A. Proof of Theorem 1

We first prove the result for the sample-splitting estimator $\hat{\theta}^{SS}$ in (2) and then discuss how to generalize for the K-fold cross fitting estimator $\hat{\theta}^{CF}$ in (3) with $\sqrt{2n}$ scaling.

For each coordinate moment function m_i, the mean value theorem and the definition of $\hat{\theta}^{SS}$ imply that

$$\frac{1}{n}\sum_{t=1}^{n}\langle\nabla_\theta m_i(Z_t,\tilde{\theta}^{(i)},\hat{h}(X_t)),\theta_0-\hat{\theta}^{SS}\rangle = \frac{1}{n}\sum_{t=1}^{n}(m_i(Z_t,\theta_0,\hat{h}(X_t))-m_i(Z_t,\hat{\theta}^{SS},\hat{h}(X_t))) = \frac{1}{n}\sum_{t=1}^{n}m_i(Z_t,\theta_0,\hat{h}(X_t)) \tag{10}$$

for some convex combination, $\tilde{\theta}^{(i)}$, of $\hat{\theta}^{SS}$ and θ_0. Hence,

$$\sqrt{n}(\theta_0-\hat{\theta}^{SS})\mathbb{I}[\det\hat{J}(\hat{h})\neq 0] = \hat{J}(\hat{h})^{-1}\mathbb{I}[\det\hat{J}(\hat{h})\neq 0]\underbrace{\frac{1}{\sqrt{n}}\sum_{t=1}^{n}m(Z_t,\theta_0,\hat{h}(X_t))}_{B}$$

$$\text{for}\quad \hat{J}(h)\triangleq\frac{1}{n}\sum_{t=1}^{n}\begin{bmatrix}\nabla_\theta m_1(Z_t,\tilde{\theta}^{(1)},h(X_t))\\ \cdots\\ \nabla_\theta m_d(Z_t,\tilde{\theta}^{(d)},h(X_t))\end{bmatrix}\in\mathbb{R}^{d\times d}.$$

We will first show in Section A.1 that $\hat{J}(\hat{h})$ converges in probability to the invertible matrix $J=\mathbb{E}\left[\nabla_\theta m(Z,\theta_0,h_0(X))\right]$. Hence, we will have $\mathbb{I}[\det\hat{J}(\hat{h})\neq 0]\xrightarrow{p}\mathbb{I}[\det J\neq 0]=1$ and $\hat{J}(\hat{h})^{-1}\mathbb{I}[\det\hat{J}(\hat{h})\neq 0]\xrightarrow{p}J^{-1}$ by the continuous mapping theorem (van der Vaart, 1998, Thm. 2.3). We will next show in Section A.2 that B converges in distribution to a mean-zero multivariate Gaussian distribution with constant covariance matrix $V=\text{Cov}(m(Z,\theta_0,h_0(X)))$. Slutsky's theorem (van der Vaart, 1998, Thm. 2.8) will therefore imply that $\sqrt{n}(\theta_0-\hat{\theta}^{SS})\mathbb{I}[\det\hat{J}(\hat{h})\neq 0]$ converges in distribution to $N(0,J^{-1}VJ^{-1})$. Finally, the following lemma, proved in Section J.1, will imply that $\sqrt{n}(\theta_0-\hat{\theta}^{SS})$ also converges in distribution to $N(0,J^{-1}VJ^{-1})$, as desired.

Lemma 11. *Consider a sequence of binary random variables $Y_n\in\{0,1\}$ satisfying $Y_n\xrightarrow{p}1$. If $X_nY_n\xrightarrow{p}X$, then $X_n\xrightarrow{p}X$. Similarly, if $X_nY_n\xrightarrow{d}X$, then $X_n\xrightarrow{d}X$.*

A.1. Convergence of $\hat{J}(\hat{h})-J$.

For each coordinate j and moment m_i and $r>0$ defined in Assumption 1.7, the mean value theorem and Cauchy-Schwarz imply that

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{J}_{ij}(\hat{h})-\hat{J}_{ij}(h_0)\right|\mathbb{I}[\tilde{\theta}^{(i)}\in\mathcal{B}_{\theta_0,r}]\mid\hat{h}\right]\mathbb{I}[\hat{h}\in\mathcal{B}_{h_0,r}]\\
&\leq\mathbb{E}\left[\left|\nabla_{\theta_j}m_i(Z_t,\tilde{\theta}^{(i)},\hat{h}(X_t))-\nabla_{\theta_j}m_i(Z_t,\tilde{\theta}^{(i)},h_0(X_t))\right|\mathbb{I}[\tilde{\theta}^{(i)}\in\mathcal{B}_{\theta_0,r}]\mid\hat{h}\right]\mathbb{I}[\hat{h}\in\mathcal{B}_{h_0,r}]\\
&=\mathbb{E}\left[\left|\langle\hat{h}(X_t)-h_0(X_t),\nabla_\gamma\nabla_{\theta_j}m_i(Z_t,\tilde{\theta}^{(i)},\tilde{h}^{(j)}(X_t))\rangle\right|\mathbb{I}[\tilde{\theta}^{(i)}\in\mathcal{B}_{\theta_0,r}]\mid\hat{h}\right]\mathbb{I}[\hat{h}\in\mathcal{B}_{h_0,r}]\\
&\leq\sqrt{\mathbb{E}\left[\|\hat{h}(X_t)-h_0(X_t)\|_2^2\mid\hat{h}\right]\sup_{h\in\mathcal{B}_{h_0,r}}\mathbb{E}\left[\sup_{\theta\in\mathcal{B}_{\theta_0,r}}\|\nabla_\gamma\nabla_{\theta_j}m_i(Z_t,\theta,h(X_t))\|_2^2\right]}
\end{aligned}$$

for $\tilde{h}^{(j)}(X_t)$ a convex combination of $h_0(X_t)$ and $\hat{h}(X_t)$. The consistency of $\hat{h}$ (Assumption 1.6) and the regularity condition Assumption 1.7b therefore imply that $\mathbb{E}[|\hat{J}_{ij}(\hat{h})-\hat{J}_{ij}(h_0)|\mathbb{I}[\tilde{\theta}^{(i)}\in\mathcal{B}_{\theta_0,r}]\mid\hat{h}]\mathbb{I}[\hat{h}\in\mathcal{B}_{h_0,r}]\xrightarrow{p}0$ and hence that $|\hat{J}_{ij}(\hat{h})-\hat{J}_{ij}(h_0)|\mathbb{I}[\hat{h}\in\mathcal{B}_{h_0,r},\tilde{\theta}^{(i)}\in\mathcal{B}_{\theta_0,r}]\xrightarrow{p}0$ by the following lemma, proved in Section J.2.

Lemma 12. *Consider a sequence of two random variables X_n,Z_n, where X_n is a finite d-dimensional random vector. Suppose that $\mathbb{E}\left[\|X_n\|_p^p|Z_n\right]\xrightarrow{p}0$ for some $p\geq 1$. Then $X_n\xrightarrow{p}0$.*

Now Assumptions 1.6 and 1.5 and the continuous mapping theorem imply that $\mathbb{I}[\hat{h}\in\mathcal{B}_{h_0,r}]\xrightarrow{p}1$. Therefore, by Lemma 11, we further have $|\hat{J}_{ij}(\hat{h})-\hat{J}_{ij}(h_0)|\mathbb{I}[\tilde{\theta}^{(i)}\in\mathcal{B}_{\theta_0,r}]\xrightarrow{p}0$.

The regularity Assumptions 1.4 and 1.7a additionally imply the uniform law of large numbers,

$$\sup_{\theta \in \mathcal{B}_{\theta_0, r}} \|\tfrac{1}{n}\textstyle\sum_{t=1}^{n} \nabla_\theta m_i(Z_t, \theta, h_0(X_t)) - \mathbb{E}_Z[\nabla_\theta m_i(Z, \theta, h_0(X))]\|_2 \xrightarrow{p} 0$$

for each moment m_i (see, e.g., Newey & McFadden, 1994, Lem. 2.4). Taken together, these conclusions yield

$$\left[\hat{J}_i(\hat{h}) - \mathbb{E}_Z[\nabla_\theta m_i(Z, \tilde{\theta}^{(i)}, h_0(X))]\right] \mathbb{I}[\tilde{\theta}^{(i)} \in \mathcal{B}_{\theta_0, r}] \xrightarrow{p} 0,$$

for each m_i, where $\hat{J}_i(\hat{h})$ denotes the i-th row of $\hat{J}(\hat{h})$.

Since $\tilde{\theta}^{(i)}$ is a convex combination of $\hat{\theta}^{SS}$ and θ_0, the consistency of $\hat{\theta}^{SS}$ implies that $\tilde{\theta}^{(i)} \xrightarrow{p} \theta_0$ and therefore that $\mathbb{I}[\tilde{\theta}^{(i)} \in \mathcal{B}_{\theta_0, r}] \xrightarrow{p} 1$ and $\mathbb{E}_Z[\nabla_\theta m_i(Z, \tilde{\theta}^{(i)}, h_0(X))] \xrightarrow{p} \mathbb{E}_Z[\nabla_\theta m_i(Z, \theta_0, h_0(X))]$ by the continuous mapping theorem. Lemma 11 therefore implies that $\hat{J}_i(\hat{h}) - \mathbb{E}_Z[\nabla_\theta m_i(Z, \tilde{\theta}^{(i)}, h_0(X))] \xrightarrow{p} 0$ and hence that $\hat{J}_i(\hat{h}) \xrightarrow{p} \mathbb{E}_Z[\nabla_\theta m_i(Z, \theta_0, h_0(X))]$, as desired.

A.2. Asymptotic Normality of B.

For a vector $\gamma \in \mathbb{R}^\ell$ and a vector $\alpha \in \mathbb{N}^\ell$, we define the shorthand $\gamma^\alpha \triangleq \prod_{i=1}^{\ell} \gamma_\ell^{\alpha_\ell}$.

To establish the asymptotic normality of B, we let $k = \max_{\alpha \in S} \|\alpha\|_1$ and apply Taylor's theorem with $k+1$-order remainder around $h_0(X_t)$ for each X_t:

$$\begin{aligned} B = {} & \underbrace{\frac{1}{\sqrt{n}} \sum_{t=1}^{n} m(Z_t, \theta_0, h_0(X_t))}_{C} + \underbrace{\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sum_{\alpha : \alpha \in S} \frac{1}{\|\alpha\|_1!} D^\alpha m(Z_t, \theta_0, h_0(X_t)) \left(\hat{h}(X_t) - h_0(X_t)\right)^\alpha}_{G} \\ & + \underbrace{\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sum_{\alpha : \|\alpha\|_1 \le k, \alpha \notin S} \frac{1}{\|\alpha\|_1!} D^\alpha m(Z_t, \theta_0, h_0(X_t)) \left(\hat{h}(X_t) - h_0(X_t)\right)^\alpha}_{E} \\ & + \underbrace{\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sum_{\alpha : \|\alpha\|_1 = k+1} \frac{1}{(k+1)!} \begin{bmatrix} D^\alpha m_1(Z_t, \theta_0, \tilde{h}^{(1)}(X_t)) \\ \cdots \\ D^\alpha m_d(Z_t, \theta_0, \tilde{h}^{(d)}(X_t)) \end{bmatrix} \left(\hat{h}(X_t) - h_0(X_t)\right)^\alpha}_{F}, \end{aligned} \tag{11}$$

where $\tilde{h}^{(i)}(X_t), i = 1, 2, \ldots, d$ are vectors which are (potentially distinct) convex combinations of $\hat{h}(X_t)$ and $h_0(X_t)$. Note that C is the sum of n i.i.d. mean-zero random vectors divided by $\sqrt{n}$ and that the covariance $V = \text{Cov}(m(Z, \theta_0, h_0(X)))$ of each vector is finite by Assumption 1.7d. Hence, the central limit theorem implies that $C \to_d N(0, V)$. It remains to show that $G, E, F \xrightarrow{p} 0$.

First we argue that the rates of first stage consistency (Assumption 1.6) imply that $E, F \xrightarrow{p} 0$. To achieve this we will show that $\mathbb{E}[|E_i| \mid \hat{h}], \mathbb{E}[|F_i| \mid \hat{h}] \xrightarrow{p} 0$, where E_i and F_i represent the i-th entries of E and F respectively. Since the number of entries d is a constant, Lemma 12 will then imply that $E, F \xrightarrow{p} 0$. First we have

$$\begin{aligned} \mathbb{E}[|E_i| \mid \hat{h}] &\le \sum_{\alpha : \|\alpha\|_1 \le k, \alpha \notin S} \frac{\sqrt{n}}{\|\alpha\|_1!} \mathbb{E}_{Z_t}[|D^\alpha m_i(Z_t, \theta_0, h_0(X_t))(\hat{h}(X_t) - h_0(X_t))^\alpha|] && \text{(triangle inequality)} \\ &\le \sum_{\alpha : \|\alpha\|_1 \le k, \alpha \notin S} \frac{\sqrt{n}}{\|\alpha\|_1!} \sqrt{\mathbb{E}[|D^\alpha m_i(Z_t, \theta_0, h_0(X_t))|^2]} \sqrt{\mathbb{E}_{X_t}[|\hat{h}(X_t) - h_0(X_t)|^{2\alpha}]} && \text{(Cauchy-Schwarz)} \\ &\le \sum_{\alpha : \|\alpha\|_1 \le k, \alpha \notin S} \frac{\sqrt{n}}{\|\alpha\|_1!} \lambda_*(\theta_0, h_0)^{1/4} \sqrt{\mathbb{E}_{X_t}[|\hat{h}(X_t) - h_0(X_t)|^{2\alpha}]} && \text{(Assumption 1.7c)} \\ &\le \max_{\alpha : \|\alpha\|_1 \le qk, \alpha \notin S} \lambda_*(\theta_0, h_0)^{1/4} \sqrt{n} \sqrt{\mathbb{E}_{X_t}[|\hat{h}(X_t) - h_0(X_t)|^{2\alpha}]} \xrightarrow{p} 0. && \text{(Assumption 1.6)} \end{aligned}$$

Since $\tilde{h}^{(i)}$ is a convex combination of $\hat{h}$ and h_0, parallel reasoning yields

$$\begin{aligned}\mathbb{E}[|F_i| \mid \hat{h}]\mathbb{I}[\hat{h} \in \mathcal{B}_{h_0,r}] &\leq \max_{\alpha:\|\alpha\|_1=k+1} \mathbb{I}[\hat{h} \in \mathcal{B}_{h_0,r}]\sqrt{\mathbb{E}_{Z_t}[|D^\alpha m_i(Z_t,\theta_0,\tilde{h}^{(i)}(X_t))|^2]}\sqrt{n}\sqrt{\mathbb{E}_{X_t}[|\hat{h}(X_t)-h_0(X_t)|^{2\alpha}]}\\ &\leq \max_{\alpha:\|\alpha\|_1=k+1} \lambda_*(\theta_0,h_0)^{1/4}\sqrt{n}\sqrt{\mathbb{E}_{X_t}[|\hat{h}(X_t)-h_0(X_t)|^{2\alpha}]} \xrightarrow{p} 0. \qquad \text{(Assumptions 1.7c and 1.6)}\end{aligned}$$

As in Section A.1, the consistency of $\hat{h}$ (Assumption 1.6) further implies that $\mathbb{E}[|F_i| \mid \hat{h}] \xrightarrow{p} 0$.

Finally, we argue that orthogonality and the rates of the first stage imply that $G \xrightarrow{p} 0$. By S-orthogonality of the moments, for $\alpha \in S$, $\mathbb{E}\left[D^\alpha m(Z_t,\theta_0,h_0(X_t))|X_t\right] = 0$ and in particular

$$\mathbb{E}\left[D^\alpha m(Z_t,\theta_0,h_0(X_t))\left(\hat{h}(X_t)-h_0(X_t)\right)^\alpha|\hat{h}\right] = \mathbb{E}\left[\mathbb{E}\left[D^\alpha m(Z_t,\theta_0,h_0(X_t))|X_t\right]\left(\hat{h}(X_t)-h_0(X_t)\right)^\alpha|\hat{h}\right] = 0. \tag{12}$$

We now show that $\mathbb{E}\left[G_i^2|\hat{h}\right] \xrightarrow{p} 0$. We have

$$\begin{aligned}\mathbb{E}\left[G_i^2|\hat{h}\right] =\ & \frac{1}{n}\sum_{t,t'=1,2,\dots,n,t\neq t'}\mathbb{E}\left[\sum_{\alpha:\|\alpha\|_1\leq k,\alpha\in S}\frac{1}{\|\alpha\|_1!}D^\alpha m_i(Z_t,\theta_0,h_0(X_t))\left(\hat{h}(X_t)-h_0(X_t)\right)^\alpha|\hat{h}\right]^2\\ &+\frac{1}{n}\sum_{t=t'=1}^{n}\mathbb{E}\left[\left(\sum_{\alpha:\|\alpha\|_1\leq k,\alpha\in S}\frac{1}{\|\alpha\|_1!}D^\alpha m_i(Z_t,\theta_0,h_0(X_t))\left(\hat{h}(X_t)-h_0(X_t)\right)^\alpha\right)^2|\hat{h}\right]\end{aligned}$$

All the cross terms are zero because of (12). Therefore:

$$\begin{aligned}\mathbb{E}\left[G_i^2|\hat{h}\right] &= \mathbb{E}\left[\left(\sum_{\alpha:\alpha\in S}\frac{1}{\|\alpha\|_1!}D^\alpha m_i(Z_t,\theta_0,h_0(X_t))\left(\hat{h}(X_t)-h_0(X_t)\right)^\alpha\right)^2|\hat{h}\right]\\ &\leq \mathbb{E}\left[\sum_{\alpha:\alpha\in S}\frac{1}{\|\alpha\|_1!}\left(D^\alpha m_i(Z_t,\theta_0,h_0(X_t))\left(\hat{h}(X_t)-h_0(X_t)\right)^\alpha\right)^2|\hat{h}\right] && \text{(Jensen's inequality)}\\ &\leq \max_{\alpha:\alpha\in S}\mathbb{E}\left[\left(D^\alpha m_i(Z_t,\theta_0,h_0(X_t))\left(\hat{h}(X_t)-h_0(X_t)\right)^\alpha\right)^2|\hat{h}\right]\\ &\leq \max_{\alpha:\alpha\in S}\sqrt{\mathbb{E}\left[(D^\alpha m_i(Z_t,\theta_0,h_0(X_t)))^4\right]}\sqrt{\mathbb{E}\left[\left(\hat{h}(X_t)-h_0(X_t)\right)^{4\alpha}|\hat{h}\right]} && \text{(Cauchy-Schwarz)}\\ &= \max_{\alpha:\alpha\in S}\sqrt{\lambda_*(\theta_0,h_0)}\sqrt{\mathbb{E}\left[\left(\hat{h}(X_t)-h_0(X_t)\right)^{4\alpha}|\hat{h}\right]} && \text{(Assumption 1.7c)}\end{aligned}$$

Given Assumption 1.5 we get that the latter converges to zero in probability. Given that the number of moments d is also a constant, we have shown that $\mathbb{E}[\|G\|_2^2|\hat{h}] \xrightarrow{p} 0$. By Lemma 12 the latter implies that $G \xrightarrow{p} 0$.

The proof for the K-fold cross fitting estimator $\hat{\theta}^{CF}$ follows precisely the same steps as the $\hat{\theta}^{SS}$ proof (with $\sqrt{2n}$ scaling instead of $\sqrt{n}$ scaling) except for the final argument concerning $G \xrightarrow{p} 0$. In this case $G = \sum_{k=1}^{K} G_k$, where, for $k = 1,\dots,K$.

$$G_k = \frac{1}{\sqrt{2n}}\sum_{t\in I_k}\sum_{\alpha:\alpha\in S}\frac{1}{\|\alpha\|_1!}D^\alpha m(Z_t,\theta_0,h_k(X_t))\left(\hat{h}_k(X_t)-h_k(X_t)\right)^\alpha.$$

K is treated as constant with respect to the other problem parameters, and therefore it suffices to show $G_k \xrightarrow{p} 0$, for all $k = 1,2,\dots,K$. Fix $k \in [K]$. By Lemma 12 it suffices to show $\mathbb{E}\left[G_k^2|\hat{h}_k\right] \xrightarrow{p} 0$. The proof of this follows exactly the same steps as proving $\mathbb{E}\left[G^2|\hat{h}\right] \xrightarrow{p} 0$ in the $\hat{\theta}^{SS}$ case. The diagonal terms can be bounded in an identical way and the cross terms are zero again because $\hat{h}_k$ is trained in the first stage on data $(X_t)_{t\in I_k^c}$ and therefore the data $(X_t)_{t\in I_k}$ remain independent given $\hat{h}_k$. Our proof is complete.

B. Proof of Theorem 2

We prove the result for the sample-splitting estimator $\hat{\theta}^{SS}$ in (2). The proof for the K-fold cross fitting estimator $\hat{\theta}^{CF}$ in (3) is analogous and follows as in (Chernozhukov et al., 2017).

Fix any compact $A \subseteq \Theta$. Our initial goal is to establish the uniform convergence

$$\sup_{\theta \in A} |\tfrac{1}{n} \textstyle\sum_{t=1}^{n} m_i(Z_t, \theta, \hat{h}(X_t)) - \mathbb{E}[m_i(Z, \theta, h_0(X))]| \xrightarrow{p} 0 \tag{13}$$

for each moment m_i. To this end, we first note that the continuity (Assumption 1.4) and domination (Assumption 1.7d) of m_i imply the uniform law of large numbers

$$\sup_{\theta \in A, h \in \mathcal{B}_{h_0,r}} |\tfrac{1}{n} \textstyle\sum_{t=1}^{n} m_i(Z_t, \theta, h(X_t)) - \mathbb{E}_Z[m_i(Z, \theta, h(X))]| \xrightarrow{p} 0$$

for each moment m_i (see, e.g., Newey & McFadden, 1994, Lem. 2.4). Moreover, the mean value theorem and two applications of Cauchy-Schwarz yield

$$\begin{aligned} |\mathbb{E}[m_i(Z, \theta, \hat{h}(X)) \mid \hat{h}] - \mathbb{E}[m_i(Z, \theta, h_0(X))]| &\le |\mathbb{E}[\langle \nabla_\gamma m_i(Z, \theta, \tilde{h}^{(i)}(X)), \hat{h}(X) - h_0(X) \rangle \mid \hat{h}]| \\ &\le |\mathbb{E}[\|\nabla_\gamma m_i(Z, \theta, \tilde{h}^{(i)}(X))\|_2 \|\hat{h}(X) - h_0(X)\|_2 \mid \hat{h}] \\ &\le \sqrt{\mathbb{E}[\|\nabla_\gamma m_i(Z, \theta, \tilde{h}^{(i)}(X))\|_2^2 \mid \hat{h}] \mathbb{E}[\|\hat{h}(X) - h_0(X)\|_2^2 \mid \hat{h}]} \end{aligned}$$

for $\tilde{h}^{(i)}$ a convex combination of h_0 and $\hat{h}$. Hence, the uniform bound on the moments of $\nabla_\gamma m_i$ (Assumption 1.7e) and the consistency of $\hat{h}$ (Assumption 1.5) imply $\sup_{\theta \in A} |\mathbb{E}[m_i(Z, \theta, \hat{h}(X)) \mid \hat{h}] - \mathbb{E}[m_i(Z, \theta, h_0(X))]| \xrightarrow{p} 0$, and therefore

$$\mathbb{I}[\hat{h} \in \mathcal{B}_{h_0,r}] \sup_{\theta \in A} |\tfrac{1}{n} \textstyle\sum_{t=1}^{n} m_i(Z_t, \theta, \hat{h}(X_t)) - \mathbb{E}[m_i(Z, \theta, h_0(X))]| \xrightarrow{p} 0$$

by the triangle inequality. Since $\mathbb{I}[\hat{h} \in \mathcal{B}_{h_0,r}] \xrightarrow{p} 1$ by the assumed consistency of $\hat{h}$, the uniform convergence (13) follows from Lemma 11. Given the uniform convergence (13), standard arguments now imply consistency given identifiability (Assumption 1.2) and either the compactness conditions of Assumption 2.1 (see, e.g., Newey & McFadden, 1994, Thm. 2.6) or the convexity conditions of Assumption 2.2 (see, e.g., Newey & McFadden, 1994, Thm. 2.7).

C. Proof of Lemma 3

We will use the inequality that for any vector of random variables $(W_1, \ldots, W_K)$,

$$\mathbb{E}\left[\textstyle\prod_{i=1}^{K} |W_i|\right] \le \textstyle\prod_{i=1}^{K} \mathbb{E}\left[|W_i|^K\right]^{\frac{1}{K}},$$

which follows from repeated application of Hölder's inequality. In particular, we have

$$\mathbb{E}_X\left[\textstyle\prod_{i=1}^{\ell} \left|\hat{h}_i(X) - h_{0,i}(X)\right|^{2\alpha_i}\right] \le \textstyle\prod_{i=1}^{\ell} \mathbb{E}_X\left[\left|\hat{h}_i(X) - h_{0,i}(X)\right|^{2\|\alpha\|_1}\right]^{\alpha_i/\|\alpha\|_1} = \textstyle\prod_{i=1}^{\ell} \|\hat{h}_i - h_{0,i}\|_{2\|\alpha\|_1}^{2\alpha_i}$$

Thus the first part follows by taking the root of the latter inequality and multiplying by $\sqrt{n}$. For the second part of the lemma, observe that under the condition for each nuisance function we have:

$$\sqrt{n} \prod_{i=1}^{\ell} \|\hat{h}_i - h_{0,i}\|_{2\|\alpha\|_1}^{\alpha_i} = n^{\frac{1}{2} - \sum_{i=1}^{\ell} \frac{\alpha_i}{\kappa_i \|\alpha\|_1}} \prod_{i=1}^{\ell} \left(n^{\frac{1}{\kappa_i \|\alpha\|_1}} \|\hat{h}_i - h_{0,i}\|_{2\|\alpha\|_1}\right)^{\alpha_i}$$

If $\frac{1}{2} - \sum_{i=1}^{\ell} \frac{\alpha_i}{\kappa_i \|\alpha\|_1} \le 0$, then all parts in the above product converge to 0 in probability.

For the second part for all $\alpha \in S$ we similarly have

$$\mathbb{E}_X\left[\textstyle\prod_{i=1}^{\ell} \left|\hat{h}_i(X) - h_{0,i}(X)\right|^{4\alpha_i}\right] \le \textstyle\prod_{i=1}^{\ell} \mathbb{E}_X\left[\left|\hat{h}_i(X) - h_{0,i}(X)\right|^{4\|\alpha\|_1}\right]^{\alpha_i/4\|\alpha\|_1} = \textstyle\prod_{i=1}^{\ell} \|\hat{h}_i - h_{0,i}\|_{4\|\alpha\|_1}^{4\alpha_i}$$

Hence to satisfy Assumption 1.5 it suffices to satisfy $\forall \alpha \in S, \forall i$, $\|\hat{h}_i - h_{0,i}\|_{4\|\alpha\|_1} \overset{p}{\to} 0$. But by Holder inequality and our hypothesis we have

$$\|\hat{h}_i - h_{0,i}\|_{4\|\alpha\|_1} \leq \|\hat{h}_i - h_{0,i}\|_{4[\max_{\alpha \in S} \|\alpha\|_1]} \overset{p}{\to} 0,$$

as we wanted.

D. Proof of Theorem 5

Suppose that the PLR model holds with the conditional distribution of η given X Gaussian. Consider a generic moment $m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))$, where $h_0(X)$ represents any additional nuisance independent of $f_0(X), g_0(X)$. We will prove the result by contradiction. Assume that m is 2-orthogonal with respect to $(f_0(X), g_0(X))$ and satisfies Assumption 1. By 0-orthogonality, we have

$$\mathbb{E}\left[m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] = 0 \tag{14}$$

for any choice of true model parameters $(\theta_0, f_0, g_0, h_0)$, so

$$\nabla_{f_0(X)} \mathbb{E}\left[m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] = \nabla_{g_0(X)} \mathbb{E}\left[m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] = 0.$$

Since m is continuously differentiable (Assumption 1.4), we may differentiate under the integral sign (Flanders, 1973) to find that

$$\begin{aligned}
0 &= \nabla_{f_0(X)} \mathbb{E}\left[m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] \\
&= \nabla_{f_0(X)} \mathbb{E}\left[m(T, \theta_0 T + f_0(X) + \epsilon, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] \\
&= \mathbb{E}\left[\nabla_2 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) + \nabla_4 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] \quad \text{and} \\
0 &= \nabla_{g_0(X)} \mathbb{E}\left[m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] \\
&= \nabla_{g_0(X)} \mathbb{E}\left[m(g_0(X) + \eta, \theta_0(g_0(X) + \eta) + f_0(X) + \eta, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] \\
&= \mathbb{E}\left[\nabla_1 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) + \nabla_2 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))\theta_0|X\right] \\
&+ \mathbb{E}\left[\nabla_5 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right].
\end{aligned}$$

Moreover, by 1-orthogonality, we have $\mathbb{E}\left[\nabla_i m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] = 0$ for $i \in \{4, 5\}$, so

$$\mathbb{E}\left[\nabla_i m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] = 0, \quad \forall i \in \{1, 2, 4, 5\} \quad \text{and} \quad \forall\,(\theta_0, f, g, h). \tag{15}$$

Hence,

$$\nabla_{g_0(X)} \mathbb{E}\left[\nabla_4 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] = \nabla_{f_0(X)} \mathbb{E}\left[\nabla_1 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] = 0,$$

and we again exchange derivative and integral using the continuity of $\nabla^2 m$ (Assumption 1.4) (Flanders, 1973) to find

$$\begin{aligned}
&\mathbb{E}\left[\nabla_{1,4} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) + \nabla_{5,4} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))\right] \\
&+ \mathbb{E}\left[\theta_0 \nabla_{2,4} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] \\
&= \mathbb{E}\left[\nabla_{4,1} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) + \nabla_{2,1} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right].
\end{aligned}$$

Since the partial derivatives of m are differentiable by Assumption 1.4, we have $\nabla_{1,4} m = \nabla_{4,1} m$ and therefore

$$\begin{aligned}
&\mathbb{E}\left[\nabla_{5,4} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] + \theta_0 \mathbb{E}\left[\nabla_{2,4} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] \\
&= \mathbb{E}\left[\nabla_{2,1} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right]
\end{aligned}$$

By 2-orthogonality, $\mathbb{E}\left[\nabla_{5,4} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] = 0$, and hence

$$\theta_0 \mathbb{E}\left[\nabla_{2,4} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] = \mathbb{E}\left[\nabla_{2,1} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right]. \tag{16}$$

Note that equality (15) also implies

$$0 = \nabla_{f_0(X)} \mathbb{E}\left[\nabla_2 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right] = \nabla_{f_0(X)} \mathbb{E}\left[\nabla_4 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))|X\right].$$

We again exchange derivative and integral using the continuity of $\nabla^2 m$ (Assumption 1.4) (Flanders, 1973) to find

$$\begin{aligned} 0 &= \mathbb{E}\left[\nabla_{2,2} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) + \nabla_{2,4} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] \\ &= \mathbb{E}\left[\nabla_{4,2} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) + \nabla_{4,4} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right]. \end{aligned} \tag{17}$$

Since the partial derivatives of m are continuous by Assumption 1.4, we have $\nabla_{2,4} m = \nabla_{4,2} m$ and therefore

$$\mathbb{E}\left[\nabla_{2,2} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] = \mathbb{E}\left[\nabla_{4,4} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right]$$

By 2-orthogonality, $\mathbb{E}\left[\nabla_{4,4} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] = 0$, and hence

$$\mathbb{E}\left[\nabla_{2,2} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] = 0 \tag{18}$$

Combining the equalities (16), (17), and (18) we find that

$$\mathbb{E}\left[\nabla_{2,1} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] = 0. \tag{19}$$

Now, the 0-orthogonality condition (14), the continuity of ∇m (Assumption 1.4), and differentiation under the integral sign (Flanders, 1973) imply that

$$\begin{aligned} 0 &= \nabla_{\theta_0} \mathbb{E}\left[m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] = \nabla_{\theta_0} \mathbb{E}\left[m(T, \theta_0 T + f_0(X) + \epsilon, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] \\ &= \mathbb{E}\left[\nabla_2 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) \cdot T + \nabla_3 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right]. \end{aligned}$$

Since $T = g_0(X) + \eta$ and $\mathbb{E}\left[\nabla_2 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] = 0$ by equality 14,

$$\mathbb{E}\left[\nabla_2 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) \cdot \eta + \nabla_3 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] = 0 \tag{20}$$

Since η is conditionally Gaussian given X, Stein's lemma (Stein, 1981), the symmetry of the partial derivatives of m, and the equality 19 imply that

$$\begin{aligned} &\mathbb{E}\left[\nabla_2 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) \cdot \eta | X\right] = \mathbb{E}\left[\nabla_2 m(g_0(X) + \eta, Y, \theta_0, f_0(X), g_0(X), h_0(X)) \cdot \eta | X\right] \\ &= \mathbb{E}\left[\nabla_{\eta,2} m(g_0(X) + \eta, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] \\ &= \mathbb{E}\left[\nabla_{1,2} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] = \mathbb{E}\left[\nabla_{2,1} m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] = 0. \end{aligned}$$

Hence the equality (20) gives $\mathbb{E}\left[\nabla_3 m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) | X\right] = 0$ which contradicts Assumption 1.3.

E. Proof of Proposition 6

Fix any moment of the form $m(T, Y, \theta, f(X), g(X), h(X))$, where h represents any nuisance in addition to (f, g). Let F be the space of all valid nuisance functions (f, g, h) and $F(X) = \{(f(X), g(X), h(X)) : (f, g, h) \in F\}$.

We prove the lemma by contradiction. Suppose m satisfies the three hypothesis of our lemma. We start by establishing that $\text{Var}\left(m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))\right) = 0$ for all $(\theta_0, f_0, g_0, h_0)$. Fix any $(\theta_0, f_0, g_0, h_0)$, and suppose $\text{Var}\left(m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))\right) > 0$. As in the beginning of the proof of Theorem 1 the mean value theorem implies

$$\hat{J}(\hat{f}, \hat{g}, \hat{h})\sqrt{n}(\theta_0 - \hat{\theta}^{SS}) = \underbrace{\frac{1}{\sqrt{n}} \sum_{t=1}^{n} m\left(T_t, Y_t, \theta_0, \hat{f}(X_t), \hat{g}(X_t), \hat{h}(X_t)\right)}_{B} \tag{21}$$

where $\hat{J}(f, g, h) \triangleq \frac{1}{n} \sum_{t=1}^{n} \nabla_\theta m(T_t, Y_t, \tilde{\theta}, f(X_t), g(X_t), h(X_t))$, for some $\tilde{\theta}$ which is a convex combination of $\hat{\theta}^{SS}, \theta_0$. In the proof of Theorem 1 we only use Assumption 1.3 to invert $J = \mathbb{E}\left[\nabla_\theta m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)\right]$ which is the in-probability limit of $\hat{J}(\hat{f}, \hat{g}, \hat{h})$. In particular, both of the following results established in the proof of the Theorem 1 remain true in our setting:

- B tends to a normal distribution with mean zero and variance $\text{Var}\left(m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))\right) > 0$.
- $\hat{J}(\hat{f}, \hat{g}, \hat{h})$ converges in probability to J.

Since in this case $J = 0$, as Assumption 1.3 is violated, and $\sqrt{n}(\theta_0 - \hat{\theta}^{SS})$ is bounded in probability, we get that $\hat{J}(\hat{f}, \hat{g}, \hat{h})\sqrt{n}(\theta_0 - \hat{\theta}^{SS})$ converges to zero in probability. By (21), this contradicts the fact that B converges to a distribution with non-zero variance. Hence, Var $(m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))) = 0$ as desired.

Now recalling that, for all $(\theta_0, f_0, g_0, h_0)$, $\mathbb{E}[m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))] = 0$, we conclude that for all $(\theta_0, f_0, g_0, h_0)$, $m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) = 0$, almost surely with respect to the random variables X, ϵ, η. Now fix $(\theta_0, f_0, g_0, h_0)$. Now suppose that for some $(a, b) \in \mathbb{R}^2$, $m(a, b, \theta_0, f_0(X), g_0(X), h_0(X)) \neq 0$. Then, since m is continuous, there exists a neighborhood $\mathcal{N}$ such that $m(a', b', \theta_0, f_0(X), g_0(X), h_0(X)) \neq 0$ for all $(a', b') \in \mathcal{N}$. Since the conditional distribution of ϵ, η has full support (a.s. X) and, given X, (T, Y) is an invertible linear function of (ϵ, η), the conditional distribution of (T, Y) given X also has full support on $\mathbb{R}^2$ (a.s. X). Hence, $\Pr(m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X)) \neq 0) \geq \Pr((T, Y) \in \mathcal{N}) > 0$. This is a contradiction as $m(T, Y, \theta_0, f_0(X), g_0(X), h_0(X))$ is a.s. zero. Therefore, for almost every X and all $a, b \in \mathbb{R}$ and $(\theta_0, f_0, g_0, h_0)$, $m(a, b, \theta_0, f_0(X), g_0(X), h_0(X)) = 0$. Since the distribution of X is independent of θ_0 and $|\Theta| \geq 2$, we therefore have

$$\mathbb{E}[m(Y, T, \theta, f_0(X), g_0(X), h_0(X))] = 0$$

for some $\theta \neq \theta_0$, which contradicts identifiability.

F. Proof of Lemma 7

Since the characteristic function of a Gaussian distribution is well-defined and finite on the whole real line, Levy's Inversion Formula implies that the Gaussian distribution is uniquely characterized by its moments (Durrett, 2010, Sec. 3.3.1).

G. Proof of Theorem 8

Smoothness follows from the fact that m is a polynomial in $(\theta, q(X), g(X), \mu_{r-1}(X))$. Non-degeneracy follows from the PLR equations (Definition 5), the property $\mathbb{E}[\eta \mid X] = 0$, and our choice of r as

$$\begin{aligned}\mathbb{E}[\nabla_\theta m(Z, \theta_0, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X])] &= -\mathbb{E}[(T - g_0(X))(\eta^r - \mathbb{E}[\eta^r|X] - r\eta\mathbb{E}[\eta^{r-1}|X])] \\ &= -\mathbb{E}[\eta(\eta^r - \mathbb{E}[\eta^r|X] - r\eta\mathbb{E}[\eta^{r-1}|X]] \\ &= -\mathbb{E}[\mathbb{E}[\eta^{r+1}|X] - r\mathbb{E}[\eta^2|X]\mathbb{E}[\eta^{r-1}|X]] \neq 0.\end{aligned}$$

We next establish 0-orthogonality using the property $\mathbb{E}[\epsilon \mid X, T] = 0$ of Definition 5:

$$\mathbb{E}\left[m(Z, \theta_0, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X]) \mid X\right] = \mathbb{E}[\epsilon\left(\eta^r - \mathbb{E}[\eta^r|X] - r\eta\mathbb{E}[\eta^{r-1}|X] \mid X\right)] = 0.$$

Our choice of r further implies identifiability as, for $\theta \neq \theta_0$,

$$\begin{aligned}\mathbb{E}\left[m(Z, \theta, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X])\right] &= (\theta_0 - \theta)\mathbb{E}[\mathbb{E}[\eta^{r+1}|X] - \mathbb{E}[\eta|X]\mathbb{E}[\eta^r|X] - rE[\eta^2|X]\mathbb{E}[\eta^{r-1}|X]] \\ &= (\theta_0 - \theta)\mathbb{E}[\mathbb{E}[\eta^{r+1}|X] - rE[\eta^2|X]\mathbb{E}[\eta^{r-1}|X]] \neq 0.\end{aligned}$$

We invoke the properties $\mathbb{E}[\eta \mid X] = 0$ and $\mathbb{E}[\epsilon \mid X, T] = 0$ of Definition 5 to derive 1-orthogonality via

$$\begin{aligned}&\mathbb{E}\left[\nabla_{q(X)} m(Z, \theta_0, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X])|X\right] = -\mathbb{E}\left[\eta^r - \mathbb{E}[\eta^r|X] - r\eta\mathbb{E}[\eta^{r-1}|X] \mid X\right] = 0, \\ &\mathbb{E}\left[\nabla_{g(X)} m(Z, \theta_0, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X])|X\right] \\ &\ = \theta_0\mathbb{E}\left[\eta^r - \mathbb{E}[\eta^r|X] - r\eta\mathbb{E}[\eta^{r-1}|X] \mid X\right] - \mathbb{E}\left[\epsilon(r\eta^{r-1} - r\mathbb{E}[\eta^{r-1}|X]) \mid X\right] = 0, \quad \text{and} \\ &\mathbb{E}\left[\nabla_{\mu_{r-1}(X)} m(Z, \theta_0, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X])\right] = -\mathbb{E}[\epsilon\, r\, \eta|X] = 0.\end{aligned}$$

The same properties also yield 2-orthogonality for the second partial derivatives of $q(X)$ via

$$\begin{aligned}&\mathbb{E}\left[\nabla^2_{q(X),q(X)} m(Z, \theta_0, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X])|X\right] = 0, \\ &\mathbb{E}\left[\nabla^2_{q(X),g(X)} m(Z, \theta_0, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X])|X\right] = \mathbb{E}\left[r\eta^{r-1} - r\mathbb{E}[\eta^{r-1}|X]|X\right] = 0, \quad \text{and} \\ &\mathbb{E}\left[\nabla^2_{q(X),\mu_{r-1}(X)} m(Z, \theta_0, q_0(X), g_0(X), \mathbb{E}[\eta^{r-1}|X])|X\right] = \mathbb{E}\left[r\, \eta|X\right] = 0,\end{aligned}$$

for the second partial derivatives of $g(X)$ via

$$\begin{aligned}
&\mathbb{E}\left[\nabla^2_{g(X),g(X)} m(Z,\theta_0,q_0(X),g_0(X),\mathbb{E}[\eta^{r-1}|X])|X\right] = \mathbb{E}\left[-(r\eta^{r-1} - r\mathbb{E}[\eta^{r-1}|X]) + \epsilon\, r(r-1)\eta^{r-2}|X\right] \\
&= r(r-2)\mathbb{E}\left[\mathbb{E}\left[\epsilon|X,T\right]\eta^{r-2}|X\right] = 0 \quad \text{and} \\
&\mathbb{E}\left[\nabla^2_{g(X),\mu_{r-1}(X)} m(Z,\theta_0,q_0(X),g_0(X),\mathbb{E}[\eta^{r-1}|X])|X\right] = -\theta_0\mathbb{E}[r\eta|X] + \mathbb{E}[\epsilon\, r|X] = 0,
\end{aligned}$$

and for the second partial derivatives of $\mu_{r-1}(X)$ via

$$\mathbb{E}\left[\nabla^2_{\mu_{r-1}(X),\mu_{r-1}(X)} m(Z,\theta_0,q_0(X),g_0(X),\mathbb{E}[\eta^{r-1}|X]) \mid X\right] = 0.$$

This establishes 2-orthogonality.

H. Proof of Theorem 9

The majority of the proof is identical to that of Theorem 8; it only remains to show that the advertised partial derivatives with respect to $\mu_r(X)$ are also mean zero given X. These equalities follow from the property $\mathbb{E}[\eta \mid X] = 0$ of Definition 5:

$$\begin{aligned}
&\mathbb{E}\left[\nabla_{\mu_r(X)} m(Z,\theta_0,q_0(X),g_0(X),\mathbb{E}[\eta^{r-1}|X],\mathbb{E}[\eta^r|X]))|X\right] = -\mathbb{E}\left[\epsilon|X\right] = 0, \\
&\mathbb{E}\left[\nabla^2_{\mu_r(X),\mu_r(X)} m(Z,\theta_0,q_0(X),g_0(X),\mathbb{E}[\eta^{r-1}|X],\mathbb{E}[\eta^r|X])) \mid X\right] = 0, \quad \text{and} \\
&\mathbb{E}\left[\nabla^2_{\mu_r(X),\mu_{r-1}(X)} m(Z,\theta_0,q_0(X),g_0(X),\mathbb{E}[\eta^{r-1}|X],\mathbb{E}[\eta^r|X])) \mid X\right] = 0.
\end{aligned}$$

I. Proof of Theorem 10

We prove the result explicitly for the excess kurtosis setting with $\mathbb{E}[\eta^4] \neq 3\mathbb{E}[\eta^2]^2$. A parallel argument yields the result for non-zero skewness ($\mathbb{E}[\eta^3] \neq 0$).

To establish $\sqrt{n}$-consistency and asymptotic normality, it suffices to check each of the preconditions of Theorems 1 and 2. Since η is independent of X and $\mathbb{E}[\eta^4] \neq 3\mathbb{E}[\eta^2]^2$, the conditions of Theorem 9 are satisfied with $r = 3$. Hence, the moments m of Theorem 9 satisfy S-orthogonality (Assumption 1.1) for $S = \{\alpha \in \mathbb{N}^4 : \|\alpha\|_1 \leq 2\} \setminus \{(1,0,0,1),(0,1,0,1)\}$ with respect to the nuisance $(\langle q_0, X\rangle, \langle \gamma_0, X\rangle, \mathbb{E}[\eta^2], \mathbb{E}[\eta^3])$, identifiability (Assumption 1.2), non-degeneracy of $\mathbb{E}\left[\nabla_\theta m(Z,\theta_0,h_0(X))\right]$ (Assumption 1.3), and continuity of ∇m^2 (Assumption 1.4). The form of m, the standard Gaussian i.i.d. components of X, and the almost sure boundedness of η and ϵ further imply that the regularity conditions of Assumption 1.7 are all satisfied for any choice of $r > 0$. Hence, it only remains to establish the first stage consistency and rate assumptions (Assumptions 1.5 and 1.6) and the convexity conditions (Assumption 2.2).

I.1. Checking Rate of First Stage (Assumption 1.6)

We begin with Assumption 1.6. Since $\{\alpha \in \mathbb{N}^4 : \|\alpha\|_1 \leq 3\} \setminus S = \{\alpha \in \mathbb{N}^4 : \|\alpha\|_1 = 3\} \cup \{(1,0,0,1),(0,1,0,1)\}$ by Lemma 3, it suffices to establish the sufficient condition (6) for $\alpha = (0,1,0,1)$ and $\alpha = (1,0,0,1)$ and the condition (7) for the α with $\|\alpha\|_1 = 3$. Hence, it suffices to satisfy

(1) $n^{\frac{1}{2}}\mathbb{E}_X[|\langle X, \hat{q} - q_0\rangle|^4]^{\frac{1}{4}} \cdot |\hat{\mu}_3 - \mathbb{E}[\eta^3]| \xrightarrow{p} 0$, which corresponds to $\alpha = (1,0,0,1)$ and condition (6),

(2) $n^{\frac{1}{2}}\mathbb{E}_X[|\langle X, \hat{\gamma} - \gamma_0\rangle|^4]^{\frac{1}{4}} \cdot |\hat{\mu}_3 - \mathbb{E}[\eta^3]| \xrightarrow{p} 0$, which corresponds to $\alpha = (0,1,0,1)$ and condition (6),

(3) $n^{\frac{1}{2}}\mathbb{E}_X[|\langle X, \hat{q} - q_0\rangle|^6]^{\frac{1}{2}} \xrightarrow{p} 0$,

(4) $n^{\frac{1}{2}}\mathbb{E}_X[|\langle X, \hat{\gamma} - \gamma_0\rangle|^6]^{\frac{1}{2}} \xrightarrow{p} 0$,

(5) $n^{\frac{1}{2}}|\hat{\mu}_2 - \mathbb{E}[\eta^2]|^3 \xrightarrow{p} 0$, and

(6) $n^{\frac{1}{2}}|\hat{\mu}_3 - \mathbb{E}[\eta^3]|^3 \xrightarrow{p} 0$,

where X a vector of i.i.d. mean-zero standard Gaussian entries, independent from the first stage, and the convergence to zero is considered in probability with respect to the first stage random variables.

We will estimate q, γ_0 using half of our first-stage sample and use our estimate $\hat{\gamma}$ to produce an estimate of the second and third moments of η based on the other half of the sample and the following lemma.

Lemma 13. *Suppose that an estimator $\hat{\gamma} \in \mathbb{R}^p$ based on n sample points satisfies $\mathbb{E}_X[|\langle X, \hat{\gamma} - \gamma_0 \rangle|^6]^{\frac{1}{2}} = O_P\left(\frac{1}{\sqrt{n}}\right)$ for X independent of $\hat{\gamma}$. If*

$$\hat{\mu}_2 := \tfrac{1}{n}\textstyle\sum_{t=1}^{n}(T'_t - \langle X'_t, \hat{\gamma}\rangle)^2 \quad \textit{and} \quad \hat{\mu}_3 := \tfrac{1}{n}\textstyle\sum_{t=1}^{n}(T'_t - \langle X'_t, \hat{\gamma}\rangle)^3 - 3\tfrac{1}{n}\textstyle\sum_{t=1}^{n}(T'_t - \langle X'_t, \hat{\gamma}\rangle)\hat{\mu}_2$$

for $(T'_t, X'_t)_{t=1}^n$ i.i.d. replicates of (T, X) independent of $\hat{\gamma}$, then

$$|\hat{\mu}_2 - \mathbb{E}[\eta^2]| = O_P\left(\frac{1}{n^{\frac{1}{3}}}\right) \quad \textit{and} \quad |\hat{\mu}_3 - \mathbb{E}[\eta^3]| = O_P\left(\frac{1}{\sqrt{n}}\right). \tag{22}$$

As a result,

$$n^{\frac{1}{2}}|\hat{\mu}_2 - \mathbb{E}[\eta^2]|^3 \xrightarrow{p} 0 \quad \textit{and} \quad n^{\frac{1}{2}}|\hat{\mu}_3 - \mathbb{E}[\eta^3]|^3 \xrightarrow{p} 0.$$

Proof. We begin with the third moment estimation. For a new datapoint (T, X) independent of $\hat{\gamma}$, define $\delta \triangleq \langle X, \gamma_0 - \hat{\gamma}\rangle$ so that $T - \langle X, \hat{\gamma}\rangle = \delta + \eta$. Since η is independent of $(X, \hat{\gamma})$, and $\mathbb{E}[\eta] = 0$, we have

$$\mathbb{E}_{X,\eta}[(\delta+\eta)^3] - 3\mathbb{E}_{X,\eta}[(\delta+\eta)]\mathbb{E}_{X,\eta}[(\delta+\eta)^2] = \mathbb{E}[\eta^3] + \mathbb{E}_X[\delta^3] - 3\mathbb{E}_X[\delta^2]\mathbb{E}_X[\delta]$$

or equivalently

$$\mathbb{E}[\eta^3] = \mathbb{E}_{X,\eta}[(\delta+\eta)^3] - 3\mathbb{E}_{X,\eta}[(\delta+\eta)]\mathbb{E}_{X,\eta}[(\delta+\eta)^2] - \mathbb{E}_X[\delta^3] + 3\mathbb{E}_X[\delta^2]\mathbb{E}_X[\delta]. \tag{23}$$

Since $\mathbb{E}_X[|\delta|^3] \leq \mathbb{E}_X[\delta^6]^{\frac{1}{2}} = O_P(1/\sqrt{n})$ by Cauchy-Schwarz and our assumption on $\hat{\gamma}$, and $|\mathbb{E}_X[\delta]\mathbb{E}_X[\delta^2]| \leq \mathbb{E}_X[|\delta|^3]$ by Holder's inequality, the equality (23) implies that

$$|\mathbb{E}[\eta^3] - (\mathbb{E}_{X,\eta}[(\delta+\eta)^3] - 3\mathbb{E}_{X,\eta}[\delta+\eta]\mathbb{E}_{X,\eta}[(\delta+\eta)^2])| = O_P(1/\sqrt{n}).$$

Since $\mathbb{E}_{X,\eta}[(\delta+\eta)^6] = O(1)$, the central limit theorem, the strong law of large numbers, and Slutsky's theorem imply that

$$\hat{\mu}_3 - (\mathbb{E}_{X,\eta}[(\delta+\eta)^3] - 3\mathbb{E}_{X,\eta}[\delta+\eta]\mathbb{E}_{X,\eta}[(\delta+\eta)^2]) = O_P(1/\sqrt{n}).$$

Therefore,

$$|\hat{\mu}_3 - \mathbb{E}[\eta^3]| = O_P(1/\sqrt{n}).$$

The second moment estimation follows similarly using the identity, $\mathbb{E}[\eta^2] = \mathbb{E}_{X,\eta}[(\delta+\eta)^2] - \mathbb{E}_X[\delta^2]$, and the fact that $\mathbb{E}_X[\delta^2] \leq \mathbb{E}_X[|\delta|^3]^{\frac{2}{3}} = O_P(n^{-\frac{1}{3}})$ by Holder's inequality. ■

In light of Lemma 13 it suffices to estimate the vectors q_0 and γ_0 using n sample points so that

- $n^{\frac{1}{2}}\mathbb{E}_X[|\langle X, \hat{q} - q_0\rangle|^4]^{\frac{1}{4}} n^{-\frac{1}{2}} \xrightarrow{p} 0 \Leftrightarrow \mathbb{E}_X[|\langle X, \hat{q} - q_0\rangle|^4]^{\frac{1}{4}} \xrightarrow{p} 0$,
- $n^{\frac{1}{2}}\mathbb{E}_X[|\langle X, \hat{\gamma} - \gamma_0\rangle|^4]^{\frac{1}{4}} n^{-\frac{1}{2}} \xrightarrow{p} 0 \Leftrightarrow \mathbb{E}_X[|\langle X, \hat{\gamma} - \gamma_0\rangle|^4]^{\frac{1}{4}} \xrightarrow{p} 0$,
- $n^{\frac{1}{2}}\mathbb{E}_X[|\langle X, \hat{q} - q_0\rangle|^6]^{\frac{1}{2}} \xrightarrow{p} 0$, and
- $n^{\frac{1}{2}}\mathbb{E}_X[|\langle X, \hat{\gamma} - \gamma_0\rangle|^6]^{\frac{1}{2}} \xrightarrow{p} 0$,

and the rest of the conditions will follow. To achieve these conclusions we use the following result on the performance of the Lasso. The following theorem is distilled from (Hastie et al., 2015, Chapter 11).

Theorem 14. *Let $p, s \in \mathbb{N}$ with $s \leq p$ and $s = o(n^{2/3}/\log p)$ and $\sigma > 0$, and suppose that we observe i.i.d. datapoints $(\tilde{Y}_i, \tilde{X}_i)_{i=1}^n$ distributed according to the model $\tilde{Y} = \langle \tilde{X}, \beta_0 \rangle + w$ for an s-sparse $\beta_0 \in \mathbb{R}^p$, $\tilde{X} \in \mathbb{R}^p$ with standard Gaussian entries, and $w \in \mathbb{R}^p$ independent mean-zero noise with $\|w\|_\infty \leq \sigma$. Suppose that p grows to infinity with n. Then with a choice of tuning parameter $\lambda_n = 2\sigma\sqrt{3\log p/n}$, the Lasso estimate $\hat{\beta}_0$ fit to this dataset satisfies $\|\hat{\beta}_0 - \beta_0\|_2 = O_P(\sqrt{s \log p/n})$.*

Proof. Using Theorem 11.1 and Example 11.2 of (Hastie et al., 2015), we know that since $\tilde{X}$ has iid $N(0,1)$ entries, if $\lambda_n = 2\sigma\sqrt{3\log(p)/n}$, we have

$$\Pr\left[\frac{\|\hat{\beta}_0 - \beta_0\|_2}{\sigma\sqrt{3s\log p/n}} > 1\right] \leq 2\exp\left\{-\frac{1}{2}\log(p)\right\}. \tag{24}$$

Since p grows unboundedly with n, for any fixed $\epsilon > 0$, we have that for some some finite N_ϵ, if $n > N_\epsilon$, the right hand side is at most ϵ. Thus we can conclude that: $\|\hat{\beta}_0 - \beta_0\|_2 = O_p\left(\sqrt{s\log p/n}\right)$. ■

Notice that for q_0 we know

$$\begin{aligned} Y &= \theta_0 T + \langle X, \beta_0 \rangle + \epsilon \\ &= \theta_0 \langle X, \gamma_0 \rangle + \theta_0 \eta + \langle X, \beta_0 \rangle + \epsilon && \text{(from the definition of } T\text{)} \\ &= \langle X, q_0 \rangle + \theta_0 \eta + \epsilon && \text{(since } q_0 = \theta_0\gamma_0 + \beta_0\text{)} \end{aligned}$$

Hence,

$$Y = \langle X, q_0 \rangle + \epsilon + \theta_0 \eta,$$

and we know that the noise term, $\epsilon + \theta_0\eta$ is almost surely bounded by $C + CM = C(M+1)$. Hence, by Theorem 14, our Lasso estimate $\hat{q}$ satisfies $\|\hat{q} - q_0\|_2 = O_P(\sqrt{s\log p/n})$. Similarly, our Lasso estimate $\hat{\gamma}$ satisfies $\|\hat{\gamma} - \gamma_0\|_2 = O_P(\sqrt{s\log p/n})$.

Now, since X has iid mean-zero standard Gaussian components, we know that for all vectors $v \in \mathbb{R}^p$ and $a \in \mathbb{N}$ it holds $\mathbb{E}[|\langle X, v\rangle|^a] = O\left(\sqrt{a}^a \|v\|_2^a\right)$. Applying this to $v = \hat{q} - q$ and $v = \hat{\gamma} - \gamma_0$ for $a \in \{4, 6\}$ we have

$$\begin{aligned} \mathbb{E}_X[|\langle X, \hat{q} - q_0\rangle|^4] &= O(\|\hat{q} - q_0\|_2^4) = O_P\left(\left[\sqrt{\tfrac{s\log p}{n}}\right]^4\right) \\ \mathbb{E}_X[|\langle X, \hat{\gamma} - \gamma_0\rangle|^4] &= O(\|\hat{\gamma} - \gamma_0\|_2^4) = O_P\left(\left[\sqrt{\tfrac{s\log p}{n}}\right]^4\right) \\ \mathbb{E}_X[|\langle X, \hat{q} - q_0\rangle|^6] &= O(\|\hat{q} - q_0\|_2^6) = O_P\left(\left[\sqrt{\tfrac{s\log p}{n}}\right]^6\right) \\ \mathbb{E}_X[|\langle X, \hat{\gamma} - \gamma_0\rangle|^6] &= O(\|\hat{\gamma} - \gamma_0\|_2^6) = O_P\left(\left[\sqrt{\tfrac{s\log p}{n}}\right]^6\right). \end{aligned}$$

Now for the sparsity level $s = o\left(\frac{n^{2/3}}{\log p}\right)$ we have $\sqrt{\frac{s\log p}{n}} = o(n^{-\frac{1}{6}})$ which implies all of the desired conditions for Assumption 1.6.

I.2. Checking Consistency of First Stage (Assumption 1.5)

Next we prove that Assumption 1.5 is satisfied. Since $\max_{\alpha \in S}\|\alpha\|_1 = 2$ it suffices by Lemma 3 to show that for our choices of $\hat{\gamma}, \hat{q}, \hat{\mu}_2$, and $\hat{\mu}_3$ we have

$$\mathbb{E}_X[|\langle X, \hat{q} - q_0\rangle|^8]^{\frac{1}{8}} \xrightarrow{p} 0 \tag{25}$$

$$\mathbb{E}_X[|\langle X, \hat{\gamma} - \gamma_0\rangle|^8]^{\frac{1}{8}} \xrightarrow{p} 0 \tag{26}$$

$$|\hat{\mu}_2 - \mathbb{E}[\eta^2]| \xrightarrow{p} 0 \tag{27}$$

$$|\hat{\mu}_3 - \mathbb{E}[\eta^3]| \xrightarrow{p} 0. \tag{28}$$

Parts (27) and (28) follow directly from Lemma 13. Since X consists of standard Gaussian entries, an analogous argument to that above implies that

$$\mathbb{E}_X[|\langle X, \hat{q} - q_0\rangle|^8]^{\frac{1}{8}} = O(\|\hat{q} - q_0\|_2) = O_P\left(\left[\sqrt{\frac{s\log p}{n}}\right]\right)$$

$$\mathbb{E}_X[|\langle X, \hat{\gamma} - \gamma_0\rangle|^8]^{\frac{1}{8}} = O(\|\hat{\gamma} - \gamma_0\|_2) = O_P\left(\left[\sqrt{\frac{s\log p}{n}}\right]\right).$$

Now for the sparsity level $s = o(\frac{n^{\frac{2}{3}}}{(M+1)^2\log p})$ we have $\sqrt{\frac{s\log p}{n}} = o(1)$ which implies also conditions (25) and (26).

I.3. Checking Convexity Conditions (Assumption 2.2)

Finally, we establish the convexity conditions (Assumption 2.2). We consider $\Theta = \mathbb{R}$, which is convex. Without loss of generality, assume $3\mathbb{E}[\eta^2]^2 > \mathbb{E}[\eta^4]$; otherwise, one can establish the convexity conditions for $-m$. Let $F_n(\theta) = \frac{1}{n}\sum_{t=1}^n m(Z_t, \theta, \hat{h}(X_t))$. Since F_n is continuously differentiable, F_n is the derivative of a convex function whenever $\nabla F_n(\theta) \geq 0$, for all $\theta \in \Theta$. Since F_n is linear in θ we have for all $\theta \in \Theta$

$$\nabla F_n(\theta) = \frac{1}{n}\sum_{t=1}^n -(T_t - \langle\hat{\gamma}, X_t\rangle)^4 + (T_t - \langle\hat{\gamma}, X_t\rangle)\hat{\mu}_3 + 3(T_t - \langle\hat{\gamma}, X_t\rangle)^2\hat{\mu}_2,$$

the established consistency of $(\hat{\gamma}, \hat{\mu}_3, \hat{\mu}_2)$ and Slutsky's theorem imply that

$$\nabla F_n(\theta) - \frac{1}{n}\sum_{t=1}^n -(T_t - \langle\gamma, X_t\rangle)^4 + (T_t - \langle\gamma, X_t\rangle)\mathbb{E}[\eta^3] + 3(T_t - \langle\gamma, X_t\rangle)^2\mathbb{E}[\eta^2] \xrightarrow{p} 0.$$

The strong law of large numbers now yields

$$\nabla F_n(\theta) - (3\mathbb{E}[\eta^2]^2 - \mathbb{E}[\eta^4]) \xrightarrow{p} 0.$$

Hence,

$$\Pr(\nabla F_n(\theta) < 0) \leq \Pr(|\nabla F_n(\theta) - (3\mathbb{E}[\eta^2]^2 - \mathbb{E}[\eta^4])| > 3\mathbb{E}[\eta^2]^2 - \mathbb{E}[\eta^4]) \to 0,$$

verifying Assumption 2.2. The proof is complete.

J. Proofs of Auxiliary Lemmata

J.1. Proof of Lemma 11

Since each Y_n is binary, and $Y_n \xrightarrow{p} 1$, for every $\epsilon > 0$,

$$\Pr[|X_n(1 - Y_n)| > \epsilon] \leq \Pr[Y_n = 0] = \Pr[|1 - Y_n| > 1/2] \to 0.$$

Hence, $X_n(1 - Y_n) \xrightarrow{p} 0$. Both advertised claims now follow by Slutsky's theorem (van der Vaart, 1998, Thm. 2.8).

J.2. Proof of Lemma 12

Let $X_{n,i}$ denote the i-th coordinate of X_n, i.e. $\|X_n\|_p^p = \sum_{i=1}^d X_{n,i}^p$. By the assumption of the lemma, we have that for every ϵ, δ, there exists $n(\epsilon, \delta)$, such that for all $n \geq n(\epsilon, \delta)$:

$$\Pr\left[\max_i \mathbb{E}\left[|X_{n,i}|^p | Z_n\right] > \epsilon\right] < \delta$$

Let $\mathcal{E}_n$ denote the event $\{\max_i \mathbb{E}\left[|X_{n,i}|^p | Z_n\right] \leq \epsilon\}$. Hence, $\Pr[\mathcal{E}_n] \geq 1 - \delta$, for any $n \geq n(\epsilon, \delta)$. By Markov's inequality, for any $n \geq n\left(\epsilon^p\delta/2d, \delta/2d\right)$, the event $\mathcal{E}_n$ implies that:

$$\Pr\left[|X_{n,i}|^p > \epsilon^p | Z_n\right] \leq \frac{\mathbb{E}\left[|X_{n,i}|^p | Z_n\right]}{\epsilon^p} \leq \frac{\delta}{2d}$$

Thus, we have:

$$\begin{aligned}\Pr[|X_{n,i}| > \epsilon] =& \ \mathbb{E}\left[\Pr\left[|X_{n,i}|^p > \epsilon^p | Z_n\right]\right]\\ =& \ \mathbb{E}\left[\Pr[|X_{n,i}|^p > \epsilon^p | Z_n] | \mathcal{E}_n\right] \cdot \Pr[\mathcal{E}_n] + \mathbb{E}\left[\Pr[|X_{n,i}|^p > \epsilon^p | Z_n] | \neg\mathcal{E}_n\right] \cdot \Pr[\neg\mathcal{E}_n] \leq \frac{\delta}{d}\end{aligned}$$

By a union bound over i, we have that $\Pr[\max_i |X_{n,i}| > \epsilon] \leq \delta$. Hence, we also have that for any ϵ, δ, for any $n \geq n({\epsilon^p \delta}/{2d}, {\delta}/{2d})$, $\Pr[\|X_n\|_\infty > \epsilon] \leq \delta$, which implies $X_n \xrightarrow{p} 0$.