Robust Estimation with Exponentially Tilted Hellinger Distance*

Bertille Antoine
Simon Fraser University
Bertille_Antoine@sfu.ca

Prosper Dovonon
Concordia University and CIREQ
Prosper.Dovonon@concordia.ca

February 16, 2016

Preliminary and incomplete

Abstract

In a recent paper, Kitamura, Otsu, and Evdokimov (2013) introduce the minimum Hellinger Distance (HD) estimator as a computationally convenient estimator that is semiparametrically efficient when the model assumption holds (correct specification) and robust to small deviations from the model (local misspecification). In this paper, we evaluate the performance of inference procedures of interest under two complementary types of misspecification, local and global.

First, we show that HD is not robust to global misspecification (or model misspecification) in the sense that HD may cease to be root n convergent when the functions defining the moment conditions are unbounded (even when their expectations are bounded). Second, in the spirit of Schennach (2007), we propose a new estimator called ETHD. Our estimator shares the same desirable asymptotic properties as HD under correct specification and local misspecification, and remains well-behaved under model misspecification. ETHD is therefore the first estimator that is efficient under correct specification, and robust to both global and local misspecification.

Keywords: misspecified models; local misspecification; higher-order asymptotics; semiparametric efficiency.

*We would like to thank Susanne Schennach and Pierre Chaussé for helpful discussions.

1 Introduction

Economic models can often be understood as simplifications of the real world and, as such, are intrinsically misspecified (see e.g. Maasoumi (1990), Hall and Inoue (2003), Schennach (2007)). As a result, the choice of an inference procedure should not solely be based on its performance under correct specification, but also on its robustness to deviations from the correct specification, so-called misspecification.

For example, modern asset pricing models rely on moment condition models that depend on a pricing kernel (or stochastic discount factor) and data. Unlike what the economic theory suggests, it is long recognized that no pricing kernel can correctly price all financial securities. As a result, the pricing kernel used in applications is the one that is the least misspecified; see e.g. Hansen and Jagannathan (1997), Kan, Robotti, and Shanken (2013), and Gospodinov, Kan, and Robotti (2014).

In this paper, we consider economic models defined by moment restrictions, and evaluate the performance of inference procedures of interest under two complementary types of misspecification: global and local misspecification. Global misspecification (also referred to as model misspecification) occurs when one cannot find a parameter value such that the population moment restriction is satisfied. Local misspecification captures the case where the population moment condition is invalid for any finite sample size, but the size of the violation is so small that it disappears asymptotically.

Economic models may indeed suffer from both types of misspecification (Hall and Inoue (2003)). However, since the extent and nature of the misspecification is always unknown in practice, it appears ideal to rely on inference procedures that are efficient in correctly specified models, and somewhat robust to both types of misspecification. To our knowledge, such an inference procedure is not currently available, and the main contribution of this paper is to fill this gap. An estimator robust to global misspecification remains asymptotically normal with the same rate of convergence as when the model is correctly specified. The appeal of such an estimator comes from the fact that its asymptotic distribution that is valid under both global misspecification and correct specification can be derived. As a result, inference immune to global misspecification becomes routinely possible. By contrast, local misspecification is only noticeable in small samples (and not at the limit). Since the true distribution of the data is expected to match the one postulated by the researcher as the sample size gets large, one can define the true parameter value as the value that solves the assumed model. An efficient estimator is robust to local misspecification when its worse mean squared error (computed over all possible small deviations of data distribution) remains the smallest compared to (admissible) competitors. Estimators that are robust to local misspecification remain consistent (for the true parameter value) so long as the true data-distribution is sufficiently close to the postulated distribution.

Maasoumi and Phillips (1982) and Gallant and White (1988) provide an early analysis of inference in globally misspecified models estimated by instrumental variables and the Generalized Method of Moments (GMM), respectively. Hall and Inoue (2003) extend their asymptotic analysis to the two-step and iterated GMM estimators. They establish that it remains convergent at the standard rate and asymptotically normal. They also derive its asymptotic distribution under global misspecification. Among competitive procedures, we highlight the following three estimators which belong to the class of Cressie-Read (CR) minimum discrepancy estimators: the Continuously Updated (CU) GMM, the Exponentially Tilting (ET), the Empirical Likelihood (EL), and the three-step Euclidean EL estimator (3S-EEL) introduced by Antoine, Bonnal and Renault[1] (2007). These estimators rely on implied probabilities to re-weight the sample observations in order to guarantee that the moment condition is exactly satisfied (in sample). The CR estimator is then chosen as the estimator that minimizes the discrepancy between the implied probabilities and the uniform weights ($1/n$). Kitamura (2000) studies ET under global misspecification and establishes its robustness, while Dovonon (2015) shows the robustness of 3S-EEL in globally misspecified models. Schennach (2007) studies EL under global misspecification and shows that it is not robust. She identifies some singularity issues in the implied probability function of EL that are responsible for its lack of robustness. Then, observing that ET's implied probabilities do not display any such singularity, she proposes the Exponentially Tilted Empirical Likelihood (ETEL) estimator that combines EL's discrepancy function with ET's implied probabilities. The ETEL estimator is quite appealing: it is efficient and shares the higher-order bias properties of EL in correctly specified models, and remains as stable as ET in globally misspecified models.

The study of inference in locally misspecified models has a long tradition in economics as it is often used to provide insights into the local power properties of test statistics. It was first applied in the context of GMM by Newey (1985). In a recent paper, Kitamura, Otsu and Evdokimov (KOE hereafter, 2013) illustrate the lack of robustness of GMM to local misspecification. Building on the pioneering work of Beran (1977a,b), they introduce a new estimator that minimizes the Hellinger Distance (HD). This estimator is quite appealing: it is easy to implement and reconciles efficiency and local robustness for moment restrictions models.

In this paper, we build on Schennach (2007) and KOE to deliver an estimator that is efficient under correct specification, and robust to both local and global misspecification. To motivate the need for a new estimator, we first show that HD is not well-behaved under global misspecification. The intuition for such a lackluster performance follows from the conjecture of Schennach (2007, p641) that connects the poor performance of estimators from the CR family to the negative value of their indexing parameter (such as HD and EL). Accordingly, the only candidate from the CR family that retains good properties under global misspecification is ET. In the hope of delivering an estimator that combines the desirable properties of ET under global misspecification and HD under both correct specification and local misspecification, we build on Schennach (2007) to show that ET and HD can easily be combined to define a new estimator called ETHD that is efficient and robust to both local and global misspecification.

This paper is organized as follows. In section 2, we briefly review the excellent properties of HD estimator under correct and local mispecification, and present a simple result that characterizes its lackluster behavior under (global) misspecification. In section 3, we introduce our estimator called exponentially tilted Hellinger distance (ETHD) that naturally combines ET and HD. The ETHD estimator is shown to be efficient when the model is correctly specified, and well-behaved under model misspecification. In section 4, we show that it is robust to local mispecification. In section 5, we show that ETHD is well-behaved and robust to global misspecification. In section 6, Monte-Carlo simulations illustrate the usefulness of our estimator. All proofs can be found in the Appendix.

[1]The 3S-EEL estimator is computationally friendly (e.g. much more than EL), and shares the same desirable bias properties as EL in correctly specified models.

2 The HD estimator under global misspecification

2.1 Definition and properties

Let θ^* be a vector of unknown parameters of interest that belongs to a compact subset $\Theta \in \mathbb{R}^p$. Let x_i be an iid vector of random variables, and $g(.)$ a vector of $m(\leq p)$ real functions such that

$$Eg(x_i, \theta^*) = 0\,.$$

The HD estimator is defined as:

$$\hat{\theta} \equiv \arg \inf_{\theta \in \Theta} \inf_{\pi} H^2(\pi, P_n), \qquad s.t. \qquad \sum_{i=1}^{n} \pi_i g(x_i, \theta) = 0, \tag{1}$$

where $H(.)$ is the Hellinger distance between two probability measures π and P_n,

$$H(\pi, P_n) = \left(\int \left(d\pi^{1/2} - dP_n^{1/2} \right)^2 \right)^{\frac{1}{2}} = \left(\sum_{i=1}^{n} \left(\sqrt{\pi_i} - 1/\sqrt{n} \right)^2 \right)^{1/2}, \tag{2}$$

and P_n is the discrete uniform distribution on $\{x_i : \; i = 1, \ldots, n\}$.

Under some mild conditions, it can be shown using arguments based on convex duality that the HD estimator is also equal to

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \max_{\gamma \in R^m} -\frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \gamma' g(x_i, \theta)}. \tag{3}$$

The definition (3) is the one adopted by Kitamura, Otsu and Evdokimov (hereafter KOE, 2013). The HD estimator also belongs to the class of generalized empirical likelihood (GEL) estimators[2] and can be characterized using the function $\rho(v) = -1/(1+v)$ on the domain $\mathcal{V} = (-1, +\infty)$; see Newey and Smith (2004). Even though the definition (3) used by KOE does not explicitly require that,

$$1 + \hat{\gamma}' g(x_i, \hat{\theta}) \geq 0 \;\text{ for } (\hat{\theta}, \hat{\gamma}) \text{ solving (3) and for all } i = 1, \ldots, n,$$

such (non-negativity) condition is however essential for the two definitions of the HD estimator, respectively (1) and (3), to be equivalent. This is due to the fact that the first order condition associated with the Lagrangian of the inner program in (1) is

$$1 - \frac{1}{\sqrt{n\pi_i}} + \gamma' g(x_i, \theta) = 0,$$

for all $i = 1, \ldots, n$ in the direction of π. Hence, admissible solutions for π exist only if

[2]The HD estimator is also a member of the empirical Cressie-Read class of estimators that is defined through the following discrepancy function indexed by a

$$h(\pi_i) = ((n\pi_i)^{a+1} - 1)/(a(a+1))\,. \tag{4}$$

The HD estimator is associated with the index $a = -1/2$.

$$1 + \hat{\gamma}' g(x_i, \hat{\theta}) \geq 0\,, \text{for all } i = 1, \ldots, n.$$

In correctly specified models, this condition can be overlooked since the Lagrange multiplier $\hat{\gamma}$ associated to $\hat{\theta}$ obtained from (3) converges sufficiently fast to 0 (under regularity conditions) to guarantee that $\hat{\gamma}' g(x_i, \hat{\theta})$ is uniformly negligible for n large enough. However, in possibly misspecified models, this condition may matter. In what follows, we shall keep it to guarantee the interpretation of the "pseudo" true value θ^* of θ as the parameter value associated with the set of induced distributions[3] the closest to the true distribution of the data in terms of Hellinger distance. In other words, and from now on, we will consider the HD estimator defined by (3) with the additional admissibility condition $1 + \gamma' g(x_i, \theta) \geq 0$ for all $i = 1, \ldots, n$; as explained above, such admissibility condition can also be seen as a restriction on the domain.

In our simulation study below (see section 6.2), we explore the consequences of the relaxation of the above admissibility condition. As already mentioned, such condition may not be innocuous under misspecification. Without any surprise, we find that the unconstrained HD (HD-unc) estimator performs slightly better than the (constrained) HD estimator under (local) misspecification. However, it loses its interpretation.

As a member of the GEL class of estimators, under Assumptions 1 and 2 of Newey and Smith (2004), Theorem 3.2 of the same paper applies to HD. Letting $G = E(\partial g(x, \theta^*)/\partial \theta')$, $\Omega = E(g(x, \theta^*) g(x, \theta^*)')$ and $\Sigma = [G' \Omega^{-1} G]^{-1}$, we have

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma)\,. \tag{5}$$

2.2 Behavior under global misspecification

In this section, we study the behavior of the HD estimator under global misspecification. In our framework, global misspecification means that one cannot find a value of the parameter such that the moment condition is satisfied, that is:

$$Eg(x_i, \theta) \neq 0, \qquad \forall \theta \in \Theta.$$

When the asymptotic distribution of an estimator derived under global misspecification coincides with its asymptotic distribution under correct specification in absence of misspecification, this estimator is said to be robust to global misspecification. Such robustness is desirable because it allows for valid and reliable inference (using the misspecification-robust asymptotic distribution) whether the model is correctly specified or not. For example, Hall and Inoue (2003) show that GMM is robust to global misspecification.

Schennach (2007, Theorem 1) shows that the empirical likelihood (EL) estimator is not robust to global misspecification by showing that the EL estimator is not $\sqrt{n}$ convergent. We now show a similar result for the HD estimator.

Theorem 2.1 *(Lack of robustness of the HD estimator under global misspecification)*
Let x_i be an iid sequence. Assume $g(x, \theta)$ to be twice continuously differentiable in θ for all x and all $\theta \in \Theta$ and such that

$$\sup_{\theta \in \Theta} E\left[|g(x_i, \theta)|^2\right] < \infty\,.$$

[3]For a given value $\theta \in \Theta$, we call "induced distribution" any distribution P satisfying $\int g(x, \theta) dP = 0$.

If

$$\inf_{\theta\in\Theta} |E[g(x_i,\theta)]| \neq 0 \quad \textit{and} \quad \sup_{x\in\mathcal{X}} u' g(x,\theta) = \infty$$

for any $\theta \in \Theta$ and any unit vector u, then there does not exist any $\theta^* \in \Theta$ such that

$$|\hat{\theta}_{HD} - \theta^*| = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Comments:
(i) The above result shows that the HD estimator is not $\sqrt{n}$ convergent. Hence, it is not robust to global misspecification. Our simulation study in section 6 illustrates this result.
(ii) The lack of robustness of the HD estimator to global misspecification is not surprising. The intuition for such a lackluster performance follows from Schennach's (2007, p641) conjecture that connects the poor performance of estimators from the Cressie-Read family to the negative value of their indexing parameter. As recalled in (4), the HD estimator is associated with index $a = -1/2$. Accordingly, the only Cressie-Read estimator that is well-behaved under global misspecification is the exponentially tilted (ET) estimator. In the next section, we combine ET and HD to define an estimator that is well-behaved in all situations.

3 The Exponentially Tilted Hellinger Distance (ETHD) estimator

In this section, we build on Schennach (2007) and KOE to deliver an estimator that is efficient under correct specification, and robust to both local and global misspecification. More specifically, we show that ET and HD can easily be combined to define our new estimator called ETHD that is efficient and robust to both forms of misspecification.

3.1 Definition and characterization of the ETHD estimator

The Exponentially Tilted Hellinger Distance (ETHD) estimator is defined as:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} H(\hat{\pi}(\theta), P_n), \tag{6}$$

where $\hat{\pi}(\theta)$ are the solution of

$$\min_{\{\pi_i\}_{i=1}^n} \sum_{i=1}^n \pi_i \log n\pi_i \left(= E_\pi \log \frac{d\pi}{dP_n}\right) \tag{7}$$

subject to

$$E_\pi g(x,\theta) = 0 \quad \text{and} \quad \sum_{i=1}^n \pi_i = 1. \tag{8}$$

The ETHD estimator combines the discrepancy function $H(.)$ of the HD estimator (defined in (2)) with the implied probabilities of the ET estimator. An alternative definition of ETHD is given by:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \Delta_n\left(\hat{\lambda}(\theta), \theta\right). \tag{9}$$

where

$$\Delta_n(\lambda,\theta) = \frac{\int \exp\left(\lambda' g(x,\theta)/2\right) dP_n}{\left(\int \exp\left(\lambda' g(x,\theta)\right) dP_n\right)^{\frac{1}{2}}}\,,$$

and

$$\hat{\lambda}(\theta) = \arg\max_{\lambda\in R^m} -\int \exp\left(\lambda' g(x,\theta)\right) dP_n\,. \tag{10}$$

The equivalence between the two definitions (6) and (9) of the ETHD estimator follows from the fact that the sequence $\{\hat{\pi}_i(\theta)\}_{i=1}^n$ defined as the solution to (7)-(8) are equal to (see Kitamura (2006)):

$$\hat{\pi}_i(\theta) = \frac{\exp\left(\hat{\lambda}(\theta)' g(x_i,\theta)\right)}{\sum_{j=1}^n \exp\left(\hat{\lambda}(\theta)' g(x_j,\theta)\right)}\,. \tag{11}$$

It follows then that

$$H^2(\hat{\pi}(\theta), P_n) = 2 - 2\frac{1}{\sqrt{n}}\sum_{i=1}^n \sqrt{\hat{\pi}_i(\theta)} = 2 - 2\frac{\frac{1}{n}\sum_{i=1}^n \exp\left(\hat{\lambda}(\theta)' g(x_i,\theta)/2\right)}{\left(\frac{1}{n}\sum_{i=1}^n \exp\left(\hat{\lambda}(\theta)' g(x_i,\theta)\right)\right)^{\frac{1}{2}}}$$

yielding the alternative definitions.

Remark 1 *Thanks to the Jensen's inequality, the concavity of the square root function implies that $\Delta_n(\lambda,\theta) \leq 1$ for all $(\lambda,\theta) \in \mathbb{R}^m \times \Theta$. This feature will be used to prove the consistency of the ETHD estimator in the next section.*

The ETHD estimator is characterized by the following first-order conditions.

Theorem 3.1 *(First-order conditions of the ETHD estimator)*
If The first-order conditions for the ETHD estimator $\hat{\theta}$ can be written as

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^n \sqrt{\pi_i(\hat{\theta})}\right)\left(\sum_{j=1}^n \frac{d(\hat{\lambda}'(\hat{\theta})g(x_j,\hat{\theta}))}{d\theta}\pi_j(\hat{\theta})\right) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \sqrt{\pi_i(\hat{\theta})}\frac{d(\hat{\lambda}'(\hat{\theta})g(x_i,\hat{\theta}))}{d\theta} = 0\,,$$

where $\hat{\lambda}(\theta)$ is such that

$$\frac{1}{n}\sum_{i=1}^n \exp(\hat{\lambda}'(\theta)g(x_i,\theta))g(x_i,\theta) = 0\,.$$

3.2 First-order asymptotic properties of ETHD

In this section, we study the first-order asymptotic properties of the ETHD estimator $\hat{\theta}$. Consistency and asymptotic normality are established for this estimator. In particular, ETHD is shown to be efficient and also enjoy some invariance properties. We also show that the minimum of $\Delta_n(\hat{\lambda}(\theta),\theta)$ can be used for moment condition specification testing.

We first show that it is consistent for the true parameter value θ^* under the following regularity conditions.

Assumption 1 *(Regularity conditions)*
(i) x_i *forms an i.i.d. sequence;*
(ii) $g(x,\theta)$ *is continuous at each* $\theta \in \Theta$ *with probability one and* Θ *compact;*
(iii) $E(g(x,\theta)) = 0 \Leftrightarrow \theta = \theta^*$*;*
(iv) $E(\sup_{\theta\in\Theta} |g(x,\theta)|^\alpha) < \infty$ *for some* $\alpha > 2$ *;*
(v) $Var(g(x,\theta))$ *is nonsingular and finite for every* $\theta \in \Theta$ *with smallest eigenvalues* $\underline{\ell}$ *bounded away from 0;*
(vi) $E\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} \exp(\lambda' g(x,\theta)\right) < \infty$*, where* Λ *is a compact subset of* $\mathbb{R}^m$ *containing an open neighborhood of 0.*

Assumptions 1(i)-(iv) are standard in the literature on inference based on moment condition models. Newey and Smith (2004) have established the consistency of the generalized empirical likelihood class of estimators under this set of assumptions. Assumption 1(v) is not particularly restrictive whereas Assumption 1(vi) is useful because of the two-step nature of our estimation procedure. Schennach (2007) has also made use of a similar assumption to establish the consistency of ETEL.

Under this assumption, we shall consider, instead of (10), the following alternative definition of $\hat{\lambda}(\theta)$:

$$\hat{\lambda}(\theta) = \arg\max_{\lambda\in\Lambda} - \int \exp\left(\lambda' g(x,\theta)\right) dP_n\,. \tag{12}$$

The definition of $\hat{\lambda}(\theta)$ in (12), is theoretically more tractable in the proof of consistency, thanks to the compactness of Λ. Besides, it does not alter the asymptotic properties of $\hat{\theta}$ since Λ contains 0 which is the population value of λ as an interior point.

Theorem 3.2 *(Consistency of the ETHD estimator)*
Under Assumption 1, we have:
(i) $\hat{\theta} \xrightarrow{P} \theta^*$*;* *(ii)* $\hat{\lambda}(\hat{\theta}) = O_P(n^{-1/2})$*; and* *(iii)* $\int g(x,\hat{\theta})dP_n = O_P(n^{-1/2})$*.*

We now turn to the derivation of the asymptotic distribution of the ETHD estimator. We make the following additional regularity assumptions for this purpose.

Assumption 2 *(Regularity assumptions 2)*
(i) $\theta^* \in int(\Theta)$*; there exists a neighborhood* $\mathcal{N}$ *around* θ^* *such that* $g(x,\theta)$ *is continuously differentiable a.s. on* $\mathcal{N}$ *and* $E\left(\sup_{\theta\in\mathcal{N}} |\partial g(x,\theta)/\partial\theta'|^2\right) < \infty$*;*
(ii) $Rank(G) = p$*, with* $G = E(\partial g(x,\theta^*)/\partial\theta')$*.*

Similarly to the two-step GMM procedure, the maximum of $\Delta_n(\hat{\lambda}(\theta),\theta)$, reached at $\hat{\theta}$ can be used to test for the validity of the moment condition model. We consider the specification test statistic:

$$S_n = 4nH^2(\hat{\pi}(\hat{\theta}), P_n) = 8n(1 - \Delta_n(\hat{\lambda}(\hat{\theta}),\hat{\theta})).$$

The asymptotic distribution of S_n along with that of the estimator are derived in the following theorem:

Theorem 3.3 *(Asymptotic distribution of the ETHD estimator)*
Let $\hat{\lambda} = \hat{\lambda}(\hat{\theta})$*. Under Assumptions 1 and 2, we have:*

(i)

$$\sqrt{n}\begin{pmatrix} \hat{\theta}-\theta^* \\ \hat{\lambda} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \Sigma & 0 \\ 0 & \Omega^{-1/2}M\Omega^{-1/2} \end{pmatrix}\right),$$

with $\Omega = E(g(x,\theta^*)g(x,\theta^*)')$, $\Sigma = \left[G'\Omega^{-1}G\right]^{-1}$ *and* $M = I_m - \Omega^{-1/2}G\Sigma G'\Omega^{-1/2}$.

(ii)

$$S_n \xrightarrow{d} \chi^2_{m-p}.$$

Comments:
The above result shows that, under correct specification, the ETHD estimator has the same limiting distribution as efficient two-step GMM, which also corresponds to the limiting distribution of the HD estimator as recalled in (5). The specification test statistic S_n has the same asymptotic distribution as the Hansen's J-test statistic. The proof actually reveals that both statistics are asymptotically equivalent under the conditions of the theorem.

4 ETHD under local misspecification

Local misspecification occurs when the true data generating process deviates from the postulated one within a distance that vanishes as the sample size grows. KOE have proposed an estimation theory for moment condition models that is robust to local misspecification in the following sense.

Consider the family of Fisher consistent and regular estimators (see Definition 1 in Appendix C). KOE derive the asymptotic minimax bound of a large class of loss functions over this family of estimators. These bounds reflect the minimum worst loss that an estimator can be exposed to, when small discrepancies exist between the true and the assumed probability distribution of the data. KOE also show that the HD estimator reaches that bound. It is therefore robust to local misspecification in the sense that it incurs the minimum worst loss following small deviations from true probability distribution, compared to other Fisher-consistent and regular estimators.

In this section, we establish a similar result for the ETHD estimator. Let $\mathcal{M}$ be the set of all probability measures on the Borel σ-field $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$ of $\mathcal{X} \subset \mathbb{R}^d$, $g: \mathcal{X}\times\Theta \to \mathbb{R}^m$ and Λ a subset of $\mathbb{R}^m$.

A natural account of the ETHD estimator suggests that we consider the functionals $T_1: \mathcal{M}\times\Theta \to \Lambda$ and $T: \mathcal{M} \to \Theta$ defined by:

$$T(P) = \arg\max_{\theta\in\Theta} \frac{\int \exp\left(T_1'(P,\theta)g(x,\theta)/2\right)dP}{\left(\int \exp\left(T_1'(P,\theta)g(x,\theta)\right)dP\right)^{\frac{1}{2}}}, \tag{13}$$

with

$$T_1(P,\theta) = \arg\max_{\lambda\in\Lambda}\left(-\int \exp(\lambda' g(x,\theta))dP\right). \tag{14}$$

With P_n the empirical measure, $T(P_n)$ shall correspond to the ETHD estimator $\hat{\theta}$. However, the definition of $T(P)$ causes some technical difficulties when $g(x,\theta)$ is unbounded for some

$\theta \in \Theta$ and $P \in \mathcal{M}$ that have been overcome by KOE via trimming. Let

$$\mathcal{X}_n = \left\{ x \in \mathcal{X} : \sup_{\theta\in\Theta} |g(x,\theta)| \leq m_n \right\}$$

and

$$g_n(x,\theta) = g(x,\theta)\mathbb{I}(x \in X_n).$$

We define:

$$\begin{aligned} \bar{T}(Q) &= \arg\max_{\theta\in\Theta} \frac{\int \exp\{T_1(Q,\theta)' g_n(x,\theta)/2\} dQ}{\left(\int \exp\{T_1(Q,\theta)' g_n(x,\theta)/2\} dQ\right)^{1/2}} \\ T_1(Q,\theta) &= \arg\max_{\lambda} -\int \exp\{\lambda' g_n(x,\theta)\} dQ. \end{aligned} \tag{15}$$

$\bar{T}(\cdot)$ then defined is simply the value of $\theta \in \Theta$ that minimizes the Hellinger distance between $P(\theta)$ and Q, where $P(\theta)$ is the distribution that minimizes the Kullback-Leibler information criterion distance between Q and the set of distributions P that satisfy $E_P\left(g_n(x,\theta)\right) = 0$; see Lemma C.1 for a proof.

As one can also expect, the definition of $T_1(Q,\theta)$ also poses some difficulties when $\exp(\lambda' g_n(x,\theta)$ is not bounded for some $\theta \in \Theta$, $\lambda \in R^m$ and $Q \in \mathcal{M}$. We overcome this by solving (15) optimizing over $\lambda \in \Lambda_n$, a convex subset of R^m containing 0 as interior point and such that $|\lambda' g_n(x,\theta)|$ is bounded over $(\lambda,\theta) \in \Lambda_n \times \Theta$. Our estimator is therefore defined by:

$$\begin{aligned} \bar{T}(Q) &= \arg\max_{\theta\in\Theta} \frac{\displaystyle\int \exp\{\bar{T}_1(Q,\theta)' g_n(x,\theta)/2\} dQ}{\left(\displaystyle\int \exp\{\bar{T}_1(Q,\theta)' g_n(x,\theta)/2\} dQ\right)^{1/2}} \\ \bar{T}_1(Q,\theta) &= \arg\max_{\lambda\in\Lambda_n} -\int \exp\{\lambda' g_n(x,\theta)\} dQ. \end{aligned} \tag{16}$$

We consider $\Lambda_n = \left\{\lambda \in R^m : |\lambda| \leq C/m_n^{1+\zeta}\right\}$, for some constant $C > 0$. The alternative estimator functional in (16) does not depart substantially from that in(15) since if we can show that $\bar{T}_1(Q,\bar{T}(Q)) = O(n^{-1/2})$ and, imposing that m_n does not grow to infinity as fast as $n^{1/2}$ makes $\bar{T}_1$ an interior solution of Λ_n, for n large enough.

It is worthwhile to mention that $\bar{T}$ in (16) is well-defined in the neighborhood of P_*, the true probability distribution. Indeed, under Assumption 1(v) and for n large enough, $\lambda \to \int \exp(\lambda' g_n(x,\theta)) dQ$ is strictly convex over Λ_n for any Q in the Hellinger ball $B_H(P_*, r/\sqrt{n})$, $r > 0$. Hence, by convexity of Λ_n, $\bar{T}_1(Q,\theta)$ exists and is unique for all $\theta \in \Theta$ and $Q \in B_H(P_*, r/\sqrt{n})$. The existence of $\bar{T}(Q)$ is guaranteed by the continuity of the map $\theta \to \frac{\int \exp\{\bar{T}_1(Q,\theta)' g_n(x,\theta)/2\} dQ}{(\int \exp\{\bar{T}_1(Q,\theta)' g_n(x,\theta)\} dQ)^{1/2}}$ for each Q. Lemma C.2 in Appendix establishes the upper-hemicontinuity of $\bar{T}(Q)$ at each Q in a neighborhood of P_*.

Let $\tau : \Theta \to \mathbb{R}$ be a possibly nonlinear transformation of the parameter that is differentiable in a neighborhood of θ^*. Following KOE, we will

(1) investigate the behavior of the bias term $\tau \circ T(Q) - \tau(\theta^*)$ in a ($\sqrt{n}$-shrinking) Hellinger ball with radius $r > 0$ around P_*:

$$B_H(P_*, r/\sqrt{n}) = \{Q \in \mathcal{M} : H(Q, P_*) \leq r/\sqrt{n}\}.$$

(2) investigate the mean squared error (MSE) of the ETHD estimator in the set

$$\bar{B}_H(P_*, r/\sqrt{n}) = B_H(P_*, r/\sqrt{n}) \cap \left\{ Q \in \mathcal{M} : E_Q \left(\sup_{\theta\in\Theta} |g(x,\theta)|^\alpha < \infty \right) \right\}.$$

(3) investigate the robustness of ETHD with respect to a general class of loss functions.

We make the following regularity assumptions:

Assumption 3 *(Regularity assumptions 3)*
(i) There exists a neighborhood $\mathcal{N}$ of θ^* such that $g(x,\theta)$ is twice continuously differentiable a.s. on $\mathcal{N}$, $E_{P_*}(\sup_{\theta\in\mathcal{N}} |g(x,\theta)|^4) < \infty$, $\sup_{x\in\mathcal{X}_n,\theta\in\mathcal{N}} |\partial g(x,\theta)/\partial\theta'| = o(n^{1/2})$ and $\sup_{x\in\mathcal{X}_n,\theta\in\mathcal{N},1\le k\le m} |\partial^2 g_k(x,\theta)/\partial\theta\partial\theta'| = o(n)$.
(ii) $\{m_n\}_{n\ge 0}$ satisfies $m_n \to \infty$, $nm_n^{-\alpha} \to 0$, and $n^{-1/2}m_n^{1+\epsilon} = O(1)$ for some $0 < \epsilon < 2$ as $n \to \infty$ and $0 < \zeta < \epsilon$.
(iii) τ is continuously differentiable at θ^*.

Theorem 4.1 *Under Assumptions 1, 2 (with expectation and variance taken under P_*), and Assumption 3, the mapping $\bar{T}$ is Fisher consistent and satisfies:*

$$\lim_{n\to\infty} \sup_{Q\in B_H(P_*, r/\sqrt{n})} n(\tau\circ\bar{T}(Q) - \tau(\theta^*))^2 = 4r^2 B^*, \tag{17}$$

for each $r > 0$, where $B^* = \left(\frac{\partial\tau(\theta^*)}{\partial\theta}\right)' \Sigma \left(\frac{\partial\tau(\theta^*)}{\partial\theta}\right)$.

This is an analogue of Theorem 3.1(ii) of KOE for the ETHD estimator. It establishes that the worst bias of $\bar{T}(Q)$ with Q lying in a Hellinger neighborhood of P_* reaches the lower bound derived by KOE for any Fisher consistent estimator.

Theorem 4.2 *Under Assumption 1(i)-(iii), (v), Assumption 2 (with expectation and variance taken under P_*), and Assumption 3, the mapping T is Fisher consistent and regular, and the ETHD estimator, $\hat{\theta} = T(P_n)$, satisfies:*

$$\lim_{b\to\infty}\lim_{n\to\infty} \sup_{Q\in\bar{B}_H(P_*, r/\sqrt{n})} \int b \wedge n(\tau\circ T(P_n) - \tau(\theta_0))^2 dQ^{\otimes n} = (1+4r^2)B^*,$$

for each $r > 0$.

Assumption 4 *The loss function $\ell : \bar{\mathbb{R}}^p \to [0,\infty]$ is (i) symmetric subconvex (i.e., for all $z \in \mathbb{R}^p$ and $c \in \mathbb{R}$, $\ell(z) = \ell(-z)$ and $\{z \in \mathbb{R}^p : \ell(z) \le c\}$ is convex); (ii) upper semicontinuous at infinity; and (iii) continuous on $\bar{\mathbb{R}}^p$.*

Theorem 4.3 *Under Assumption 1(i)-(iii), (v), Assumption 2, with expectation and variance they contain taken under P_*, Assumption 3, and 4, the mapping T is Fisher consistent and the ETHD estimator, $\hat{\theta} = T(P_n)$, satisfies:*

$$\lim_{b\to\infty}\lim_{r\to\infty}\lim_{n\to\infty} \sup_{Q\in\bar{B}_H(P_*, r/\sqrt{n})} \int b \wedge \ell\left(\sqrt{n}(\tau\circ T(P_n) - \tau\circ T(Q)\right) dQ^{\otimes n} = \int \ell dN(0, B^*).$$

for each $r > 0$.

Theorems 4.2 and 4.3 are the analogues for ETHD estimator of Theorems 3.2(ii) and 3.3(ii) of KOE. Theorem 4.2 derives the worst possible MSE of ETHD when the true distribution is in any $\sqrt{n}$-shrinking vicinity of the postulated distribution P_*. The derived quantity amounts to the minimum reachable by any Fisher consistent and regular estimator as obtained by Theorem 3.2(ii) of KOE. This result makes ETHD locally asymptotically minimax optimal in terms of MSE just as is MHDE.

Theorem 4.3 extends Theorem 4.2 to a more general class of loss functions. Theorem 3.3(i) of KOE establishes the mean loss of $N(0, B^*)$ as the smallest maximum loss of any Fisher consistent estimator in large samples. We establish here that ETHD does not incur more loss asymptotically making once more this estimator asymptotically minimax optimal with respect to these loss functions. This result makes ETHD equivalent to MHDE in that respect.

5 ETHD under global misspecification

Our preliminary simulation study (see section 6.1 below) reveals that HD is much more affected by global misspecification than other estimators such as the one proposed in this paper ETHD, and standard estimators (e.g. GMM, ET, and ETEL).

We now derive the asymptotic distribution of ETHD under global misspecification, and under the following regularity assumptions.

Assumption 5 *(Regularity conditions under global misspecification)*
(i) x_i forms an i.i.d. sequence;
(ii) The objective function $\Delta_n(\theta, \lambda(\theta))$ is maximized at a unique "pseudo-true" value θ^* with $\theta^* \in int(\Theta)$ and Θ compact; in addition, $\lambda^* \equiv \lambda(\theta^*)$;
(iii) $g(x, \theta)$ is twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ^*;
(iv) $E\left(\sup_{\theta\in\Theta,\lambda\in\Lambda} \exp(\lambda' g(x, \theta))\right) < \infty$ where Λ is a compact subset of $\mathbb{R}^m$ such that $\lambda^* \in int(\Lambda)$;
(v) $\inf_{\theta\in\Theta} |E[g(x_i, \theta)]| \neq 0$.

Theorem 5.1 *(Asymptotics under global misspecification)*
Under regularity assumption 5, we have

$$\sqrt{n}\begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\lambda} - \lambda^* \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, R^{-1}\Omega^* R^{-1}).$$

when R and Ω^* are explicitly defined in the appendix in the proof.

As discussed in section 2.2, an estimator is said to be robust to global misspecification when its asymptotic distribution derived under global misspecification coincides with its asymptotic distribution under correct specification. In our case, it remains to show that the above asymptotic variance corresponds to the one of Theorem 3.3. With the explicit formulas provided in the appendix, it is straightforward to show such result.

(To be completed.)

6 Monte-Carlo simulations

We now present the results of our Monte-Carlo simulations that illustrate the finite sample properties of the different estimators considered in this paper.

6.1 Study under correct specification and global misspecification

We use the experimental design suggested in Schennach (2007), where we wish to estimate the mean while imposing a known variance. The two moment conditions are

$$g(x_i, \theta) = \left[x_i - \theta \quad (x_i - \theta)^2 - 1 \right]' ,$$

where x_i is drawn from either a correctly specified model C, or a misspecified model M,

$$\begin{array}{ll} x_i \sim \mathcal{N}(0,1) & \text{(for model C)} \\ x_i \sim \mathcal{N}(0,s^2) & \text{(with } s \neq 1 \text{ for model M)} \end{array}$$

As explained in Schennach (2007), this experiment is carefully designed such that the pseudo-true value ($\theta^* = 0$) for the misspecified model is the same for the estimators of interest, thus enabling a meaningful comparison of their variances.
In table 1, we compute the standard deviations of the HD, EL, ET, ETEL, and ETHD estimators of θ for a sample of size 1,000 and a sample of size 5,000, evaluated with 5,000 replications. Under correct specification, all the estimators perform equally well. This is not the case under global misspecification. The variability of HD is clearly larger than that of the other estimators. In addition, the standard deviation of HD does not decrease much when the sample size increases. By contrast, ETHD displays a much lower variance that decreases when the sample size increases for 1,000 to 5,000: it may not be shrinking exactly by the expected factor of $\sqrt{5}$, but much closely it is much better behaved than HD. Figure 1 shows the ratio of standard deviations for sample sizes 1,000 and 5,000 over a grid of misspecification parameters s.

6.2 Study under local misspecification

We use the experimental design suggested in KOE to explore the robustness of estimators to local misspecification. Consider $x = (x_1,\ x_2)' \sim \mathcal{N}(0, 0.4^2 I_2)$. This normal law corresponds to the true DGP P_0. The associated moment condition g is

$$g(x,\theta) = [\exp(-0.72 - \theta(x_1 + x_2) + 3x_2) - 1] \begin{pmatrix} 1 \\ x_2 \end{pmatrix} .$$

The moment condition is uniquely solved at $\theta_0 = 3$. The goal is to estimate this value using the above specification of g from contaminated data where we use

$$x \sim \mathcal{N}(0, \Sigma_{(\delta,\rho)} \quad \text{with} \quad \Sigma_{(\delta,\rho)} = 0.4^2 \begin{pmatrix} (1+\delta)^2 & \rho(1+\delta) \\ \rho(1+\delta) & 1 \end{pmatrix} .$$

The unperturbed case corresponds to $\delta = \rho = 0$. In the simulation, we set $\rho = 0.1\sqrt{2}\cos(2\pi w)$ and $\delta = 0.1\sin(2\pi w)$ where we let w vary over $w_j = 1/64$ with $j = 0, 1, \cdots, 63$. This yields 64 different designs and, for each of them, 5,000 replications are performed. We consider the following estimator: ETHD, ET, EL, and HD.
Figure 3 shows the RMSE and $Pr\left(\left|\hat{\theta} - \theta_0\right| > 0.5\right)$ for several estimators of interest, while the bias is displayed in figure 4. We see that GMM is affected by perturbations much more than EL, HD, ET and ETHD, except for the values of w between 0.4 and 0.6. The performance of the other four estimators are rather close. In particular, our ETHD estimator remains well-behaved

throughout the simulation designs, whereas ET seems to perform a little worse than ETHD, HD and EL.

Figure 5 compares the performance of two version of the HD estimator (as discussed in section 2): the estimator HD and the estimator HD-unc where the constraint maintaining the positivity of the implied probability is relaxed. Overall, the performance of both estimators are very close to each other; however, it is interesting to point out for the values of w between 0.5 and 0.7, HD is closer to ETHD, and performs slightly better than HD-unc which is closer to EL.

A Results of the Monte-Carlo study

A.1 Study under correct specification and global misspecification

Model C with s = 1.0						
	GMM	HD	EL	ET	ETEL	ETHD
Sample size T=1000	0.0320	0.0320	0.0320	0.0320	0.0320	0.0320
Sample size T=5000	0.0139	0.0139	0.0139	0.0139	0.0139	0.0139
Model M with s = 0.75						
	GMM	HD	EL	ET	ETEL	ETHD
Sample size T=1000	0.0486	0.0485	0.0748	0.0332	0.0466	0.0409
Sample size T=5000	0.0212	0.0377	0.0732	0.0151	0.0257	0.0216

Table 1: Standard deviations of the GMM, HD, ET, EL, ETEL, ETHD estimators for models C and M with 5,000 replications

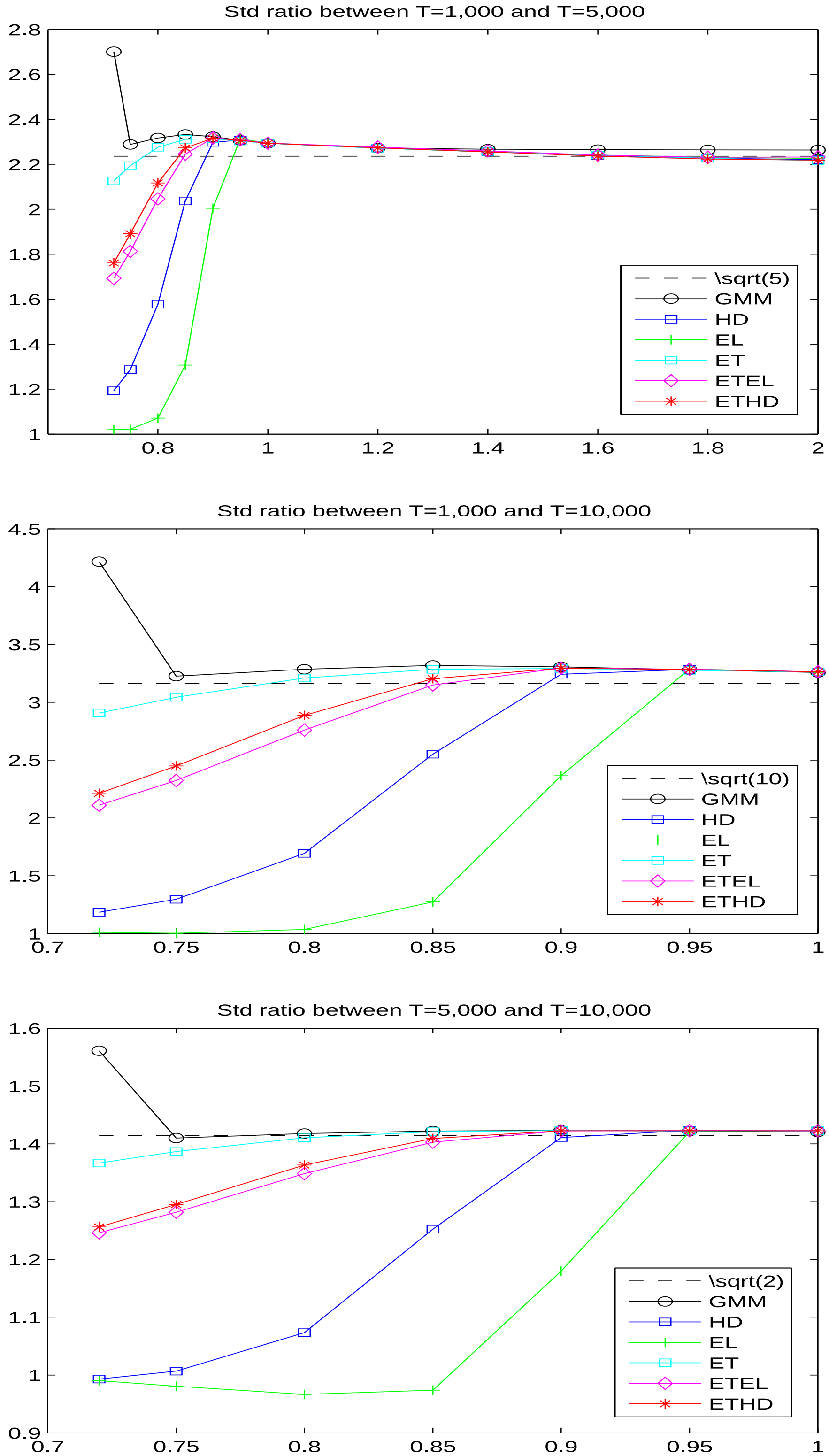

Figure 1: Ratio of standard deviations for sample sizes 1,000 - 5,000; 1,000 - 10,000 and 5,000 - 10,000 over a grid of misspecification parameters s

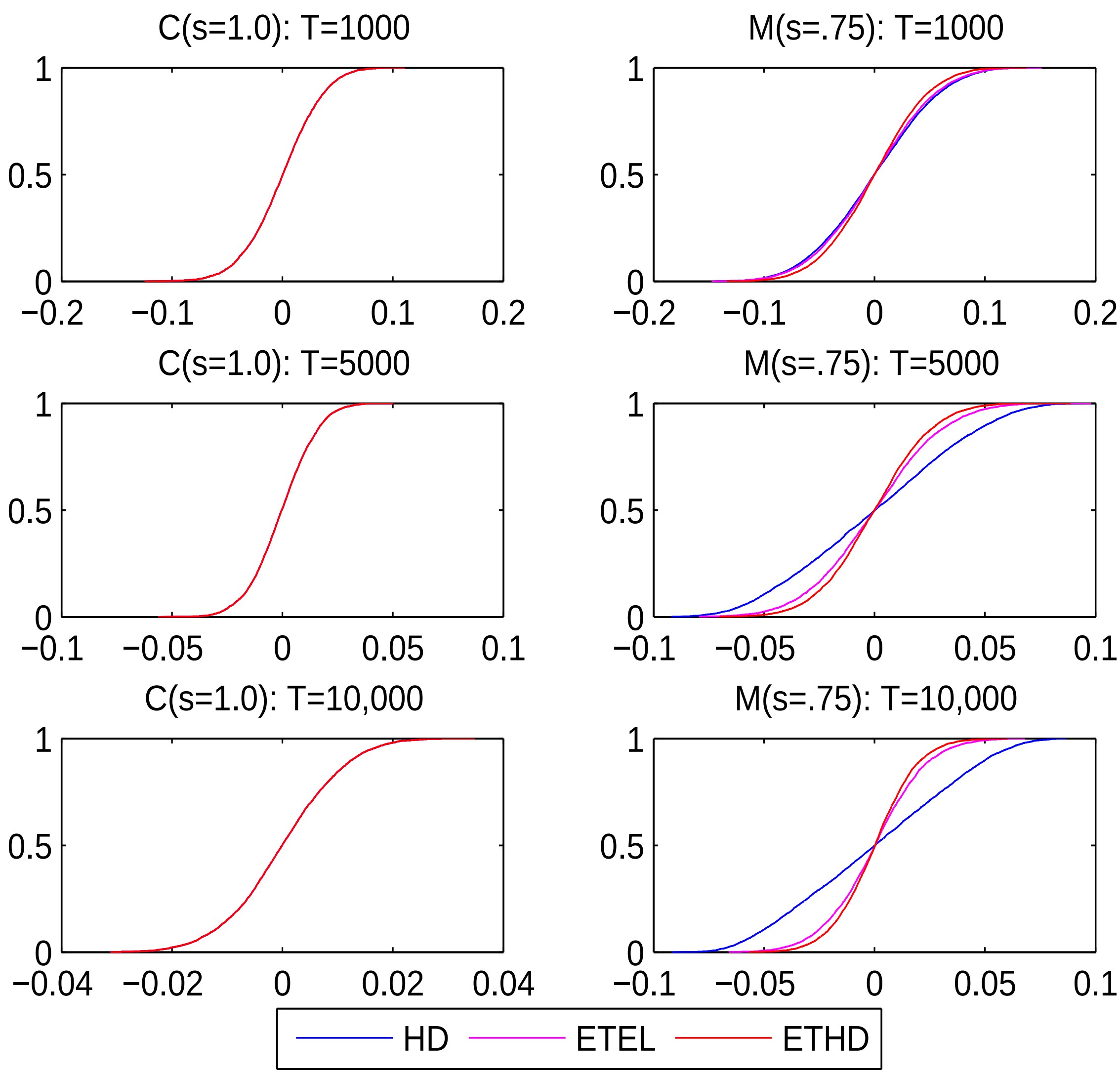

Figure 2: Simulated cumulative distribution of HD, ETEL and ETHD under correct specification (C(s=1.0)) and global misspecification (M(s=.75))

A.2 Study under local misspecification

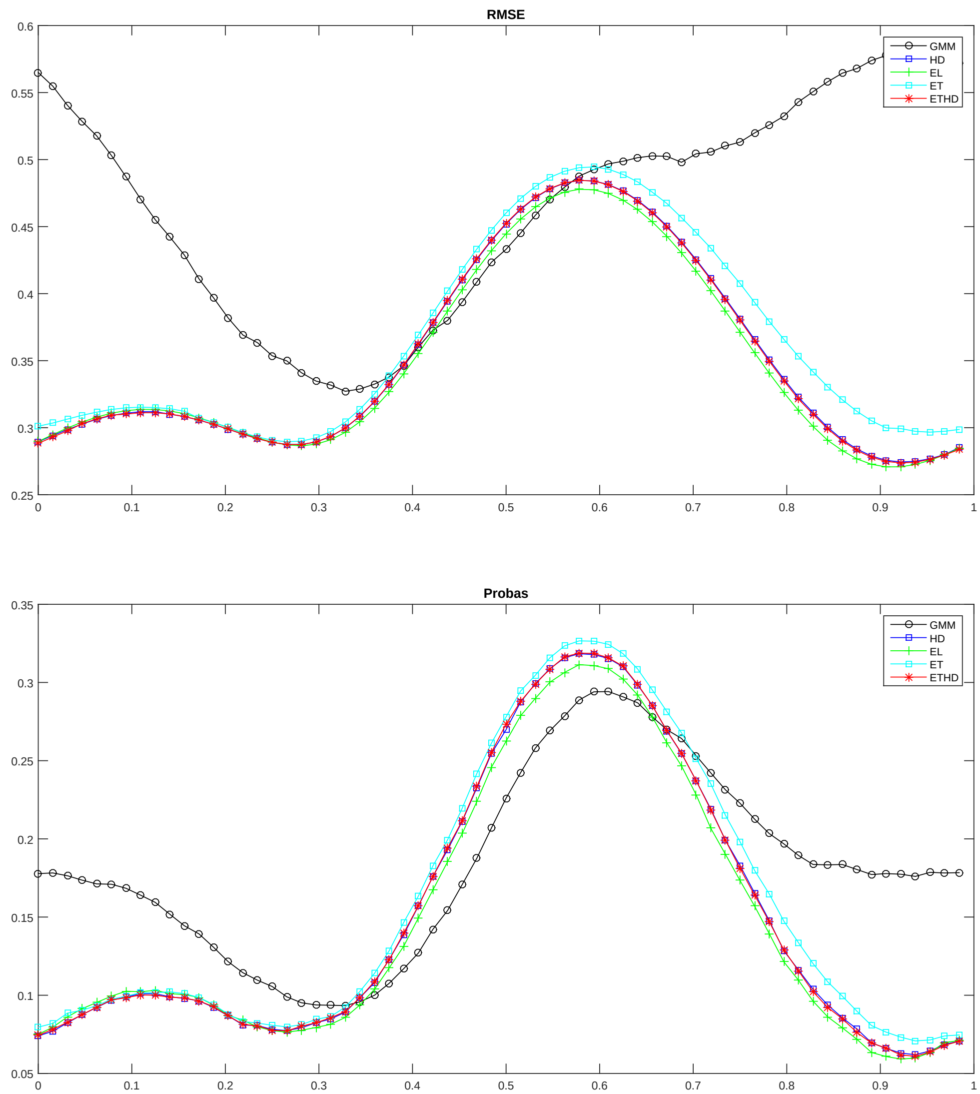

Figure 3: Local neighborhood of the true model: RMSE (top); Probas (bottom) denotes $Pr\left(\left|\hat{\theta}-\theta_0\right|>0.5\right)$.

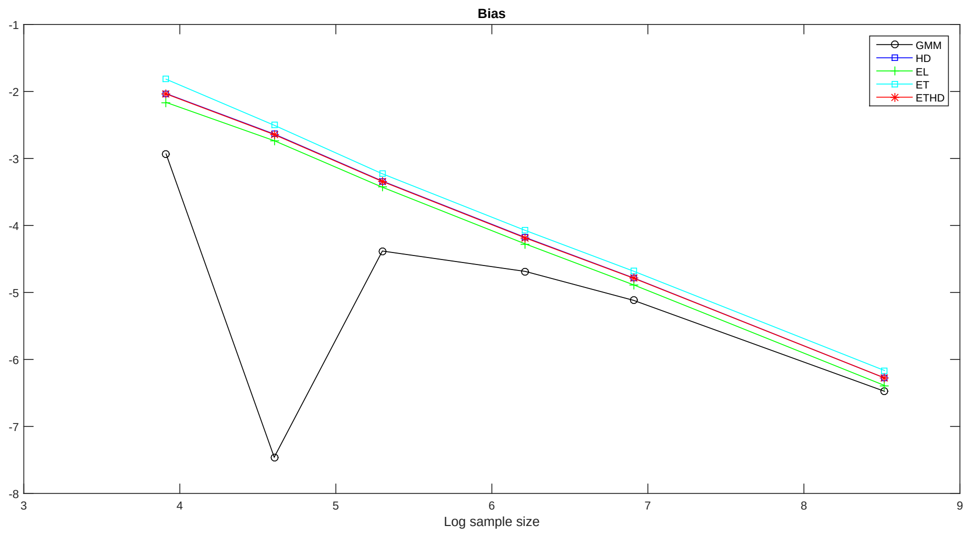

Figure 4: Local neighborhood of the true model: Bias.

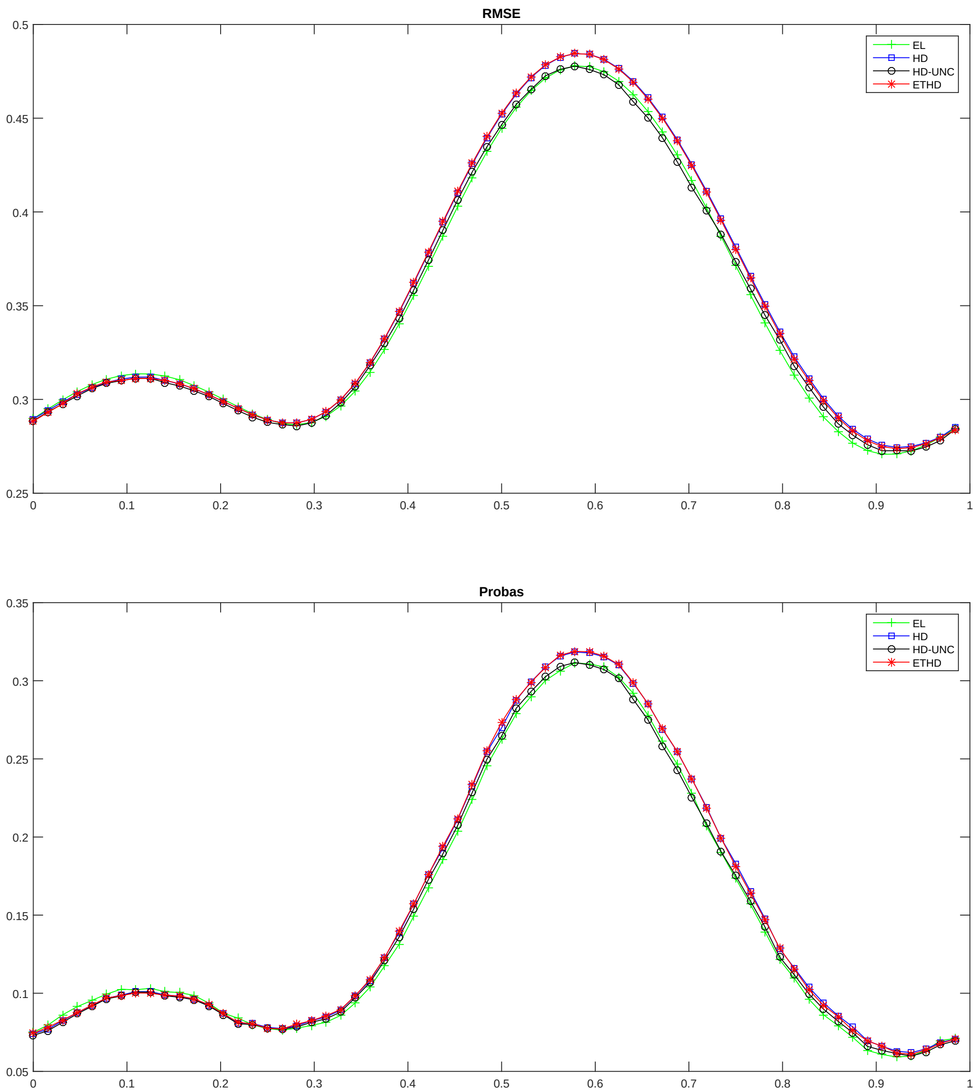

Figure 5: Local neighborhood of the true model: RMSE (top); Probas (bottom) denotes $Pr\left(\left|\hat{\theta}-\theta_0\right|>0.5\right)$. The estimator D-unc denotes the estimator HD with the additional positivity constraint for the implied probabilities as discussed in section 2.

B Proofs of the theoretical results

Proof of Theorem 2.1: Our proof closely follows the steps of the proof of Theorem 1 in Schennach (2007).
We start from the interpretation of HD estimator as a GEL estimator (see Newey and Smith (2004)) and KOE (2013, p1191).

$$\hat{\theta}_{HD} = \arg\min_{\theta}\max_{\gamma} -\frac{1}{n}\sum_{i=1}^{n}\frac{2}{(1-\gamma' g(x_i,\theta)/2)}\,.$$

The FOC wrt θ and γ write, respectively:

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n}\frac{\hat{G}_i'\hat{\gamma}}{[1-\hat{\gamma}' g(x_i,\hat{\theta})/2]^2} = 0 \quad \text{where} \quad \hat{G}_i = \frac{\partial g(x_i,\hat{\theta})}{\partial\theta'}$$
$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n}\frac{g(x_i,\hat{\theta})}{[1-\hat{\gamma}' g(x_i,\hat{\theta})/2]^2} = 0\,.$$

The asymptotic properties of GEL-type estimators are well known:

$$\sqrt{n}\left[\begin{pmatrix}\hat{\theta}\\ \hat{\gamma}\end{pmatrix} - \begin{pmatrix}\theta^*\\ \gamma^*\end{pmatrix}\right] \xrightarrow{d} \mathcal{N}(0, H_k^{-1}S_k H_k^{-1})$$

with

$$S_k = E[\phi(\theta^*,\gamma^*)\phi(\theta^*,\gamma^*)'] = \begin{pmatrix} E[\tau_i^4 G_i'\gamma\gamma' G_i] & E[\tau_i^4 G_i'\gamma g_i'] \\ E[\tau_i^4 g_i\gamma' G_i] & E[\tau_i^4 g_i g_i'] \end{pmatrix}$$

and

$$\tau_i = \frac{1}{1-\gamma' g_i/2}$$
$$\phi(\theta,\gamma) = \begin{pmatrix} \frac{G_i'\gamma}{(1-\gamma' g_i/2)^2} \\ \frac{g_i}{(1-\gamma' g_i/2)^2} \end{pmatrix}$$
$$H_k = E\left(\frac{\partial\phi'(\theta^*,\gamma^*)}{\partial[\theta'\ \gamma']'}\right) = E\begin{pmatrix} \tau_i^3 G_i\gamma\gamma' G_i + \tau_i^2\frac{\partial(G_i'\gamma)}{\partial\theta'} & \tau_i^3 G_i'\gamma g_i' + \tau_i^2 G_i' \\ \tau_i^3 g_i\gamma' G_i + \tau_i^2 G_i & \tau_i^3 g_i g_i' \end{pmatrix}$$

From the calculations in the dual problem, we have:

$$\sqrt{\pi}_i = \frac{1}{\sqrt{n}(1-\gamma' g_i/2)} > 0 \;\Rightarrow\; \frac{1}{(1-\gamma' g_i/2)} > 0 \tag{18}$$

Since $\{g(x,\theta_k^*), x\in\mathcal{X}\}$ is unbounded in every direction, the set $\{g(x,\theta_k^*)\in C_k\}$ becomes unbounded in every direction as $k\to\infty$, Hence, the only way to have (18) is to have $\gamma_k^*\to 0$ as $k\to\infty$. Since $\gamma_k^*\to 0$ as $k\to\infty$, S_k and H_k can be simplified by noting that when $(H_k^{-1}S_kH_k^{-1})$ is calculated: any term containing γ_k^* will be dominated by terms not containing it. We get:

$$S_k \to \begin{pmatrix} 0 & 0 \\ 0 & E(\tau_i^4 g_i g_i') \end{pmatrix}$$

and

$$H_k^{-1} \to \begin{pmatrix} 0 & E(\tau_i^2 G_i') \\ E(\tau_i^2 G_i) & E(\tau_i^3 g_i g_i') \end{pmatrix}^{-1} \equiv \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

Define Σ_k as the (p, p) top-left submatrix of $(H_k^{-1} S_k H_k^{-1})$, that is

$$\Sigma_k = B_{12} E(\tau_i^4 g_i g_i') B_{21}$$

Recall $\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1}$ top-right corner term is $-F^{-1}BD^{-1}$ with $F = A - BD^{-1}C$. Thus:

$$B_{12} = \left[E(\tau_i^2 G_i') \left(E(\tau_i^3 g_i g_i') \right)^{-1} E(\tau_i^2 G_i) \right]^{-1} E(\tau_i^2 G_i') \left(E(\tau_i^3 g_i g_i') \right)^{-1} = B_{21}'$$

To show that Σ_k diverges, we show the following three properties:

(i) $E(\tau_i^4 g_i g_i')$ has a divergent eigenvalue;

(ii) $|E(\tau_i^2 G_i)| = o\left(\left[E(\tau_i^4 |g_i g_i'|) \right]^{1/2} \right)$;

(iii) $|B_{12}| \left[E(\tau_i^4 |g_i g_i'|) \right]^{1/2}$ diverges.

(i) First, we show that $E(\tau_i^4 g_i g_i')$ has a divergent eigenvalue:

$$\begin{aligned} g_i(1 - \gamma' g_i/2)^2 &= g_i(1 - \gamma' g_i + (\gamma' g_i)^2/4) \\ &= g_i - g_i g_i' \gamma + g_i g_i' \gamma g_i' \gamma/4 \\ &= g_i - g_i g_i' \gamma/2(2 - g_i' \gamma/2) \\ &= g_i - g_i g_i' \gamma/2 - g_i g_i' \gamma/2(1 - g_i' \gamma/2) \end{aligned}$$

$$\begin{aligned} \Rightarrow \quad & g_i = \frac{g_i}{(1 - \gamma' g_i/2)^2} - \frac{g_i(g_i'\gamma)/2}{(1 - \gamma' g_i/2)^2} - \frac{g_i(g_i'\gamma)/2}{(1 - \gamma' g_i/2)} \\ \Rightarrow \quad & E(g_i) = 0 - \left\{ e\left[\frac{g_i g_i'}{(1 - \gamma' g_i/2)^2} \right] + E\left[\frac{g_i g_i'}{(1 - \gamma' g_i/2)} \right] \right\} \frac{\gamma}{2} \\ \Rightarrow \quad & E(g_i) \equiv -(\Omega_1 + \Omega_2) \frac{\gamma}{2} \end{aligned}$$

Since $\inf_{k \geq \bar{k}} E(g(x_i, \theta_k^*)) > 0$ some $\bar{k} \in \mathbb{N}$, the only way to have $\gamma_k \to 0$ is if $(\Omega_1 + \Omega_2)$ has a divergent eigenvalue. Let v be a unit eigenvector associated with such eigenvalue:

$$\begin{aligned} v'\Omega_1 v &= E\left(\frac{v' g_i}{(1 - \gamma' g_i/2)^2} v' g_i \right) \leq \left[E\left(\frac{v' g_i}{(1 - \gamma' g_i/2)^2} \right)^2 \right]^{1/2} \\ v'\Omega_2 v &= E\left(\frac{v' g_i}{(1 - \gamma' g_i/2)} v' g_i \right) \leq \left[E\left(\frac{v' g_i}{(1 - \gamma' g_i/2)} \right)^2 \right]^{1/2} = (v'\Omega_1 v)^{1/2} [E(v' g_i)^2]^{1/2} \end{aligned}$$

Hence,

$$v'\Omega v \equiv v'\Omega_1 v + v'\Omega_2 v \leq [E(v' g_i)^2]^{1/2} \left\{ \left[E\left(\frac{v' g_i}{(1 - \gamma' g_i/2)^2} \right)^2 \right]^{1/2} + (v'\Omega_1 v)^{1/2} \right\}$$

Since

a) $\quad E(v' g(x_i, \theta_k^*))^2 \leq \sup_{\theta \in \Theta} E|g(x_i, \theta)|^2 < \infty$ by assumption,

b) $\quad v'\Omega_1 v \leq \left[E\left(\frac{v' g_i}{(1 - \gamma' g_i/2)^2} \right)^2 E(v' g_i)^2 \right]^{1/2}$ diverges as shown above,

c) $\quad E\left(\frac{v' g_i}{(1 - \gamma' g_i/2)^2} \right)^2 = E[\tau_i^4 (v' g_i)^2]$, we conclude that $E(\tau_i^4 g_i g_i')$ has a divergent eigenvalue.

(ii) We now show that $|E(\tau_i^2 G_i)| = o\left(\left[E(\tau_i^4|g_i g_i'|)\right]^{1/2}\right)$.

$$\begin{aligned}
\tau_i^2 G_i &= \frac{1}{(1-\gamma' g_i/2)^2} G_i = \left[1 + \tau_i^2 \gamma' g_i - \tau_i^2 \left(\frac{\gamma' g_i}{2}\right)^2\right]^2 G_i \\
|E(\tau_i^2 G_i)| &= |E\left[(1 + \tau_i^2 \gamma' g_i - \tau_i^2 \left(\frac{\gamma' g_i}{2}\right)^2) G_i\right]| \\
&\leq E|G_i| + E|\tau_i^2 \gamma' g_i G_i| + E|\tau_i^2 \left(\frac{\gamma' g_i}{2}\right)^2 G_i| \\
E\tau_i^2 |\gamma' g_i G_i| &= E\left(\tau_i^2 |g_i||G_i|\right)|\gamma| \\
&\leq \left[E(\tau_i^4 |g_i|^2)\right]^{1/2} \left[|G_i|^2\right]^{1/2} |\gamma|
\end{aligned}$$

where the last inequality follows from CS. Then,

$$\frac{E(\tau_i^2|\gamma' g_i G_i|)}{[E(\tau_i^4|g_i|^2)]^{1/2}} \to 0 \quad \Rightarrow \quad E(\tau_i^2|\gamma' g_i G_i|) = o\left([E(\tau_i^4|g_i|^2)]^{1/2}\right)$$

$$E|\tau_i^2 \left(\frac{\gamma' g_i}{2}\right)^2 G_i| \;=\; E(\tau_i^2|g_i|^2|G_i|)|\gamma|^2 \;\leq\; \left[E(\tau_i^4|g_i|^2)\right]^{1/2}\left[E|g_i|^2|G_i|^2\right]^{1/2}|\gamma|^2$$

$$\Rightarrow \quad |E(\tau_i^2 G_i)| = o\left(\left[E(\tau_i^4|g_i|^2)\right]^{1/2}\right) \;=\; o\left(\left[E(\tau_i^4|g_i g_i'|)\right]^{1/2}\right) \;=\; o\left(\left[E(\tau_i^4 v' g_i g_i' v)\right]^{1/2}\right)$$

(iii) Finally, we show that $|B_{12}|\left[E(\tau_i^4|g_i g_i'|)\right]^{1/2} \to \infty$.
First, it follows from CS that:

$$|B_{12}|\left[E(\tau_i^4|g_i g_i'|)\right]^{1/2} \geq |B_{12}E(\tau_i^2 G_i)| \frac{\left[E(\tau_i^4|g_i g_i'|)\right]^{1/2}}{|E(\tau_i^2 G_i)|}$$

Then, from the definition of B_{12}, we have:

$$B_{12}E(\tau_i^2 G_i) = I_p \quad \Rightarrow \quad |B_{12}E(\tau_i^2 G_i)| = O_p(1)$$

Finally, we showed in (ii) above that

$$\begin{aligned}
|E(\tau_i^2 G_i)| = o\left(\left[E(\tau_i^4|g_i g_i'|)\right]^{1/2}\right) \quad &\Rightarrow \quad \frac{|E(\tau_i^2 G_i)|}{\left[E(\tau_i^4|g_i g_i'|)\right]^{1/2}} \to 0 \\
&\Rightarrow \quad \frac{\left[E(\tau_i^4|g_i g_i'|)\right]^{1/2}}{|E(\tau_i^2 G_i)|} \to \infty
\end{aligned}$$

The rest of the proof follows from the proof of Theorem 1 in Schennach (2007). □

Proof of Theorem 3.1: To simplify the notation, we make the dependence of all quantities on $\hat{\theta}$ implicit and introduce the following notations: $\hat{\pi}_i = \hat{\pi}_i(\hat{\theta})$, $\hat{\lambda} = \hat{\lambda}(\hat{\theta})$, $g_i = g_i(x, \hat{\theta})$. In addition, $\sum_i = \sum_{i=1}^n$.

Let us start with the following preliminary computation:

$$
\begin{aligned}
\frac{d\hat{\pi}_i}{d\theta} &= \frac{d}{d\theta}\left[\frac{\exp(\hat{\lambda}' g_i)}{\sum_j \exp(\hat{\lambda}' g_j)}\right] \\
&= \frac{1}{\left[\sum_j \exp(\hat{\lambda}' g_j)\right]^2}\left[\frac{d(\exp(\hat{\lambda}' g_i)))}{d\theta}\sum_j \exp(\hat{\lambda}' g_j) - \exp(\hat{\lambda}' g_i)\sum_j \frac{d}{d\theta}\exp(\hat{\lambda}' g_j)\right] \\
&= \frac{1}{\left[\sum_j \exp(\hat{\lambda}' g_j)\right]^2}\left[\frac{d(\hat{\lambda}' g_i)}{d\theta}\sum_j \exp(\hat{\lambda}' g_i)\sum_j \exp(\hat{\lambda}' g_i) - \exp(\hat{\lambda}' g_i)\sum_j \frac{d(\hat{\lambda}' g_j)}{d\theta}\exp(\hat{\lambda}' g_j)\right] \\
&= \frac{\exp(\hat{\lambda}' g_i)}{\sum_j \exp(\hat{\lambda}' g_j) \times \sum_k \exp(\hat{\lambda}' g_k)}\left[\frac{d(\hat{\lambda}' g_i)}{d\theta}\sum_j \exp(\hat{\lambda}' g_j) - \sum_j \frac{d(\hat{\lambda}' g_j)}{d\theta}\exp(\hat{\lambda}' g_j)\right] \\
&= \hat{\pi}_i\left[\frac{d(\hat{\lambda}' g_i)}{d\theta} - \sum_j \frac{d(\hat{\lambda}' g_j)}{d\theta}\frac{\exp(\hat{\lambda}' g_j)}{\sum_k \exp(\hat{\lambda}' g_k)}\right] \\
&= \hat{\pi}_i\left[\frac{d(\hat{\lambda}' g_i)}{d\theta} - \sum_j \hat{\pi}_j\frac{d(\hat{\lambda}' g_j)}{d\theta}\right]
\end{aligned}
$$

We can now proceed from

$$
H^2(\hat{\pi}, P_n) = 2 - \frac{2}{\sqrt{n}}\sum_i \sqrt{\hat{\pi}_i}
$$

The differentiation wrt θ gives:

$$
\begin{aligned}
\frac{dH^2}{d\theta} &= -\frac{1}{\sqrt{n}}\sum_i \frac{d\hat{\pi}_i}{d\theta}\hat{\pi}_i^{-1/2} \\
&= -\frac{1}{\sqrt{n}}\sum_i \left(\sqrt{\hat{\pi}_i}\frac{d(\hat{\lambda}' g_i)}{d\theta} - \sqrt{\hat{\pi}_i}\sum_j \frac{d(\hat{\lambda}' g_j)}{d\theta}\hat{\pi}_j\right) \\
&= \frac{1}{\sqrt{n}}\sum_i \sqrt{\hat{\pi}_i}\sum_j \frac{d(\hat{\lambda}' g_j)}{d\theta}\hat{\pi}_j - \frac{1}{\sqrt{n}}\sum_i \sqrt{\hat{\pi}_i}\frac{d(\hat{\lambda}' g_i)}{d\theta} \\
&= 0
\end{aligned}
$$

From (10), the FOC for $\hat{\lambda}$ is:

$$
\sum_i g_i \exp(\hat{\lambda}' g_i) = 0\,.
$$

□

Lemma B.1 *Let $(\hat{\lambda}, \hat{\theta})$ be a random sequence of $\Lambda \times \Theta$. If Assumption 1-(i), (ii), (iii), (v) and (vi) hold and*

$$
\frac{\int \exp\left(\hat{\lambda}' g(x, \hat{\theta})/2\right) dP_n}{\left(\int \exp\left(\hat{\lambda}' g(x, \hat{\theta})\right) dP_n\right)^{1/2}} \xrightarrow{P} 1,
$$

then $\hat{\lambda} \xrightarrow{P} 0$.

Proof of Lemma B.1: From Assumption 1-(i), (ii) and (vi),

$$\int \exp\left(\lambda' g(x,\theta)\right) dP_n \quad \text{and} \quad \int \exp\left(\lambda' g(x,\theta)/2\right) dP_n$$

converges in probability towards $E\left(\exp\left(\lambda' g(x,\theta)\right)\right)$ and $E\left(\exp\left(\lambda' g(x,\theta)/2\right)\right)$, uniformly over $\Lambda \times \Theta$, respectively.

In addition, $E\left(\exp\left(\lambda' g(x,\theta)\right)\right)$ is positive and continuous in (λ, θ). Hence, its minimum is reached on that set and is positive. Therefore, letting:

$$\hat{f}(\lambda,\theta) = \int \exp\left(\lambda' g(x,\theta)/2\right) dP_n; \quad \hat{k}(\lambda,\theta) = \left(\int \exp\left(\lambda' g(x,\theta)\right) dP_n\right)^{1/2},$$

$$f(\lambda,\theta) = E\left(\exp\left(\lambda' g(x,\theta)/2\right)\right); \quad \text{and} \quad k(\lambda,\theta) = \left(E\left(\exp\left(\lambda' g(x,\theta)\right)\right)\right)^{1/2},$$

we can use the fact that:

$$\sup_{\theta\in\Theta,\lambda\in\Lambda}\left|\frac{\hat{f}(\lambda,\theta)}{\hat{k}(\lambda,\theta)} - \frac{f(\lambda,\theta)}{k(\lambda,\theta)}\right| = \sup_{\theta\in\Theta,\lambda\in\Lambda}\left|\frac{(\hat{f}(\lambda,\theta)-f(\lambda,\theta))k(\lambda,\theta)-(\hat{k}(\lambda,\theta)-k(\lambda,\theta))f(\lambda,\theta)}{(\hat{k}(\lambda,\theta)-k(\lambda,\theta))k(\lambda,\theta)+k(\lambda,\theta)^2}\right|$$

to confirm that $\frac{\hat{f}(\lambda,\theta)}{\hat{k}(\lambda,\theta)}$ converges in probability towards $\frac{f(\lambda,\theta)}{k(\lambda,\theta)}$, uniformly over $\Lambda \times \Theta$. Thus,

$$\frac{\hat{f}(\hat{\lambda},\hat{\theta})}{\hat{k}(\hat{\lambda},\hat{\theta})} = \frac{f(\hat{\lambda},\hat{\theta})}{k(\hat{\lambda},\hat{\theta})} + o_P(1) \equiv \frac{E\left(\exp\left(\hat{\lambda}' g(x,\hat{\theta})/2\right)\right)}{\left(E\left(\exp\left(\hat{\lambda}' g(x,\hat{\theta})\right)\right)\right)^{1/2}} + o_P(1).$$

Hence,

$$\frac{E\left(\exp\left(\hat{\lambda}' g(x,\hat{\theta})/2\right)\right)}{\left(E\left(\exp\left(\hat{\lambda}' g(x,\hat{\theta})\right)\right)\right)^{1/2}} \to 1, \tag{19}$$

in probability as $n \to \infty$.

We now show that $\hat{\lambda} \xrightarrow{P} 0$. For this, let $\epsilon > 0$, $\mathcal{N} = \{\lambda : |\lambda| < \epsilon\}$ and $\bar{\mathcal{N}}$ its complement. Note that, by the Jensen's inequality, $\frac{E(\exp(\lambda' g(x,\theta)/2))}{(E(\exp(\lambda' g(x,\theta))))^{1/2}} \leq 1$ for all λ and θ with equality occurring only if $\exp\left(\lambda' g(x,\theta)\right)$ is constant. Under Assumption 1-(v), this is the case only when $\lambda = 0$.

By continuity of its objective function and the compactness of the maximization set, there exists $\bar{\theta} \in \Theta$ and $\bar{\lambda} \in \bar{\mathcal{N}} \cap \Lambda$ such that:

$$\max_{\theta\in\Theta,\lambda\in\Lambda\cap\bar{\mathcal{N}}} \frac{E\left(\exp\left(\lambda' g(x,\theta)/2\right)\right)}{\left(E\left(\exp\left(\lambda' g(x,\theta)\right)\right)\right)^{1/2}} = \frac{E\left(\exp\left(\bar{\lambda}' g(x,\bar{\theta})/2\right)\right)}{\left(E\left(\exp\left(\bar{\lambda}' g(x,\bar{\theta})\right)\right)\right)^{1/2}} \equiv A_\epsilon.$$

Since $\bar{\lambda} \neq 0$, $A_\epsilon < 1$. From (19), $\frac{E\left(\exp\left(\hat{\lambda}' g(x,\hat{\theta})/2\right)\right)}{\left(E\left(\exp\left(\hat{\lambda}' g(x,\hat{\theta})\right)\right)\right)^{1/2}} > A_\epsilon$ w.p.a.1 for n large enough. Therefore, $\hat{\lambda} \notin \bar{\mathcal{N}}$ w.p.a.1. Hence, $\hat{\lambda} \in \mathcal{N}$ w.p.a.1. This establishes that $\hat{\lambda} \xrightarrow{P} 0$. □

Lemma B.2 *If Assumption 1 holds and $\hat{\theta}$ is the ETHD estimator, then*

$$\Delta_n(\hat{\lambda}(\hat{\theta}),\hat{\theta}) = 1 + O_P(n^{-1}), \quad \hat{\lambda}(\hat{\theta}) = O_P(n^{-1/2}) \quad \textit{and} \quad \int g(x,\hat{\theta})dP_n = O_P(n^{-1/2}).$$

Proof of Lemma B.2: We proceed in three steps. In Step 1, we show that $\Delta_n(\hat{\lambda}(\hat{\theta}), \hat{\theta}) = 1 + O_P(n^{-1})$. This allows, thanks to Lemma B.1 to deduce that $\hat{\lambda}(\hat{\theta}) = o_P(1)$. In Step 2, we derive the order of magnitude of $\hat{\lambda}(\hat{\theta})$ and that of $\int g(x, \hat{\theta}) dP_n$ in Step 3.

Step 1: We first show that $\Delta_n(\hat{\lambda}(\hat{\theta}), \hat{\theta}) = 1 + O_P(n^{-1})$.

By definition of $\hat{\theta}$, we have:

$$\Delta_n\left(\hat{\lambda}(\theta^*), \theta^*\right) \leq \Delta_n\left(\hat{\lambda}(\hat{\theta}), \hat{\theta}\right) \leq 1. \tag{20}$$

Also, by the central limit theorem, $\sqrt{n} \int g(x, \theta^*) dP_n = O_P(1)$. We can therefore apply Lemma A2 of Newey and Smith (2004) to the constant sequence $\bar{\theta} = \theta^*$ and claim that $\hat{\lambda}(\theta^*) = O_P(n^{-1/2})$ and $\int \exp\left(\hat{\lambda}(\theta^*)' g(x, \theta^*)\right) dP_n \geq 1 + O_P(n^{-1})$.

Since $\int \exp\left(\lambda' g(x, \theta^*)\right) dP_n$ is minimized at $\hat{\lambda}(\theta^*)$ over $\Lambda \ni 0$, we can claim that:

$$1 + O_P(n^{-1}) \leq \int \exp\left(\hat{\lambda}(\theta^*)' g(x, \theta^*)\right) dP_n \leq 1.$$

Thus

$$\varepsilon_n \equiv \int \exp\left(\hat{\lambda}(\theta^*)' g(x, \theta^*)\right) dP_n - 1 = O_P(n^{-1}).$$

By definition of $\hat{\lambda}(\theta^*)$, $\int \exp\left(\hat{\lambda}(\theta^*)' g(x, \theta^*)\right) dP_n \leq \int \exp\left(\hat{\lambda}(\theta^*)' g(x, \theta^*)/2\right) dP_n$. Hence,

$$\left(\int \exp\left(\hat{\lambda}(\theta^*)' g(x, \theta^*)\right) dP_n\right)^{1/2} \leq \Delta_n\left(\hat{\lambda}(\theta^*), \theta^*\right) \leq 1.$$

But, $\left(\int \exp\left(\hat{\lambda}(\theta^*)' g(x, \theta^*)\right) dP_n\right)^{1/2} = (1 + \varepsilon_n)^{1/2} = 1 + \frac{1}{2}\varepsilon_n + O(\varepsilon_n^2) = 1 + O_P(n^{-1})$. We deduce that $\Delta_n\left(\hat{\lambda}(\theta^*), \theta^*\right) = 1 + O_P(n^{-1})$, ensuring also that $\Delta_n\left(\hat{\lambda}(\hat{\theta}), \hat{\theta}\right) = 1 + O_P(n^{-1})$ via (20).

Step 2: Let us show that $\hat{\lambda}(\hat{\theta}) = O_P(n^{-1/2})$. By a second order Taylor expansion of $\Delta_n(\hat{\lambda}(\hat{\theta}), \hat{\theta})$ around $\lambda = 0$ with a Lagrange remainder, we have:

$$\Delta_n(\hat{\lambda}(\hat{\theta}), \hat{\theta}) = \Delta_n(0, \hat{\theta}) + \frac{\partial \Delta_n(0, \hat{\theta})}{\partial \lambda'} \hat{\lambda}(\hat{\theta}) + \frac{1}{2} \hat{\lambda}(\hat{\theta})' \frac{\partial^2 \Delta_n(\dot{\lambda}, \hat{\theta})}{\partial \lambda \partial \lambda'} \hat{\lambda}(\hat{\theta}), \tag{21}$$

with $\dot{\lambda} \in (0,\, \hat{\lambda}(\hat{\theta}))$. Letting g denote $g(x, \hat{\theta})$ in the next two equations, we have:

$$\begin{aligned} \frac{\partial \Delta_n(\lambda, \hat{\theta})}{\partial \lambda} &= \frac{1}{2}\left(\int \exp(\lambda' g) dP_n\right)^{-3/2} \times \\ &\left(\int g \exp(\lambda' g/2) dP_n \int \exp(\lambda' g) dP_n - \int g \exp(\lambda' g) dP_n \int \exp(\lambda' g/2) dP_n\right) \end{aligned}$$

and

$$
\begin{aligned}
\frac{\partial^2\Delta_n(\lambda,\hat\theta)}{\partial\lambda\partial\lambda'} &= \tfrac{1}{2}\left(\int \exp(\lambda' g)dP_n\right)^{-3/2}\times \\
&\left(\tfrac{1}{2}\int gg'\exp(\lambda' g/2)dP_n\int \exp(\lambda' g)dP_n-\int gg'\exp(\lambda' g)dP_n\int \exp(\lambda' g/2)dP_n\right) \\
&+\tfrac{1}{2}\left(\int \exp(\lambda' g)dP_n\right)^{-5/2}\times \\
&\left(\tfrac{3}{2}\int \exp(\lambda' g/2)dP_n\int g\exp(\lambda' g)dP_n\int g'\exp(\lambda' g)dP_n\right. \\
&-\tfrac{1}{2}\int \exp(\lambda' g)dP_n\int g\exp(\lambda' g)dP_n\int g'\exp(\lambda' g/2)dP_n \\
&\left.-\tfrac{1}{2}\int \exp(\lambda' g)dP_n\int g\exp(\lambda' g/2)dP_n\int g'\exp(\lambda' g)dP_n\right).
\end{aligned}
$$

Hence, $\frac{\partial\Delta_n(0,\hat\theta)}{\partial\lambda}=0$. We also have that:

$$
\frac{\partial^2\Delta_n(\dot\lambda,\hat\theta)}{\partial\lambda\partial\lambda'}=-\frac{1}{4}Var(g(x,\hat\theta))+o_P(1). \tag{22}
$$

To see this, we observe that, thanks to Assumption 1-(i), (ii) and (vi), we can claim, relying on Lemma 2.4 of Newey and McFadden (1994), that $\int \exp\left(\lambda' g(x,\theta)\right)dP_n$ converges in probability towards $E\left(\exp\left(\lambda' g(x,\theta)\right)\right)$, uniformly over $\Lambda\times\Theta$. Thus

$$
\int \exp\left(\dot\lambda' g(x,\hat\theta)\right)dP_n = E\left(\exp\left(\dot\lambda' g(x,\hat\theta)\right)\right)+o_P(1).
$$

Assumption 1(vi) ensures that $\exp\left(\dot\lambda' g(x,\hat\theta)\right)$ is dominated by an integrable random variable. We can therefore apply the Lebesgue dominated convergence theorem. First, observe that, thanks to Assumption 1-(iv), $g(x,\hat\theta)=O_P(1)$ and since $\dot\lambda\overset{P}{\to}0$, we have $\dot\lambda' g(x,\hat\theta)=o_P(1)$. Thus, we can claim that $E\left(\exp\left(\dot\lambda' g(x,\hat\theta)\right)\right)\to 1$ in probability as $n\to\infty$. Hence,

$$
\int \exp\left(\dot\lambda' g(x,\hat\theta)\right)dP_n\overset{P}{\to}1.
$$

We can also claim that

$$
\int g(x,\hat\theta)\exp\left(\dot\lambda' g(x,\hat\theta)\right)dP_n = E\left(g(x,\hat\theta)\exp\left(\dot\lambda' g(x,\hat\theta)\right)\right)+o_P(1)=E\left(g(x,\hat\theta)\right)+o_P(1).
$$

To see this, let $\mathcal{N}\subset\mathbb{R}^m$ be a small neighborhood of 0. For λ closed to 0, we have

$$
|g(x,\theta)\exp\left(\lambda' g(x,\theta)\right)|\le \sup_{\theta\in\Theta}|g(x,\theta)|\sup_{\theta\in\Theta,\lambda\in\mathcal{N}}\exp\left(\lambda' g(x,\theta)\right).
$$

Applying the Holder inequality with β: $1/\alpha+1/\beta=1$, have:

$$
\begin{aligned}
&E\left(\sup_{\theta\in\Theta}|g(x,\theta)|\sup_{\theta\in\Theta,\lambda\in\mathcal{N}}\exp\left(\lambda' g(x,\theta)\right)\right) \\
\le\ &\left(E\sup_{\theta\in\Theta}|g(x,\theta)|^{\alpha}\right)^{\frac{1}{\alpha}}\left(E\sup_{\theta\in\Theta,\lambda\in\mathcal{N}}\exp\left(\beta\lambda' g(x,\theta)\right)\right)^{\frac{1}{\beta}} \\
\le\ &\left(E\sup_{\theta\in\Theta}|g(x,\theta)|^{\alpha}\right)^{\frac{1}{\alpha}}\left(E\sup_{\theta\in\Theta,\lambda\in\Lambda}\exp\left(\lambda' g(x,\theta)\right)\right)^{\frac{1}{\beta}}<\infty.
\end{aligned}
$$

(We use in this conclusion Assumption 1(iv).) The claim follows.

We can proceed the same way to show that:

$$\int g(x,\hat{\theta})g(x,\hat{\theta})' \exp\left(\dot{\lambda}' g(x,\theta)\right) dP_n = E\left(g(x,\hat{\theta})g(x,\hat{\theta})'\right) + o_P(1);$$

$$\int g(x,\hat{\theta})g(x,\hat{\theta})' \exp\left(\dot{\lambda}' g(x,\theta)/2\right) dP_n = E\left(g(x,\hat{\theta})g(x,\hat{\theta})\right) + o_P(1);$$

$$\int g(x,\hat{\theta}) \exp\left(\dot{\lambda}' g(x,\theta)/2\right) dP_n = E\left(g(x,\hat{\theta})\right) + o_P(1); \quad \text{and} \quad \int \exp\left(\dot{\lambda}' g(x,\theta)/2\right) dP_n = 1 + o_P(1)$$

and (22) follows.

Therefore, (21) can be written:

$$\Delta_n(\hat{\lambda}(\hat{\theta}),\hat{\theta}) = 1 - \frac{1}{8}\hat{\lambda}(\hat{\theta})' Var(g(x,\hat{\theta}))\hat{\lambda}(\hat{\theta}) + o_P(1)|\hat{\lambda}(\hat{\theta})|^2. \tag{23}$$

Thus

$$\frac{1}{8}\hat{\lambda}(\hat{\theta})' Var(g(x,\hat{\theta}))\hat{\lambda}(\hat{\theta}) + o_P(1)|\hat{\lambda}(\hat{\theta})|^2 = O_P(n^{-1}).$$

From Assumption 1(v), this implies that:

$$\underline{\ell}|\hat{\lambda}(\hat{\theta})|^2/8 + o_P(1)|\hat{\lambda}(\hat{\theta})|^2 \leq \frac{1}{8}\hat{\lambda}(\hat{\theta})' Var(g(x,\hat{\theta}))\hat{\lambda}(\hat{\theta}) + o_P(1)|\hat{\lambda}(\hat{\theta})|^2 = O_P(n^{-1})$$

for some $\underline{\ell} > 0$ and we can conclude that

$$|\hat{\lambda}(\hat{\theta})|^2(1 + o_P(1)) = O_P(n^{-1})$$

implying that

$$|\hat{\lambda}(\hat{\theta})|^2 = O_P(n^{-1})$$

or, equivalently, $\hat{\lambda}(\hat{\theta}) = O_P(n^{-1/2})$, concluding Step 2.

Step 3: Now, we show that $\int g(x,\hat{\theta})dP_n = O_P(n^{-1/2})$. Let $\tilde{\lambda} = -\frac{\int g(x,\hat{\theta})dP_n}{\sqrt{n}|\int g(x,\hat{\theta})dP_n|} + \hat{\lambda}(\hat{\theta})$. By definition,

$$\int \exp\left(\hat{\lambda}(\hat{\theta})' g(x,\hat{\theta})\right) dP_n \leq \int \exp\left(\tilde{\lambda}' g(x,\hat{\theta})\right) dP_n.$$

A second order Taylor expansion of each side around 0 with a Lagrange remainder gives:

$$\int \exp\left(\hat{\lambda}(\hat{\theta})' g(x,\hat{\theta})\right) dP_n = 1 + \hat{\lambda}(\hat{\theta})' \int g(x,\hat{\theta})dP_n + \frac{1}{2}\hat{\lambda}(\hat{\theta})' \int g(x,\hat{\theta})g(x,\hat{\theta})' \exp\left(\dot{\lambda}' g(x,\hat{\theta})\right) dP_n \hat{\lambda}(\hat{\theta})$$

and

$$\begin{aligned} \int \exp\left(\tilde{\lambda}' g(x,\hat{\theta})\right) dP_n &= 1 + \hat{\lambda}(\hat{\theta})' \int g(x,\hat{\theta})dP_n - n^{-1/2}\left|\int g(x,\hat{\theta})dP_n\right| \\ &\quad + \tfrac{1}{2}\tilde{\lambda}' \int g(x,\hat{\theta})g(x,\hat{\theta})' \exp\left(\ddot{\lambda}' g(x,\hat{\theta})\right) dP_n \tilde{\lambda}, \end{aligned}$$

with $\dot{\lambda} \in (0, \hat{\lambda}(\hat{\theta}))$ and $\ddot{\lambda} \in (0, \tilde{\lambda})$. Since $\hat{\lambda}(\hat{\theta})$ and $\tilde{\lambda}$ are both $O_P(n^{-1/2})$, so are $\dot{\lambda}$ and $\ddot{\lambda}$ and, as a result, the quadratic terms in both expansions are of order $O_P(n^{-1})$. Thus:

$$1 + \hat{\lambda}(\hat{\theta})' \int g(x,\hat{\theta})dP_n + O_P(n^{-1}) \leq 1 + \hat{\lambda}(\hat{\theta})' \int g(x,\hat{\theta})dP_n - n^{-1/2}\left|\int g(x,\hat{\theta})dP_n\right| + O_P(n^{-1})$$

and we can conclude that: $\left|\int g(x,\hat{\theta})dP_n\right| = O_P(n^{-1/2})$. $\square$

Proof of Theorem 3.2:
Proofs of (ii) and (iii) follow from Lemma B.2. We show (i). We have

$$\int g(x,\hat{\theta})dP_n = \int g(x,\hat{\theta})dP_n - E(g(x,\hat{\theta})) + E(g(x,\hat{\theta})).$$

By uniform convergence in probability of $\int g(x,\theta)dP_n$ towards $E(g(x,\theta))$ over Θ, we have:

$$\int g(x,\hat{\theta})dP_n = E(g(x,\hat{\theta})) + o_P(1).$$

From (iii), we can deduce that $E(g(x,\hat{\theta})) \to 0$ in probability as $n \to \infty$. Since $E(g(x,\theta)) = 0$ is solved only at θ^*, the fact that $\theta \to E(g(x,\theta))$ is continuous and Θ compact allows us to conclude that $\hat{\theta} \xrightarrow{P} \theta^*$. □

Proof of Theorem 3.3: (i) We essentially rely on mean-value expansions of the first order optimality conditions for $\hat{\theta}$ and $\hat{\lambda}$. Since $\hat{\theta}$ converges in probability to θ^* which is an interior point, with probability approaching 1, $\hat{\theta}$ is an interior solution and solves the first order condition:

$$\frac{d\Delta_n(\hat{\lambda}(\theta),\theta)}{d\theta} = \frac{N_1(\hat{\lambda}(\theta),\theta)}{D_1(\hat{\lambda}(\theta),\theta)} - \frac{N_2(\hat{\lambda}(\theta),\theta)}{D_2(\hat{\lambda}(\theta),\theta)} = 0, \tag{24}$$

with

$$N_1(\lambda,\theta) = \frac{1}{2}\int \left(\frac{d\hat{\lambda}'}{d\theta}(\theta)g(x,\theta) + \frac{\partial g(x,\theta)'}{\partial\theta}\lambda\right) \exp\left(\lambda' g(x,\theta)/2\right) dP_n,$$

$$N_2(\lambda,\theta) = \frac{1}{2}\int \left(\frac{d\hat{\lambda}'}{d\theta}(\theta)g(x,\theta) + \frac{\partial g(x,\theta)'}{\partial\theta}\lambda\right) \exp\left(\lambda' g(x,\theta)\right) dP_n \times \int \exp\left(\lambda' g(x,\theta)/2\right) dP_n,$$

$$D_1(\lambda,\theta) = \left(\int \exp(\lambda' g(x,\theta)dP_n\right)^{1/2}, \quad D_2(\lambda,\theta) = \left(\int \exp(\lambda' g(x,\theta)dP_n\right)^{3/2}.$$

Also, the fact that $\hat{\lambda}(\hat{\theta})$ converges in probability to 0 makes it an interior solution so that it solves in λ the first-order condition:

$$\int g(x,\hat{\theta}) \exp\left(\lambda' g(x,\hat{\theta})\right) dP_n = 0. \tag{25}$$

We will consider the left hand sides of (24) and (25) and carry out their mean-value expansions around $(0,\theta^*)$. Regarding (24), we have:

$$N_1(0,\theta^*) = N_2(0,\theta^*) = \frac{1}{2}\frac{d\hat{\lambda}(\theta^*)'}{d\theta}\int g(x,\theta^*)dP_n, \quad D_1(0,\theta^*) = D_2(0,\theta^*) = 1,$$

so that the first term in the expansion is nil. Hence, the mean-value expansion of (24) is:

$$0 = \frac{\partial}{\partial\theta'}\left(\frac{N_1(\lambda,\theta)}{D_1(\lambda,\theta)} - \frac{N_2(\lambda,\theta)}{D_2(\lambda,\theta)}\right)\Bigg|_{(\dot{\lambda},\dot{\theta})} (\hat{\theta}-\theta^*) + \frac{\partial}{\partial\lambda'}\left(\frac{N_1(\lambda,\theta)}{D_1(\lambda,\theta)} - \frac{N_2(\lambda,\theta)}{D_2(\lambda,\theta)}\right)\Bigg|_{(\dot{\lambda},\dot{\theta})} \hat{\lambda}, \tag{26}$$

where $\dot{\lambda} \in (0, \hat{\lambda})$ and $\dot{\theta} \in (\theta^*, \hat{\theta})$ and both may vary from row to row. We have:

$$\begin{aligned}
\tfrac{\partial N_1(\lambda,\theta)}{\partial \theta'} &= \tfrac{1}{2} \int \Bigg(\sum_{k=1}^{m} \frac{d^2 \hat{\lambda}_k(\theta)}{d\theta d\theta'} g_k(x,\theta) + \sum_{k=1}^{m} \frac{\partial^2 g_k(x,\theta)}{\partial\theta\partial\theta'} \lambda_k + \frac{d\hat{\lambda}(\theta)'}{d\theta} \frac{\partial g(x,\theta)}{\partial \theta'} \\
&\quad + \tfrac{1}{2} \left(\tfrac{d\hat{\lambda}(\theta)'}{d\theta} g(x,\theta) + \tfrac{\partial g(x,\theta)'}{\partial\theta} \lambda \right) \lambda' \tfrac{\partial g(x,\theta)}{\partial\theta'} \Bigg) \exp\left(\lambda' g(x,\theta)/2\right) dP_n \\
\tfrac{\partial N_2(\lambda,\theta)}{\partial \theta'} &= \tfrac{1}{2} \int \Bigg(\sum_{k=1}^{m} \frac{d^2 \hat{\lambda}_k(\theta)}{d\theta d\theta'} g_k(x,\theta) + \sum_{k=1}^{m} \frac{\partial^2 g_k(x,\theta)}{\partial\theta\partial\theta'} \lambda_k + \frac{d\hat{\lambda}(\theta)'}{d\theta} \frac{\partial g(x,\theta)}{\partial \theta'} \\
&\quad + \left(\tfrac{d\hat{\lambda}(\theta)'}{d\theta} g(x,\theta) + \tfrac{\partial g(x,\theta)'}{\partial\theta} \lambda \right) \lambda' \tfrac{\partial g(x,\theta)}{\partial\theta'} \Bigg) \exp\left(\lambda' g(x,\theta)\right) dP_n \times \int \exp\left(\lambda g(x,\theta)/2\right) dP_n \\
&\quad + \tfrac{1}{4} \int \left(\frac{d\hat{\lambda}(\theta)'}{d\theta} g(x,\theta) + \frac{\partial g(x,\theta)'}{\partial\theta} \lambda \right) \exp\left(\lambda g(x,\theta)\right) dP_n \times \int \lambda' \frac{\partial g(x,\theta)}{\partial\theta'} \exp\left(\lambda' g(x,\theta)/2\right) dP_n \\
\tfrac{\partial D_1(\lambda,\theta)}{\partial \theta'} &= \tfrac{1}{2} \int \lambda' \frac{\partial g(x,\theta)}{\partial\theta'} \exp\left(\lambda' g(x,\theta)\right) dP_n \times \left(\int \exp\left(\lambda' g(x,\theta)\right) dP_n \right)^{-1/2} \\
\tfrac{\partial D_2(\lambda,\theta)}{\partial \theta'} &= \tfrac{3}{2} \int \lambda' \frac{\partial g(x,\theta)}{\partial\theta'} \exp\left(\lambda' g(x,\theta)\right) dP_n \times \left(\int \exp\left(\lambda' g(x,\theta)\right) dP_n \right)^{1/2}.
\end{aligned}$$

Also,

$$\begin{aligned}
\tfrac{\partial N_1(\lambda,\theta)}{\partial \lambda'} &= \tfrac{1}{2} \int \left(\frac{\partial g(x,\theta)'}{\partial\theta} + \frac{1}{2} \left(\frac{d\hat{\lambda}(\theta)'}{d\theta} g(x,\theta) + \frac{\partial g(x,\theta)'}{\partial\theta} \lambda \right) g(x,\theta)' \right) \exp\left(\lambda' g(x,\theta)/2\right) dP_n \\
\tfrac{\partial N_2(\lambda,\theta)}{\partial \lambda'} &= \tfrac{1}{2} \int \left(\frac{\partial g(x,\theta)'}{\partial\theta} + \left(\frac{d\hat{\lambda}(\theta)'}{d\theta} g(x,\theta) + \frac{\partial g(x,\theta)'}{\partial\theta} \lambda \right) g(x,\theta)' \right) \exp\left(\lambda' g(x,\theta)\right) dP_n \\
&\quad \times \textstyle\int \exp\left(\lambda' g(x,\theta)/2\right) dP_n \\
&\quad + \tfrac{1}{4} \int \left(\frac{d\hat{\lambda}(\theta)'}{d\theta} g(x,\theta) + \frac{\partial g(x,\theta)'}{\partial\theta} \lambda \right) \exp\left(\lambda' g(x,\theta)\right) dP_n \times \int g(x,\theta)' \exp\left(\lambda' g(x,\theta)/2\right) dP_n \\
\tfrac{\partial D_1(\lambda,\theta)}{\partial \lambda'} &= \tfrac{1}{2} \int g(x,\theta)' \exp\left(\lambda' g(x,\theta)\right) dP_n \times \left(\int \exp\left(\lambda' g(x,\theta)\right) dP_n \right)^{-1/2} \\
\tfrac{\partial D_2(\lambda,\theta)}{\partial \lambda'} &= \tfrac{3}{2} \int g(x,\theta)' \exp\left(\lambda' g(x,\theta)\right) dP_n \times \left(\int \exp\left(\lambda' g(x,\theta)\right) dP_n \right)^{1/2}.
\end{aligned}$$

Since $\dot{\lambda} \in (0, \hat{\lambda}(\hat{\theta}))$, we have that $\dot{\lambda} = O_P(n^{-1/2})$. Hence, by Lemma B1 of Newey and Smith (2004), $\max_{1 \le i \le n} |\dot{\lambda}' g(x_i, \dot{\theta})| = o_P(1)$. Therefore, we can claim that $\int f(x) \exp\left(\dot{\lambda}' g(x,\dot{\theta})\right) dP_n = \int f(x) dP_n + o_P(1)$ for any f such that $E(f(x))$ exists. Also, under our assumptions, $\frac{d\hat{\lambda}(\dot{\theta})}{d\theta'} = O_P(1)$ as well as $\frac{d^2 \hat{\lambda}_k(\dot{\theta})}{d\theta d\theta'} = O_P(1)$, for all $k = 1, \ldots, m$. Furthermore, since $\hat{\theta} - \theta^* = O_P(n^{-1/2})$, a mean-value expansion of $\int g(x,\dot\theta)dP_n$ around θ^* ensures that $\int g(x,\dot\theta)dP_n = O_P(n^{-1/2})$. Under these observations, we have:

$$\begin{aligned}
\tfrac{\partial N_1(\dot\lambda,\dot\theta)}{\partial\theta'} &= \tfrac{1}{2}\tfrac{d\hat\lambda(\dot\theta)'}{d\theta}\int \frac{\partial g(x,\dot\theta)}{\partial\theta'}dP_n + o_P(1), \quad \frac{\partial N_2(\dot\lambda,\dot\theta)}{\partial\theta'} = \frac{1}{2}\frac{d\hat\lambda(\dot\theta)'}{d\theta}\int \frac{\partial g(x,\dot\theta)}{\partial\theta'}dP_n + o_P(1),\\
D_1(\dot\lambda,\dot\theta) &= 1+o_P(1), \quad D_2(\dot\lambda,\dot\theta) = 1+o_P(1), \quad \tfrac{\partial D_1(\dot\lambda,\dot\theta)}{\partial\theta'} = o_P(1), \quad \tfrac{\partial D_2(\dot\lambda,\dot\theta)}{\partial\theta'} = o_P(1).
\end{aligned}$$

Also,

$$\begin{aligned}
\tfrac{\partial N_1(\dot\lambda,\dot\theta)}{\partial\lambda'} &= \tfrac{1}{2}\int \frac{\partial g(x,\dot\theta)'}{\partial\theta}dP_n + \frac{1}{4}\frac{d\hat\lambda(\dot\theta)'}{d\theta}\int g(x,\dot\theta)g(x,\dot\theta)'dP_n + o_P(1),\\
\tfrac{\partial N_2(\dot\lambda,\dot\theta)}{\partial\lambda'} &= \tfrac{1}{2}\int \frac{\partial g(x,\dot\theta)'}{\partial\theta}dP_n + \frac{1}{2}\frac{d\hat\lambda(\dot\theta)'}{d\theta}\int g(x,\dot\theta)g(x,\dot\theta)'dP_n + o_P(1),\\
\tfrac{\partial D_1(\dot\lambda,\dot\theta)}{\partial\lambda'} &= o_P(1), \quad \tfrac{\partial D_2(\dot\lambda,\dot\theta)}{\partial\lambda'} = o_P(1).
\end{aligned}$$

As a result,

$$\begin{aligned}
\tfrac{\partial}{\partial\theta'}\left(\tfrac{N_1(\lambda,\theta)}{D_1(\lambda,\theta)} - \tfrac{N_2(\lambda,\theta)}{D_2(\lambda,\theta)}\right)\Big|_{(\dot\lambda,\dot\theta)} &= o_P(1)\\
\tfrac{\partial}{\partial\lambda'}\left(\tfrac{N_1(\lambda,\theta)}{D_1(\lambda,\theta)} - \tfrac{N_2(\lambda,\theta)}{D_2(\lambda,\theta)}\right)\Big|_{(\dot\lambda,\dot\theta)} &= -\tfrac{1}{4}\tfrac{d\hat\lambda(\dot\theta)'}{d\theta}\int g(x,\dot\theta)g(x,\dot\theta)'dP_n + o_P(1).
\end{aligned}$$

Note that

$$\begin{aligned}
\tfrac{d\hat\lambda(\theta)}{d\theta'} &= -\left(\int g(x,\theta)g(x,\theta)'\exp\left(\hat\lambda(\theta)'g(x,\theta)\right)dP_n\right)^{-1}\\
&\quad \times\int\left(\frac{\partial g(x,\theta)}{\partial\theta'} + g(x,\theta)\hat\lambda(\theta)'\frac{\partial g(x,\theta)}{\partial\theta'}\right)\exp\left(\hat\lambda(\theta)'g(x,\theta)\right)dP_n.
\end{aligned}$$

Again, since $\dot\theta = \theta^* + O_P(n^{-1/2})$ and $\int g(x,\dot\theta)dP_n = O_P(n^{-1/2})$, Lemma A2 of Newey and Smith (2004) ensures that $\hat\lambda(\dot\theta) = O_P(n^{-1/2})$. Thus, thanks to their Lemma A1, we also have $\max_{1\le i\le n}|\hat\lambda(\dot\theta)'g(x_i,\dot\theta)| = o_P(1)$. We can therefore claim that:

$$\frac{d\hat\lambda(\dot\theta)}{d\theta'} = -\left(\int g(x,\theta)g(x,\theta)'dP_n\right)^{-1}\int\frac{\partial g(x,\theta)}{\partial\theta'}dP_n + o_P(1) = -\Omega^{-1}G + o_P(1).$$

Hence,

$$\frac{\partial}{\partial\lambda'}\left(\frac{N_1(\lambda,\theta)}{D_1(\lambda,\theta)} - \frac{N_2(\lambda,\theta)}{D_2(\lambda,\theta)}\right)\Bigg|_{(\dot\lambda,\dot\theta)} = \frac{1}{4}G' + o_P(1)$$

and the expansion of (24) yields:

$$\sqrt{n}G'\hat\lambda = o_P(1). \tag{27}$$

The expansion of (25) around $(\theta^*,0)$ yields:

$$\begin{aligned}
0 &= \int g(x,\theta^*)dP_n + \int\left(\frac{\partial g(x,\dot\theta)}{\partial\theta'} + g(x,\dot\theta)\dot\lambda'\frac{\partial g(x,\dot\theta)}{\partial\theta'}\right)\exp\left(\dot\lambda'g(x,\dot\theta)\right)dP_n(\hat\theta-\theta^*)\\
&\quad + \textstyle\int g(x,\dot\theta)g(x,\dot\theta)'\exp\left(\dot\lambda'g(x,\dot\theta)\right)dP_n\hat\lambda,
\end{aligned}$$

with $(\dot{\lambda}, \dot{\theta}) \in (0, \hat{\lambda}(\hat{\theta})) \times (\theta^*, \hat{\theta})$ and may differ from row to row. By similar arguments to those previously made, this expression reduces to:

$$G\sqrt{n}(\hat{\theta} - \theta^*) + \Omega\sqrt{n}\hat{\lambda} = -\sqrt{n}\int g(x, \theta^*)dP_n + o_P(1). \tag{28}$$

Together, (27) and (28) yield:

$$\begin{pmatrix} \Omega & G \\ G' & 0 \end{pmatrix} \sqrt{n} \begin{pmatrix} \hat{\lambda} \\ \hat{\theta} - \theta^* \end{pmatrix} = \begin{pmatrix} -\sqrt{n}\int g(x, \theta^*)dP_n \\ 0 \end{pmatrix} + o_P(1) \tag{29}$$

By the standard partitioned inverse matrix formula (see Magnus and Neudecker (1999, p.11)), we have

$$\begin{pmatrix} \Omega & G \\ G' & 0 \end{pmatrix}^{-1} = \begin{pmatrix} \Omega^{-1/2}M\Omega^{-1/2} & \Omega^{-1}G\Sigma \\ \Sigma G'\Omega^{-1} & -\Sigma \end{pmatrix}.$$

Hence,

$$\sqrt{n} \begin{pmatrix} \hat{\lambda} \\ \hat{\theta} - \theta^* \end{pmatrix} = - \begin{pmatrix} \Omega^{-1/2}M\Omega^{-1/2} \\ \Sigma G'\Omega^{-1} \end{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(x_i, \theta^*) + o_P(1)$$

and the statement (i) of the theorem follows easily.

To establish (ii), we use the fact that

$$\sqrt{n}\hat{\lambda} = -\Omega^{-1/2}M\Omega^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i, \theta^*) + o_P(1)$$

and Equation (23). This equation implies that

$$8n\left(1 - \Delta_n(\hat{\lambda}, \hat{\theta})\right) = n\hat{\lambda}'\Omega\hat{\lambda} + o_P(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i, \theta^*)'\Omega^{-1/2}M\Omega^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i, \theta^*) + o_P(1)$$

and the result follows since $\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Omega^{-1/2}g(x_i, \theta^*)\right)' M \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Omega^{-1/2}g(x_i, \theta^*)\right)$ is asymptotically distributed as a χ^2_{m-p}.

C Local misspecification

This section first introduces the definition of (asymptotic) Fisher consistency and then provides proofs to the main results in Section 4 of the main text. The following definition of Fisher consistency and regularity can be found in KOE (2013, Definition 3.1).

Let $T_a(P_n)$ be an estimator of θ^* based on a mapping $T_a : \mathcal{M} \to \Theta$. Let $\mathcal{P}$ be the set of all probability measures P for which there exists $\theta \in \Theta$ satisfying $E_P(g(x, \theta)) = 0$ and let $P_{\theta,\zeta}$ be a regular parametric submodel of $\mathcal{P}$ such that $P_{\theta^*,0} = P_*$ and such that $P_{\theta^*+t/\sqrt{n},\zeta_n} \in B_H(P_*, r/\sqrt{n})$ holds for $\zeta_n = O(n^{-1/2})$ eventually.

Definition 1 ***(Fisher consistent and regular estimator)***

(i) T_a is asymptotically Fisher consistent if for every $(P_{\theta^*+t/\sqrt{n},\zeta_n})_{n\in\mathbb{N}}$ and $t \in \mathbb{R}^p$,

$$\sqrt{n}\left(T_a(P_{\theta^*+t/\sqrt{n},\zeta_n}) - \theta^*\right) \to t.$$

(ii) T_a *is regular for* θ^* *if, for every* $(P_{\theta_n,\zeta_n})_{n\in\mathbb{N}}$ *with* $\theta_n = \theta + O(n^{-1/2})$ *and* $\zeta_n = O(n^{-1/2})$*, there exists a probability measure* M *such that:*

$$\sqrt{n}(T_a(P_n) - T_a(P_{\theta_n,\zeta_n})) \xrightarrow{d} M, \quad \textit{under} \quad P_{\theta_n,\zeta_n},$$

where the measure M *does not depend on the sequence* $(\theta_n\zeta_n)$*.*

Proof of Theorem 4.1: The proof follows similar lines as those of Theorem 3.1(ii) in KOE (2013). To establish Fisher consistency, let $P_{\theta,\zeta}$ be a regular sub-model such that for $t \in \mathbb{R}^p$, $P_{\theta_n,\zeta_n} \in B_H(P_0, r/\sqrt{n})$ for n large enough, with $\theta_n = \theta^* + t/\sqrt{n}$ and $\zeta_n = O(n^{-1/2})$. We have to show that

$$\sqrt{n}(\bar{T}(P_{\theta_n,\zeta_n}) - \theta^*) \to t,$$

as $n \to \infty$. From Lemma C.5,

$$\sqrt{n}(\bar{T}(P_{\theta_n,\zeta_n}) - \theta^*) = -\Sigma G'\Omega^{-1}\sqrt{n}\int g_n(x,\theta^*)dP_{\theta_n,\zeta_n} + o(1).$$

By a mean-value expansion, we have:

$$\sqrt{n}\int g_n(x,\theta^*)dP_{\theta_n,\zeta_n} = \sqrt{n}\int g_n(x,\theta_n)dP_{\theta_n,\zeta_n} - \int \frac{\partial g_n(x,\dot{\theta})}{\partial \theta'}dP_{\theta_n,\zeta_n}t,$$

with $\dot{\theta} \in (\theta^*, \theta_n)$ and may vary from row to row. Noting that $\int g(x,\theta_n)dP_{\theta_n,\zeta_n} = 0$, by the similar argument to (A.16) of KOE (2013, proofs), we have $\int g_n(x,\theta_n)dP_{\theta_n,\zeta_n} = o(n^{-1/2})$. The convergence follows by applying their Lemma A.4(i) to $\int \frac{\partial g_n(x,\dot{\theta})}{\partial\theta'}dP_{\theta_n,\zeta_n}$.

We next show 17. Let $F = \frac{\partial\tau(\theta_0)}{\partial\theta'}\Sigma G'\Omega^{-1}$. Thanks to Lemma C.2(ii), $\bar{T}_{Q_n} \to \theta_0$ as $n \to \infty$ and Lemma C.5 guarantees that $\bar{T}_{Q_n} - \theta^* = O(n^{-1/2})$. A Taylor expansion of $\tau(\bar{T}_{Q_n})$ around θ^* ensures that:

$$\sqrt{n}\left(\tau\circ\bar{T}_{Q_n} - \tau(\theta^*)\right) = -\sqrt{n}F\int g_n(x,\theta^*)dQ_n + o(1)$$

$$= \quad -\sqrt{n}F\int g_n(x,\theta^*)\left(dQ_n^{1/2} - dP_*^{1/2}\right)dQ_n^{1/2} - \sqrt{n}F\int g_n(x,\theta^*)\left(dQ_n^{1/2} - dP_*^{1/2}\right)dP_*^{1/2} + o(1).$$

By the triangle inequality, we have

$$n(\left(\tau\circ\bar{T}_{Q_n} - \tau(\theta^*)\right)^2 \le n(A_1 + A_2 + 2A_3) + o(1),$$

with

$$A_1 = \left|F\int g_n(x,\theta^*)\left(dQ_n^{1/2} - dP_*^{1/2}\right)dQ_n^{1/2}\right|^2, \quad A_1 = \left|F\int g_n(x,\theta^*)\left(dQ_n^{1/2} - dP_*^{1/2}\right)dP_*^{1/2}\right|^2$$

and $A_3 = \sqrt{A_1 \cdot A_2}$. By the Cauchy-Schwarz inequality and then by Lemma A.5(i) of KOE (2013, proofs), we have:

$$A_1 \le \left|F\int g_n(x,\theta^*)g_n(x,\theta^*)'dQ_n\right|\int\left(dQ_n^{1/2} - dP_*^{1/2}\right)^2 \le B^*\frac{r^2}{n} + o(n^{-1}).$$

By the same way, we have $A_2 \le B^*\frac{r^2}{n} + o(n^{-1})$ and we can deduce that $A_3 \le B^*\frac{r^2}{n} + o(n^{-1})$. Therefore,

$$n\left(\tau\circ\bar{T}_{Q_n} - \tau(\theta^*)\right)^2 \le 4r^2B^* + o(1), \tag{30}$$

Besides, we know from Lemma C.2(i) that, for all $n \geq n_0$, $\bar{T}_Q$ exists for all $Q \in B_H(P_*, r/\sqrt{n})$. Since $\bar{T}_Q \in \Theta$ compact, there exists $C > 0$ such that

$$L_n \equiv \sup_{Q \in B_H(P_*, r/\sqrt{n})} n \left(\tau \circ \bar{T}_{Q_n} - \tau(\theta^*)\right)^2 \leq C\dot{n} < \infty.$$

Let Q_n be a sequence such that $Q_n \in B_H(P_*, r/\sqrt{n})$ for all $nn \geq n_0$ and

$$L_n \leq n \left(\tau \circ \bar{T}_{Q_n} - \tau(\theta^*)\right)^2 + \frac{1}{2^n}.$$

we have:

$$\limsup_{n\to\infty} L_n \leq \limsup_{n\to\infty} n \left(\tau \circ \bar{T}_{Q_n} - \tau(\theta^*)\right)^2$$

and using (30), we deduce that $\limsup_{n\to\infty} L_n \leq 4r^2 B^*$. Theorem 3.1(i) of KOE (2013) guarantees that $\liminf_{n\to\infty} L_n \geq 4r^2 B^*$ leading to (17). $\square$

Proof of Theorem 4.2: This proof also follows similar lines to the proof of Theorem 3.2(ii) of KOE (2013). We have:

$$\begin{aligned} & \limsup_{n\to\infty} \sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} \int b \wedge n \left(\tau \circ T(P_n) - \tau(\theta^*)\right)^2 dQ^{\otimes n} \\ = \ & \limsup_{n\to\infty} \sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} \int b \wedge n \left((\tau \circ T(P_n) - \tau \circ \bar{T}(P_n)) + (\bar{T}(P_n) - \tau(\theta^*))\right)^2 dQ^{\otimes n} \\ \leq \ & A_1 + 2A_2 + A_3, \end{aligned}$$

with

$$\begin{aligned} A_1 &= \limsup_{n\to\infty} \sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} \int b \wedge n \left(\tau \circ T(P_n) - \tau \circ \bar{T}(P_n)\right)^2 dQ^{\otimes n}, \\ A_2 &= \limsup_{n\to\infty} \sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} \int b \wedge n \left|\tau \circ T(P_n) - \tau \circ \bar{T}(P_n)\right| \left|\tau \circ \bar{T}(P_n) - \tau(\theta^*)\right| dQ^{\otimes n}, \\ A_3 &= \limsup_{n\to\infty} \sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} \int b \wedge n \left(\tau \circ \bar{T}(P_n) - \tau(\theta^*)\right)^2 dQ^{\otimes n}. \end{aligned}$$

We show that $A_1 = A_2 = 0$. Note that $T(P_n) = \bar{T}(P_n)$ if $[(x_1, \dots, x_n) \in \mathcal{X}_n^n] \cap [\bar{T}_1(P_n) \in int(\Lambda_n)]$. Thus,

$$\begin{aligned} A_1 &\leq b \times \limsup_{n\to\infty} \sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} \int_{[(x_1,\dots,x_n) \notin \mathcal{X}_n^n] \cup [\bar{T}_1(P_n) \notin int(\Lambda_n)]} dQ^{\otimes n} \\ &\leq b \times \limsup_{n\to\infty} \sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} \int_{(x_1,\dots,x_n) \notin \mathcal{X}_n^n} dQ^{\otimes n} + b \times \limsup_{n\to\infty} \sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} \int_{\bar{T}_1(P_n) \notin int(\Lambda_n)} dQ^{\otimes n} \\ &\equiv A_{11} + A_{12}. \end{aligned}$$

From KOE(2013, proofs), Equation (A.5), $A_{11} = 0$. We show that

$$\limsup_{n\to\infty} \sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} Q^{\otimes n} \left(\bar{T}_1(P_n) \geq C/m_n^{1+\zeta}\right) = 0.$$

Since $Q^{\otimes n}$ is a probability measure, we have $0 \leq Q^{\otimes n} \left(\bar{T}_1(P_n) \geq C/m_n^{1+\zeta}\right) \leq 1$. Hence,

$$\sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} Q^{\otimes n} \left(\bar{T}_1(P_n) \geq C/m_n^{1+\zeta}\right)$$

is finite. Let Q_n be the sequence of probability measures in $\bar{B}_H(P_*, r/\sqrt{n})$ such that:

$$\sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} Q^{\otimes n}\left(\bar{T}_1(P_n) \geq C/m_n^{1+\zeta}\right) \leq Q_n^{\otimes n}\left(\bar{T}_1(P_n) \geq C/m_n^{1+\zeta}\right) + \frac{1}{2^n}.$$

We have

$$\limsup_{n\to\infty} \sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} Q^{\otimes n}\left(\bar{T}_1(P_n) \geq C/m_n^{1+\zeta}\right) \leq \limsup_{n\to\infty} Q_n^{\otimes n}\left(\bar{T}_1(P_n) \geq C/m_n^{1+\zeta}\right).$$

But, thanks to Lemma C.8, $\bar{T}_1(P_n) = O_P(n^{-1/2})$ under Q_n. Thus, under Assumption 3 (iii), we have: $m_n^{1+\zeta}\bar{T}_1(P_n) = o_P(1)$ under Q_n. It results that $Q_n^{\otimes n}\left(\bar{T}_1(P_n) \geq C/m_n^{1+\zeta}\right) \to 0$ as $n \to \infty$ showing that $\limsup_{n\to\infty} \sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} Q^{\otimes n}\left(\bar{T}_1(P_n) \geq C/m_n^{1+\zeta}\right) = 0$.

Consider A_3. Note that $\sup_{Q \in \bar{B}_H(P_*, r/\sqrt{n})} \int b \wedge n\left(\tau \circ \bar{T}(P_n) - \tau(\theta^*)\right)^2 dQ^{\otimes n} \leq b < \infty$. Therefore, there exists $\bar{Q}_n \in \bar{B}_H(P_*, r/\sqrt{n})$ such that

$$A_3 \leq \limsup_{n\to\infty} \int b \wedge n\left(\tau \circ \bar{T}(P_n) - \tau(\theta^*)\right)^2 d\bar{Q}_n^{\otimes n}.$$

Note that, thanks to Lemma C.8, $\sqrt{n}(\tau \circ \bar{T}(P_n) - \tau \circ \bar{T}(\bar{Q}_n))$ converges in distribution towards $N(0, B^*)$ under $\bar{Q}_n$. Let $\int b \wedge n\left(\tau \circ \bar{T}(P_n) - \tau(\theta^*)\right)^2 d\bar{Q}_n^{\otimes n}$ be a subsequence of this sequence that converge to the lim sup (we keep n to denote the subsequence for simplicity). This has a further subsequence along which $\sqrt{n}(\tau \circ \bar{T}(\bar{Q}_n) - \tau(\theta_0))$ converges towards its lim sup, say $\tilde{t}$. Thanks to Theorem 4.1, $|\tilde{t}|$ is finite. Hence, along this final subsequence,

$$\sqrt{n}(\tau \circ \bar{T}(P_n) - \tau(\theta^*)) = \sqrt{n}(\tau \circ \bar{T}(P_n) - \tau \circ \bar{T}(\bar{Q}_n)) + \sqrt{n}(\tau \circ \bar{T}(\bar{Q}_n) - \tau(\theta^*))$$

converges in distribution towards $N(\tilde{t}, B^*)$ under $\bar{Q}_n$.

We can deduce that:

$$A_3 \leq \int b \wedge (Z + \tilde{t})^2 dN(0, B^*) \leq B^* + \tilde{t}^2 \leq B^* + \limsup_{n\to\infty} n(\tau \circ \bar{T}(\bar{Q}_n) - \tau(\theta^*))^2 \leq B^* + 4r^2 B^*,$$

where the lim sup is taking over the initial sequence and the last inequality follows from Theorem 4.1. This concludes the proof. □

Proof of Theorem 4.3: To be completed. □

Lemma C.1 *Let* $Q \in \mathcal{M}$, $\mathscr{P}_\theta = \{P \in \mathcal{M} : \int g(x,\theta)dP = 0\}$ *with* $\theta \in \Theta$ *and* $P(\theta)$ *solution to* $\min_{P \in \mathscr{P}_\theta} \int \log\left(\frac{dP}{dQ}\right) dP$. *We have*

$$\arg\min_{\theta \in \Theta} H(P(\theta), Q) = \arg\max_{\theta \in \Theta} \frac{\int \exp(\lambda(\theta)' g(x,\theta)/2) dQ}{\left(\int \exp(\lambda(\theta)' g(x,\theta)) dQ\right)^{1/2}},$$

with $\lambda(\theta) = \arg\min_{\lambda} \int \exp\left(\lambda(\theta)' g(x,\theta)\right) dQ$.

Proof of Lemma C.1 From Kitamura and Stutzer (1997), the solution $P(\theta)$ to $\min_{P \in \mathscr{P}_\theta} \int \log\left(\frac{dP}{dQ}\right) dP$ has the Gibbs canonical density with respect to Q given by:

$$\frac{dP(\theta)}{dQ} = \frac{\exp\left(\lambda(\theta)' g(x,\theta)\right)}{\int \exp\left(\lambda(\theta)' g(x,\theta)\right) dQ}.$$

We can conclude using the fact that:

$$H(P(\theta), Q) = \left(2 - 2\int dP(\theta)^{1/2} dQ^{1/2}\right)^{1/2} = \left(2 - 2\int \left(\frac{dP(\theta)}{dQ}\right)^{1/2} dQ\right)^{1/2}.$$

□

Lemma C.2 *If Assumptions 1, 2 (with expectation and variance taken under P_*), and Assumption 3 hold, then:*

(i) For each $r > 0$, there exists n_0 such that $\bar{T}(Q)$ exists and is upper hemi-continuous at each $Q \in B_H(P_*, r/\sqrt{n})$ under the Hellinger metric for all $n \geq n_0$.

(ii) $\bar{T}_{Q_n} \to \theta^*$ as $n \to \infty$ for each $r > 0$ and sequence $Q_n \in B_H(P_*, r/\sqrt{n})$.

Proof of Lemma C.2 (i) By the boundedness of the functions $(\lambda, \theta) \to \exp(\lambda' g_n(x, \theta))$ and $(\lambda, \theta) \to \exp(\lambda' g_n(x, \theta)/2)$ over $\Lambda_n \times \Theta$, the function:

$$(\lambda, \theta, Q) \to \int \exp\left(\lambda' g_n(x, \theta)\right) dQ \quad \text{and} \quad (\lambda, \theta, Q) \to \int \exp\left(\lambda' g_n(x, \theta)/2\right) dQ$$

are continuous in their arguments on $\Lambda_n \times \Theta \times \mathcal{M}$ under the Levy metric for $\mathcal{M}$. Since Λ_n is compact, $\lambda \to -\int \exp(\lambda' g_n(x, \theta))\, dQ$ reaches its maximum at $\bar{T}_1(\theta, Q) \in \Lambda_n$.

Let $r > 0$. Assume for now that this function is strictly concave for any $Q \in B_H(P_*, r/\sqrt{n_0})$ and $\theta \in \Theta$, for some n_0. (This will be established later.) By inclusion, this function is also strictly concave for any $Q \in B_H(P_*, r/\sqrt{n})$, for any $n \geq n_0$. In this case, the convexity of Λ_n makes $\bar{T}_1(\theta, Q)$ unique maximizer. The maximum theorem of Berge guarantees that $\bar{T}_1(\theta, Q)$ is upper hemi-continuous meaning that $\bar{T}_1(\theta, Q)$, as a function, is continuous under the Levy metric for $B_H(P_*, r/\sqrt{n})$. As a result,

$$(\theta, Q) \to \frac{\int \exp\left(\bar{T}_1(\theta, Q)' g_n(x, \theta)/2\right) dQ}{\int \left(\exp\left(\bar{T}_1(\theta, Q)' g_n(x, \theta)\right) dQ\right)^{1/2}}$$

is also continuous under the Levy metric. Since Θ is compact, this function reaches its maximum over Θ at $\bar{T}(Q) \in \Theta$ and the maximum theorem of Berge guarantees that $\bar{T}(Q)$ is upper hemi-continuous with respect to the Levy metric. Since the Levy metric is dominated by the Hellinger metric, $\bar{T}(Q)$ is also upper hemi-continuous under the Hellinger metric on $B_H(P_*, r/\sqrt{n})$.

To conclude the proof, we show that there exists n_0 such that, for all $n \geq n_0$, $\lambda \to -\int \exp(\lambda' g_n(x, \theta) dQ$ is strictly concave on Λ_n for any $Q \in B_H(P_*, r/\sqrt{n})$ and $\theta \in \Theta$. By the boundedness of $\lambda' g_n(x, \theta)$ on Λ_n, it suffices to find such n_0 such that, for all $n \geq n_0$, $\int g_n(x, \theta) g_n(x, \theta) dQ$ is positive definite for all θ and Q in the named sets. For this, it suffices to show that:

$$\left|\int g_n(x, \theta) g_n(x, \theta)' dQ - \int g_n(x, \theta) g_n(x, \theta) dP_*\right| \quad \text{and} \quad \left|\int g_n(x, \theta) g_n(x, \theta)' dP_* - \int g(x, \theta) g(x, \theta) dP_*\right|$$

converge to 0, uniformly in $\theta \in \Theta$ and $Q \in B_H(P_*, r/\sqrt{n})$ and Assumption 1(v) will allow us to conclude.

$$
\begin{aligned}
& \left|\int g_n(x,\theta)g_n(x,\theta)'dQ - \int g_n(x,\theta)g_n(x,\theta)'dP_*\right| \\
= & \left|\int g_n(x,\theta)g_n(x,\theta)'\left[\left(dQ^{1/2}-dP_*^{1/2}\right)^2 - 2dP_*^{1/2}\left(dP_*^{1/2}-dQ^{1/2}\right)\right]\right| \\
\leq & \int \left|g_n(x,\theta)g_n(x,\theta)'\right|\left(dQ^{1/2}-dP_*^{1/2}\right)^2 + 2\left(\int \left|g_n(x,\theta)g_n(x,\theta)'\right|^2 dP_*\right)^{1/2}\left(\int \left(dQ^{1/2}-dP_*^{1/2}\right)^2\right)^{1/2} \\
\leq & m_n^2\tfrac{r^2}{n} + 2m_n^{(4-\alpha)/2}\tfrac{r}{\sqrt{n}}\int \sup_{\theta\in\Theta}|g(x,\theta)|^\alpha dP_* \leq m_n^2\frac{r^2}{n} + 2m_n\frac{r}{\sqrt{n}}\int \sup_{\theta\in\Theta}|g(x,\theta)|^\alpha dP_* \\
= & o(1).
\end{aligned}
$$

The first inequality follows from the triangle and the Cauchy-Schwarz inequalities. The second inequality uses the definitions of g_n and Hellinger balls. The last one uses the fact that $\alpha > 2$ and the order of magnitude follows from Assumptions 1(iv) and 3(ii).

$$
\begin{aligned}
& \left|\int g_n(x,\theta)g_n(x,\theta)'dP_* - \int g(x,\theta)g(x,\theta)dP_*\right| = \left|\int g(x,\theta)g(x,\theta)'\mathbb{I}(x\notin\mathcal{X}_n)dP_*\right| \\
\leq & \int |g(x,\theta)|^2\,\mathbb{I}(x\notin\mathcal{X}_n)dP_* \leq \left(\int |g(x,\theta)|^\alpha\, dP_*\right)^{2/\alpha}\left(\int \mathbb{I}(x\notin\mathcal{X}_n)dP_*\right)^{1-2/\alpha} \\
\leq & \left(\int \sup_{\theta\in\Theta}|g(x,\theta)|^\alpha dP_*\right)^{2/\alpha}\left(P_*\left(\sup_{\theta\in\Theta}|g(x,\theta)|\geq m_n\right)\right)^{(\alpha-2)/\alpha} \\
\leq & C\left(\frac{1}{m_n^\alpha}E_{P_*}\left(\sup_{\theta\in\Theta}|g(x,\theta)|^\alpha\right)\right)^{(\alpha-2)/\alpha} = o(1)
\end{aligned}
$$

The second inequality follows from the Holder inequality and the third one from the Markov inequality. The order of magnitude follows from Assumption 3(ii); $C > 0$ is a generic constant.

(ii) Let $r > 0$ and $Q_n \in B_H(P_*, r/\sqrt{n})$. Along the same lines as KOE's (2013) proof of their Lemma 7.1(ii), we can show that:

$$\sup_{\theta\in\Theta}|E_{Q_n}(g_n(x,\theta)) - E_{P_*}(g(x,\theta))| \to 0,$$

as $n\to\infty$. Also, from Lemma C.3, we have $E_{Q_n}(g_n(x,\bar{T}_{Q_n})) = O(n^{-1/2})$. From the fact that

$$\left|E_{P_*}(g(x,\bar{T}_{Q_n}))\right| \leq \left|E_{P_*}(g(x,\bar{T}_{Q_n})) - E_{Q_n}(g_n(x,\bar{T}_{Q_n}))\right| + \left|E_{Q_n}(g_n(x,\bar{T}_{Q_n}))\right|,$$

we deduce that $E_{P_*}(g(x,\bar{T}_{Q_n}))\to 0$ as $n\to\infty$. Since $\theta\to E_{P_*}(g(x,\theta))$ is continuous and Θ is compact, Assumption 1(iii) allows us to conclude that $\bar{T}_{Q_n}\to\theta^*$ as $n\to\infty$. $\square$

Let

$$\mathbf{\Delta}_Q(\lambda,\theta) = \frac{\int \exp(\lambda' g(x,\theta)/2)dQ}{\left(\int \exp(\lambda' g(x,\theta))dQ\right)^{1/2}}.$$

Lemma C.3 *If Assumptions 1, 2 (with expectation and variance taken under P_*), and Assumption 3 hold, then: for each $r > 0$ and any sequence $Q_n \in B_H(P_*, r/\sqrt{n})$,*

(i) $\quad \mathbf{\Delta}_{Q_n}(\bar{T}_1(\bar{T}_{Q_n}, Q_n), \bar{T}_{Q_n}) = 1 + O(n^{-1})$,

(ii) $\quad \bar{T}_1(\bar{T}_{Q_n}, Q_n) = O(n^{-1/2})$, *and*

(iii) $\quad E_{Q_n}(g_n(x, \bar{T}_{Q_n})) = O(n^{-1/2})$.

Proof of Lemma C.3

(i) We have $\mathbf{\Delta}_{Q_n}(\bar{T}_1(\theta^*, Q_n), \theta^*) \leq \mathbf{\Delta}_{Q_n}(\bar{T}_1(\bar{T}_{Q_n}, Q_n), \bar{T}_{Q_n}) \leq 1$ where the first inequality follows by definition and the second by the Jensen's inequality and the convexity of $x \to \exp(x)$. We deduce (i) from Lemma C.4.

(ii) Next, we show that $\hat{\lambda}_n \equiv \bar{T}_1(\bar{T}_{Q_n}, Q_n) = O(n^{-1/2})$. By a second-order Taylor expansion of $\mathbf{\Delta}_{Q_n}(\bar{T}_1(\bar{T}_{Q_n}, Q_n), \bar{T}_{Q_n})$ in the direction of its first component around $\lambda = 0$, we have:

$$\mathbf{\Delta}_{Q_n}(\hat{\lambda}_n, \bar{T}_{Q_n}) = \mathbf{\Delta}_{Q_n}(0, \hat{\theta}_n) + \frac{\partial \mathbf{\Delta}_{Q_n}(0, \hat{\theta}_n)}{\partial \lambda'}\hat{\lambda}_n + \frac{1}{2}\hat{\lambda}_n' \frac{\partial^2 \mathbf{\Delta}_{Q_n}(\dot{\lambda}, \hat{\theta}_n)}{\partial \lambda \partial \lambda'}\hat{\lambda}_n, \tag{31}$$

with $\hat{\theta}_n \equiv \bar{T}_{Q_n}$ and $\dot{\lambda} \in (0, \hat{\lambda}_n)$. This expansion is actually the same as (21) with $\mathbf{\Delta}_{Q_n}$, $\hat{\lambda}_n$ and $\hat{\theta}_n$ replacing Δ_n, $\hat{\lambda}$ and $\hat{\theta}$, respectively. We have $\frac{\partial \mathbf{\Delta}_{Q_n}(0,\hat{\theta}_n)}{\partial \lambda'} = 0$ and $\frac{\partial^2 \mathbf{\Delta}_{Q_n}(\lambda,\hat{\theta}_n)}{\partial \lambda \partial \lambda'}$ is analogue to $\frac{\partial^2 \Delta_n(\lambda,\hat{\theta})}{\partial \lambda \partial \lambda'}$ as derived in the proof of Lemma B.2 with Q_n, $\hat{\theta}_n$ and $g_n(\cdot)$ replacing P_n, $\hat{\theta}$ and $g(\cdot)$, respectively.

We observe that

$$\frac{\partial^2 \mathbf{\Delta}_{Q_n}(\dot{\lambda}, \hat{\theta}_n)}{\partial \lambda \partial \lambda'} = -\frac{1}{4}Var_{P_*}(g_n(x, \hat{\theta}_n)) + o(1). \tag{32}$$

To see this, observe that, for any $\lambda \in \Lambda_n$,

$$\int \exp(\lambda' g_n(x, \theta) dQ_n = \int \exp(\lambda' g_n(x, \theta))(dQ_n - dP_*) + \int \exp(\lambda' g_n(x, \theta))dP_* \equiv (1) + (2).$$

By the Lebesgue dominated convergence theorem, (2) converges to 1 uniformly over $\lambda \in \Lambda_n$. Also, by a similar treatment as the one applied to $g_n(x, \theta)g_n(x, \theta)'$ in the proof of Lemma C.2, we have

$$\begin{aligned} & \int \exp(\lambda' g_n(x, \theta))dQ_n \\ = & \int \exp(\lambda' g_n(x, \theta)) \left\{ \left(dQ_n^{1/2} - dP_*^{1/2}\right)^2 - 2dP_*^{1/2}\left(dP_*^{1/2} - dQ_n^{1/2}\right) \right\} \leq C\left(\frac{r^2}{n} + 2\frac{r}{\sqrt{n}}\right), \end{aligned}$$

where C is a positive constant. Thus $\int \exp(\dot{\lambda}' g_n(x, \theta))dQ_n \to 1$ as $n \to \infty$.

We also have

$$\int g_n(x, \hat{\theta}_n) \exp(\dot{\lambda}' g_n(x, \hat{\theta}_n))dQ_n = \int g_n(x, \hat{\theta}_n)dP_* + o(1).$$

To see this, write:

$$\begin{aligned} & \int g_n(x, \hat{\theta}_n) \exp(\dot{\lambda}' g_n(x, \hat{\theta}_n))dQ_n \\ = & \int g_n(x, \hat{\theta}_n) \exp(\dot{\lambda}' g_n(x, \hat{\theta}_n))(dQ_n - dP_*) + \int g_n(x, \hat{\theta}_n) \exp(\dot{\lambda}' g_n(x, \hat{\theta}_n))dP_* \equiv (1') + (2'). \end{aligned}$$

By similar expression of $(dQ_n - dP_*)$ as in previous derivations, we can see that $(1') = o(1)$. Regarding $(2')$, write

$$(2') - \int g_n(x, \hat{\theta}_n)dP_* = \int g_n(x, \hat{\theta}_n)\left(\exp(\dot{\lambda}' g_n(x, \hat{\theta}_n)) - 1\right)dP_*$$

which, by the Lebesgue dominated convergence theorem converges to 0. Along similar lines, we also have that

$$\int g_n(x,\hat{\theta}_n)g_n(x,\hat{\theta}_n)'\exp(\dot{\lambda}'g_n(x,\hat{\theta}_n))dQ_n = \int g_n(x,\hat{\theta}_n)g_n(x,\hat{\theta}_n)'dP_* + o(1).$$

This completes the justification of (32). Thus, the right hand side of (32) yields:

$$1 - \frac{1}{8}\hat{\lambda}_n' Var_{P_*}(g_n(x,\hat{\theta}_n))\hat{\lambda}_n + o(|\hat{\lambda}_n|^2) = 1 + O(n^{-1}). \tag{33}$$

It is not hard to see that $Var_{P_*}(g_n(x,\hat{\theta}_n)) = Var_{P_*}(g(x,\hat{\theta}_n)) + o(1)$. Hence, (33) implies that:

$$\hat{\lambda}_n' Var_{P_*}(g(x,\hat{\theta}_n))\hat{\lambda}_n + |\hat{\lambda}_n|^2 o(1) = O(n^{-1}).$$

We conclude, using Assumption 1(v) that

$$(\underline{\ell} + o(1))|\hat{\lambda}_n|^2 = O(n^{-1}),$$

that is $|\hat{\lambda}_n| = O(n^{-1/2})$, or $\bar{T}_1(\bar{T}_{Q_n}, Q_n) = O(n^{-1/2})$.

(iii) This is obtained along the same lines as Step 3 in the proof of Lemma B.2 with Q_n, $\hat{\lambda}_n$ and $\bar{T}_{Q_n}$ replacing P_n, $\hat{\lambda}(\hat{\theta})$ and $\hat{\theta}$, respectively. □

Lemma C.4 *If Assumptions 1, 2 (with expectation and variance taken under P_*), and Assumption 3 hold, then: for each $r > 0$ and any sequence $Q_n \in B_H(P_*, r/\sqrt{n})$,*

(i) $\bar{T}_1(\theta^*, Q_n) = O(n^{-1/2})$,

(ii) $\mathbf{\Delta}_{Q_n}(\bar{T}_1(\theta^*, Q_n), \theta^*) = 1 + O(n^{-1})$.

Proof of Lemma C.4

(i) Let $\lambda_n^* = \bar{T}_1(\theta^*, Q_n)$. By a second-order mean-value expansion and by definition, we have:

$$\begin{aligned} & \int \exp(\lambda_n^{*'} g_n(x,\theta^*))dQ_n \\ = \quad & -1 - \lambda_n^{*'}\int g_n(x,\theta^*)dQ_n - \frac{1}{2}\lambda_n^{*'}\int g_n(x,\theta^*)g_n(x,\theta^*)'\exp(\dot{\lambda}'g_n(x,\theta^*))dQ_n\lambda_n^* \geq -1, \end{aligned}$$

with $\dot{\lambda}_n \in (0, \lambda_n^*)$. Hence,

$$\lambda_n^{*'}\int g_n(x,\theta^*)g_n(x,\theta^*)'\exp(\dot{\lambda}'g_n(x,\theta^*))dQ_n\lambda_n^* \leq 2|\lambda_n^*|\left|\int g_n(x,\theta^*)dQ_n\right|.$$

Using a similar argument as that in the proof of Lemma C.3, we have

$$\int g_n(x,\theta^*)g_n(x,\theta^*)'\exp(\dot{\lambda}'g_n(x,\theta^*))dQ_n = \int g_n(x,\theta^*)g_n(x,\theta^*)'dP_* + o(1) \succeq Var_{P_*}(g_n(x,\theta^*)) + o(1).$$

Thus we have

$$(\underline{\ell} + o(1))|\lambda_n^*|^2 \leq |\lambda_n^*|\left|\int g_n(x,\theta^*)dQ_n\right|.$$

To conclude (i), we just need to show that $\left|\int g_n(x,\theta^*)dQ_n\right| = O(n^{-1/2})$. This can readily be deduce from the proof of Lemma 7.1(i) of KOE (2013).

(ii) This is obtained by a second-order Taylor expansion of $\mathbf{\Delta}_{Q_n}(\bar{T}_1(\theta^*, Q_n), \theta^*)$ in its first argument around $\lambda = 0$. The first term in this expansion is 1, the second term in this expansion is nil whereas the second derivative is, similarly to (32), equal to $Var_{P_*}(g(x,\theta^*)) + o(1)$ which is finite. Using (i), we can deduce the claim. □

Lemma C.5 *If Assumptions 1, 2 (with expectation and variance taken under P_*), and Assumption 3 hold, then: for each $r > 0$ and any sequence $Q_n \in B_H(P_*, r/\sqrt{n})$,*

$$\sqrt{n}(\bar{T}_{Q_n} - \theta^*) = -\Sigma G' \Omega^{-1} \sqrt{n} \int g_n(x, \theta^*) dQ_n + o(1). \tag{34}$$

Proof of Lemma C.5 Let $\hat{\theta}_n \equiv \bar{T}_{Q_n}$, $\hat{\lambda}_n(\theta) \equiv \bar{T}_1(\theta, Q_n)$ and, for economy of notation, $\hat{\lambda}_n \equiv \bar{T}_1(\hat{\theta}_n, Q_n)$. Since $\hat{\theta}_n \to \theta^*$, Lemma C.6 ensures that, $\hat{\theta}_n$ satisfies the first order optimality condition:

$$\frac{d}{d\theta} \mathbf{\Delta}_{Q_n}(\hat{\lambda}_n(\theta), \theta)\bigg|_{\theta = \hat{\theta}_n} = 0,$$

that is

$$\frac{N_{1n}(\hat{\lambda}_n, \hat{\theta}_n)}{D_{1n}(\hat{\lambda}_n, \hat{\theta}_n)} - \frac{N_{2n}(\hat{\lambda}_n, \hat{\theta}_n)}{D_{2n}(\hat{\lambda}_n, \hat{\theta}_n)} = 0, \tag{35}$$

with $N_{1n}(\lambda, \theta)$, $D_{1n}(\lambda, \theta)$, $N_{2n}(\lambda, \theta)$, $D_{2n}(\lambda, \theta)$ defined similarly to $N_1(\lambda, \theta)$, $D_1(\lambda, \theta)$, $N_2(\lambda, \theta)$, $D_2(\lambda, \theta)$ in Equation (24) with $\hat{\lambda}(\theta)$, P_n and g replaced by $\hat{\lambda}_n(\theta)$, Q_n and g_n, respectively.

Note that $N_{1n}(0, \theta^*) = N_{2n}(0, \theta^*) = \frac{1}{2}\frac{d\hat{\lambda}_n(\theta^*)}{d\theta} \int g_n(x, \theta^*) dQ_n$ and $D_{1n}(0, \theta^*) = D_{2n}(0, \theta^*) = 1$. Hence, a mean-value expansion of (35) around $(0, \theta^*)$ yields

$$0 = \frac{\partial}{\partial \theta'} \left(\frac{N_{1n}(\lambda, \theta)}{D_{1n}(\lambda, \theta)} - \frac{N_{2n}(\lambda, \theta)}{D_{2n}(\lambda, \theta)} \right)\bigg|_{(\dot{\lambda}, \dot{\theta})} (\hat{\theta}_n - \theta^*) + \frac{\partial}{\partial \lambda'} \left(\frac{N_{1n}(\lambda, \theta)}{D_{1n}(\lambda, \theta)} - \frac{N_{2n}(\lambda, \theta)}{D_{2n}(\lambda, \theta)} \right)\bigg|_{(\dot{\lambda}, \dot{\theta})} \hat{\lambda}_n, \tag{36}$$

with $(\dot{\lambda}, \dot{\theta}) \in (0, \hat{\lambda}_n) \times (\theta^*, \hat{\theta}_n)$ and may differ from row to row. The expressions of

$$\frac{\partial N_{jn}}{\partial \theta'}, \quad \frac{\partial D_{jn}}{\partial \theta'}, \quad \frac{\partial N_{jn}}{\partial \lambda'}, \quad \frac{\partial D_{jn}}{\partial \lambda'},$$

$j = 1, 2$are analogue to the expressions of the partial derivatives of N_j and D_j as given following (25) with, again, $\hat{\lambda}(\theta)$, g and P_n replaced by $\hat{\lambda}_n(\theta)$, g_n and Q_n, respectively. Thanks to Lemma C.7, we have, for $j = 1, 2$:

$$\frac{\partial N_{jn}}{\partial \theta'}(\dot{\lambda}, \dot{\theta}) = \frac{1}{2}\frac{d\hat{\lambda}_n(\dot{\theta})'}{d\theta} \int \frac{\partial g_n}{\partial \theta'}(x, \dot{\theta}) dQ_n + o(1),$$

$$D_{jn}(\dot{\lambda}, \dot{\theta}) = 1 + o(1), \quad \frac{\partial D_{jn}}{\partial \theta'}(\dot{\lambda}, \dot{\theta}) = o(1), \quad \frac{\partial D_{jn}}{\partial \lambda'}(\dot{\lambda}, \dot{\theta}) = o(1),$$

$$\frac{\partial N_{1n}}{\partial \lambda'}(\dot{\lambda}, \dot{\theta}) = \frac{1}{2} \int \frac{\partial g_n'}{\partial \theta}(x, \dot{\theta}) dQ_n + \frac{1}{4}\frac{d\hat{\lambda}_n(\dot{\theta})'}{d\theta} \int g_n(x, \dot{\theta}) g_n(x, \dot{\theta})' dQ_n + o(1),$$

and

$$\frac{\partial N_{2n}}{\partial \lambda'}(\dot{\lambda}, \dot{\theta}) = \frac{1}{2} \int \frac{\partial g_n'}{\partial \theta}(x, \dot{\theta}) dQ_n + \frac{1}{2}\frac{d\hat{\lambda}_n(\dot{\theta})'}{d\theta} \int g_n(x, \dot{\theta}) g_n(x, \dot{\theta})' dQ_n + o(1).$$

As a result,

$$\frac{\partial}{\partial \theta'} \left(\frac{N_{1n}(\lambda, \theta)}{D_{1n}(\lambda, \theta)} - \frac{N_{2n}(\lambda, \theta)}{D_{2n}(\lambda, \theta)} \right)\bigg|_{(\dot{\lambda}, \dot{\theta})} = o(1)$$

and

$$\frac{\partial}{\partial \lambda'} \left(\frac{N_{1n}(\lambda, \theta)}{D_{1n}(\lambda, \theta)} - \frac{N_{2n}(\lambda, \theta)}{D_{2n}(\lambda, \theta)} \right)\bigg|_{(\dot{\lambda}, \dot{\theta})} = -\frac{1}{4}\frac{d\hat{\lambda}_n(\dot{\theta})'}{d\theta} \int g_n(x, \dot{\theta}) g_n(x, \dot{\theta})' dQ_n + o(1).$$

Also, from Lemma C.6,

$$\frac{d\hat{\lambda}_n(\dot{\theta})}{d\theta'} = -\left(\int g_n(x,\dot{\theta})g_n(x,\dot{\theta})'dQ_n\right)^{-1}\int \frac{\partial g_n(x,\dot{\theta})}{\partial\theta'}dQ_n + o(1).$$

The expansion in (36) becomes:

$$G'\sqrt{n}\hat{\lambda}_n = o(1) + o(\sqrt{n}|\hat{\theta}_n - \theta^*|). \tag{37}$$

Turning to the first order condition for $\hat{\lambda}_n$, we know that $\hat{\lambda}_n = O(n^{-1/2})$ and therefore is interior to to Λ_n for n large enough. Therefore, $\hat{\lambda}_n$ solves:

$$\int g_n(x,\hat{\theta}_n)\exp\left(\hat{\lambda}_n' g_n(x,\hat{\theta}_n)\right)dQ_n = 0.$$

A first-order mean-value expansion of this equation around $(0,\theta^*)$ yields:

$$\begin{aligned} 0 \;=\; & \int g_n(x,\theta^*)dQ_n + \int\left(I_m + g_n(x,\dot{\theta})\dot{\lambda}'\right)\frac{\partial g_n}{\partial\theta'}(x,\dot{\theta})\exp\left(\dot{\lambda}'g_n(x,\dot{\theta})\right)dQ_n(\hat{\theta}_n - \theta^*) \\ & + \int g_n(x,\dot{\theta})g_n(x,\dot{\theta})'\exp\left(\dot{\lambda}'g_n(x,\dot{\theta})\right)dQ_n\hat{\lambda}_n, \end{aligned}$$

with $(\dot{\lambda},\dot{\theta}) \in (0,\hat{\lambda}_n)\times(\theta^*,\hat{\theta}_n)$ and may differ from row to row. By similar arguments as previously made, we get:

$$G\sqrt{n}(\hat{\theta}_n - \theta^*) + \Omega\sqrt{n}\hat{\lambda}_n = -\sqrt{n}\int g_n(x,\theta^*)dQ_n + o(1) + o(|\sqrt{n}(\hat{\theta}_n - \theta^*)|). \tag{38}$$

Using (37) and (38), we get

$$\sqrt{n}(\hat{\theta}_n - \theta^*) + o(|\sqrt{n}(\hat{\theta}_n - \theta^*)|) = -\sqrt{n}\Sigma G'\Omega^{-1}\int g_n(x,\theta^*)dQ_n + o(1)$$

which is sufficient to deduce the result. □

Lemma C.6 *If Assumptions 1, 2 (with expectation and variance taken under P_*), and Assumption 3 hold, then: for each $r > 0$, for n large enough and any sequence $Q_n \in B_H(P_*, r/\sqrt{n})$, the functions $\theta \to \bar{T}_1(\theta, Q_n)$ and $\theta \to \mathbf{\Delta}_{Q_n}(\bar{T}_1(\theta,Q_n),\theta)$ are continuously differentiable in a neighborhood of θ^* contained in Λ_n and on that neighborhood,*

$$\frac{\partial\bar{T}_1(\theta,Q_n)}{\partial\theta'} = -\left(\int \partial g_n(x,\theta)g_n(x,\theta)'dQ_n\right)^{-1}\int\frac{g_n(x,\theta)}{\partial\theta'}dQ_n + o(1).$$

Proof of Lemma C.6 From Lemma C.4, $\bar{T}_1(\theta^*,Q_n) = O(n^{-1/2})$ therefore, $\bar{T}_1(\theta^*,Q_n)$ is an interior point to Λ_n for n large enough. The fact that $|\lambda' g_n(x,\theta)|$ is bounded and Assumption 1(iv) ensure that $\lambda \to \int\exp(\lambda' g_n(x,\theta))dQ_n$ is differentiable on Θ and we can claim that $(\bar{T}_1(\theta^*,Q_n),\theta^*)$ solves the first order condition:

$$H_n(\lambda,\theta) \equiv \int g_n(x,\theta)\exp(\lambda' g_n(x,\theta))dQ_n = 0.$$

From the proof of Lemma C.2(i), we know that $\int g_n(x,\theta^*)g_n(x,\theta^*)'\exp(\lambda' g_n(x,\theta^*))dQ_n$ is positive definite. The implicit function theorem guarantees that the $H_n(\lambda,\theta) = 0$ defines an implicit function $\lambda_n(\theta)$ in the neighborhood of θ^* contained in Λ_n. The fact that $\lambda \to \int\exp(\lambda' g_n(x,\theta))dQ_n$ is strictly convex (see proof of Lemma C.2(i)) means that its global minimum solves $H_n(\lambda,\theta) = 0$. Thus, $\lambda_n(\theta) =$

$\bar{T}_1(\theta, Q_n)$ on that neighborhood. Note that under Assumptions 1(iv) and 2(i), this neighborhood can be so chosen that $(\theta, \lambda) \to H_n(\lambda, \theta)$ is continuously differentiable, property inherited by the implicit function $\lambda_n(\theta)$, therefore by $\theta \to \bar{T}_1(\theta, Q_n)$. We have:

$$\begin{aligned}\frac{\partial \bar{T}_1(\theta,Q_n)}{\partial \theta'} = & -\left(\int g_n(x,\theta) g_n(x,\theta)' \exp(\lambda' g_n(x,\theta)) dQ_n\right)^{-1} \times \\ & \int \left(I_m + g_n(x,\theta)\lambda'\right) \frac{\partial g_n(x,\theta)}{\partial \theta'} \exp(\lambda' g_n(x,\theta)) dQ_n,\end{aligned}$$

with $\lambda = \bar{T}_1(\theta, Q_n)$. But since $|\lambda' g_n(x,\theta)| \to 0$ over $(\lambda, \theta) \in \Lambda_n \times \Theta$, the term in the brackets and the second term in the product are equal, up to $o(1)$, to $\int g_n(x,\theta) g_n(x,\theta)' dQ_n$ and $\int \frac{\partial g_n(x,\theta)}{\partial \theta'} dQ_n$, respectively; yielding the expected result.

The differentiability of $\theta \to \mathbf{\Delta}_{Q_n}(\bar{T}_1(\theta, Q_n), \theta)$ follows from the fact that of $(\lambda, \theta) \to \mathbf{\Delta}_{Q_n}(\lambda, \theta)$ and $\theta \to \bar{T}_1(\theta, Q_n)$. $\square$

Lemma C.7 *Let $h(x,\theta)$ be a function measurable on $\mathcal{X}$ for each $\theta \in \Theta$ taking value in $\mathbb{R}^\ell$ and let $h_n(x,\theta) = h(x,\theta)\mathbb{I}(x \in \mathcal{X}_n)$. If $\sup_{\theta \in \mathcal{N}, x \in \mathcal{X}_n} |h(x,\theta)| = o(n)$ and $E_{P_*}\left(\sup_{\theta \in \mathcal{N}} |h(x,\theta)|^2\right) < \infty$ for $\mathcal{N} \subset \Theta$, we have:*

$$\sup_{\theta \in \mathcal{N}} \left| \int h_n(x,\theta) \exp(\lambda' g_n(x,\theta)) dQ_n - \int h_n(x,\theta) dP_* \right| = o(1),$$

for any $\lambda \in \Lambda_n$ and any sequence $Q_n \in B_H(P_*, r/\sqrt{n})$.

Proof of Lemma C.7 We have:

$$\begin{aligned}
& \left| \int h_n(x,\theta) \exp(\lambda' g_n(x,\theta)) dQ_n - \int h_n(x,\theta) dP_* \right| \\
= & \left| \int h_n(x,\theta) \exp(\lambda' g_n(x,\theta)) (dQ_n - dP_*) + \int h_n(x,\theta) (\exp(\lambda' g_n(x,\theta)) - 1) dP_* \right| \\
\leq & \left| \int h_n(x,\theta) \exp(\lambda' g_n(x,\theta)) (dQ_n - dP_*) \right| + \left| \int h_n(x,\theta) (\exp(\lambda' g_n(x,\theta)) - 1) dP_* \right| \equiv (1) + (2).
\end{aligned}$$

Since $|\lambda' g_n(x,\theta)| \to 0$, $\exp(\lambda' g_n(x,\theta))$ is bounded and we have:

$$\begin{aligned}
(1) &= C \left| \int h_n(x,\theta)(dQ_n - dP_*) \right| \\
&\leq C \left| \int h_n(x,\theta) \left\{ \left(dQ_n^{1/2} - dP_*^{1/2}\right)^2 + 2 dP_*^{1/2} \left(dQ_n^{1/2} - dP_*^{1/2}\right) \right\} \right| \\
&\leq C \sup_{\theta \in \mathcal{N}, x \in \mathcal{X}_n} |h_n(x,\theta)| \int \left(dQ_n^{1/2} - dP_*^{1/2}\right)^2 \\
&\quad + 2C \left(\int \sup_{\theta \in \mathcal{N}} |h_n(x,\theta)|^2 dP_* \right)^{1/2} \left(\int \left(dQ_n^{1/2} - dP_*^{1/2}\right)^2 \right)^{1/2} \\
&\leq o(n) \frac{r^2}{n} + \text{const.} \frac{r}{\sqrt{n}} = o(1).
\end{aligned}$$

Again, since $|\lambda' g_n(x,\theta)| \to 0$, we have $|h_n(x,\theta)(\exp(\lambda' g_n(x,\theta)) - 1)| \leq C \sup_{\theta \in \mathcal{N}} |h_n(x,\theta)|$ which has finite expectation under P_*. We can therefore deduce by the Lebesgue dominated theorem that $(2) = o(1)$. $\square$

Lemma C.8 *Let $r > 0$ and Q_n be a sequence contained in $B_H(P_*, r/\sqrt{n})$. If Assumptions 1, 2 (with expectation and variance taken under P_*), and Assumption 3 hold, then we have:*

$$\sqrt{n}(\bar{T}_{P_n} - \theta^*) \quad = \quad -\Sigma G'\Omega^{-1}\sqrt{n}\int g_n(x,\theta^*)dP_n + o_P(1) \quad \textit{under } Q_n$$

$$\sqrt{n}(\bar{T}_{P_n} - \bar{T}_{Q_n}) \ \stackrel{d}{\to} \ N(0,\Sigma), \quad \textit{under } Q_n$$

Proof of Lemma C.8: To be completed. □

Lemma C.9 *Let $r > 0$ and Q_n be a sequence contained in $B_H(P_*, r/\sqrt{n})$. If Assumptions 1, 2 (with expectation and variance taken under P_*), and Assumption 3 hold, then, the following statements hold under Q_n:*

(i) $\bar{T}_1(\theta^*, P_n) = O_P(n^{-1/2})$,

(ii) $E_{P_n}\left(g_n(x,\bar{T}_{P_n})\right) = O_P(n^{-1/2})$, $\quad E_{P_n}\left(g_n(x,\bar{T}_{P_n})g_n(x,\bar{T}_{P_n})'\right) = \Omega + O_P(n^{-1/2})$, *and* $E_{P_n}\left(\frac{\partial g_n}{\partial \theta'}(x,\bar{T}_{P_n})\right) = G + o_P(1)$,

(iii) $\bar{T}_1(\bar{T}_{P_n}, P_n) = O_P(n^{-1/2})$.

Proof of Lemma C.9: To be completed. □

D Global misspecification

Proof of Theorem 5.1:

The proof is split into two parts: in (i), we show the consistency of $\hat{\theta}$ and $\lambda(\hat{\theta})$; in (ii), we derive the asymptotic distribution of the estimators.

(i) First, we show the consistency of $\hat{\lambda}$ and $\hat{\theta}$.

Let Λ^* be an open neighborhood of Λ that contains λ^*. Its complement, $\Lambda \setminus \Lambda^*$ is a closed subset of a compact set, and so it is compact as well. The strict convexity of the function $E(\exp(\lambda' g(x,\theta)))$ implies that, for every neighborhood Λ^* (as above), there exists a constant $\kappa > 0$ such that

$$\inf_{\lambda\in\Lambda(\theta)\setminus\Lambda^*} E\left(\exp(\lambda' g(x,\theta)\right) > E\left(\exp(\lambda^{*\prime} g(x,\theta))\right) + \kappa\,,$$

uniformly in $\theta \in \Theta$; the uniformity follows from the fact that $\Lambda \setminus \Lambda^*$ is compact. Since,

$$\hat{\lambda} \in \Lambda \setminus \Lambda^* \quad \Rightarrow \quad E\left(\exp(\hat{\lambda}' g(x,\theta))\right) > E\left(\exp(\lambda^{*\prime} g(x,\theta))\right) + \kappa$$

we have,

$$P\left(\hat{\lambda} \in \Lambda \setminus \Lambda^*\right) \quad \leq \quad P\left(E\left[\exp(\hat{\lambda}' g(x,\theta))\right] > E\left[\exp(\lambda^{*\prime} g(x,\theta))\right] + \kappa\right)\,.$$

We now show that the above probability on the RHS converges to zero, which implies that $\hat{\lambda}$ converges to λ^* since the proof was done for any neighborhood Λ^* containing λ^*.

$$
\begin{aligned}
P\left(E\left[\exp(\hat{\lambda}'g(x,\theta))\right] > E\left[\exp(\lambda^{*\prime}g(x,\theta))\right] + \kappa\right) \quad = \quad & P\Bigg(E\left[\exp(\hat{\lambda}'g(x,\theta))\right] - \int \exp(\hat{\lambda}'g(x,\theta))dP_n \\
& + \int \exp(\hat{\lambda}'g(x,\theta))dP_n - \int \exp(\lambda^{*\prime}g(x,\theta))dP_n \\
& + \int \exp(\lambda^{*\prime}g(x,\theta))dP_n - E\left[\exp(\lambda^{*\prime}g(x,\theta))\right] > \kappa\Bigg) \\
\stackrel{n}{\to} \quad & 0
\end{aligned}
$$

The convergence to zero follows from the uniform convergence of $\int \exp(\lambda'g(x,\theta))dP_n$ to $E(\exp(\lambda'g(x,\theta)))$ (which follows from 5(iv); see also the proof of Lemma B.1) and from the definition of $\hat{\lambda}$.

To prove the consistency of $\hat{\theta}$, we will make use of the consistency of $\hat{\lambda}$. Similarly to the proof of Lemma B.1, we can justify a uniform convergence of the objective function $\Delta_n(\lambda(\theta),\theta)$ over (Λ,Θ) which implies that:

$$
\begin{aligned}
& \forall\epsilon > 0 \ \lim_n P\left(|\Delta_n(\lambda(\hat{\theta}),\hat{\theta}) - \Delta(\lambda(\hat{\theta}),\hat{\theta})| < \epsilon/3\right) = 1 \\
\Rightarrow \quad & \forall\epsilon > 0 \ \lim_n P\left(\Delta_n(\lambda(\hat{\theta}),\hat{\theta}) < \Delta(\lambda(\hat{\theta}),\hat{\theta}) + \epsilon/3\right) = 1
\end{aligned} \tag{39}
$$

Similarly, we can show that

$$
\forall\epsilon > 0 \ \lim_n P\left(\Delta(\lambda(\theta^*),\theta^*) < \Delta_n(\lambda(\theta^*),\theta^*) + \epsilon/3\right) = 1 \tag{40}
$$

By definition of $\hat{\theta}$, we have:

$$
\forall\epsilon > 0 \ \lim_n P\left(\Delta_n(\lambda(\theta^*),\theta^*) < \Delta_n(\lambda(\hat{\theta}),\hat{\theta}) + \epsilon/3\right) = 1 \tag{41}
$$

From equations (39) and (41), we get:

$$
\forall\epsilon > 0 \ \lim_n P\left(\Delta_n(\lambda(\theta^*),\theta^*) < \Delta(\lambda(\hat{\theta}),\hat{\theta}) + 2\epsilon/3\right) = 1 \tag{42}
$$

We can now use equation (40) to deduce:

$$
\forall\epsilon > 0 \ \lim_n P\left(\Delta(\lambda(\theta^*),\theta^*) < \Delta(\lambda(\hat{\theta}),\hat{\theta}) + \epsilon\right) = 1 \tag{43}
$$

We now use the identification assumption and the definition of $\hat{\theta}$ to deduce that, for every neighborhood $\mathcal{N}^*$ of θ^*, there exists a constant $\eta > 0$ such that

$$
\exists\eta > 0 \ / \ \sup_{\theta\in\Theta\backslash\mathcal{N}^*} \Delta(\lambda(\theta),\theta) + \eta < \Delta(\lambda(\theta^*),\theta^*) \,.
$$

Then, we have:

$$
\begin{aligned}
\hat{\theta} \in \Theta \setminus \mathcal{N}^* \quad \Rightarrow \quad & \Delta(\lambda(\hat{\theta}),\hat{\theta}) \leq \sup_{\theta\in\Theta\backslash\mathcal{N}^*} \Delta(\lambda(\theta),\theta) + \eta < \Delta(\lambda(\theta^*),\theta^*) \\
\Rightarrow \quad & P\left(\hat{\theta} \in \Theta \setminus \mathcal{N}^*\right) \leq P\left(\Delta(\lambda(\hat{\theta}),\hat{\theta}) < \Delta(\lambda(\theta^*),\theta^*)\right) \\
\stackrel{n}{\to} \quad & 0
\end{aligned}
$$

where the convergence to 0 follows directly from equation (43) above.

(ii) To derive the asymptotic distribution of ETHD estimator under global misspecification, we follow the proof of Theorem 3.3 and write a mean-value expansion of the FOC around (θ^*, λ^*).

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} N_1(\lambda^*, \theta^*)/D_1(\lambda^*, \theta^*) - N_2(\lambda^*, \theta^*)/D_2(\lambda^*, \theta^*) \\ \int g(x, \theta^*) \exp(\lambda^{*\prime} g(x, \theta^*)) dP_n \end{pmatrix} + \overline{R}_n \begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\lambda} - \lambda^* \end{pmatrix} \tag{44}$$

where, with $\overline{\theta} \in (\theta^*, \hat{\theta})$ and $\overline{\lambda} \in (\lambda^*, \hat{\lambda})$,

$$\overline{R}_n = \begin{pmatrix} R_{\theta,\theta}(\overline{\theta}, \overline{\lambda}) & R_{\theta,\lambda}(\overline{\theta}, \overline{\lambda}) \\ R_{\lambda,\theta}(\overline{\theta}, \overline{\lambda}) & R_{\lambda,\lambda}(\overline{\theta}, \overline{\lambda}) \end{pmatrix},$$

$$\begin{aligned}
R_{\theta,\theta}(\theta, \lambda) &= \frac{\partial}{\partial \theta} \left(\frac{N_1(\lambda, \theta)}{D_1(\lambda, \theta)} - \frac{N_2(\lambda, \theta)}{D_2(\lambda, \theta)} \right) \\
R_{\theta,\lambda}(\theta, \lambda) &= \frac{\partial}{\partial \lambda} \left(\frac{N_1(\lambda, \theta)}{D_1(\lambda, \theta)} - \frac{N_2(\lambda, \theta)}{D_2(\lambda, \theta)} \right) \\
R_{\lambda,\theta}(\theta, \lambda) &= \int \left[\frac{\partial g'(x, \theta)}{\partial \theta} + \frac{\partial g'(x, \theta)}{\partial \theta} \lambda g(x, \theta)' \right] \exp(\lambda' g(x, \theta)) dP_n \\
R_{\lambda,\lambda}(\theta, \lambda) &= \int g(x, \theta) g'(x, \theta) \exp(\lambda' g(x, \theta)) dP_n
\end{aligned}$$

and D_i, N_i and the above derivatives have been defined and computed in the proof of Theorem 3.3. We then get:

$$\begin{aligned}
R\sqrt{n} \begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\lambda} - \lambda^* \end{pmatrix} &= -\sqrt{n} \begin{pmatrix} N_1(\lambda^*, \theta^*)/D_1(\lambda^*, \theta^*) - N_2(\lambda^*, \theta^*)/D_2(\lambda^*, \theta^*) \\ \int g(x, \theta^*) \exp(\lambda^{*\prime} g(x, \theta^*)) dP_n \end{pmatrix} + o_p(1) \\
&\equiv \sqrt{n} A_n^* + o_p(1)
\end{aligned} \tag{45}$$

with $\texttt{Plim}\left[\overline{R}_n\right] = R$ and

$$\begin{aligned}
& \frac{N_1(\lambda^*, \theta^*)}{D_1(\lambda^*, \theta^*)} - \frac{N_2(\lambda^*, \theta^*)}{D_2(\lambda^*, \theta^*)} \\
= \; & \frac{1}{2(\int \exp(\lambda^{*\prime} g(x, \theta^*)) dP_n)^{1/2}} \left[\frac{d\hat{\lambda}^{*\prime}}{d\theta} \int g(x, \theta^*) \exp(\lambda^{*\prime} g(x, \theta^*)/2) dP_n + \lambda^{*\prime} \int \frac{\partial g(x, \theta^*)'}{\partial \theta} \exp(\lambda^{*\prime} g(x, \theta^*)/2) dP_n \right. \\
& \left. - \frac{d\hat{\lambda}^{*\prime}}{d\theta} \int g(x, \theta^*) \frac{\exp(\lambda^{*\prime} g(x, \theta^*))}{\int \exp(\lambda^{*\prime} g(x, \theta^*)) dP_n} dP_n - \lambda^{*\prime} \int \frac{\partial g(x, \theta^*)'}{\partial \theta} \frac{\exp(\lambda^{*\prime} g(x, \theta^*))}{\int \exp(\lambda^{*\prime} g(x, \theta^*)) dP_n} dP_n \right]
\end{aligned}$$

And by CLT on A_n, we have:

$$\sqrt{n} \begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\lambda} - \lambda^* \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, R^{-1} \Omega^* R^{-1}) \qquad \text{with} \qquad \Omega^* = \texttt{AVar}\left(A_n^*\right)$$

The expected result directly follows.

We now show that under correct specification, the expansion (45) coincides with (29), that is:

$$\begin{pmatrix} G' & 0 \\ \Omega & G \end{pmatrix} \sqrt{n} \begin{pmatrix} \hat{\theta} - \theta^* \\ \hat{\lambda} \end{pmatrix} = \begin{pmatrix} 0 \\ -\sqrt{n} \int g(x, \theta^*) dP_n \end{pmatrix} + o_p(1)$$

After replacing λ^* by 0, we easily get that

$$\frac{N_1(\lambda^*, \theta^*)}{D_1(\lambda^*, \theta^*)} - \frac{N_2(\lambda^*, \theta^*)}{D_2(\lambda^*, \theta^*)} = 0$$

It remains to show that

$$\texttt{Plim}\overline{R}_n = \begin{pmatrix} 0 & G \\ G' & \Omega \end{pmatrix}$$

After replacing λ^* by 0, we easily get that

$$\begin{aligned}
R_{\lambda,\theta}(\theta^*,\lambda^*) &= \int \frac{\partial g(x,\theta^*)}{\partial\theta} dP_n \\
R_{\lambda,\lambda}(\theta^*,\lambda^*) &= \int g(x,\theta^*)g'(x,\theta^*)dP_n \\
R_{\theta,\lambda}(\theta^*,\lambda^*) &= \frac{\partial N_1(\lambda,\theta)}{\partial\lambda} - \frac{\partial N_1(\lambda,\theta)}{\partial\lambda} \\
&\quad \text{since } D_1(\lambda^*,\theta^*) = D_2(\lambda^*,\theta^*) = 1 \text{ and } \partial D_1(\lambda^*,\theta^*)/\partial\lambda = \partial D_2(\lambda^*,\theta^*)/\partial\lambda = 0 \\
&= -\frac{1}{2}\int \frac{d\hat{\lambda}'}{d\theta} g(x,\theta^*)g'(x,\theta^*)dP_n \\
R_{\theta,\theta}(\theta^*,\lambda^*) &= \frac{\partial N_1(\lambda,\theta)}{\partial\theta} - \frac{\partial N_1(\lambda,\theta)}{\partial\theta} \\
&\quad \text{since } D_1(\lambda^*,\theta^*) = D_2(\lambda^*,\theta^*) = 1 \text{ and } \partial D_1(\lambda^*,\theta^*)/\partial\theta = \partial D_2(\lambda^*,\theta^*)/\partial\theta = 0 \\
&= 0
\end{aligned}$$

And the expected result follows readily.