "Lasso" estimations also supports this hypothesis, because the higher number of observations in the auxiliary sample reduces the degree of extrapolation.

In Table 2, we report the median ATE estimation results for the same models to check the robustness of our results to the outliers due to sample splitting. We see that both the coefficients and standard errors are similar to the "Mean ATE" estimates and standard errors. The similarity between the "Mean ATE" and "Median ATE" suggests that the distribution of the ATE across different splits is approximately symmetric and relatively thin-tailed.

The effect of Unemployment Insurance Bonus on Unemployment Duration. As a further example, we re-analyze the Pennsylvania Reemployment Bonus experiment which was conducted by the US Department of Labor in the 1980s to test the incentive effects of alternative compensation schemes for unemployment insurance (UI). This experiment has been previously studied by Bilias (2000) and Bilias and Koenker (2002). In these experiments, UI claimants were randomly assigned either to a control group or one of five treatment groups.[2] In the control group the standard rules of the UI applied. Individuals in the treatment groups were offered a cash bonus if they found a job within some pre-specified period of time (qualification period), provided that the job was retained for a specified duration. The treatments differed in the level of the bonus, the length of the qualification period, and whether the bonus was declining over time in the qualification period; see Bilias and Koenker (2002) for further details on data.

In our empirical example, we focus only on the most generous compensation scheme, treatment 4, and drop all individuals who received other treatments. In this treatment, the bonus amount is high and qualification period is long compared to other treatments, and claimants are eligible to enroll in a workshop. Our treatment variable, D, is an indicator variable for the treatment 4, and the outcome variable, Y, is the log of duration of unemployment for the UI claimants. The vector of covariates, Z, consists of age group dummies, gender, race, number of dependents, quarter of the experiment, location within the state, existence of recall expectations, and type of occupation.

We report estimates of the ATE on unemployment duration both in the partially linear model and in the interactive model. We again consider the same methods with the same tuning choices, with one exception, for estimating the nuisance functions as in the previous example and so do not repeat details for brevity. The one exception is that we implement neural networks with 2 hidden layers and a decay parameter of 0.02 in this example which yields better prediction performance. In "Lasso" estimation, we use a set of 96 control variables formed by taking nonlinear functions and interactions of the raw set of covariates. For the remaining approaches, we use only the 14 raw control variables listed above.

Table 3 presents estimates of the "Mean ATE" on unemployment duration using the partially linear model and interactive model in panel A and B, respectively. To reduce the disproportionate impact of extreme propensity score weights in the interactive model, we trim the propensity scores at 0.01 and 0.99 as in the previous example. For both the partially linear model and the interactive model, we report estimates obtained using 2-fold cross-fitting and 5-fold cross-fitting.

The estimation results are consistent with the findings of previous studies which have analyzed the Pennsylvania Bonus Experiment. The ATE on unemployment duration is negative and significant across all estimation methods at the 5% level with the exception of the estimate of the ATE obtained from the interactive model using random forests, which is significant at the 10% level. When looking at standard errors it is useful to remember that they include both sampling variation and variation

[2]There are six treatment groups in the experiments. Following Bilias (2000). we merge the groups 4 and 6.

due to random sample splitting. It is reassuring to see that the variation due to sample splitting does not change the conclusion. It is also interesting to see that, similar to the result in the first empirical example, the "Mean ATE" estimates are broadly similar across different estimation models. Finally in Table 4 we report the "Median ATE" estimates. The median estimates are close to the mean estimates, giving further evidence for the stability of estimation across different random splits.

In conclusion, we want to emphasize some important observations that can be drawn from these empirical examples. First, for both examples the choice of the machine learning method in estimating nuisance functions does not substantively change the conclusion, and we obtained broadly consistent results regardless of which method we employ. Second, the similarity between the median and mean estimates suggests that the results are robust to the particular sample split used in estimation in these examples.

TABLE 1. Estimated Mean ATE of 401(k) Eligibility on Net Financial Assets

	Lasso	Reg. Tree	Random Forest	Boosting	Neural Net.	Ensemble	Best
A. Interactive Model							
ATE (2 fold)	6331	7581	7966	7826	7805	7617	7800
	(2712)	(1374)	(1549)	(1345)	(1688)	(1299)	(1325)
ATE (5 fold)	6964	8023	8104	7699	7772	7658	7890
	(1654)	(1311)	(1364)	(1223)	(1324)	(1204)	(1198)
B. Partially Linear Model							
ATE (2 fold)	7718	8745	9180	8768	9040	9043	9106
	(1796)	(1488)	(1526)	(1451)	(1494)	(1432)	(1430)
ATE (5 fold)	8182	8913	9248	9092	9038	9186	9214
	(1578)	(1440)	(1402)	(1380)	(1394)	(1381)	(1361)

Notes: Estimated Mean ATE and standard errors (in parentheses) from a linear model (Panel B) and heterogeneous effect model (Panel A) based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Further details about the methods are provided in the main text.

TABLE 2. Estimated Median ATE of 401(k) Eligibility on Net Financial Assets

	Lasso	Reg. Tree	Random Forest	Boosting	Neural Net.	Ensemble	Best
A. Interactive Model							
ATE (2 fold)	6725	7557	8034	7820	7800	7620	7800
	(1612)	(1283)	(1400)	(1199)	(1474)	(1198)	(1185)
ATE (5 fold)	7133	8046	8099	7690	7795	7668	7876
	(1420)	(1242)	(1296)	(1179)	(1290)	(1180)	(1149)
B. Partially Linear Model							
ATE (2 fold)	7707	8770	9204	8746	9104	9061	9129
	(1785)	(1424)	(1392)	(1391)	(1388)	(1343)	(1342)
ATE (5 fold)	8202	8894	9252	9089	9065	9199	9232
	(1581)	(1440)	(1400)	(1378)	(1393)	(1379)	(1359)

Notes: Estimated Median ATE and standard errors (in parentheses) from a linear model (Panel B) and heterogeneous effect model (Panel A) based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Further details about the methods are provided in the main text.

TABLE 3. Estimated Mean ATE of Cash Bonus on Unemployment Duration

	Lasso	Reg. Tree	Random Forest	Boosting	Neural Net.	Ensemble	Best
A. Interactive Model							
ATE (2 fold)	-0.081	-0.084	-0.072	-0.078	-0.073	-0.079	-0.078
	(0.036)	(0.037)	(0.042)	(0.036)	(0.041)	(0.036)	(0.036)
ATE (5 fold)	-0.081	-0.084	-0.070	-0.076	-0.072	-0.079	-0.076
	(0.036)	(0.037)	(0.040)	(0.036)	(0.038)	(0.036)	(0.036)
B. Partially Linear Model							
ATE (2 fold)	-0.081	-0.083	-0.076	-0.076	-0.073	-0.076	-0.076
	(0.036)	(0.037)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)
ATE (5 fold)	-0.080	-0.084	-0.075	-0.075	-0.074	-0.075	-0.075
	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)

Notes: Estimated Mean ATE and standard errors (in parentheses) from a linear model (Panel B) and heterogeneous effect model (Panel A) based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Further details about the methods are provided in the main text.

TABLE 4. Estimated Median ATE of Cash Bonus on Unemployment Duration

	Lasso	Reg. Tree	Random Forest	Boosting	Neural Net.	Ensemble	Best
A. Interactive Model							
ATE (2 fold)	-0.081	-0.084	-0.073	-0.078	-0.074	-0.079	-0.078
	(0.036)	(0.036)	(0.041)	(0.036)	(0.039)	(0.036)	(0.036)
ATE (5 fold)	-0.081	-0.085	-0.069	-0.076	-0.072	-0.079	-0.076
	(0.036)	(0.037)	(0.039)	(0.036)	(0.038)	(0.036)	(0.036)
B. Partially Linear Model							
ATE (2 fold)	-0.081	-0.084	-0.077	-0.076	-0.074	-0.076	-0.076
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
ATE (5 fold)	-0.079	-0.084	-0.076	-0.075	-0.073	-0.075	-0.075
	(0.036)	(0.037)	(0.036)	(0.035)	(0.036)	(0.035)	(0.035)

Notes: Estimated Median ATE and standard errors (in parentheses) from a linear model (Panel B) and heterogeneous effect model (Panel A) based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Further details about the methods are provided in the main text.

2. Proofs

Notation. The symbols P and E denote probability and expectation operators with respect to a generic probability measure. If we need to signify the dependence on a probability measure P, we use P as a subscript in P_P and E_P. Note also that we use capital letters such as W to denote random elements and use the corresponding lower case letters such as w to denote fixed values that these random elements can take in the set $\mathcal{W}$. In what follows, we use $\|\cdot\|_{P,q}$ to denote the $L^q(P)$ norm; for example, for measurable $f : \mathcal{W} \to \mathbb{R}$, we denote

$$\|f(W)\|_{P,q} := \left(\int |f(w)|^q dP(w)\right)^{1/q}.$$

Define the empirical process $\mathbb{G}_{n,I}(\psi(W))$ as a linear operator acting on measurable functions $\psi : \mathcal{W} \to \mathbb{R}$ such that $\|\psi(W)\|_{P,2} < \infty$ via,

$$\mathbb{G}_{n,I}(\psi(W)) := \frac{1}{\sqrt{n}} \sum_{i \in I} f(W_i) - \int f(w) dP(w).$$

Analogously, we defined the empirical expectation and probability as:

$$\mathbb{E}_{n,I}(\psi(W)) := \frac{1}{n} \sum_{i \in I} f(W_i); \quad \mathbb{P}_{n,I}(A) := \frac{1}{n} \sum_{i \in I} 1(W_i \in A).$$

Proof of Theorem II.1. We will demonstrate the result for the case of ATE estimator, which uses the score:

$$\psi(W, \theta, \eta) := g(1, Z) - g(0, Z) + \tfrac{D(Y - g(1,Z))}{m(Z)} - \tfrac{(1-D)(Y - g(0,Z))}{1 - m(Z)} - \theta,$$

and the result for ATTE follows similarly. Choose any sequence $\{P_n\} \in \mathcal{P}$.

Step 1: (Main Step). Letting $\check{\theta}_{0,k} = \check{\theta}_0(I_k, I_k^c)$, write

$$\sqrt{n}(\check{\theta}_k - \theta_0) = \mathbb{G}_{n,I_k} \psi(W; \theta_0, \widehat{\eta}_0(I_k^c)) + \sqrt{n} \int \psi(w, \theta_0, \widehat{\eta}_0(I_k^c)) dP(w).$$

Steps 2, 3, and 4 below demonstrate that for each $k = 1, ..., K$

$$\int (\psi(w, \theta_0, \widehat{\eta}_0(I_k^c)) - \psi(w, \theta_0, \eta_0))^2 dP_n(w) = o_{P_n}(1), \tag{2.1}$$

$$\sqrt{n} \int (\psi(w, \theta_0, \widehat{\eta}_0(I_k^c)) - \psi(w, \theta_0, \eta_0)) dP_n(w) = o_{P_n}(1), \tag{2.2}$$

$$\widehat{\sigma}^2 - \sigma^2 = o_{P_n}(1), \tag{2.3}$$

where σ^2 is bounded away and from above by assumptions. These equations are the minimal conditions needed on the estimators of the nuisance parameters, and could be used to replace the more primitive conditions stated in the text.

Assertion (2.1) implies that

$$\mathbb{G}_{n,I_k}(\psi(W; \theta_0, \widehat{\eta}_0(I_k^c)) - \psi(W; \theta_0, \eta_0)) = o_{P_n}(1),$$

since the quantity in the display converges in probability conditionally on the data $(W_i)_{i\in I_k^c}$ by (2.1) and Chebyshev inequality, which in turn implies the unconditional convergence in probability, as noted in the following simple lemma.

Lemma 2.1. *Let $\{X_m\}$ and $\{Y_m\}$ be a sequence of random vectors. If for any $\epsilon > 0$, $\mathrm{P}(\|X_m\| > \epsilon \mid Y_m) \to_{\mathrm{P}} 0$, then $\mathrm{P}(\|X_m\| > \epsilon) \to 0$. In particular, this occurs if $\mathrm{E}[\|X_m\|^q \mid Y_m] \to_{\mathrm{P}} 0$ for some $q \geqslant 1$, by Chebyshev inequality.*

Proof. For any $\epsilon > 0$ $\mathrm{P}(\|X_m\| > \epsilon) \leqslant \mathrm{E}[\mathrm{P}(\|X_m\| > \epsilon \mid Y_m)] \to 0$, since the sequence $\{\mathrm{P}(\|X_m\| > \epsilon \mid Y_m)$ is uniformly integrable. ∎

Using independence of data blocks $(W_i)_{i\in I_k}$, $k = 1, ..., K$, the application of the Lindeberg-Feller theorem and the Cramer-Wold device, we conclude that

$$\left(\sigma^{-1}\sqrt{n}(\check{\theta}_{0,k} - \theta_0)\right)_{k=1}^K = \left(\sigma^{-1}\mathbb{G}_{n,I_k}\psi(W;\theta_0,\eta_0)\right)_{k=1}^K + o_{P_n}(1) \rightsquigarrow (\mathcal{N}_k)_{k=1}^K,$$

where $(\mathcal{N}_k)_{k=1}^K$ is a Gaussian vector with independent $N(0,1)$ coordinates. Therefore,

$$\begin{aligned}\sigma^{-1}\sqrt{nK}(\tilde{\theta}_0 - \theta_0) &= \sigma^{-1}\sqrt{nK}\left(\frac{1}{K}\sum_{k=1}^K \check{\theta}_{0,k} - \theta_0\right) \\ &= \frac{1}{\sqrt{K}}\sum_{k=1}^K \sigma^{-1}\mathbb{G}_{n,I_k}\psi(W;\theta_0,\eta_0) \rightsquigarrow \frac{1}{\sqrt{K}}\sum_{k=1}^K \mathcal{N}_k = N(0,1),\end{aligned}$$

where the last line uses the sum-stability of the normal distribution. Moreover, the result continues to hold if σ is replaced by $\widehat{\sigma}$ in view of (2.3) and σ bounded away from zero and from above.

The above claim implies that $\mathrm{CI}_n = [\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\widehat{\sigma}/\sqrt{N}]$ obeys

$$\mathrm{P}_{P_n}(\theta_0 \in \mathrm{CI}_n) \to (1-\alpha).$$

The last two claims hold under any sequence $\{P_n\} \in \mathcal{P}$, which implies that these claims hold uniformly in $P \in \mathcal{P}$. Indeed, for example, choose $\{P_n\}$ such that, for some $\epsilon_n \to 0$

$$\sup_{P\in\mathcal{P}} |\mathrm{P}_P(\theta_0 \in \mathrm{CI}_n) - (1-\alpha)| \leqslant |\mathrm{P}_{P_n}(\theta_0 \in \mathrm{CI}_n) - (1-\alpha)| + \epsilon_n.$$

The right side converges to zero, which implies the uniform convergence.

Step 2: This step demonstrates the assertion (2.1). Elementary calculations and the repeated use of Holder's inequality give

$$\begin{aligned}\|\psi(W;\theta_0,\widehat{\eta}_0(I_k^c)) - \psi(W;\theta_0;\eta_0)\|_{P,2} &\leqslant C_\varepsilon \max_{d\in\{0,1\}} \|\widehat{g}_0(d,Z;I_k^c) - g_0(d,Z)\|_{P,2} \\ &+ C_\varepsilon\left(\|Y\|_{P,4} + \max_{d\in\{0,1\}}\|g_0(d,Z)\|_{P,4}\right)\cdot\sqrt{\|\widehat{m}_0(Z;I_k^c) - m_0(Z)\|_{P,2}} \\ &\leqslant C_\varepsilon(\delta_n + 2C\sqrt{\delta_n}) \to 0,\end{aligned}$$

with P-probability no less than $1-\Delta_n$ for all $P \in \mathcal{P}$, where C_ε depends on ε and $\mathcal{P}$.

Step 3: This step demonstrates the assertion (2.2). Rewrite (2.2) as

$$\sqrt{n}\int\left(\left\{\frac{\widehat{m}_0(z;I_k^c)-m_0(z)}{\widehat{m}_0(z;I_k^c)}\right\}\{\widehat{g}_0(1,z;I_k^c)-g_0(1,z)\}\right.$$
$$\left.-\left\{\frac{\widehat{m}_0(z;I_k^c)-m(z)}{1-\widehat{m}_0(z;I_k^c)}\right\}\{\widehat{g}_0(0,z;I_k^c)-g_0(0,z)\}\right)dP(z).$$

Using the Cauchy-Schwarz inequality and the assumption that $P(\varepsilon \leqslant m_0(Z) \leqslant 1-\varepsilon)=1$ and that $P(\varepsilon \leqslant \widehat{m}_0(Z;I_k^c) \leqslant 1-\varepsilon)=1$ with P-probability at least $1-\Delta_n$, uniformly for all $P\in\mathcal{P}$, this quantity is bounded with the same probability by

$$\sqrt{n}\frac{2}{\varepsilon}\|\widehat{g}_0(d,Z;I_k^c)-g_0(d,Z)\|_{P,2}\cdot\|\widehat{m}_0(Z;I_k^c)-m_0(Z)\|_{P,2}\leqslant\frac{2\delta_n}{\varepsilon}\to 0.$$

Step 4: This step demonstrates the assertion (2.3). Here $\to_{\mathrm{P}}$ is meant to be convergence uniformly in $P\in\mathcal{P}$. We can write

$$\widehat{\sigma}^2=\frac{1}{K}\sum_{k=1}^K\widehat{\sigma}_k^2,\quad \widehat{\sigma}_k^2:=\mathbb{E}_{n,I_k}\psi^2(W;\widehat{\theta}_k,\widehat{\eta}_k(I_k^c));\quad \sigma^2=\mathrm{E}_P\psi^2(W;\theta_0,\eta_0).$$

We claim that for each $k=1,...,K$,

$$\mathbb{E}_{n,I_k}\psi^2(W;\widehat{\theta}_0,\widehat{\eta}_0(I^c))-\mathbb{E}_{n,I_k}\psi^2(W;\theta_0,\eta_0)\to_{\mathrm{P}}0,\quad \mathbb{E}_{n,I_k}\psi^2(W;\theta_0,\eta_0)-\sigma^2\to_{\mathrm{P}}0.$$

The latter property holds by the Chebyshev Inequality. Further, letting I denote a generic I_k, the relation $a^2-b^2=(a-b)(a+b)$, and the Cauchy-Schwarz and triangle inequalities yield:

$$|\mathbb{E}_{n,I}\{\psi^2(W;\widehat{\theta}_0,\widehat{\eta}_0(I^c))-\psi^2(W;\theta_0,\eta_0)\}|\leqslant r_n\times\left(2\|\psi(W;\theta_0,\eta_0)\|_{\mathbb{P}_{n,I},2}+r_n\right)$$

where

$$r_n:=\|\psi(W;\widehat{\theta}_0,\widehat{\eta}_0(I^c))-\psi(W;\theta_0,\eta_0)\|_{\mathbb{P}_{n,I},2}.$$

Since $\|\psi(W;\theta_0,\eta_0)\|^2_{\mathbb{P}_{n,I},2}-\sigma^2\to_{\mathrm{P}}0$ as noted above, and σ^2 is bounded above by assumption, the claim follows, provided $r_n\to_{\mathrm{P}}0$.

Indeed, we have that

$$r_n\leqslant\|\tilde{\theta}_0-\theta_0\|+C_\varepsilon\max_{d\in\{0,1\}}\|\widehat{g}_0(d,Z;I^c)-g_0(d,Z)\|_{\mathbb{P}_{n,I},2}$$
$$+C_\varepsilon\left(\|Y\|_{\mathbb{P}_{n,I},4}+\max_{d\in\{0,1\}}\|g_0(d,Z)\|_{\mathbb{P}_{n,I},4}\right)\cdot\sqrt{\|\widehat{m}_0(Z;I^c)-m_0(Z)\|_{\mathbb{P}_{n,I},2}},$$

with P-probability no less than $1-\Delta_n$ for all $P\in\mathcal{P}$, where C_ε depends on ϵ and $\mathcal{P}$. We have that by Markov inequality:

$$\|Y\|^4_{\mathbb{P}_{n,I},4}+\max_{d\in\{0,1\}}\|g_0(d,Z)\|^4_{\mathbb{P}_{n,I},4}\to_{\mathrm{P}}\|Y\|^4_{P,4}+\max_{d\in\{0,1\}}\|g_0(d,Z)\|^4_{P,4};$$

and, with probability at least $1-\Delta_n$,

$$\mathrm{E}[\|\widehat{g}_0(d,Z;I^c)-g_0(d,Z)\|^2_{\mathbb{P}_{n,I},2}\mid(W_i)_{i\in I^c}]=\|\widehat{g}_0(d,Z;I^c)-g_0(d,Z)\|^2_{P,2}\leqslant\delta_n,$$
$$\mathrm{E}[\|\widehat{m}_0(Z;I^c)-m_0(Z)\|^2_{\mathbb{P}_{n,I},2}\mid(W_i)_{i\in I^c}]=\|\widehat{m}_0(Z;I^c)-m_0(Z)\|^2_{P,2}\leqslant\delta_n,$$

which implies by Lemma 2.1 that

$$\|\widehat{g}_0(d, Z; I^c) - g_0(d, Z)\|_{\mathbb{P}_{n,I},2} + \|\widehat{m}_0(Z; I^c) - m_0(Z)\|_{\mathbb{P}_{n,I},2} \to_{\mathrm{P}} 0.$$

Conclude that $r_n \to_{\mathrm{P}} 0$. ∎

References

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: 2369–2429. ArXiv, 2010.

Bilias, Y. 2000. "Sequential testing of duration data: the case of the Pennsylvania'reemployment bonus' experiment." *Journal of Applied Econometrics*, 15: 575–594.

Bilias, Y., and R. Koenker. 2002. "Quantile regression for duration data: a reappraisal of the Pennsylvania reemployment bonus experiments." *Economic Applications of Quantile Regression 2002*, 199–220.

Chernozhukov, V., and C. Hansen. 2004. "The effects of 401 (k) participation on the wealth distribution: an instrumental quantile regression analysis." *Review of Economics and statistics*, 86(3): 735–751.

Poterba, J. M., S. F. Venti, and D. A. Wise. 1994*a*. "401(k) plans and tax-deferred savings." In *Studies in the Economics of Aging.* , ed. D. Wise, 105–142. Chicago:University of Chicago Press.

Poterba, J. M., S. F. Venti, and D. A. Wise. 1994*b*. "Do 401(k) contributions crowd out other personal saving?" *Journal of Public Economics*, 58: 1–32.