APPENDIX

A Main results

In this section of the appendix we provide the proofs for the main results of Section 2. At the end of the section we give some background on asymptotically linear estimators.

A.1 Proof of Theorem 1

A.1.1 Notation and assumptions

In all our applications Θ is either a vector space or an affine space. Let $T(\Theta)$ and $T^*(\Theta)$ be the tangent and co-tangent spaces of Θ.[19] Thus, for $\theta_1, \theta_2 \in \Theta$ we have $(\theta_1 - \theta_2) \in T(\Theta)$, and $T^*(\Theta)$ is the set of linear maps $u : T(\Theta) \to \mathbb{R}$. For $v \in T(\Theta)$ and $u \in T^*(\Theta)$ we use the bracket notation $\langle v, u \rangle \in \mathbb{R}$ to denote their scalar product. Our squared distance measure $d(\theta_0, \theta(\eta))$ on Θ induces a norm on the tangent space $T(\Theta)$, namely for $v \in T(\Theta)$,

$$\|v\|^2_{\mathrm{ind},\eta} = \lim_{\epsilon \to 0} \frac{d\big(\theta(\eta) + \epsilon^{1/2} v, \theta(\eta)\big)}{\epsilon}.$$

For every $\eta \in \mathcal{B}$ we assume that there exists a map $\Omega_\eta : T(\Theta) \to T^*(\Theta)$ such that, for all $v \in T(\Theta)$,

$$\|v\|^2_{\mathrm{ind},\eta} = \langle v, \Omega_\eta v \rangle .$$

We assume that Ω_η is invertible, and write $\Omega_\eta^{-1} : T^*(\Theta) \to T(\Theta)$ for its inverse.

For a scalar function on Θ, such as $\delta : \Theta \mapsto \mathbb{R}$, we have $\nabla_\theta \delta_\theta \in T^*(\Theta)$; that is, the typical element of $T^*(\Theta)$ is a gradient. Conversely, for a map to Θ, such as $\eta \mapsto \theta(\eta)$, we have $\frac{\partial \theta(\eta)}{\partial \eta_k} \in T(\Theta)$. The two versions of the Jacobian $G'_\eta : \mathbb{R}^{\dim \eta} \to T(\Theta)$ and $G_\eta : T^*(\Theta) \to \mathbb{R}^{\dim \eta}$ are defined by

$$G'_\eta : q \mapsto \sum_{k=1}^{\dim \eta} q_k \frac{\partial \theta(\eta)}{\partial \eta_k}, \qquad G_\eta : u \mapsto \left(\left\langle \frac{\partial \theta(\eta)}{\partial \eta_k}, u \right\rangle \right)_{k=1,\ldots,\dim \eta},$$

where $q \in \mathbb{R}^{\dim \eta}$ and $u \in T^*(\Theta)$. Similarly, the Hessian $H_{\theta(\eta)} : T(\Theta) \to T^*(\Theta)$ is defined by

$$H_{\theta(\eta)} : v \mapsto \mathbb{E}_{\theta(\eta)} \left[\left\langle v, \nabla_\theta \log f_{\theta(\eta)}(Y) \right\rangle \nabla_\theta \log f_{\theta(\eta)}(Y) \right].$$

[19] If Θ is a more general manifold (not just an affine space), then the tangent and co-tangent spaces depend on the particular value of $\theta \in \Theta$. We then need a connection on the manifold that provides a map between the tangent spaces at $\theta(\eta)$ and $\theta_0 \in \Gamma_\epsilon(\eta)$. All the proofs can be extended to that case, as long as the underlying connection on the manifold is sufficiently smooth. However, this additional formalism is unnecessary to deal with the models discussed in this paper.

The definitions of the projected Hessian $\widetilde{H}_{\theta(\eta)} : T(\Theta) \to T^*(\Theta)$ and the projected gradient operator $\widetilde{\nabla}_\theta$ are then as in the main text, namely $\widetilde{H}_{\theta(\eta)} = H_{\theta(\eta)} - H_{\theta(\eta)} G'_\eta H_\eta^{-1} G_\eta H_{\theta(\eta)}$, and $\widetilde{\nabla}_\theta = \nabla_\theta - H_{\theta(\eta)} G'_\eta H_\eta^{-1} \nabla_\eta$. We have $\widetilde{\nabla}_\theta \delta_{\theta(\eta)} \in T^*(\Theta)$.

The dual norm for $u \in T^*(\Theta)$ was defined in the main text. We have

$$\|u\|_\eta = \sup_{v \in T(\Theta)\setminus\{0\}} \frac{\langle v, u \rangle}{\|v\|_{\text{ind},\eta}}, \qquad\qquad \|u\|_\eta^2 = \left\langle \Omega_\eta^{-1} u, u \right\rangle .$$

$\|\cdot\|_\eta$ is also the norm on $T^*(\Theta)$ that is naturally induced by $d(\theta_0, \theta(\eta))$. We use the shorter notation $\|\cdot\|_\eta$ for that norm, because it also appears in the main text. Notice also that $\|\cdot\|_{\text{ind},\eta}$, $\|\cdot\|_\eta$, Ω_η, and Ω_η^{-1} could all be defined for general $\theta \in \Theta$, but since we use them only at the reference values $\theta = \theta(\eta)$ we index them simply by η.

Throughout we assume that $\dim \eta$ is finite. For vectors $w \in \mathbb{R}^{\dim \eta}$ we use the standard Euclidean norm $\|w\|$, and for $\dim \eta \times \dim \eta$ matrices we use the spectral matrix norm, which we also denote by $\|\cdot\|$.

The vector norms $\|\cdot\|_{\text{ind},\eta}$, $\|\cdot\|_\eta$, $\|.\|$ on $T(\Theta)$, $T^*(\Theta)$, $\mathbb{R}^{\dim \eta}$ immediately induce norms on any maps between $T(\Theta)$, $T^*(\Theta)$, $\mathbb{R}^{\dim \eta}$, and $\mathbb{R}$. With a slight abuse of notation we denote all those norms simply by $\|.\|_\eta$. For example, for $H_{\theta(\eta)} : T(\Theta) \to T^*(\Theta)$ we have

$$\left\| H_{\theta(\eta)} \right\|_\eta := \sup_{v \in T(\Theta)\setminus\{0\}} \frac{\|H_{\theta(\eta)} v\|_\eta}{\|v\|_{\text{ind},\eta}} = \sup_{v,w \in T(\Theta)\setminus\{0\}} \frac{\left\langle w, H_{\theta(\eta)} v \right\rangle}{\|v\|_{\text{ind},\eta} \|w\|_{\text{ind},\eta}}.$$

Our first set of assumptions is as follows.

Assumption A1. *We assume that $Y_i \sim$ i.i.d.f_{θ_0}. In addition, we impose the following regularity conditions:*

(i) *We consider $n \to \infty$ and $\epsilon \to 0$ such that $\epsilon n \to c$, for some constant $c \in (0, \infty)$.*

(ii) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left| \delta_{\theta_0} - \delta_{\theta(\eta)} - \left\langle \theta_0 - \theta(\eta), \nabla_\theta \delta_{\theta(\eta)} \right\rangle \right| = o(\epsilon^{1/2})$.

(iii) $$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) \right]^2 dy = o(1),$$
$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left\| \nabla_\theta \log f_{\theta(\eta)}(y) \right\|_\eta^2 \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) \right]^2 dy = o(1),$$
$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) - \left\langle \theta_0 - \theta(\eta), \nabla_\theta f_{\theta(\eta)}^{1/2}(y) \right\rangle \right]^2 dy = o(\epsilon).$$

(iv) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \epsilon^{-1/2} \|\theta_0 - \theta(\eta)\|_{\text{ind},\eta} = 1 + o(1)$. *Furthermore, for $u(\eta) \in T^*(\Theta)$ with $\sup_{\eta\in\mathcal{B}} \|u(\eta)\|_\eta = O(1)$ we have*

$$\sup_{\eta\in\mathcal{B}} \left| \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \epsilon^{-1/2} \left\langle \theta_0 - \theta(\eta), u(\eta) \right\rangle - \|u(\eta)\|_\eta \right| = o(1).$$

(v) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \|\nabla_\theta \delta_{\theta_0}\|_\eta = O(1)$, $\sup_{\eta\in\mathcal{B}} \left\|H_\eta^{-1}\right\| = O(1)$, $\sup_{\eta\in\mathcal{B}} \|G_\eta\|_\eta = O(1)$, $\sup_{\eta\in\mathcal{B}} \left\|\Omega_\eta^{-1}\right\|_\eta = O(1)$, $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} \|\nabla_\theta \log f_{\theta_0}(Y)\|_\eta^{2+\nu} = O(1)$, *for some* $\nu > 0$.

Part (i) of Assumption A1 describes our asymptotic framework, where the assumption $\epsilon n \to c$ is required to ensure that the squared worst-case bias (of order ϵ) and the variance (of order $1/n$) of the estimators for δ_{θ_0} are asymptotically of the same order, so that MSE provides a meaningful balance between bias and variance also asymptotically. Part (ii) requires δ_{θ_0} to be sufficiently smooth in θ_0, so that a first-order Taylor expansion provides a good local approximation of δ_{θ_0}. Part (iii) imposes similar smoothness assumption on $f_{\theta_0}(y)$ in θ_0. The first condition in part (iii) is just continuity in Hellinger distance, and the second condition is very similar, but also involves the score of the model. The last condition in part (iii) is a standard condition of differentiability in quadratic mean (see, e.g., equation (5.38) in Van der Vaart, 2000). Part (iv) of the assumption requires that our distance measure $d(\theta, \theta(\eta))$ converges to the associated norm for small values ϵ in a smooth fashion. Finally, part (v) requires invertibility of H_η^{-1} and Ω_η^{-1}, and uniform boundedness of various derivatives and of the $(2+\nu)$-th moment of $\nabla_\theta \log f_{\theta(\eta)}(y)$. Notice that invertibility of $H_{\theta(\eta)}$ is *not* required for our results.

For many of the proofs (specifically, all results below up to Proposition A1) we only need the regularity conditions in Assumption A1. However, in order to describe the properties of our Minimum-MSE estimator $\widehat{\delta}_\epsilon^{\mathrm{MMSE}} = \delta_{\theta(\widehat{\eta})} + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i, \widehat{\eta})$ we also need to account for the fact that $\widehat{\eta}$ is itself already an estimator. It turns our that the leading-order asymptotic properties of $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ are actually independent of whether η is known or estimated in the construction of $\widehat{\delta}_\epsilon^{\mathrm{MMSE}}$ (see Lemma A3 below), but formally showing this requires some additional assumptions, which we present next.

For a given η, let $\mathcal{H}(\eta)$ be the set of functions $h = h(\cdot, \eta)$ that satisfy the constraints (2) and (4). The minimization problem (12) in the main text can then be rewritten as

$$\begin{aligned} Q_\epsilon^{\mathrm{MMSE}}(\eta) &:= \min_{h\in\mathcal{H}(\eta)} \left[b_\epsilon(h,\eta)^2 + \frac{\mathrm{Var}_{\theta(\eta)}(h(Y,\eta))}{n} \right] \\ &= b_\epsilon(h_\epsilon^{\mathrm{MMSE}}, \eta)^2 + \frac{\mathrm{Var}_{\theta(\eta)}(h_\epsilon^{\mathrm{MMSE}}(Y,\eta))}{n}. \end{aligned} \tag{A1}$$

The optimal $h_\epsilon^{\mathrm{MMSE}}(\cdot, \eta) \in \mathcal{H}(\eta)$ can be expressed as

$$h_\epsilon^{\mathrm{MMSE}}(y, \eta) = \langle v_\epsilon^{\mathrm{MMSE}}(\eta), \nabla_\theta \log f_{\theta(\eta)}(y) \rangle, \tag{A2}$$

with

$$v_\epsilon^{\text{MMSE}}(\eta) := G_\eta' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} + \left[\mathbb{I} - G_\eta' H_\eta^{-1} G_\eta H_{\theta(\eta)}\right] \left[\widetilde{H}_{\theta(\eta)} + (\epsilon n)^{-1} \Omega_\eta\right]^{-1} \widetilde{\nabla}_\theta \delta_{\theta(\eta)}, \quad \text{(A3)}$$

where $v_\epsilon^{\text{MMSE}}(\eta) \in T(\Theta)$, and $\mathbb{I}$ denotes the identity operator on $T(\Theta)$. It is easy to verify that $h_\epsilon^{\text{MMSE}}(y, \eta)$ in (A2) indeed satisfies the first-order conditions of problem (12).

Assumption A2. *We assume that*

(i) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left(\mathbb{E}_{\theta_0} \|\widehat{\eta} - \eta\|^4\right)^{1/4} = o(n^{-1/4})$.

(ii) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} \left\|\nabla_\eta h_\epsilon^{\text{MMSE}}(Y, \eta)\right\|^2 = O(1)$.

(iii) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} \sup_{\tilde{\eta}\in B(\eta,r_\epsilon)} \left\|\frac{1}{n}\sum_{i=1}^n \nabla_{\eta\eta'} h_\epsilon^{\text{MMSE}}(Y_i, \tilde{\eta})\right\|^2 = O(1)$, *for a Euclidean ball* $B(\eta, r_\epsilon)$ *around* η *with radius* $r_\epsilon = o(1)$.

Part (i) of Assumption A2 requires $\widehat{\eta}$ to converge at a rate faster than $n^{1/4}$, although in most applications we actually expect it to converge at rate $n^{1/2}$.[20] Part (ii) of Assumption A2 requires a uniformly bounded second moment for $\nabla_\eta h_\epsilon^{\text{MMSE}}(y, \eta)$. Since (A2) and (A3) give an explicit expression for $h_\epsilon^{\text{MMSE}}(y, \eta)$, we could replace Assumption A2(ii) by appropriate assumptions on the model primitives $f_{\theta_0}(y)$, δ_{θ_0} and Ω_η, but for the sake of brevity we state the assumption in terms of $h_\epsilon^{\text{MMSE}}(y, \eta)$. The same is true for part (iii) of Assumption A2. Notice that this last part of the assumption involves a supremum over $\tilde{\eta}$ inside of an expectation – in order to verify it, one either requires a uniform Lipschitz bound on the dependence of $h_\epsilon^{\text{MMSE}}(Y_i, \eta)$ on η, or some empirical process method to control the entropy of that function (e.g., a bracketing argument). But since η is a finite-dimensional parameter these are all standard arguments.

Remark. We found that $h_\epsilon^{\text{MMSE}}(y, \eta)$ can be expressed in the form $\langle v(\eta), \nabla_\theta \log f_{\theta(\eta)}(y)\rangle$, thus automatically satisfying the constraint (2). By choosing $v(\eta) = G_\eta' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)} + \widetilde{v}(\eta)$, where $G_\eta H_{\theta(\eta)} \widetilde{v}(\eta) = 0$, the constraint (4) is also satisfied. Using this one can alternatively represent the worst-case MSE problem as

$$Q_\epsilon^{\text{MMSE}}(\eta) = \min_{\widetilde{v}\in T(\Theta)} \left[\epsilon \left\|\widetilde{\nabla}_\theta \delta_{\theta(\eta)} - \widetilde{H}_{\theta(\eta)} \widetilde{v}\right\|_\eta^2 + \frac{1}{n} \left\langle \widetilde{v}, \widetilde{H}_{\theta(\eta)} \widetilde{v} \right\rangle\right] + \frac{1}{n} \left(\nabla_\eta \delta_{\theta(\eta)}\right)' H_\eta^{-1} \nabla_\eta \delta_{\theta(\eta)}.$$

[20] By slightly modifying the proof of Lemma A3 below one could replace Assumption A2(i) by $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left(\mathbb{E}_{\theta_0} \|\widehat{\eta} - \eta\|^2\right)^{1/2} = o(n^{-1/2})$ – i.e., convergence in ℓ^2 only, but at a faster rate – although this would require slightly different versions of parts (ii) and (iii) of that assumption as well.

This concise expression for the leading order worst-case MSE highlights the terms of order ϵ (from squared bias) and of order $1/n$ (from variance terms). This representation also shows that instead of solving for the optimal influence function $h(y,\eta)$ we can alternatively solve for an optimal vector $\widetilde{v} \in T(\Theta)$, which is particularly convenient in models where the dimension of y exceeds that of θ.

A.1.2 Proof

In the following, as in the rest of the paper, we always implicitly assume that all functions of y are measurable, and that correspondingly all expectations and integrals over y are well-defined.

Lemma A1. *Let Assumption A1 and the conditions on $h_\epsilon(\cdot,\eta)$ in Theorem 1 hold. Then,*

(i) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left|\mathbb{E}_{\theta_0} h_\epsilon^2(Y,\eta) - \mathbb{E}_{\theta(\eta)} h_\epsilon^2(Y,\eta)\right| = o(1).$

(ii) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left|\mathbb{E}_{\theta_0} h_\epsilon(Y,\eta) - \mathbb{E}_{\theta(\eta)} h_\epsilon(Y,\eta) - \ \langle \theta_0 - \theta(\eta), \mathbb{E}_{\theta(\eta)} h_\epsilon(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y)\rangle\right| = o(\epsilon^{1/2}).$

Proof of Lemma A1. # Part (i): Without loss of generality we may assume that $\kappa \leq 4$, since if $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} |h_\epsilon(Y,\eta)|^\kappa = O(1)$ holds for $\kappa > 4$, then it also holds for $\kappa \leq 4$. Let $\xi = \kappa/(\kappa-2) \geq 2$. We then have

$$\int_{\mathcal{Y}} \left| f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right|^\xi dy \leq \int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) \right]^2 dy,$$

where we used that $|a-b| \leq |a^c - b^c|^{1/c}$, for any $a, b \geq 0$ and $c \geq 1$, and plugged in $a = f_{\theta_0}^{1/\xi}(y)$, $b = f_{\theta(\eta)}^{1/\xi}(y)$, and $c = \xi/2$. Thus, the first part of Assumption A1(iii) also implies

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left| f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right|^\xi dy = o(1). \tag{A4}$$

Next, we find

$$
\begin{aligned}
&\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left|\mathbb{E}_{\theta_0} h_\epsilon^2(Y,\eta) - \mathbb{E}_{\theta(\eta)} h_\epsilon^2(Y,\eta)\right| \\
&= \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left| \int_{\mathcal{Y}} h_\epsilon^2(y,\eta) \frac{f_{\theta_0}(y) - f_{\theta(\eta)}(y)}{f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y)} \left[f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right] dy \right| \\
&\leq \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} |h_\epsilon(y,\eta)|^{\frac{2\xi}{\xi-1}} \left| \frac{f_{\theta_0}(y) - f_{\theta(\eta)}(y)}{f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y)} \right|^{\frac{\xi}{\xi-1}} dy \right\}^{\frac{\xi-1}{\xi}} \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left| f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right|^{\xi} dy \right\}^{\frac{1}{\xi}} \\
&\leq \xi \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} |h_\epsilon(y,\eta)|^{\frac{2\xi}{\xi-1}} \left| f_{\theta_0}(y) + f_{\theta(\eta)}(y) \right| dy \right\}^{\frac{\xi-1}{\xi}} \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left| f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right|^{\xi} dy \right\}^{\frac{1}{\xi}} \\
&\leq \xi \left\{ 2 \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} |h_\epsilon(Y,\eta)|^{\kappa} \right\}^{\frac{\xi-1}{\xi}} \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left| f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y) \right|^{\xi} dy \right\}^{\frac{1}{\xi}} = o(1),
\end{aligned}
$$

where the first inequality is an application of Hölder's inequality, the second inequality uses that $\left| \frac{f_{\theta_0}(y) - f_{\theta(\eta)}(y)}{f_{\theta_0}^{1/\xi}(y) - f_{\theta(\eta)}^{1/\xi}(y)} \right|^{\xi/(\xi-1)} \leq \xi^{\xi/(\xi-1)} \left[f_{\theta_0}(y) + f_{\theta(\eta)}(y) \right]$,[21] the last line uses that $\kappa = 2\xi/(\xi-1)$, and the final conclusion follows from our assumptions and (A4).

Part (ii): We have

$$
\begin{aligned}
&\mathbb{E}_{\theta_0} h_\epsilon(Y,\eta) - \mathbb{E}_{\theta(\eta)} h_\epsilon(Y,\eta) - \left\langle \theta_0 - \theta(\eta), \mathbb{E}_{\theta(\eta)} h_\epsilon(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y) \right\rangle \\
&= \int_{\mathcal{Y}} h_\epsilon(y,\eta) \left[f_{\theta_0}(y) - f_{\theta(\eta)}(y) - \left\langle \theta_0 - \theta(\eta), \nabla_\theta \log f_{\theta(\eta)}(y) \right\rangle f_{\theta(\eta)}(y) \right] dy \\
&= \underbrace{\int_{\mathcal{Y}} h_\epsilon(y,\eta) \left[f_{\theta_0}^{1/2}(y) + f_{\theta(\eta)}^{1/2}(y) \right] \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) - \frac{1}{2} \left\langle \theta_0 - \theta(\eta), \nabla_\theta \log f_{\theta(\eta)}(y) \right\rangle f_{\theta(\eta)}^{1/2}(y) \right] dy}_{=: a_{\eta,\theta_0,q}^{(1)}} \\
&\quad + \frac{1}{2} \underbrace{\int_{\mathcal{Y}} h_\epsilon(y,\eta) f_{\theta(\eta)}^{1/2}(y) \left\langle \theta_0 - \theta(\eta), \nabla_\theta \log f_{\theta(\eta)}(y) \right\rangle \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) \right] dy}_{=: a_{\eta,\theta_0,q}^{(2)}}.
\end{aligned}
$$

Applying the Cauchy-Schwarz inequality and our assumptions we find that

$$
\begin{aligned}
&\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left| a_{\eta,\theta_0,q}^{(1)} \right|^2 \\
&\leq 4 \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} h_\epsilon^2(Y,\eta) \right\} \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) - \left\langle \theta_0 - \theta(\eta), \nabla_\theta f_{\theta(\eta)}^{1/2}(y) \right\rangle \right]^2 dy \right\} \\
&= o(\epsilon),
\end{aligned}
$$

[21] For $a, b \geq 0$ there exists $c \in [a,b]$ such that by the mean value theorem we have $(a^\xi - b^\xi)/(a-b) = \xi c^{\xi-1} \leq \xi \max(a^{\xi-1}, b^{\xi-1})$, and therefore $[(a^\xi - b^\xi)/(a-b)]^{\xi/(\xi-1)} \leq \xi^{\xi/(\xi-1)} \max(a^\xi, b^\xi) \leq \xi^{\xi/(\xi-1)}(a^\xi + b^\xi)$, which we apply here with $a = f_{\theta_0}^{1/\xi}(y)$ and $b = f_{\theta(\eta)}^{1/\xi}(y)$.

and

$$\begin{aligned}
&\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left|a^{(2)}_{\eta,\theta_0,q}\right|^2 \\
&\leq \left\{\mathbb{E}_{\theta(\eta)} h^2_\epsilon(Y,\eta)\right\} \left\{ \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \|\theta_0-\theta(\eta)\|^2_{\text{ind},\eta} \int_{\mathcal{Y}} \left\|\nabla_\theta \log f_{\theta(\eta)}(y)\right\|^2_\eta \left[f^{1/2}_{\theta_0}(y) - f^{1/2}_{\theta(\eta)}(y)\right]^2 dy \right\} \\
&= o(\epsilon).
\end{aligned}$$

Combining this gives the statement in the lemma. ∎

Let $\Delta_{\eta,\theta_0} := \delta_{\theta_0} - \delta_{\theta(\eta)}$. For a function $h = h(y,\eta)$ we define

$$\begin{aligned}
Q_\epsilon(h,\eta,\theta_0) &:= \mathbb{E}_{\theta_0} \left(\frac{1}{n}\sum_{i=1}^n h(Y_i,\eta) - \Delta_{\eta,\theta_0} \right)^2 \\
&= \left[\mathbb{E}_{\theta_0} h(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 + \frac{1}{n}\text{Var}_{\theta_0}\left[h(Y,\eta) - \Delta_{\eta,\theta_0}\right] \\
&= \frac{n-1}{n}\left[\mathbb{E}_{\theta_0} h(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 + \frac{1}{n}\mathbb{E}_{\theta_0}\left[h(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2. \qquad \text{(A5)}
\end{aligned}$$

Also, recall the definition of the worst-case bias in (8) of the main text:

$$b_\epsilon(h,\eta) = \epsilon^{\frac{1}{2}} \left\|\nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)}\, h(Y,\eta)\, \nabla_\theta \log f_{\theta(\eta)}(Y)\right\|_\eta.$$

Lemma A2. *Let Assumption A1 and the conditions on $h_\epsilon(\cdot,\eta)$ in Theorem 1 hold. Then,*

$$\sup_{\eta\in\mathcal{B}} \left| \sup_{\theta_0\in\Gamma_\epsilon(\eta)} Q_\epsilon(h_\epsilon,\eta,\theta_0) - b_\epsilon(h_\epsilon,\eta)^2 - \frac{\text{Var}_{\theta(\eta)}(h_\epsilon(Y,\eta))}{n} \right| = o(\epsilon).$$

Proof of Lemma A2. Using the Cauchy-Schwarz inequality and our assumptions we find that

$$\begin{aligned}
&\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left| \left\langle \theta_0 - \theta(\eta), \mathbb{E}_{\theta(\eta)} h_\epsilon(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y) \right\rangle \right| \\
&\leq \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left\{ \|\theta_0-\theta(\eta)\|_{\text{ind},\eta} \left[\mathbb{E}_{\theta(\eta)} h^2_\epsilon(Y,\eta)\right]^{1/2} \left[\mathbb{E}_{\theta(\eta)} \left\|\nabla_\theta \log f_{\theta(\eta)}(Y)\right\|^2_\eta\right]^{1/2} \right\} = o(1), \qquad \text{(A6)}
\end{aligned}$$

and similarly

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \left| \left\langle \theta_0 - \theta(\eta), \nabla_\theta \delta_{\theta(\eta)} \right\rangle \right| = o(1). \qquad \text{(A7)}$$

Lemma A1(ii) and (A6) imply that $\mathbb{E}_{\theta_0} h_\epsilon(Y,\eta) = \mathbb{E}_{\theta(\eta)} h_\epsilon(Y,\eta) + o(1)$, uniformly in $(\theta_0,\eta) \in \Gamma_\epsilon$. In turn, Assumption A1(ii) guarantees that $\Delta_{\eta,\theta_0} = o(1)$, uniformly in $(\theta_0,\eta) \in \Gamma_\epsilon$. Combining with Lemma A1(i) we thus obtain

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[h_\epsilon(Y,\eta) - \Delta_{\eta,\theta_0}\right]^2 &= \mathbb{E}_{\theta_0}\left[h_\epsilon(Y,\eta)\right]^2 - 2\,\Delta_{\eta,\theta_0}\,\mathbb{E}_{\theta_0} h_\epsilon(Y,\eta) + \Delta^2_{\eta,\theta_0} \\
&= \mathbb{E}_{\theta(\eta)}\left[h_\epsilon(Y,\eta)\right]^2 + o(1) = \text{Var}_{\theta(\eta)}(h_\epsilon(Y,\eta)) + o(1),
\end{aligned}$$

uniformly in $(\theta_0, \eta) \in \Gamma_\epsilon$, where in the last step we have also used that $h_\epsilon(y, \eta)$ satisfies the unbiasedness constraint (2). Using that constraint again, as well as Lemma A1(ii) and Assumptions A1(ii) and A1(iv) we find

$$\begin{aligned}
&\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} |\mathbb{E}_{\theta_0} h_\epsilon(Y, \eta) - \Delta_{\eta, \theta_0}| \\
&\quad = \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \left| \left\langle \theta_0 - \theta(\eta), \mathbb{E}_{\theta(\eta)} h_\epsilon(Y, \eta) \nabla_\theta \log f_{\theta(\eta)}(Y) - \nabla_\theta \delta_{\theta(\eta)} \right\rangle \right| + o(\epsilon^{1/2}) \\
&\quad = \epsilon^{1/2} \left\| \mathbb{E}_{\theta(\eta)} h_\epsilon(Y, \eta) \nabla_\theta \log f_{\theta(\eta)}(Y) - \nabla_\theta \delta_{\theta(\eta)} \right\|_\eta + o(\epsilon^{1/2}) = b_\epsilon(h_\epsilon, \eta) + o(\epsilon^{1/2}),
\end{aligned}$$

uniformly in $\eta \in \mathcal{B}$. The results in the previous two displays together with the last expression for $Q_\epsilon(h, \eta, \theta_0)$ in equation (A5) yield the statement of the lemma. ∎

Proposition A1. *Let Assumption A1 and the conditions on $h_\epsilon(\cdot, \eta)$ in Theorem 1 hold. Then,*

$$\sup_{\eta \in \mathcal{B}} \left\{ Q_\epsilon^{\text{MMSE}}(\eta) - \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_\epsilon - \delta_{\theta_0} \right)^2 \right] \right\} \leq o(\epsilon).$$

Proof of Proposition A1. Using (20), the definition of $Q_\epsilon(h, \eta, \theta_0)$, and also applying Lemma A2, we find that

$$\begin{aligned}
&\sup_{\eta \in \mathcal{B}} \left\{ Q_\epsilon^{\text{MMSE}}(\eta) - \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\widehat{\delta}_\epsilon - \delta_{\theta_0} \right)^2 \right] \right\} \\
&= \sup_{\eta \in \mathcal{B}} \left\{ Q_\epsilon^{\text{MMSE}}(\eta) - \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0} \left[\left(\frac{1}{n} \sum_{i=1}^n h_\epsilon(Y_i, \eta) + \delta_{\theta(\eta)} - \delta_{\theta_0} \right)^2 \right] \right\} + o(\epsilon) \\
&= \sup_{\eta \in \mathcal{B}} \left[Q_\epsilon^{\text{MMSE}}(\eta) - \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} Q_\epsilon(h_\epsilon, \eta, \theta_0) \right] + o(\epsilon) \\
&= \sup_{\eta \in \mathcal{B}} \left[Q_\epsilon^{\text{MMSE}}(\eta) - b_\epsilon(h_\epsilon, \eta)^2 - \frac{\text{Var}_{\theta(\eta)}(h_\epsilon(Y, \eta))}{n} \right] + o(\epsilon).
\end{aligned}$$

Moreover, by the definition of $Q_\epsilon^{\text{MMSE}}(\eta)$ in (A1) we have

$$Q_\epsilon^{\text{MMSE}}(\eta) \leq b_\epsilon(h_\epsilon, \eta)^2 + \frac{\text{Var}_{\theta(\eta)}(h_\epsilon(Y, \eta))}{n}.$$

Combining the last two displays gives the statement of the proposition. ∎

Recall that $\widehat{\delta}_\epsilon^{\,\text{MMSE}} = \delta_{\theta(\widehat{\eta})} + \frac{1}{n} \sum_{i=1}^n h_\epsilon^{\text{MMSE}}(Y_i, \widehat{\eta})$. The following lemma shows that the fact that η is being estimated in the construction of $\widehat{\delta}_\epsilon^{\,\text{MMSE}}$ can be neglected to first order. Notice that this result requires the additional regularity conditions in Assumption A2, which were not required for any of the previous results.

Lemma A3. *Under Assumptions A1 and A2 we have*

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0}\left[\widehat{\delta}_\epsilon^{\,\mathrm{MMSE}} - \delta_{\theta(\eta)} - \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta)\right]^2 = o(\epsilon).$$

Proof of Lemma A3. By a Taylor expansion in η we find that

$$\begin{aligned}
\widehat{\delta}_\epsilon^{\,\mathrm{MMSE}} &= \delta_{\theta(\widehat{\eta})} + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i,\widehat{\eta}) \\
&= \delta_{\theta(\eta)} + \frac{1}{n}\sum_{i=1}^n h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta) + \underbrace{(\widehat{\eta}-\eta)'\left[\nabla_\eta \delta_{\theta(\eta)} + \mathbb{E}_{\theta(\eta)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\right]}_{=r^{(1)}_{\eta,\theta_0}} \\
&\quad + \underbrace{(\widehat{\eta}-\eta)'\frac{1}{n}\sum_{i=1}^n\left[\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta) - \mathbb{E}_{\theta_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta)\right]}_{=r^{(2)}_{\eta,\theta_0}} \\
&\quad + \underbrace{(\widehat{\eta}-\eta)'\left[\mathbb{E}_{\theta_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta) - \mathbb{E}_{\theta(\eta)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\right]}_{=r^{(3)}_{\eta,\theta_0}} \\
&\quad + \underbrace{\frac{1}{2}(\widehat{\eta}-\eta)'\left[\frac{1}{n}\sum_{i=1}^n \nabla_{\eta\eta'} h_\epsilon^{\mathrm{MMSE}}(Y_i,\widetilde{\eta})\right](\widehat{\eta}-\eta)}_{=r^{(4)}_{\eta,\theta_0}},
\end{aligned} \tag{A8}$$

where $\widetilde{\eta}$ is a value between $\widehat{\eta}$ and η. Our constraints (2) and (4) guarantee that $\nabla_\eta \delta_{\theta(\eta)} + \mathbb{E}_{\theta(\eta)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta) = 0$, that is, we have $r^{(1)}_{\eta,\theta_0} = 0$. Using Assumption A2 we furthermore find

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left|r^{(2)}_{\eta,\theta_0}\right|^2 &\le \mathbb{E}_{\theta_0}\|\widehat{\eta}-\eta\|^2 \;\mathbb{E}_{\theta_0}\left\|\frac{1}{n}\sum_{i=1}^n\left[\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta) - \mathbb{E}_{\theta_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y_i,\eta)\right]\right\|^2 \\
&\le \mathbb{E}_{\theta_0}\|\widehat{\eta}-\eta\|^2 \;\frac{1}{n}\,\mathbb{E}_{\theta_0}\left\|\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\right\|^2 = o(n^{-1/2})O(n^{-1}) = o(\epsilon^{3/2}) = o(\epsilon),
\end{aligned}$$

uniformly in $(\theta_0,\eta)\in\Gamma_\epsilon$, where in the second step we have used the independence of Y_i across i. Similarly, we have

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left|r^{(3)}_{\eta,\theta_0}\right|^2 &\le \mathbb{E}_{\theta_0}\|\widehat{\eta}-\eta\|^2\left\|\mathbb{E}_{\theta_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta) - \mathbb{E}_{\theta(\eta)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\right\|^2 \\
&= o(n^{-1/2})O(\epsilon) = o(\epsilon^{3/2}) = o(\epsilon),
\end{aligned}$$

uniformly in $(\theta_0,\eta)\in\Gamma_\epsilon$, where we have used that

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\left\|\mathbb{E}_{\theta_0}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta) - \mathbb{E}_{\theta(\eta)}\nabla_\eta h_\epsilon^{\mathrm{MMSE}}(Y,\eta)\right\| = O(\epsilon^{1/2}),$$

which follows from Assumptions A1(iii) and A2(ii) by using the proof strategy of part (ii) of Lemma A1. Finally, we have

$$\mathbb{E}_{\theta_0}\left|r^{(4)}_{\eta,\theta_0}\right|^2 \leq \mathbb{E}_{\theta_0}\left\|\widehat{\eta}-\eta\right\|^4 \mathbb{E}_{\theta_0}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{\eta\eta'}h^{\mathrm{MMSE}}_{\epsilon}(Y_i,\widetilde{\eta})\right\|^2 = o(n^{-1}) = o(\epsilon),$$

uniformly in $(\theta_0,\eta) \in \Gamma_\epsilon$, where we have used Assumption A2(iii). We have thus shown that

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\mathbb{E}_{\theta_0}\left|r^{(1)}_{\eta,\theta_0}+r^{(2)}_{\eta,\theta_0}+r^{(3)}_{\eta,\theta_0}+\frac{1}{2}r^{(4)}_{\eta,\theta_0}\right|^2 = o(\epsilon),$$

which together with (A8) gives the statement of the lemma. ∎

Proposition A2. *Under Assumptions A1 and A2 we have*

$$\sup_{\eta\in\mathcal{B}}\left|\sup_{\theta_0\in\Gamma_\epsilon(\eta)}\mathbb{E}_{\theta_0}\left[\left(\widehat{\delta}^{\mathrm{MMSE}}_{\epsilon}-\delta_{\theta_0}\right)^2\right]-Q^{\mathrm{MMSE}}_{\epsilon}(\eta)\right| = o(\epsilon).$$

Proof of Proposition A2. Applying Lemma A3 together with the definition of $Q_\epsilon(h,\eta,\theta_0)$ in (A5) we obtain

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\left\{\mathbb{E}_{\theta_0}\left[\left(\widehat{\delta}^{\mathrm{MMSE}}_{\epsilon}-\delta_{\theta_0}\right)^2\right]-Q_\epsilon(h^{\mathrm{MMSE}}_{\epsilon},\eta,\theta_0)\right\} = o(\epsilon).$$

Assumptions A1(i) and A1(v) imply that $\sup_{\eta\in\mathcal{B}}\left\|v^{\mathrm{MMSE}}_{\epsilon}(\eta)\right\|_{\mathrm{ind},\eta} = O(1)$.[22] From the explicit solution for $h^{\mathrm{MMSE}}_{\epsilon}(y,\eta)$ in (A2) and (A3) together with Assumption A2 we conclude

$$\begin{aligned}\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\mathbb{E}_{\theta_0}\left[h^{\mathrm{MMSE}}_{\epsilon}(Y,\eta)\right]^{2+\nu} &= \sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\mathbb{E}_{\theta_0}\langle v^{\mathrm{MMSE}}_{\epsilon}(\eta),\nabla_\theta\log f_{\theta(\eta)}(y)\rangle^{2+\nu}\\ &\leq \sup_{\eta\in\mathcal{B}}\left\|v^{\mathrm{MMSE}}_{\epsilon}(\eta)\right\|^{2+\nu}_{\mathrm{ind},\eta}\sup_{(\theta_0,\eta)\in\Gamma_\epsilon}\mathbb{E}_{\theta_0}\left\|\nabla_\theta\log f_{\theta_0}(Y)\right\|^{2+\nu}_{\eta} = O(1).\end{aligned}$$

Thus, $h^{\mathrm{MMSE}}_{\epsilon}(y,\eta)$ satisfies the regularity conditions for $h_\epsilon(y,\eta)$ in Theorem 1 with $\kappa = 2+\nu$. We can therefore apply Lemma A2 with $h_\epsilon(y,\eta) = h^{\mathrm{MMSE}}_{\epsilon}(y,\eta)$ to find

$$\sup_{\eta\in\mathcal{B}}\left|\sup_{\theta_0\in\Gamma_\epsilon(\eta)}Q_\epsilon(h^{\mathrm{MMSE}}_{\epsilon},\eta,\theta_0)-Q^{\mathrm{MMSE}}_{\epsilon}(\eta)\right| = o(\epsilon).$$

Combining the last two displays gives the statement of the proposition. ∎

Proof of Theorem 1. Combining Propositions A1 and A2 gives the the statement of the theorem. ∎

[22] Notice that $\sup_{\eta\in\mathcal{B}}\left\|H_{\theta(\eta)}\right\|_{\eta} = O(1)$ follows from the bounded moment condition on the score $\nabla_\theta\log f_{\theta(\eta)}(y)$ in part (v) of Assumption A1.

A.2 Proof of Corollary 1

Let $q(\eta)$ denote the MSE difference in the curly brackets in (21). Corollary 1 then immediately follows from Theorem 1 and $\int_{\mathcal{B}} q(\eta)w(\eta)d\eta \leq \left[\int_{\mathcal{B}} w(\eta)d\eta\right]\left[\sup_{\eta\in\mathcal{B}} q(\eta)\right]$.

A.3 Proof of Theorem 2

Assumption A3.

(i) *We consider* $n \to \infty$ *and* $\epsilon \to 0$ *such that* $\epsilon n \to c$, *for some constant* $c \in (0, \infty)$.

(ii) $\widehat{\delta} - \delta_{\theta(\eta)} - \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \eta) = o_{P_{\theta_0}}(n^{-\frac{1}{2}})$, *uniformly in* $(\theta_0, \eta) \in \Gamma_\epsilon$.

(iii) *Let* $\sigma_h^2(\theta_0, \eta) = \mathrm{Var}_{\theta_0} h(Y, \eta)$. *We assume that there exists a constant* c, *independent of* ϵ, *such that* $\inf_{(\theta_0,\eta)\in\Gamma_\epsilon} \sigma_h(\theta_0, \eta) \geq c > 0$. *Furthermore, for all sequences* $a_n = c_{1-\mu/2} + o(1)$ *we have*

$$\inf_{(\theta_0,\eta)\in\Gamma_\epsilon} \Pr_{\theta_0}\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{h(Y_i, \eta) - \mathbb{E}_{\theta_0} h(Y, \eta)}{\sigma_h(\theta_0, \eta)}\right| \leq a_n\right] \geq 1 - \mu + o(1).$$

(iv) $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0}\|\widehat{\eta} - \eta\|^2 = o(1)$, $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0}[\widehat{\sigma}_h - \sigma_h(\theta_0, \eta)]^2 = o(1)$.

(v) $\sup_{\eta\in\mathcal{B}} \|\nabla_\eta b_\epsilon(h, \eta)\| = O(\epsilon^{\frac{1}{2}})$.

Part (ii) is weaker than the local regularity of the estimator $\widehat{\delta}$ that we assumed when analyzing the minimum-MSE estimator, see equation (20). In turn, related to but differently from the conditions we used for Theorem 1, part (iii) requires a form of local asymptotic normality of the estimator.

Proof of Theorem 2. Let $\widehat{\delta}$ be an estimator and $h(y, \eta)$ be the corresponding influence function such that part (ii) in Assumption A3 holds. Define $\widehat{R}_\eta := \widehat{\delta} - \delta_{\theta(\eta)} - \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \eta)$. We then have

$$\begin{aligned}\widehat{\delta} - \delta_{\theta_0} &= \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \eta) + \delta_{\theta(\eta)} - \delta_{\theta_0} + \widehat{R}_\eta \\ &= \frac{1}{n}\sum_{i=1}^{n} [h(Y_i, \eta) - \mathbb{E}_{\theta_0} h(Y, \eta)] - \left[\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y, \eta)\right] + \widehat{R}_\eta,\end{aligned}$$

and therefore

$$\underbrace{\frac{|\widehat{\delta} - \delta_{\theta_0}| - b_\epsilon(h, \widehat{\eta}) - \widehat{\sigma}_h\, c_{1-\mu/2}/\sqrt{n}}{\sigma_h(\theta_0, \eta)/\sqrt{n}}}_{=\text{lhs}} \leq \underbrace{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{h(Y_i, \eta) - \mathbb{E}_{\theta_0} h(Y, \eta)}{\sigma_h(\theta_0, \eta)}\right|}_{=\text{rhs}} - c_{1-\mu/2} + \widehat{r}_{\eta,\theta_0}, \quad \text{(A9)}$$

where

$$\widehat{r}_{\eta,\theta_0} := c_{1-\mu/2} + \frac{\left|\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y,\eta)\right| + \left|\widehat{R}_\eta\right| - b_\epsilon(h,\widehat{\eta}) - \widehat{\sigma}_h\, c_{1-\mu/2}/\sqrt{n}}{\sigma_h(\theta_0,\eta)/\sqrt{n}}$$
$$= \frac{\sqrt{n}}{\sigma_h(\theta_0,\eta)} \left\{ |\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y,\eta)| + |\widehat{R}_\eta| - b_\epsilon(h,\widehat{\eta}) - \frac{\widehat{\sigma}_h - \sigma_h(\theta_0,\eta)}{\sqrt{n}} c_{1-\mu/2} \right\}$$

From (A9) we conclude that the event rhs ≤ 0 implies the event lhs ≤ 0, and therefore $\Pr_{\theta_0}(\text{lhs} \leq 0) \geq \Pr_{\theta_0}(\text{rhs} \leq 0)$, which we can also write as

$$\Pr_{\theta_0}\left[|\widehat{\delta} - \delta_{\theta_0}| \leq b_\epsilon(h,\widehat{\eta}) + \frac{\widehat{\sigma}_h}{\sqrt{n}} c_{1-\mu/2} \right]$$
$$\geq \Pr_{\theta_0}\left[\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{h(Y_i,\eta) - \mathbb{E}_{\theta_0} h(Y,\eta)}{\sigma_h(\theta_0,\eta)} \right| \leq c_{1-\mu/2} - \widehat{r}_{\eta,\theta_0} \right] \tag{A10}$$

By part (v) in Assumption A3 there exists a constant $C > 0$ such that $\sup_{\eta\in\mathcal{B}} \|\nabla_\eta b_\epsilon(h,\eta)\| \leq C\epsilon^{\frac{1}{2}}$, and therefore

$$\sup_{\eta\in\mathcal{B}} |b_\epsilon(h,\widehat{\eta}) - b_\epsilon(h,\eta)| \leq C\,\epsilon^{\frac{1}{2}}\, \|\widehat{\eta} - \eta\|.$$

Using this we find that

$$|\widehat{r}_{\eta,\theta_0}| \leq \frac{\sqrt{n}}{\sigma_h(\theta_0,\eta)} \left\{ \left| |\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y,\eta)| - b_\epsilon(h,\eta) \right| \right.$$
$$\left. + \frac{|\widehat{\sigma}_h - \sigma_h(\theta_0,\eta)|}{\sqrt{n}} c_{1-\mu/2} + C\,\epsilon^{\frac{1}{2}}\, \|\widehat{\eta} - \eta\| + |\widehat{R}_\eta| \right\}.$$

Parts (ii) and (iii) in Assumption A3 imply that, uniformly in $(\theta_0,\eta) \in \Gamma_\epsilon$, we have

$$\frac{\sqrt{n}}{\sigma_h(\theta_0,\eta)} \widehat{R}_\eta = o_{P_{\theta_0}}(1),$$

and analogously we find from the conditions in Assumption A3 that

$$\frac{\widehat{\sigma}_h - \sigma_h(\theta_0,\eta)}{\sigma_h(\theta_0,\eta)} = o_{P_{\theta_0}}(1), \qquad \frac{\sqrt{n}}{\sigma_h(\theta_0,\eta)} \epsilon^{\frac{1}{2}}\, \|\widehat{\eta} - \eta\| = o_{P_{\theta_0}}(1),$$

uniformly in $(\theta_0,\eta) \in \Gamma_\epsilon$. Finally, since we also impose Assumption A1 and $\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \mathbb{E}_{\theta_0} h^2(Y,\eta) = O(1)$ we obtain, analogously to the proof of Lemma A1(ii) above, that[23]

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \frac{\sqrt{n}}{\sigma_h(\theta_0,\eta)} \left| |\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y,\eta)| - b_\epsilon(h,\eta) \right| = o(1).$$

We thus conclude that $\widehat{r}_{\eta,\theta_0} = o_{P_{\theta_0}}(1)$, uniformly in $(\theta_0,\eta) \in \Gamma_\epsilon$. Together with (A10) and part (iii) in Assumption A3 this implies (23), hence Theorem 2. ∎

[23] Notice that the proof of part (ii) of Lemma A1 only requires a bounded second moment of $h(y,\eta)$.

A.4 Asymptotically linear estimators

In this subsection we provide some background on the asymptotically linear representation (1), and we give several examples. See, e.g., Bickel *et al.* (1993) and Rieder (1994) on local asymptotic expansions of regular estimators.

Consider an asymptotically linear estimator $\widehat{\delta}$ which has the following representation under f_{θ_0}, for $\theta_0 \in \Theta$,

$$\widehat{\delta} = \delta^*_{\theta_0} + \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, \theta_0) + o_{P_{\theta_0}}(n^{-\frac{1}{2}}), \tag{A11}$$

where $\delta^*_{\theta_0}$ is the probability limit of $\widehat{\delta}$ under f_{θ_0}, and $\phi(y, \theta_0)$ is its influence function. The pseudo-true value $\delta^*_{\theta_0}$ generally differs from the true parameter value δ_{θ_0}. The influence function is assumed to satisfy

$$\mathbb{E}_{\theta_0}\phi(Y, \theta_0) = 0, \quad \nabla_\theta \delta^*_{\theta_0} + \mathbb{E}_{\theta_0}\nabla_\theta \phi(Y, \theta_0) = 0, \quad \text{for all } \theta_0 \in \Theta. \tag{A12}$$

The first condition in (A12) requires that the estimator be asymptotically unbiased for the pseudo-true value $\delta^*_{\theta_0}$. The second condition is a version of the generalized information identity.[24] Expansion (A11) and conditions (A12) are satisfied for a large class of estimators, see below for examples.

Furthermore, suppose that

$$\delta^*_{\theta(\eta)} = \delta_{\theta(\eta)}, \quad \text{for all } \eta \in \mathcal{B}. \tag{A13}$$

Condition (A13) requires that $\widehat{\delta}$ be asymptotically unbiased for $\delta_{\theta(\eta)}$ under $f_{\theta(\eta)}$, that is, under correct specification of the reference model. Note that, under mild regularity conditions, the function

$$h(y, \eta) = \phi(y, \theta(\eta))$$

will then be automatically "locally robust" with respect to η, as defined in Chernozhukov *et al.* (2016). Indeed,

$$\begin{aligned}\mathbb{E}_{\theta(\eta)}\nabla_\eta h(Y, \eta) &= \mathbb{E}_{\theta(\eta)}\nabla_\eta \phi(y, \theta(\eta)) = \nabla_\eta \theta(\eta)\, \mathbb{E}_{\theta(\eta)}\nabla_\theta \phi(y, \theta(\eta)) \\ &= -\nabla_\eta \theta(\eta)\nabla_\theta \delta^*_{\theta(\eta)} = -\nabla_\eta \delta^*_{\theta(\eta)} = -\nabla_\eta \delta_{\theta(\eta)},\end{aligned}$$

where we have used (A12) at $\theta_0 = \theta(\eta)$, and that, by (A13), $\nabla_\eta \delta^*_{\theta(\eta)} = \nabla_\eta \delta_{\theta(\eta)}$.

[24] The generalized information identity can alternatively be written in terms of the influence function and the score of the model (or any parametric sub-model in semi-parametric settings); see, e.g., Newey (1990).

To relate (1), which is taken around $\delta_{\theta(\eta)}$, to expansion (A11), which is taken around $\delta^*_{\theta_0}$, note that by an expansion around $\theta(\eta)$, and making use of the second identity in (A12), (A11) will imply (1) provided $\frac{1}{n}\sum_{i=1}^n \nabla_\theta \phi(Y_i, \widetilde{\theta}) - \mathbb{E}_{\widetilde{\theta}} \nabla_\theta \phi(Y, \widetilde{\theta})$ is $o_{P_{\theta_0}}(1)$, uniformly in $\eta \in \mathcal{B}$, $\theta_0 \in \Gamma_\epsilon(\eta)$, $\widetilde{\theta} \in \Gamma_\epsilon(\eta)$.

Examples. As a first example, consider an estimator $\widehat{\delta}$ solving $\sum_{i=1}^n m(Y_i, \widehat{\delta}) = 0$, where m is a smooth scalar moment function. The pseudo-true value solves $\mathbb{E}_{\theta_0} m(Y, \delta^*_{\theta_0}) = 0$ for all $\theta_0 \in \Theta$. Expanding the moment condition around $\delta^*_{\theta_0}$ implies that (A11) holds under mild conditions on m, with

$$\phi(y, \theta_0) = \left[-\mathbb{E}_{\theta_0} \nabla_\delta m(Y, \delta^*_{\theta_0})\right]^{-1} m(y, \delta^*_{\theta_0}).$$

It is easy to see that (A12) is satisfied. Moreover, (A13) is satisfied when the moment restriction is satisfied under the reference model; that is, whenever $\mathbb{E}_{\theta(\eta)} m(Y, \delta_{\theta(\eta)}) = 0$ for all $\eta \in \mathcal{B}$.

As a second example, consider an estimator $\widehat{\delta}$ solving $\sum_{i=1}^n m(Y_i, \widehat{\delta}, \widehat{\eta}) = 0$, where $\widehat{\eta}$ is a preliminary estimator which solves $\sum_{i=1}^n q(Y_i, \widehat{\eta}) = 0$, for smooth moment functions m (scalar) and q (vector-valued). In this case (A11) holds under regularity conditions on m and q, with

$$\begin{aligned}\phi(y, \theta_0) =& [\mathbb{E}_{\theta_0}(-\nabla_\delta m(Y, \delta^*_{\theta_0}, \eta^*_{\theta_0}))]^{-1} \quad m(y, \delta^*_{\theta_0}, \eta^*_{\theta_0}) \\ &+ \mathbb{E}_{\theta_0}(\nabla_\eta m(Y, \delta^*_{\theta_0}, \eta^*_{\theta_0}))' [\mathbb{E}_{\theta_0}(-\nabla_\eta q(Y, \eta^*_{\theta_0}))]^{-1} q(y, \eta^*_{\theta_0}) \Big),\end{aligned}$$

where $\eta^*_{\theta_0}$ and $\delta^*_{\theta_0}$ satisfy $\mathbb{E}_{\theta_0} q(Y, \eta^*_{\theta_0}) = 0$ and $\mathbb{E}_{\theta_0} m(Y, \delta^*_{\theta_0}, \eta^*_{\theta_0}) = 0$ for all $\theta_0 \in \Theta$. It can be verified that (A12) holds. Moreover, (A13) holds provided the moment restrictions for η and $\delta_{\theta(\eta)}$ are satisfied under the reference model, that is, whenever $\mathbb{E}_{\theta(\eta)} q(Y, \eta) = 0$ and $\mathbb{E}_{\theta(\eta)} m(Y, \delta_{\theta(\eta)}, \eta) = 0$ for all $\eta \in \mathcal{B}$.

As a third example, consider the (non-random) estimator $\widehat{\delta} = \delta_{\theta(\eta)}$, where η is a known, fixed parameter (i.e., $\mathcal{B} = \{\eta\}$). In this case $\phi(y, \theta_0) = \delta_{\theta(\eta)} - \delta^*_{\theta_0} = 0$. It follows that both (A12) and (A13) hold.

As a last example, consider the estimator $\widehat{\delta} = \delta_{\theta(\widehat{\eta})}$, where as above the preliminary estimator $\widehat{\eta}$ solves $\sum_{i=1}^n q(Y_i, \widehat{\eta}) = 0$. In this case (A11) will hold, with

$$\phi(y, \theta_0) = (\nabla_\eta \delta_{\theta(\eta^*_{\theta_0})})' \left[\mathbb{E}_{\theta_0}(-\nabla_\eta q(Y, \eta^*_{\theta_0}))\right]^{-1} q(y, \eta^*_{\theta_0}),$$

where $\eta^*_{\theta_0}$ solves $\mathbb{E}_{\theta_0} q(Y, \eta^*_{\theta_0}) = 0$. It is easy to see that (A12) is satisfied. Moreover, (A13) holds provided $\mathbb{E}_{\theta(\eta)} q(Y, \eta) = 0$ for all $\eta \in \mathcal{B}$.

B Semi-parametric models

In this section of the appendix we provide results and additional examples for the semi-parametric setting of Section 4.

B.1 Dual of the Kullback-Leibler divergence

Let A be a random variable with domain $\mathcal{A}$, reference distribution $f_*(a)$ and "true" distribution $f_0(a)$. We use notation $f_*(a)$ and $f_0(a)$ as if those were densities, but point masses are also allowed. Twice the Kullback-Leibler (KL) divergence reads

$$d(f_0, f_*) = -\,2\,\mathbb{E}_0 \log \frac{f_*(A)}{f_0(A)},$$

where $\mathbb{E}_0$ is the expectation under f_0. Let $\mathcal{F}$ be the set of all distributions, in particular, $f \in \mathcal{F}$ implies $\int_{\mathcal{A}} f(a)da = 1$. Let $q : \mathcal{A} \to \mathbb{R}$ be a real valued function. For given $f_* \in \mathcal{F}$ and $\epsilon > 0$ we define

$$\|q\|_{*,\epsilon} := \max_{\{f_0 \in \mathcal{F} \,:\, d(f_0, f_*) \le \epsilon\}} \frac{\mathbb{E}_0\, q(A) - \mathbb{E}_*\, q(A)}{\sqrt{\epsilon}},$$

where $\mathbb{E}_*$ is the expectation under f_*.

We have the following result.

Lemma B4. *For $q : \mathcal{A} \to \mathbb{R}$ and $f_* \in \mathcal{F}$ we assume that the moment-generating function $m_*(t) = \mathbb{E}_* \exp(t\, q(A))$ exists for $t \in (\delta_-, \delta_+)$ and some $\delta_- < 0$ and $\delta_+ > 0$.[25] For $\epsilon \in (0, \delta_+^2)$ we then have*

$$\|q\|_{*,\epsilon} = \sqrt{\mathrm{Var}_*(q(A))} + O(\epsilon^{\frac{1}{2}}).$$

Proof. Let the cumulant-generating function of the random variable $q(A)$ under the reference measure f_* be $k_*(t) = \log m_*(t)$. We assume existence of $m_*(t)$ and $k_*(t)$ for $t \in (\delta_-, \delta_+)$. This also implies that all derivatives of $m_*(t)$ and $k_*(t)$ exist in this interval. We denote the p-th derivative of $m_*(t)$ by $m_*^{(p)}(t)$, and analogously for $k_*(t)$.

[25] Existence of $m_*(t)$ in an open interval around zero is equivalent to having an exponential decay of the tails of the distribution of the random variable $Q = q(A)$. If $q(a)$ is bounded, then $m_*(t)$ exists for all $t \in \mathbb{R}$.

In the following we denote the maximizing f_0 in the definition of $\|q\|_{*,\epsilon}$ simply by f_0. Applying standard optimization method (Karush-Kuhn-Tucker) we find the well-known exponential tilting result

$$f_0(a) = c\, f_*(a)\, \exp(t\, q(a)),$$

where the constants $c, t \in (0, \infty)$ are determined by the constraints $\int f_0(a)da = 1$ and $d(f_0, f_*) = \epsilon$. Using the constraint $\int f_0(a)da = 1$ we can solve for c to obtain

$$f_0(a) = \frac{f_*(a)\exp(t\, q(a))}{\mathbb{E}_* \exp(t\, q(A))} = \frac{f_*(a)\, \exp(t\, q(a))}{m_*(t)}.$$

Using this we find that

$$\begin{aligned} d(t) &:= d(f_0, f_*) \\ &= 2\, \mathbb{E}_* \frac{f_0(A)}{f_*(A)} \log \frac{f_0(A)}{f_*(A)} \\ &= \frac{2\, t}{m_*(t)}\, \mathbb{E}_* \exp(t\, q(A)) q(A) - \frac{2 \log m_*(t)}{m_*(t)}\, \mathbb{E}_* \exp(t\, q(A)) \\ &= \frac{2\, t\, m_*^{(1)}(t)}{m_*(t)} - 2 \log m_*(t). \\ &= 2\left[t\, k_*^{(1)}(t) - k_*(t)\right] \end{aligned}$$

We have $d(0) = 0$, $d^{(1)}(0) = 0$, $d^{(2)}(0) = 2k_*^{(2)}(0) = 2\mathrm{Var}_*(q(A))$, $d^{(3)}(t) = 4k_*^{(3)}(t) + 2tk_*^{(4)}(t)$. A mean-value expansion thus gives

$$d(t) = \mathrm{Var}_*(q(A))t^2 + \frac{t^3}{6}\left[4\, k_*^{(3)}(\tilde{t}) + 2\, \tilde{t}\, k_*^{(4)}(\tilde{t})\right]$$

where $0 \le \tilde{t} \le t \le \delta_+$. The value t that satisfies the constraint $d(t) = \epsilon$ therefore satisfies

$$t = \frac{\epsilon^{\frac{1}{2}}}{\sqrt{\mathrm{Var}_*(q(A))}} + O(\epsilon).$$

Next, using that $\|q\|_{*,\epsilon} = \epsilon^{-\frac{1}{2}}\, \mathbb{E}_* \left[\left(\frac{f_0(A)}{f_*(A)} - 1\right) q(A)\right]$ we find

$$\|q\|_{*,\epsilon} = \epsilon^{-\frac{1}{2}} \left[k_*^{(1)}(t) - k_*^{(1)}(0)\right]$$

Again using that $k_*^{(2)}(0) = \mathrm{Var}_*(q(A))$ and applying a mean value expansion we obtain

$$\begin{aligned} \|q\|_{*,\epsilon} &= \epsilon^{-\frac{1}{2}} \left[t\, k_*^{(2)}(t) + \frac{1}{2}\, t^2\, k_*^{(3)}(\bar{t})\right] \\ &= \epsilon^{-\frac{1}{2}} \left[t\, \mathrm{Var}_*(q(A)) + \frac{1}{2}\, t^2\, k_*^{(3)}(\bar{t})\right] \\ &= \sqrt{\mathrm{Var}_*(q(A))} + O(\epsilon^{\frac{1}{2}}), \end{aligned}$$

where $\bar{t} \in [0, t]$. ∎

B.2 Mapping between the setup of Section 2 and the semi-parametric case

In order to link the formulas in Section 2 to the ones we derived in Section 4 for semi-parametric models, let us focus for simplicity on the case where η is known and $\theta = \pi$ is the density of A. In this case, elements v of the tangent space satisfy $\int_{\mathcal{A}} v(a)da = 0$, and the corresponding squared norm is $\|v\|^2_{\text{ind},\eta} = \int_{\mathcal{A}} \frac{v(a)^2}{\pi_\gamma(a)} da$. Hence Ω_η is such that, for any two elements of the tangent space, $\langle w, \Omega_\eta v\rangle = \int_{\mathcal{A}} w(a) \frac{v(a)}{\pi_\gamma(a)} da$.

In turn, elements u of the co-tangent space satisfy $\int_{\mathcal{A}} u(a)\pi_\gamma(a)da = 0$. The squared dual norm is $\|u\|^2_\eta = \int_{\mathcal{A}} u^2(a)\pi_\gamma(a)da = \text{Var}_\gamma(u(A))$ (see Subsection B.1), and Ω_η^{-1} is such that $\langle \Omega_\eta^{-1} u, s\rangle = \int_{\mathcal{A}} u(a)s(a)\pi_\gamma(a)da = \text{Cov}_\gamma(u(A), s(A))$.

Next, $\nabla_\theta \delta_{\theta(\eta)}$ is an element of the co-tangent space such that, for all tangents v,

$$\langle v, \nabla_\theta \delta_{\theta(\eta)}\rangle = \int_{\mathcal{A}} v(a)(\Delta(a) - \delta_{\theta(\eta)})da.$$

We identify $\nabla_\theta \delta_{\theta(\eta)}$ with $\Delta - \delta_{\theta(\eta)}$. In turn, $\nabla_\theta \log f_{\theta(\eta)}(y)$ is an element of the co-tangent space such that, for all tangents v,

$$\langle v, \nabla_\theta \log f_{\theta(\eta)}(y)\rangle = \frac{\int_{\mathcal{A}} g_\beta(y \,|\, a)v(a)da}{\int_{\mathcal{A}} g_\beta(y \,|\, a)\pi_\gamma(a)da} - \int_{\mathcal{A}} v(a)da = \mathbb{E}_{\beta,\gamma}\left(\frac{v(A)}{\pi_\gamma(A)} \,|\, y\right) - \mathbb{E}_\gamma\left(\frac{v(A)}{\pi_\gamma(A)}\right)$$

We identify $\nabla_\theta \log f_{\theta(\eta)}(y)$ with $\frac{g_\beta(y \,|\, \cdot)}{\int_{\mathcal{A}} g_\beta(y \,|\, a)\pi_\gamma(a)da} - 1$.

For any tangent v, $H_{\theta(\eta)} v$ is a co-tangent element such that, for all tangents w,

$$\begin{aligned}\langle w, H_{\theta(\eta)} v\rangle =& \mathbb{E}_{\theta(\eta)} \langle v, \nabla_\theta \log f_{\theta(\eta)}(Y)\rangle \langle w, \nabla_\theta \log f_{\theta(\eta)}(Y)\rangle \\ =& \text{Cov}_{\beta,\gamma}\left[\mathbb{E}_{\beta,\gamma}\left(\frac{v(A)}{\pi_\gamma(A)} \,|\, Y\right), \mathbb{E}_{\beta,\gamma}\left(\frac{w(A)}{\pi_\gamma(A)} \,|\, Y\right)\right]\end{aligned}$$

In particular, it follows that defining h_ϵ^{MMSE} as in (29) gives the same expression as in (41) since, for all y,

$$\begin{aligned} h_\epsilon^{\text{MMSE}}(y) &= \langle \left[H_{\theta(\eta)} + (\epsilon n)^{-1}\Omega_\eta\right]^{-1} \nabla_\theta \delta_{\theta(\eta)}, \nabla_\theta \log f_{\theta(\eta)}(y)\rangle \\ &= \mathbb{E}_{\beta,\gamma}\left[\frac{1}{\pi_\gamma(A)}\left[H_{\theta(\eta)} + (\epsilon n)^{-1}\Omega_\eta\right]^{-1} [\nabla_\theta \delta_{\theta(\eta)}](A) \,|\, y\right] \\ &= \mathbb{E}_{\beta,\gamma}\left[\left[\mathbb{I}_{\mathcal{A}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}\right]^{-1} [\Delta - \delta](A) \,|\, y\right] \\ &= \mathbb{E}_{\mathcal{A}|\mathcal{Y}}\left[\mathbb{I}_{\mathcal{A}} + (\epsilon n)^{-1}\mathbb{I}_{\mathcal{A}}\right]^{-1} [\Delta - \delta](y).\end{aligned}$$

We also briefly want to discuss when the conditions in Assumption A1(iii) are satisfied for this model. Firstly, we have

$$\int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y)\right]^2 dy = 2\, H^2(f_{\theta_0}, f_{\theta(\eta)}) \leq 2\, D_{\text{KL}}(f_{\theta_0} || f_{\theta(\eta)}) \leq 2\, D_{\text{KL}}(\pi_0 || \pi_\gamma),$$

where the first inequality is the general relation $H^2(f_{\theta_0}, f_{\theta(\eta)}) \leq D_{\mathrm{KL}}(f_{\theta_0}||f_{\theta(\eta)})$ between the squared Hellinger distance H^2 and the Kullback-Leibler divergence D_{KL}, and the second inequality is sometimes called the "chain rule" for the Kullback-Leibler divergence, which can be derived by an application of Jensen's inequality. Finally, recall that we defined our distance measure $d(\theta_0, \theta(\eta))$ in the semi-parametric case to be twice the Kullback-Leibler divergence $2D_{\mathrm{KL}}(\pi_0||\pi_\gamma) = 2\int_{\mathcal{A}} \log\left(\frac{\pi_0(a)}{\pi_\gamma(a)}\right)\pi_0(a)da$. We therefore find that

$$\sup_{(\theta_0,\eta)\in\Gamma_\epsilon} \int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y)\right]^2 dy \leq \sup_{(\theta_0,\eta)\in\Gamma_\epsilon} d(\theta_0, \theta(\eta))) = \epsilon = o(1),$$

that is, the first condition in Assumption A1(iii) is satisfied here. The second condition in that assumption follows if, for example, we assume that $\sup_{y\in\mathcal{Y}} \mathrm{Var}_\gamma\left[g_\beta(y \mid A)\right] / \left[\mathbb{E}_\gamma g_\beta(y \mid A)\right]^2 = O(1)$,[26] because an upper bound on $\left\|\nabla_\theta \log f_{\theta(\eta)}(y)\right\|_\eta^2 = \mathrm{Var}_\gamma\left[g_\beta(y \mid A)\right] / \left[\mathbb{E}_\gamma g_\beta(y \mid A)\right]^2$ can then simply be taken out of the integral over $y \in \mathcal{Y}$.

Regarding the last condition of Assumption A1(iii), we first note that since f_θ is linear in $\theta = \pi$ here we have $f_{\theta_0}(y) - f_{\theta(\eta)}(y) = \langle \theta_0 - \theta(\eta), \nabla_\theta f_{\theta(\eta)}(y)\rangle$, and therefore

$$\left\langle \theta_0 - \theta(\eta), \nabla_\theta f_{\theta(\eta)}^{1/2}(y)\right\rangle = \frac{f_{\theta_0}(y) - f_{\theta(\eta)}(y)}{2f_{\theta(\eta)}^{1/2}(y)}.$$

Using this, we obtain

$$\begin{aligned}
&\int_{\mathcal{Y}} \left[f_{\theta_0}^{1/2}(y) - f_{\theta(\eta)}^{1/2}(y) - \left\langle \theta_0 - \theta(\eta), \nabla_\theta f_{\theta(\eta)}^{1/2}(y)\right\rangle\right]^2 dy \\
&= \frac{1}{4}\int_{\mathcal{Y}} \left[\left(\frac{f_{\theta_0}(y)}{f_{\theta(\eta)}(y)}\right)^{1/2} - 1\right]^4 f_{\theta(\eta)}(y)dy \leq \frac{1}{64}\int_{\mathcal{Y}} \left[\frac{f_{\theta_0}(y)}{f_{\theta(\eta)}(y)} - 1\right]^4 f_{\theta(\eta)}(y)dy \\
&= \frac{1}{64}\int_{\mathcal{Y}} \frac{\left[f_{\theta_0}(y) - f_{\theta(\eta)}(y)\right]^4}{\left[f_{\theta(\eta)}(y)\right]^3}dy = \frac{1}{64}\int_{\mathcal{Y}} \left[\left\langle \theta_0 - \theta(\eta), \nabla_\theta \log f_{\theta(\eta)}(y)\right\rangle\right]^4 f_{\theta(\eta)}(y)dy \\
&\leq \frac{1}{64}\left\|\theta_0 - \theta(\eta)\right\|_{\mathrm{ind},\eta}^4 \mathbb{E}_{\theta(\eta)}\left\|\nabla_\theta \log f_{\theta(\eta)}(Y)\right\|_\eta^4,
\end{aligned}$$

where the first inequality follows from $\sqrt{a} - 1 \leq (a-1)/2$ for $a \geq 0$. This shows that the last part of Assumption A1(iii) holds, provided $\mathbb{E}_{\theta(\eta)}\left\|\nabla_\theta \log f_{\theta(\eta)}(Y)\right\|_\eta^4 = O(1)$.

[26] If γ is not assumed known, then one should also take the supremum over γ in that condition.

to a multinomial logit demand model. Note that in this particular case π is parameter-free. A widely echoed concern in the literature on demand analysis is that properties of the logit, in particular independence of irrelevant alternatives (IIA), may have undesirable consequences for the estimation of δ_{θ_0}; see Anderson *et al.* (1992), for example.

Assuming that β and γ are known for simplicity, in this example we have, by (36) and (38),

$$b_\epsilon(h,\beta,\gamma) = \epsilon^{\frac{1}{2}} \sqrt{\mathbb{E}_\gamma \left[\left(\Delta(A,X,\beta) - \mathbb{E}_\gamma \Delta(\widetilde{A},X,\beta) - \sum_{j=1}^{J} q_j(A,X,\beta) h(j,X) \right)^2 \right]},$$

where

$$q_j(a,x,\beta) = \mathbf{1}\left\{ x_j'\beta_j + a_j \geq x_k'\beta_k + a_k \text{ for all } k \neq j \right\}.$$

Moreover, we have, for all $k = 1, ..., K$ and x,

$$\mathbb{E}_{\beta,\gamma}\left[\sum_{j=1}^{J} q_j(A,x,\beta) h_\epsilon^{\text{MMSE}}(j,x,\beta) \mid Y = k, X = x \right] + (\epsilon n)^{-1} h_\epsilon^{\text{MMSE}}(k,x,\beta)$$
$$= \mathbb{E}_{\beta,\gamma}\left[\Delta(A,x,\beta) \mid Y = k, X = x \right] - \mathbb{E}_\gamma \Delta(A,x,\beta).$$

B.5 Individual effects in panel data (continued)

In this subsection we consider panel data models where $g_\beta(y \,|\, a, x)$ may be misspecified. Let us start with the case where neither g_β nor π_γ are correctly specified. We treat β and γ as known for simplicity. We have

$$b_\epsilon(h,\beta,\gamma) = \epsilon^{\frac{1}{2}} \sqrt{\widehat{\text{Var}}_{\beta,\gamma}\left[\Delta(A,X) - h(Y,X) \right]}.$$

In this case, there is a unique h function which minimizes the bias (to first-order), which corresponds to the *empirical Bayes* h function; that is,

$$h^{\text{EB}}(y,x,\beta,\gamma) = \mathbb{E}_{\beta,\gamma}\left[\Delta(A,X) \,|\, Y = y, X = x \right] - \mathbb{E}_\gamma\left[\Delta(A,X) \,|\, X = x \right], \quad \text{for all } y, x.$$

Note that here there is no scope for achieving fixed-T or even large-T identification (except in the trivial case where $\Delta(A,X) = \Delta(X)$ does not depend on A).

Consider next the case where π_γ is correctly specified, but g_β may be misspecified. We have

$$b_\epsilon(h,\beta,\gamma) = \epsilon^{\frac{1}{2}} \sqrt{\widehat{\text{Var}}_{\beta,\gamma}\left[\Delta(A,X) - h(Y,X) - \mathbb{E}_\beta\left[\Delta(A,X) - h(\widetilde{Y},X) \,|\, A, X \right] \right]}.$$

Example. Consider again the OLS/IV example of Subsection 3.2, but now drop the Gaussian assumptions on the distributions. For known Π, the set of moment conditions corresponds to the moment functions

$$\Psi(y,x,z,\theta) = \begin{pmatrix} x(y - x'\beta - \rho'(x - \Pi z)) \\ z(y - x'\beta) \end{pmatrix}$$

In this case, letting $W = (X', Z')'$ we have

$$K_\eta = -\mathbb{E}_{f_0}(XW'),\ K_{\theta(\eta)} = -\mathbb{E}_{f_0}\begin{pmatrix} XX' & XZ' \\ (X - \Pi Z)X' & 0 \end{pmatrix}\quad V_{\theta(\eta)} = \mathbb{E}_{f_0}\left((Y - X'\beta)^2 WW'\right)$$

Given a preliminary estimator $\widetilde{\beta}$, $V_{\theta(\eta)}$ can be estimated as $\frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\widetilde{\beta})^2 W_i W_i'$, whereas K_η and $K_{\theta(\eta)}$ can be estimated as sample means. The estimator based on (C16) then interpolates nonlinearly between the OLS and IV estimators, similarly as in the likelihood case.

Remarks. If the researcher is willing to specify a complete parametric model f_{θ_0} compatible with the moment conditions (48), the choice of ϵ can then be based on the approach described in Subsection 2.5. Alternatively, the choice of ϵ can be based on specification testing ideas which do not require full specification, such as a test of exogeneity in the OLS/IV example above.

Lastly, the approach outlined here can be useful in fully specified structural models when the likelihood function, score and Hessian of the model are difficult to compute. Given a set of moment conditions implied by the structural model, instead of implementing (28) one may compute the optimal a vector though (C16), which only involves the moment functions and their derivatives. When the moments are computed by simulation, their derivatives can be approximated using numerical differentiation. Note that this minimum-MSE estimator has a different interpretation (and a larger mean squared error) compared to the estimator in (28) that relies on the full likelihood structure.

C.2 Bayesian interpretation

A different approach to account for misspecification of the reference model would be to specify a prior on the parameter θ_0. A Bayesian decision maker could then compute the posterior mean $\mathbb{E}[\delta_{\theta_0} \mid Y_1, \ldots, Y_n]$. As we discuss in C.2.1 below, in the parametric case of Section 3, when θ_0 is endowed with the Gaussian prior $\mathcal{N}(\theta(\eta), \epsilon\Omega^{-1})$ and η is endowed with

a non-dogmatic prior, this posterior mean coincides with our minimum-MSE estimator up to smaller-order terms; that is,

$$\mathbb{E}\left[\delta_{\theta_0} \,|\, Y_1, \ldots, Y_n\right] = \widehat{\delta}_{\epsilon}^{\text{MMSE}} + o_P(\epsilon^{\frac{1}{2}}) + o_P\left(n^{-\frac{1}{2}}\right) \tag{C17}$$

A related question is the interpretation of our minimax estimator in terms of a least-favorable prior distribution. As we discuss in C.2.2 below, in the parametric case a least-favorable prior for θ_0 given η concentrated on the neighborhood $\Gamma_\epsilon(\eta)$ puts all mass at the boundary of $\Gamma_\epsilon(\eta)$.

C.2.1 Gaussian prior

Consider the known η case to start with. To see that (C17) holds, note that, under sufficient regularity conditions,

$$\mathbb{E}\left[\delta_{\theta_0} \,|\, Y_1, \ldots, Y_n, \eta\right] = \delta_{\theta(\eta)} + (\nabla_\theta \delta_{\theta(\eta)})' \mathbb{E}\left[\theta_0 - \theta(\eta) \,|\, Y_1, \ldots, Y_n, \eta\right] + o_P(\epsilon^{\frac{1}{2}}), \tag{C18}$$

where

$$\begin{aligned}\mathbb{E}\left[\theta_0 - \theta(\eta) \,|\, Y_1, \ldots, Y_n, \eta\right] &= \frac{\int (\theta_0 - \theta(\eta)) \prod_{i=1}^n f_{\theta_0}(Y_i) \exp\left(-\frac{1}{2\epsilon}(\theta_0 - \theta(\eta))'\Omega(\theta_0 - \theta(\eta))\right) d\theta_0}{\int \prod_{i=1}^n f_{\theta_0}(Y_i) \exp\left(-\frac{1}{2\epsilon}(\theta_0 - \theta(\eta))'\Omega(\theta_0 - \theta(\eta))\right) d\theta_0} \\ &= \epsilon^{\frac{1}{2}} \frac{\int u \prod_{i=1}^n f_{\theta(\eta)+\epsilon^{\frac{1}{2}}u}(Y_i) \exp\left(-\frac{1}{2}u'\Omega u\right) du}{\int \prod_{i=1}^n f_{\theta(\eta)+\epsilon^{\frac{1}{2}}u}(Y_i) \exp\left(-\frac{1}{2}u'\Omega u\right) du}.\end{aligned}$$

Now, since, up to smaller terms,

$$\prod_{i=1}^n f_{\theta(\eta)+\epsilon^{\frac{1}{2}}u}(Y_i) \approx \prod_{i=1}^N f_{\theta(\eta)}(Y_i) \exp\left(\epsilon^{\frac{1}{2}} u' \sum_{i=1}^n \nabla_\theta \log f_{\theta(\eta)}(Y_i) - \frac{1}{2}\epsilon n u' H_{\theta(\eta)} u\right)$$

we have

$$\begin{aligned}&\mathbb{E}\left[\theta_0 - \theta(\eta) \,|\, Y_1, \ldots, Y_n, \eta\right] \\ &= \epsilon^{\frac{1}{2}} \frac{\int u \exp\left(\epsilon^{\frac{1}{2}} u \sum_{i=1}^n \nabla_\theta \log f_{\theta(\eta)}(Y_i) - \frac{1}{2}u'[\Omega + \epsilon n H_{\theta(\eta)}]u\right) du}{\int \exp\left(\epsilon^{\frac{1}{2}} u \sum_{i=1}^n \nabla_\theta \log f_{\theta(\eta)}(Y_i) - \frac{1}{2}u'[\Omega + \epsilon n H_{\theta(\eta)}]u\right) du} + o_P(\epsilon^{\frac{1}{2}}) + o_P\left(n^{-\frac{1}{2}}\right) \\ &= \epsilon n [\Omega + \epsilon n H_{\theta(\eta)}]^{-1} \frac{1}{n} \sum_{i=1}^n \nabla_\theta \log f_{\theta(\eta)}(Y_i) + o_P(\epsilon^{\frac{1}{2}}) + o_P\left(n^{-\frac{1}{2}}\right).\end{aligned}$$

Lastly, in the case where η is estimated, let us endow it with a non-dogmatic prior. Under regularity conditions, taking expectations in (C18) with respect to the posterior distribution of η implies that (C17) holds.

C.2.2 Least favorable prior

Consider the known η case, in the parametric setting with weighted Euclidean norm. Consider the minimax problem

$$\inf_h \sup_\rho \int\!\!\int_{\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[(\widehat{\delta}_{h,\eta} - \delta_{\theta_0})^2\right] \rho(\theta_0) d\theta_0,$$

where ρ belongs to a class of priors supported on $\Gamma_\epsilon(\theta(\eta))$.

Assuming that the order of the infimum and supremum can be reversed, a least-favorable prior ρ^{LF} solves

$$\sup_\rho \inf_h \int\!\!\int_{\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[(\widehat{\delta}_{h,\eta} - \delta_{\theta_0})^2\right] \rho(\theta_0) d\theta_0.$$

For given h the integral is equal to

$$\begin{aligned}
&\int\!\!\int_{\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[(\widehat{\delta}_{h,\eta} - \delta_{\theta_0})^2\right] \rho(\theta_0) d\theta_0 \\
&= \int_{\Gamma_\epsilon(\eta)} \left(\frac{\mathrm{Var}_{\theta(\eta)}\, h(Y,\eta)}{n} + \left(\delta_{\theta(\eta)} + \mathbb{E}_{\theta_0} h(Y,\eta) - \delta_{\theta_0}\right)^2 \right) \rho(\theta_0) d\theta_0 + o(\epsilon) + o\left(n^{-1}\right) \\
&= \frac{\mathrm{Var}_{\theta(\eta)}\, h(Y,\eta)}{n} \\
&+ \left(\mathbb{E}_{\theta(\eta)} h(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y) - \nabla_\theta \delta_{\theta(\eta)}\right)' \Omega^{-\frac{1}{2}} V_\Omega(\rho) \Omega^{-\frac{1}{2}} \left(\mathbb{E}_{\theta(\eta)} h(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y) - \nabla_\theta \delta_{\theta(\eta)}\right) \\
&+ o(\epsilon) + o\left(n^{-1}\right),
\end{aligned}$$

where

$$V_\Omega(\rho) = \int\!\!\int_{\Gamma_\epsilon(\eta)} \Omega^{\frac{1}{2}} \left(\theta_0 - \theta(\eta)\right) \left(\theta_0 - \theta(\eta)\right)' \Omega^{\frac{1}{2}} \rho(\theta_0) d\theta_0.$$

This quantity (net of the lower-order terms) is minimized, subject to the unbiasedness restriction, at h^* which solves

$$h^*(y,\eta) = n\, \nabla_\theta \log f_{\theta(\eta)}(y)' \Omega^{-\frac{1}{2}} V_\Omega(\rho) \Omega^{-\frac{1}{2}} \left(\nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)} h^*(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y)\right).$$

Let now

$$v = \Omega^{-1} \left(\nabla_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)} h_\epsilon^{\mathrm{MMSE}}(Y,\eta) \nabla_\theta \log f_{\theta(\eta)}(Y)\right),$$

and consider a prior ρ^{LF} that puts all mass at $\theta(\eta) + \epsilon^{\frac{1}{2}} v / \|v\|_\Omega$, say. Note that ρ^{LF} puts all mass at the boundary of $\Gamma_\epsilon(\eta)$ (see also footnote 7).

Then

$$V_\Omega(\rho^{\mathrm{LF}}) = \epsilon \frac{\Omega^{\frac{1}{2}} v v' \Omega^{\frac{1}{2}}}{v' \Omega v}.$$

Moreover, it can be checked that, for $\rho = \rho^{\mathrm{LF}}$,

$$h^*(\cdot, \eta) = h_\epsilon^{\mathrm{MMSE}}(\cdot, \eta),$$

and that ρ^{LF} is least-favorable.

In the case where η is estimated, consider the following problem, for a given prior w on η and a preliminary estimator $\widehat{\eta}$,

$$\inf_h \sup_\rho \int_{\mathcal{B}} \iint_{\Gamma_\epsilon(\eta)} \mathbb{E}_{\theta_0}\left[(\widehat{\delta}_{h,\widehat{\eta}} - \delta_{\theta_0})^2\right] \rho(\theta_0 \,|\, \eta) w(\eta) d\theta_0 d\eta,$$

where $\rho(\cdot \,|\, \eta)$ belongs to a class of priors supported on $\Gamma_\epsilon(\theta(\eta))$ for all η. Note that this formulation provides a Bayesian interpretation for the weight function w appearing in (19).

Applying the above arguments to the estimated-η case, one can derive a related least-favorable prior that satisfies

$$V_\Omega(\rho^{\mathrm{LF}}(\cdot|\eta)) = \epsilon \frac{\Omega^{\frac{1}{2}} v v' \Omega^{\frac{1}{2}}}{v'\Omega v}, \quad \text{for} \quad v = \Omega^{-1}\left(\widetilde{\nabla}_\theta \delta_{\theta(\eta)} - \mathbb{E}_{\theta(\eta)} h_\epsilon^{\mathrm{MMSE}}(Y, \eta) \widetilde{\nabla}_\theta \log f_{\theta(\eta)}(Y)\right).$$

For such a prior, the implied optimal $h^*(\cdot, \eta)$ is again equal to $h_\epsilon^{\mathrm{MMSE}}(\cdot, \eta)$.

C.3 Partial identification

Here we discuss how our approach relates to a partial identification analysis. We focus on the general setup described in Section 2, for a given reference model indexed by a known η. Consider the following *restricted identified set* for δ_{θ_0}, where f_0 denotes the population distribution of Y,

$$\mathcal{S}_{\epsilon,\eta} = \{\delta_{\theta_0} \,:\, \theta_0 \in \Theta,\, f_{\theta_0} = f_0,\, d(\theta_0, \theta(\eta)) \leq \epsilon\}.$$

$\mathcal{S}_{\epsilon,\eta}$ is equal to the intersection of the identified set for δ_{θ_0} with the image by δ of the neighborhood $\Gamma_\epsilon(\eta)$.

Proposition C3. *For any $\epsilon \geq 0$ we have*

$$\operatorname{diam} \mathcal{S}_{\epsilon,\eta} \leq 2 \inf_h b_\epsilon(h, \eta), \tag{C19}$$

where $\operatorname{diam} \mathcal{S}_{\epsilon,\eta} = \sup_{(\delta_1,\delta_2)\in\mathcal{S}_{\epsilon,\eta}^2} |\delta_2 - \delta_1|$ denotes the diameter of the restricted identified set, and the infimum is taken over any function h such that $\mathbb{E}_{f_0} h(Y)$ exists. Moreover, (C19) holds with equality whenever

$$\sup_{\theta_0 \in \Gamma_\epsilon(\eta)} \delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y, \eta) = \sup_{\theta_0 \in \Gamma_\epsilon(\eta)} -\left(\delta_{\theta_0} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_0} h(Y, \eta)\right) = b_\epsilon(h, \eta). \tag{C20}$$

Note that (C20) is satisfied when $\Gamma_\epsilon(\eta)$ is symmetric around $\theta(\eta)$ and $\delta_{\theta_0} - \mathbb{E}_{\theta_0} h(Y, \eta)$ is linear in θ_0. In addition, (C20) approximately holds – up to lower-order terms – when ϵ tends to zero.

Proof. Let h such that $\mathbb{E}_{f_0} h(Y)$ exists. Let $(\delta_1, \delta_2) \in \mathcal{S}^2_{\epsilon,\eta}$, with $\delta_1 = \delta_{\theta_1}$ and $\delta_2 = \delta_{\theta_2}$. Then $\mathbb{E}_{\theta_1} h(Y) = \mathbb{E}_{\theta_2} h(Y) = \mathbb{E}_{f_0} h(Y)$, so

$$\begin{aligned}|\delta_2 - \delta_1| &= |\delta_{\theta_2} - \delta_{\theta_1}| \\ &\leq |\delta_{\theta_2} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_2} h(Y) + \mathbb{E}_{\theta(\eta)} h(Y)| + |\delta_{\theta_1} - \delta_{\theta(\eta)} - \mathbb{E}_{\theta_1} h(Y) + \mathbb{E}_{\theta(\eta)} h(Y)| \leq 2 b_\epsilon(h, \eta).\end{aligned}$$

This shows (C19).

To see when (C19) holds with equality, note that the problem

$$\sup_{(\delta_1,\delta_2)\in\mathcal{S}^2_{\epsilon,\eta}} \delta_{\theta_2} - \delta_{\theta_1}$$

can equivalently be written as

$$\sup_{(\theta_1,\theta_2)\in\Gamma_\epsilon(\eta)^2} \delta_{\theta_2} - \delta_{\theta_1} + \int_{\mathcal{Y}} \lambda_1(y) f_{\theta_1}(y) dy + \int_{\mathcal{Y}} \lambda_2(y) f_{\theta_2}(y) dy, \tag{C21}$$

where λ_1 and λ_2 are the functional Lagrange multipliers associated with the restrictions $f_{\theta_1} = f_0$ and $f_{\theta_2} = f_0$, respectively. Hence, (C21) is equal to

$$\begin{aligned}&\sup_{\theta_1\in\Gamma_\epsilon(\eta)} \left(-\delta_{\theta_1} + \delta_{\theta(\eta)} + \int_{\mathcal{Y}} \lambda_1(y) f_{\theta_1}(y) dy\right) + \sup_{\theta_2\in\Gamma_\epsilon(\eta)} \left(\delta_{\theta_2} - \delta_{\theta(\eta)} + \int_{\mathcal{Y}} \lambda_2(y) f_{\theta_2}(y) dy\right) \\ &\quad = b_\epsilon(\lambda_1, \eta) + b_\epsilon(-\lambda_2, \eta) \geq 2 \inf_h b_\epsilon(h, \eta),\end{aligned}$$

where we have used (C20).

∎

C.4 Different approaches

Distance function. Consider again the setup of Section 3, now equipped with the distance measure $d(\theta_0, \theta) = (\max_{k=1,\ldots,\dim\theta} |\theta_k - \theta_{0k}|)^2$. In this case,

$$\|u\|_{\eta,\epsilon} = \|u\|_\eta = \sum_{k=1}^{\dim\theta} |u_k|$$

is the ℓ^1 norm of the vector u. Hence, computing $h_\epsilon^{\mathrm{MMSE}}(\cdot, \eta)$ in (11) requires minimizing a convex function which combines a quadratic objective function with an ℓ^1 penalty, similarly as in the LASSO (Tibshirani, 1996).

Choice of epsilon. While in the paper we focus on a model detection error approach as in Hansen and Sargent (2008), other rules could be used to set ϵ. For example, an alternative calibration strategy is to target a maximal percentage increase in variance relative to the estimate based on the parametric reference model. Specifically, one may set $\epsilon(k)$ such that the variance of $\widehat{\delta}^{\text{MMSE}}_{\epsilon(k)}$ is lower than k times the variance of $\delta_{\theta(\widehat{\eta}^{\text{MLE}})}$, for any given constant $k \geq 1$, where $\widehat{\eta}^{\text{MLE}}$ is the MLE based on the reference model. If k is kept fixed as n tends to infinity, ϵn will be constant in the limit. For example, in the parametric case of Section 3, by (28) and given a preliminary estimator $\widehat{\eta}$, $\epsilon = \epsilon(k)$ can be chosen such that:

$$(\widetilde{\nabla}_\theta \delta_{\theta(\widehat{\eta})})'[\widetilde{H}_{\theta(\widehat{\eta})} + (\epsilon n)^{-1}\Omega]^{-1}\widetilde{H}_{\theta(\widehat{\eta})}[\widetilde{H}_{\theta(\widehat{\eta})} + (\epsilon n)^{-1}\Omega]^{-1}\widetilde{\nabla}_\theta \delta_{\theta(\widehat{\eta})} = (k-1)\,(\nabla_\eta \delta_{\theta(\widehat{\eta})})' H_\eta^{-1} \nabla_\eta \delta_{\theta(\widehat{\eta})}.$$

Role of the unbiasedness constraint (2). The asymptotic unbiasedness restriction (2) on the candidate h functions is motivated by the aim to focus on an estimator which performs well under the reference model, while in addition providing some robustness away from the reference model. Interestingly, in the case with known η and a weighted Euclidean norm, (29) remains valid when (2) is dropped. In this case our minimax objective coincides with a minimax regret criterion.

Loss function. While we focus on a quadratic loss function other losses are compatible with our approach. In fact, for any loss function $L(a, b)$ that is strictly convex and smooth in its first argument, minimizing the maximum value of

$$\mathbb{E}_{\theta_0}\left[L\left(\widetilde{\delta}_{h,\widehat{\eta}}, \delta_{\theta_0}\right)\right]$$

on Γ_ϵ will lead to the same expressions for the minimum-MSE h function. This is due to our focus on a local asymptotic approach, and the fact that $L(a, b) \approx c|a - b|^2$ when $|a - b| \approx 0$.

Fixed-ϵ bias. In this paper we rely on a small-ϵ asymptotic. The tractability of our results relies crucially on a local approach. Nevertheless, in some models it is possible to provide relatively simple bias formulas for fixed ϵ. To see this, let us consider the setup of Section 4 for known β and γ. We have the following result.

Proposition C4. *For any $\epsilon > 0$ we have*

$$b_\epsilon(h, \beta, \gamma) = \left| C\mathbb{E}_\gamma\left[\left(\widetilde{\Delta}_\gamma(A, \beta) - \mathbb{E}_\beta\left(h(Y) \mid A\right)\right) \exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\gamma(A, \beta) - \mathbb{E}_\beta\left(h(Y) \mid A\right)\right)\right)\right]\right|, \tag{C22}$$

for $\widetilde{\Delta}_\gamma(a,\beta) = \Delta(a,\beta) - \mathbb{E}_\gamma \Delta(A,\beta)$*, and* $C > 0$ *and* λ_2 *two constants which satisfy equations (C23)-(C24) given in the proof.*

Proposition C4 provides an explicit expression for the bias, for *any* $\epsilon > 0$. Note that both C and λ_2 depend on ϵ. When ϵ tends to zero one can show that $1/\lambda_2$ tends to zero, and the bias converges to the expression in (36).

While it would be theoretically possible to follow a fixed-ϵ approach throughout the analysis, instead of the local approach we advocate, proceeding in that way would face several challenges. First, the bias in (C22) depends on parameters C and λ_2 which need to be recovered given ϵ, increasing computational cost. Second, simple fixed-ϵ derivations seem to be limited to settings where the parameter θ_0 (that is, π_0 in the present setting) enters the likelihood function linearly. Under linearity, similar derivations have been used in other contexts, see Schennach (2013) for an example. The third and main challenge is that characterizing mean squared errors and confidence intervals would become less tractable, while as we have seen those remain simple calculations under a local approximation. Lastly, note that the local approach allows us to provide insights into the form of the solution, as shown by our discussion of the panel data example.

Proof. Let us omit the reference to β, γ for conciseness, and denote $\pi = \pi_\gamma$. Consider the maximization of $\left|\delta_{\pi_0} - \delta_\pi - \int h(y) f_{\pi_0}(y) dy\right|$ with respect to π_0. Let $\widetilde{\Delta}_\pi(a) = \Delta(a) - \delta_\pi$. The corresponding Lagrangian is

$$\mathcal{L} = \iint_{\mathcal{Y}\times\mathcal{A}} \left(\widetilde{\Delta}_\pi(a) - h(y)\right) g(y \mid a)\pi_0(a) dy da + \lambda_1 \int_{\mathcal{A}} \pi_0(a) da + 2\lambda_2 \int_{\mathcal{A}} \log\left(\frac{\pi_0(a)}{\pi(a)}\right) \pi_0(a) da.$$

The first-order conditions with respect to π_0 are then

$$\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y) g(y \mid a) dy + [\lambda_1 + 2\lambda_2] + 2\lambda_2 \log\left(\frac{\pi_0(a)}{\pi(a)}\right) = 0.$$

Hence, using that π_0 integrates to one,

$$\pi_0(a) = C \exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y,x) g(y \mid a) dy\right)\right) \pi(a),$$

where

$$C^{-1} = \int_{\mathcal{A}} \exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}} h(y,x) g(y \mid a) dy\right)\right) \pi(a) da. \tag{C23}$$

Since, at the least-favorable π_0, $2\int_{\mathcal{A}}\left(\log\left(\frac{\pi_0(a)}{\pi(a)}\right)\right)\pi_0(a)da = \epsilon$, we have

$$\epsilon = 2\log C - \frac{C}{\lambda_2}\int_{\mathcal{A}}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}}\left(h(y,x)g(y\,|\,a)dy\right)\right)\times \exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}}\left(h(y,x)g(y\,|\,a)dy\right)\right)\right)\pi(a)da. \tag{C24}$$

It follows that

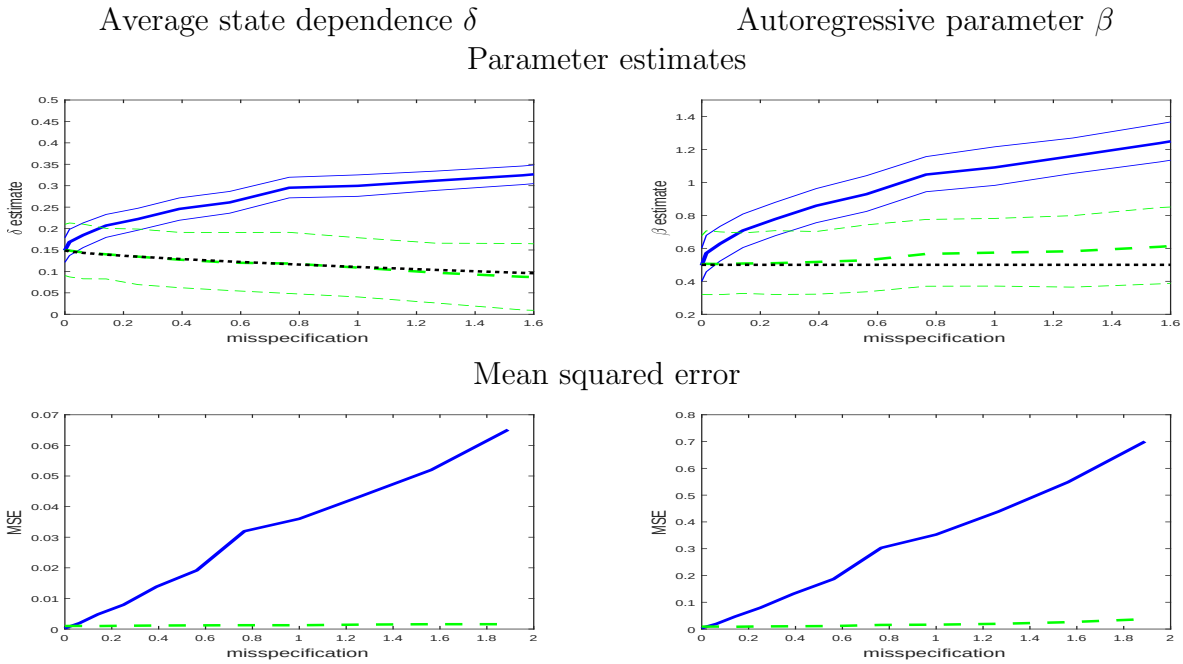
$$b_\epsilon(h) = \left|C\int_{\mathcal{A}}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}}\left(h(y,x)g(y\,|\,a)dy\right)\times \exp\left(-\frac{1}{2\lambda_2}\left(\widetilde{\Delta}_\pi(a) - \int_{\mathcal{Y}}\left(h(y,x)g(y\,|\,a)dy\right)\right)\right)\pi(a)da\right|,$$

where C and λ_2 satisfy (C23)-(C24).

Hence (C22) follows.

∎

Figure C1: Estimates and mean squared error of random-effects and minimum-MSE estimators under varying amount of misspecification, $p = 10^{-10}$



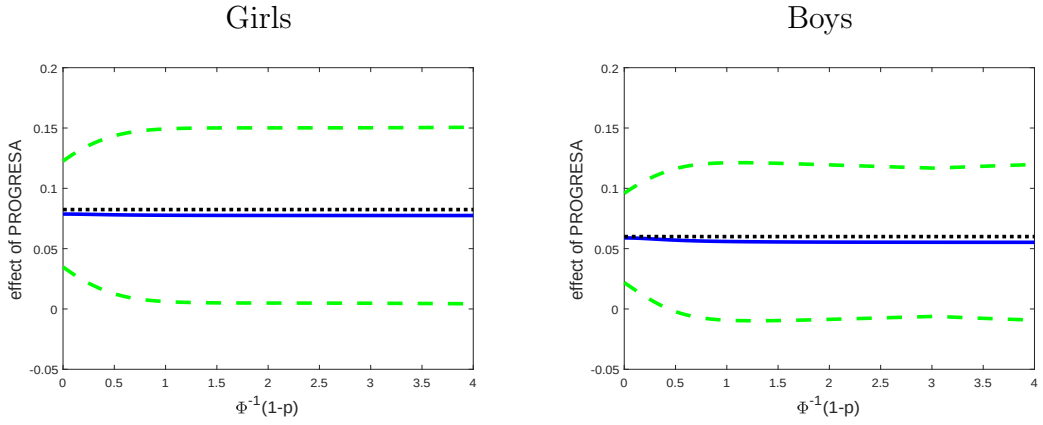
Notes: Random-effects (solid) and minimum-MSE (dashed) for δ (left graphs) and β (right graphs). True parameter values are shown in dotted. $n = 500$, $T = 5$. The reference specification for π is normal with mean $-.25 + .5Y_{i0}$ and standard deviation .8, whereas the true π_0 is normal with the same standard deviation and mean $-.25+\nu+.5Y_{i0}$. On the x-axis we report twice the KL divergence; that is, $\nu^2/.64$. Top panel: mean and 95% interval. Bottom panel: mean squared error. ϵ is chosen according to (43) for a detection error probability $p = 10^{-10}$. (μ, σ) are treated as known.

Table C1: Effect of the PROGRESA subsidy and counterfactual reforms, reference model estimated on both controls and treated

	Model-based		Minimum-MSE		Experimental	
	PROGRESA impacts					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.082	.060	.078	.055	.087	.050
non-robust CI	(.026,.139)	(.018,.102)	-	-	-	-
robust CI	(-.012,.177)	(-.058,.178)	(.005,.150)	(-.008,.119)	-	-
	Counterfactual 1: doubling subsidy					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.154	.112	.147	.105	-	-
robust CI	(-.008,.315)	(-.091,.315)	(.025,.270)	(-.004,.214)	-	-
	Counterfactual 2: unconditional transfer					
	Girls	Boys	Girls	Boys	Girls	Boys
estimate	.007	.000	.003	-.012	-	-
robust CI	(-.542,.557)	(-.478,.478)	(-.201,.207)	(-.193,.169)	-	-

Notes: Sample from Todd and Wolpin (2006). $p = .01$. CI are 95% confidence intervals. The unconditional transfer amounts to 5000 pesos in a year.

Figure C2: Effect of the PROGRESA subsidy as a function of the detection error probability, reference model estimated on both controls and treated



Notes: Sample from Todd and Wolpin (2006). $\epsilon(p)$ is chosen according to (32), with $\Phi^{-1}(1-p)$ reported on the x-axis. The minimum-MSE estimates of the effect of PROGRESA on school attendance are shown in solid. 95% confidence intervals based on those estimates are in dashed. The dotted line shows the unadjusted model-based prediction. Girls (left) and boys (right).