Unit Averaging for Heterogeneous Panels

Christian Brownlees[†] Vladislav Morozov[‡,*]

February 15, 2023

Abstract

In this work we introduce a unit averaging procedure to efficiently recover unit-specific parameters in a heterogeneous panel model. The procedure consists in estimating the parameter of a given unit using a weighted average of all the unit-specific parameter estimators in the panel. The weights of the average are determined by minimizing an MSE criterion. We analyze the properties of the minimum MSE unit averaging estimator in a local heterogeneity framework inspired by the literature on frequentist model averaging. The analysis of the estimator covers both the cases in which the cross-sectional dimension of the panel is fixed and large. In both cases we obtain the local asymptotic distribution of the minimum MSE unit averaging estimators and of the associated weights. A GDP nowcasting application for a panel of European countries showcases the benefits of the procedure.

Keywords: heterogeneous panels, frequentist model averaging, prediction

JEL: C33, C52, C53

1 Introduction

Estimation of unit-specific parameters in panel data models with heterogeneous parameters is a topic of active research in econometrics (Maddala, Trost, Li, and Joutz, 1997; Pesaran,

[†] Department of Economics and Business, Universitat Pompeu Fabra and Barcelona School of Economics; e-mail: christian.brownlees@upf.edu.
[‡] Department of Economics and Business, Universitat Pompeu Fabra and Barcelona School of Economics; e-mail: vladislav.morozov@upf.edu.
[*] Corresponding author.
We thank Jan Ditzen, Kirill Evdokimov, Geert Mesters, Luca Neri, Katerina Petrova, Barbara Rossi, Wendun Wang, and the participants at 26th International Panel Data Conference, 7th RCEA Time Series Workshop, 9th SIDE WEEE, and the 2021 Econometric Research in Finance Workshop for comments and discussion. Christian Brownlees acknowledges support from the Spanish Ministry of Science and Technology (Grant MTM2012-37195) and the Spanish Ministry of Economy and Competitiveness through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2011-0075).

Shin, and Smith, 1999; Wang, Zhang, and Paap, 2019; Liu, Moon, and Schorfheide, 2020). Estimation of unit-specific parameters is relevant, for instance, when interest lies in constructing forecasts for the individual units in the panel (Baltagi, 2013; Zhang, Zou, and Liang, 2014; Wang et al., 2019; Liu et al., 2020), which typically arises in the analysis of international panels of macroeconomic time series (Marcellino, Stock, and Watson, 2003). Other unit-specific parameters of interest include individual slopes (Maddala et al., 1997; Maddala, Li, and Srivastava, 2001; Wang et al., 2019) and long-run effects of a change in a covariate (Pesaran and Smith, 1995; Pesaran et al., 1999).

There are three natural strategies for estimating unit-specific parameters (Baltagi, Bresson, and Pirotte, 2008). The simplest approach consists in estimating each unit-specific parameter from its individual time series. While this strategy typically leads to approximately unbiased estimation, the resulting estimators suffer from large estimation variability when the time dimension is small. In the second approach, an assumption of parameter homogeneity is imposed and a common panel-wide estimator is used for all unit-specific parameters. This strategy leads to small variability; however, it suffers from large bias in the presence of heterogeneity. The third strategy is a compromise between the first two. It combines pooled and individual approaches in the attempt to obtain an estimator with favorable risk properties (Maddala et al., 2001; Wang et al., 2019; Liu et al., 2020). This is appealing when the time dimension is moderate and there is a nontrivial bias-variance trade-off between individual-specific and panel-wide estimation.

In this paper we propose a novel unit-specific compromise estimator that we call the unit averaging estimator. The estimator is fairly general and is designed for possibly nonlinear panel models estimated by M-estimation. We are concerned with the estimation of a unit-specific "focus" parameter. Focus parameters considered include the examples mentioned above as well as other smooth transformations of the unit-specific parameter vector. The unit averaging estimator for the unit-specific focus parameter is then defined as a weighted average of all the unit-specific focus parameter estimators in the panel. The

weights of the average are chosen by minimizing a unit-specific MSE criterion. Specifically, we introduce two MSE criteria that differ in whether the cross-sectional dimension of the panel is treated as fixed (fixed-N) or large (large-N). The minimum MSE weights solve a quadratic optimization problem that is straightforward to compute in both approaches.

We analyze the theoretical properties of the proposed unit averaging methodology. The analysis is carried out using a local heterogeneity assumption, in which we assume that individual coefficients are local in the time dimension to a common mean. This is inspired by and builds upon the notion of local misspecification used in the frequentist model averaging literature (Hjort and Claeskens, 2003; Claeskens and Hjort, 2008; Hansen, 2008). Local heterogeneity may be interpreted as a theoretical device to emulate panels where the time dimension is moderate. In such a setting each unit carries information about the other units in the panel and thus may improve estimation accuracy. Using the local heterogeneity framework we derive the local asymptotic MSE of the unit averaging estimator as well as estimators for this quantity. The minimum MSE weights minimize the local asymptotic MSE estimators. As we show, these minimum MSE weights minimize an appropriately defined notion of the asymptotic MSE contaminated by a noise component that we characterize explicitly. Finally, we obtain the limiting distribution of the minimum MSE unit averaging estimator similarly to Liu (2015).

In a simulation study, we assess the finite sample properties of the proposed methodology. We compare our unit averaging estimator against the unit-specific estimator, the mean group estimator as well as alternative unit averaging estimators based on AIC weights, BIC weights and Mallows weights (Buckland, Burnham, and Augustin, 1997; Hansen, 2007; Wan, Zhang, and Zou, 2010). The simulation study shows that the proposed methodology performs favorably relative to these benchmarks. The improvement in MSE is strongest for those units whose individual parameter is sufficiently distant from both the mean of the distribution and the endpoints of the support.

An application to GDP nowcasting for a panel of European countries showcases the

methodology (Marcellino and Schumacher, 2010; Schumacher, 2016). GDP prediction is a natural application of the unit averaging methodology since the literature documents both evidence of heterogeneity between countries and the benefits of pooling data (Garcia-Ferrer, Highfield, Palm, and Zellner, 1987; Hoogstrate, Palm, and Pfann, 2000; Marcellino et al., 2003). We find that unit averaging using minimum MSE weights improves prediction accuracy and that the magnitude of the improvement is larger for shorter panels.

This paper is related to different strands of the literature. First, it is related to the literature on frequentist model averaging. Important contributions in this area include Hjort and Claeskens (2003), Hansen (2007), Hansen (2008), Wan et al. (2010), Hansen and Racine (2012), Liu (2015), and Gao, Zhang, Wang, and Zou (2016), among others. Gao et al. (2016); Yin, Liu, and Lin (2021) deal with model averaging estimators specifically tailored for panel models. The main difference with respect to these contributions is that we focus on averaging different units with the same model whereas these papers average different models. Second, this paper is related to the literature on unit-specific estimation using compromise estimators. Important contributions in this area include Zhang et al. (2014), Wang et al. (2019), Issler and Lima (2009) and Liu et al. (2020). The main difference with respect to these contributions is that we focus on a setting where the time dimension is moderate (as opposed to either large or small) and that we do not require strict exogeneity. Moreover, the existing literature largely focuses on linear models (Baltagi et al., 2008; Wang et al., 2019) whereas our framework allows for a large class of nonlinear models.

The rest of the paper is structured as follows. Section 2 introduces the unit averaging methodology. Section 3 studies the theoretical properties of the procedure. Section 4 contains the simulation study. Section 5 contains the empirical application. Concluding remarks follow in section 6. All proofs are collected in the proof appendix. Additional results are collected in the online appendix, available from the authors' websites.

2 Methodology

We introduce our unit averaging methodology within the framework of a fairly general class of panel data models with heterogeneous parameters. Let $\{\boldsymbol{z}_{it}\}$ with $i = 1, \ldots, N$ and $t = 1, \ldots, T$ denote a panel where $\boldsymbol{z}_{it}$ denotes a random vector of observations taking values in $\mathcal{Z} \subset \mathbb{R}^d$. For each unit in the panel, we define the unit-specific parameter $\boldsymbol{\theta}_i \in \Theta \subset \mathbb{R}^p$ as

$$\boldsymbol{\theta}_i = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\left(\frac{1}{T} \sum_{t=1}^{T} m(\boldsymbol{\theta}, \boldsymbol{z}_{it}) \right) ,$$

where $m : \Theta \times \mathcal{Z} \to \mathbb{R}$ denotes a smooth criterion function.

Our interest lies in estimating the unit-specific "focus" parameter $\mu(\boldsymbol{\theta}_i)$ for a fixed unit i with minimal MSE, where $\mu : \Theta \to \mathbb{R}$ denotes a smooth function (similarly to the setup in Hjort and Claeskens (2003)). For example, $\mu(\boldsymbol{\theta}_i)$ may denote a component of $\boldsymbol{\theta}_i$, the conditional mean of a response variable given the covariates, or the long-run effect of a covariate. To simplify exposition and without loss of generality, we focus on the problem of estimating the focus parameter $\mu(\boldsymbol{\theta}_1)$ for unit 1. In this paper we consider the case in which the focus function μ takes values in $\mathbb{R}$. It is straightforward to generalize the framework to a focus function taking values in $\mathbb{R}^q$ for some $q > 1$.

To estimate $\mu(\boldsymbol{\theta}_1)$ we consider the class of unit averaging estimators given by

$$\hat{\mu}(\boldsymbol{w}) = \sum_{i=1}^{N} w_i \mu(\hat{\boldsymbol{\theta}}_i) , \tag{1}$$

where $\boldsymbol{w} = (w_i)$ is a N-vector such that $w_i \geq 0$ for all i and $\sum_{i=1}^{N} w_i = 1$, and $\hat{\boldsymbol{\theta}}_i$ for $i = 1, \ldots, N$ is the unit-specific estimator given by

$$\hat{\boldsymbol{\theta}}_i = \arg\max_{\boldsymbol{\theta} \in \Theta} \frac{1}{T} \sum_{t=1}^{T} m(\boldsymbol{\theta}, \boldsymbol{z}_{it}) . \tag{2}$$

The class of estimators in (1) is fairly broad and contains a number of important special cases. It includes the individual estimator of unit 1 $\hat{\mu}_1 = \mu(\hat{\boldsymbol{\theta}}_1)$ and the mean group

estimator $\hat{\mu}_{MG} = N^{-1}\sum_{i=1}^{N}\mu(\hat{\boldsymbol{\theta}}_i)$.[1] Other important special cases include estimators based on smooth AIC/BIC weights (Buckland et al., 1997) and Mallows weights (Hansen, 2007; Wan et al., 2010), as well as a Stein-type estimator (Maddala et al., 1997) given by $\mu_{\text{Stein}} = \lambda\mu(\hat{\boldsymbol{\theta}}_1) + (1-\lambda)N^{-1}\sum_{i=1}^{N}\mu(\hat{\boldsymbol{\theta}}_i)$ where $\lambda \in [0,1]$.

The class of estimators in (1) may be motivated by the following representation for the individual parameters $\boldsymbol{\theta}_i$. Assume that $\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \boldsymbol{\eta}_i$ holds for each $i = 1,\ldots,N$, where $\boldsymbol{\theta}_0$ is a common mean component and $\boldsymbol{\eta}_i$ is a zero-mean random idiosyncratic component. In such a setup all units in the panel carry information on $\boldsymbol{\theta}_0$, and so all units may be useful for estimating $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \boldsymbol{\eta}_1$. The vector of weights $\boldsymbol{w}$ controls the balance between bias and variance of estimator (1). Assigning a large weight to unit 1 leads to low bias but may also lead to excessive variability. Alternatively, assigning large weights to units other than unit 1 induces bias but may substantially reduce variability. Such a trade-off is particularly relevant in a moderate-T setting, defined as the range of values of T for which the variability of the individual estimators $\hat{\boldsymbol{\theta}}_i$ is of the same order of magnitude as $\boldsymbol{\eta}_i$.

In this work we introduce a weighting scheme called minimum-MSE unit averaging weights. These weights seek to strike a balance between the bias and variance of the unit averaging estimator. We introduce two unit averaging schemes which differ in whether the cross-sectional dimension N is treated as fixed or large. In both regimes the weights are chosen by minimizing an estimator of the local asymptotic approximation to the MSE (LA-MSE) of the unit averaging estimator. The LA-MSE provides an approximation to the moderate-T MSE of the unit averaging estimator and is justified in detail in the next section.

In the fixed-N regime we average over a fixed finite collection of $\bar{N}$ units. In this regime we allow all units to have a non-negligible weight. Let $\boldsymbol{w}^{\bar{N}} = (w_i^{\bar{N}})$ be a $\bar{N}$-vector such that $w_i^{\bar{N}} \geq 0$ for all i and $\sum_{i=1}^{\bar{N}} w_i^{\bar{N}} = 1$. The fixed-$N$ LA-MSE estimator associated

[1] An alternative mean group estimation approach consists in setting $\hat{\boldsymbol{\theta}}_{MG} = N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i$ and defining $\hat{\mu}_{MG} = \mu(\hat{\boldsymbol{\theta}}^{MG})$. As follows from lemma 1 and theorem 2, the two approaches have identical asymptotic properties in our setup. The two definitions are also numerically identical if μ is affine.

with $\boldsymbol{w}^{\bar{N}}$ is given by

$$\widehat{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) = \sum_{i=1}^{\bar{N}} \sum_{j=1}^{\bar{N}} w_i^{\bar{N}} [\hat{\boldsymbol{\Psi}}_{\bar{N}}]_{i\,j} w_j^{\bar{N}} \ , \tag{3}$$

where $\hat{\boldsymbol{\Psi}}_{\bar{N}} \in \mathbb{R}^{\bar{N}\times\bar{N}}$ with entries $[\hat{\boldsymbol{\Psi}}_{\bar{N}}]_{i\,i} = \nabla\mu(\hat{\boldsymbol{\theta}}_1)'(T(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1)(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1)' + \hat{\boldsymbol{V}}_i)\nabla\mu(\hat{\boldsymbol{\theta}}_1)$ and $[\hat{\boldsymbol{\Psi}}_N]_{i\,j} = \nabla\mu(\hat{\boldsymbol{\theta}}_1)'T(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1)(\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}_1)'\nabla\mu(\hat{\boldsymbol{\theta}}_1)$ when $i \neq j$, and $\hat{\boldsymbol{V}}_i$ is an estimator of the asymptotic variance of $\hat{\boldsymbol{\theta}}_i$. We remark that $\nabla\mu(\hat{\boldsymbol{\theta}}_1)'T(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1)(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1)'\nabla\mu(\hat{\boldsymbol{\theta}}_1)$ and $\nabla\mu(\hat{\boldsymbol{\theta}}_1)'\hat{\boldsymbol{V}}_i\nabla\mu(\hat{\boldsymbol{\theta}}_1)$ are estimators of, respectively, the squared bias and variance of $\mu(\hat{\boldsymbol{\theta}}_i)$ as estimators of $\mu(\boldsymbol{\theta}_1)$. The fixed-$N$ minimum MSE weights are

$$\hat{\boldsymbol{w}}^{\bar{N}} = \underset{\boldsymbol{w}\in\Delta^{\bar{N}}}{\arg\min}\, \widehat{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}) \ , \tag{4}$$

where $\Delta^{\bar{N}} = \{\boldsymbol{w} \in \mathbb{R}^{\bar{N}} : \sum_{i=1}^{\bar{N}} w_i = 1, w_i \geq 0, i = 1, \ldots, \bar{N}\}$.

In the large-N regime we average an arbitrarily large collection of units, and we model N as diverging to infinity. When N is large, some units are mechanically constrained to have a small weight since weights are non-negative and sum to unity. Accordingly, in our framework we partition units into two sets, a set of $\bar{N} \geq 0$ unrestricted units that are allowed to have a non-negligible weight, and a set of remaining $N - \bar{N}$ units that are restricted to have a negligible weight in the limit. In the large-N regime we show that the LA-MSE is determined by the weights assigned to the $\bar{N}$ unrestricted units. Let $\boldsymbol{w}^{N,\infty} = (w_i^{N,\infty})$ be an N-vector and assume, without loss of generality, that the weights of the unrestricted units are placed in the first $\bar{N}$ positions. The vector of weights $\boldsymbol{w}^{N,\infty}$ is such that $w_i^{N,\infty} \geq 0$ for all i, $\sum_{i=1}^{N} w_i^{N,\infty} = 1$, and the weights of the restricted units ($i > \bar{N}$) are given by $w_i^{N,\infty} = (1 - \sum_{j=1}^{\bar{N}} w_i^{N,\infty})/(N - \bar{N})$. We remark that other weighting schemes are allowed for the restricted units. However the negligibility restriction implies that any admissible sequence of weights leads to the same LA-MSE and thus, for simplicity, we opt for equal weights here. Let $\boldsymbol{w}^{\bar{N},\infty} = (w_i^{\bar{N},\infty})$ be a $\bar{N}$-vector such that

$w_i^{\bar{N},\infty} = w_i^{N,\infty}$ for $i = 1, \ldots, \bar{N}$. The large-N LA-MSE estimator associated with $\boldsymbol{w}^{N,\infty}$ is controlled by $\boldsymbol{w}^{\bar{N},\infty}$ and given by

$$\begin{aligned}
&\widehat{LA\text{-}MSE}_\infty(\boldsymbol{w}^{\bar{N},\infty}) \qquad (5)\\
&= \sum_{i=1}^{\bar{N}}\sum_{j=1}^{\bar{N}} w_i^{\bar{N},\infty}[\hat{\boldsymbol{\Psi}}_{\bar{N}}]_{ij} w_j^{\bar{N},\infty} + \left[\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\right)\left(\sqrt{T}\nabla\mu(\hat{\boldsymbol{\theta}}_1)'\left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i\right)\right)\right.\\
&\quad \left. -2\sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\nabla\mu(\hat{\boldsymbol{\theta}}_1)\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1\right)\right]\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\right)\left(\sqrt{T}\nabla\mu(\hat{\boldsymbol{\theta}}_1)'\left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i\right)\right).
\end{aligned}$$

Notice that the restricted units may have an impact on the bias but not the variance of the unit averaging estimator. The large-N minimum MSE weights $\hat{\boldsymbol{w}}^{N,\infty} = (\hat{w}_i^{N,\infty})$ are given by

$$\hat{w}_i^{N,\infty} = \begin{cases} \hat{w}_i^{\bar{N},\infty} & i \leq \bar{N} \\ \left(1 - \sum_{j=1}^{\bar{N}} w_j^{\bar{N},\infty}\right)(N - \bar{N})^{-1} & i > \bar{N} \end{cases} \qquad (6)$$

where

$$\hat{\boldsymbol{w}}^{\bar{N},\infty} = \underset{\boldsymbol{w}\in\tilde{\Delta}^{\bar{N}}}{\arg\min}\, \widehat{LA\text{-}MSE}_\infty(\boldsymbol{w})$$

with $\tilde{\Delta}^{\bar{N}} = \{\boldsymbol{w} \in \mathbb{R}^{\bar{N}} : w_i \geq 0, \sum_{i=1}^{N} w_i \leq 1\}$. It is important to emphasize that the optimization problem defining $\hat{\boldsymbol{w}}^{\bar{N},\infty}$ is $\bar{N}$-dimensional and can be solved by standard quadratic programming methods.

Two remarks are in order before we proceed. First, the fixed- and large-N regimes cover the cases of practical importance.[2] If each unit is potentially important and we do not wish to restrict any weight, then we can apply the fixed-N approximation. The fixed-N regime is agnostic in this sense. Alternatively, if some units can only make an individually negligible contribution to the average, we can apply the large-N regime. Using the large-N approximation requires choosing $\bar{N}$. In principle, $\bar{N}$ can be chosen arbitrarily, though choosing $\bar{N} > N$ effectively results in a fixed-N approximation. An appropriate choice

[2]We also derive LA-MSE in case N is arbitrarily large and an infinite number of units have a non-zero weight in the limit. However, this case appears to be of limited interest in practice, and we do not study properties of the data-dependent weights in this case.

of $\bar{N}$ might be implied by economic logic. For example, in a macroeconomic application using country-level data, we might order the countries so that immediate neighbors of the country of interest, its important trading partners, and the large economies of the world are placed in the first $\bar{N}$ positions. The other countries are judged to contribute relatively little individually, and are placed in positions beyond $\bar{N}$.

Second, the fixed- and large-N LA-MSE estimators have the appealing property of being applicable both when the amount of time series information in the panel is moderate or large. When the amount of time series information is moderate, the LA-MSE approximates the infeasible population problem of minimizing the MSE, along with uncertainty about individual parameters (see the discussion following theorem 3). When the amount of time series information is large, the bias term in the MSE dominates, and the unit averaging estimator based on the minimum MSE weights converges to the individual estimator $\mu(\hat{\boldsymbol{\theta}}_1)$.

3 Theory

3.1 Assumptions

Our focus is on a moderate-T setting. To emulate it and the trade-off between unit-specific and panel-wide information, we introduce what we call the *local heterogeneity* assumption. This is inspired and is analogous to the local misspecification assumption used in the frequentist model averaging literature (Hjort and Claeskens, 2003; Hansen, 2016).

A.1 (Local Heterogeneity)**.** *The sequence of unit-specific parameters* $\{\boldsymbol{\theta}_i\}$ *is such that*

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 + \frac{\boldsymbol{\eta}_i}{\sqrt{T}} \ ,$$

where $\{\boldsymbol{\eta}_i\}$ *is a sequence of i.i.d. random vectors such that* $\mathbb{E}_{\boldsymbol{\eta}}(\boldsymbol{\eta}_i) = \mathbf{0}$ *and* $\mathbb{E}_{\boldsymbol{\eta}} \|\boldsymbol{\eta}_i\|^{12} < \infty$.[3]

[3]Here and below $\|\cdot\|$ means the 2-norm, unless specifically labeled otherwise.

All analysis is done conditional on $\sigma(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots)$ *and all distributional statements below are conditional on* $\sigma(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots)$ *unless specifically stated otherwise.*

A number of remarks on this assumption are in order. First, the scaling by $\sqrt{T}$ allows us to approximate a moderate-T setting using asymptotic theory techniques by creating a nontrivial asymptotic bias-variance trade-off. Intuitively, as T becomes larger, the signal strength becomes proportionally weaker so that the amount of information in each time series remains constant. This is a standard technique to approximate finite-sample properties of model selection and averaging estimators (see, for example, Hjort and Claeskens (2003); Liu (2015); Yin et al. (2021)). Second, since the focus of the analysis lies in recovering individual effects, all probability statements are implicitly conditional on $\sigma(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots)$. Such a conditioning is natural, given our focus on estimating $\mu(\boldsymbol{\theta}_1)$ and typical when individual parameters are of interest (Vaida and Blanchard, 2005; Donohue, Overholser, Xu, and Vaida, 2011; Zhang et al., 2014). Importantly, all the results we establish are shown to hold with $\boldsymbol{\eta}$-probability 1, that is, for almost any realization of $\{\boldsymbol{\eta}_i\}$.

In this paper we assume that the cross-sectional units are independent.

A.2 (Independence)**.** *For each* $i, j_1, \dots, j_k, k$ *such that* $i \neq j_1, \dots, j_k$ $\{\{\boldsymbol{z}_{it}\}_{t=0}^{\infty}, \boldsymbol{\eta}_i\}$ *and* $\{\{\{\boldsymbol{z}_{j_1 t}\}_{t=0}^{\infty}, \boldsymbol{\eta}_{j_1}\}, \dots, \{\{\boldsymbol{z}_{j_k t}\}_{t=0}^{\infty}, \boldsymbol{\eta}_{j_k}\}\}$ *are independent.*

The unit-specific estimators are assumed to satisfy a number of regularity conditions.

A.3 (Individual Objective Function)**.**

(i) The parameter space Θ *is convex.*

(ii) The function $m(\boldsymbol{\theta}, \boldsymbol{z}) : \Theta \times \mathcal{Z} \to \mathbb{R}$ *is twice continuously differentiable in* $\boldsymbol{\theta}$ *for each value of* $\boldsymbol{z}$*.* $m(\boldsymbol{\theta}, \boldsymbol{z})$ *is measurable as a function of* $\boldsymbol{z}$ *for every value of* $\boldsymbol{\theta}$*.*

(iii) There exists a positive finite constant T_0 *(which does not depend on* i*) such that for all* i *and* $T > T_0$ *it holds that the unit-specific estimator satisfies* $\hat{\boldsymbol{\theta}}_i \in \text{int}(\Theta)$ *a.s..*

(iv) *The gradient of the unit-specific objective function satisfies*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i,\boldsymbol{z}_{it}) \Rightarrow N(0,\boldsymbol{\Sigma}_i)\ ,$$

where $\boldsymbol{\Sigma}_i = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\mathbb{E}\left[\left(\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i,\boldsymbol{z}_{it})\right)\left(\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i,\boldsymbol{z}_{it})\right)'\right]$.

(v) *There exist a positive finite constant* $C_{\nabla m}$ *(which does not depend on* i *or* T*) such that, for all* i *and all* $T > T_0$ *and for some* $\delta \in (0,1)$*, it holds that*

$$\mathbb{E}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i,\boldsymbol{z}_{it})\right\|^{2(1+\delta)} \leq C_{\nabla m}\ .$$

(vi) *The Hessian of the unit-specific objective function satisfies*

$$\sup_{\boldsymbol{\theta}\in[\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i]}\left\|\frac{1}{T}\sum_{t=1}^{T}\nabla^2 m(\boldsymbol{\theta},\boldsymbol{z}_{it}) - \boldsymbol{H}_i\right\| \xrightarrow{p} 0\ ,$$

where $\boldsymbol{H}_i = \lim_{T\to\infty}\mathbb{E}(T^{-1}\sum_{t=1}^{T}\nabla^2 m(\boldsymbol{\theta}_i,\boldsymbol{z}_{it}))$.

(vii) *Let* $D_{iT} = \sup_{\boldsymbol{\theta}\in[\boldsymbol{\theta}_i,\hat{\boldsymbol{\theta}}_i]}\left\|\left(T^{-1}\sum_{t=1}^{T}\nabla^2 m(\boldsymbol{\theta},\boldsymbol{z}_{it})\right)\boldsymbol{H}_i^{-1} - \boldsymbol{I}\right\|_\infty$. $D_{iT} < 1$ *a.s. for all* i *and all* $T > T_0$*. There exists a positive constant* $C_{\nabla^2 m}$ *such that, for all* i *and all* $T > T_0$ *and for* δ *as in (v), it holds that*

$$\mathbb{E}\left[\left(\frac{D_{iT}}{1-D_{iT}}\right)^{\frac{2(2+\delta)(1+\delta)}{\delta}}\right] \leq C_{\nabla^2 m}.$$

(viii) *The matrices* $\boldsymbol{\Sigma}_i$ *and* $\boldsymbol{H}_i$ *satisfy* $\underline{\lambda}_{\boldsymbol{\Sigma}} \leq \lambda_{\min}(\boldsymbol{\Sigma}_i) \leq \lambda_{\max}(\boldsymbol{\Sigma}_i) \leq \overline{\lambda}_{\boldsymbol{\Sigma}}$ *and* $\underline{\lambda}_{\boldsymbol{H}} \leq \lambda_{\min}(\boldsymbol{H}_i) \leq \lambda_{\max}(\boldsymbol{H}_i) \leq \overline{\lambda}_{\boldsymbol{H}}$ *where* $\underline{\lambda}_{\boldsymbol{\Sigma}}$*,* $\overline{\lambda}_{\boldsymbol{\Sigma}}$*,* $\underline{\lambda}_{\boldsymbol{H}}$ *and* $\overline{\lambda}_{\boldsymbol{H}}$ *are positive constants that do not depend on* i*.*

(ix) *Let* $\boldsymbol{V}_i = \boldsymbol{H}_i^{-1}\boldsymbol{\Sigma}_i\boldsymbol{H}_i^{-1}$*. Then, there is a sequence of estimators* $\{\hat{\boldsymbol{V}}_i\}$ *such that, for all* i*,* $\hat{\boldsymbol{V}}_i$ *is consistent for* $\boldsymbol{V}_i$*, and, for all* $T > T_0$*,* $\lambda_{\min}(\hat{\boldsymbol{V}}_i) > 0$ *holds almost surely.*

Assumption A.3 requires the unit-specific estimators to be consistent, asymptotically normal and to satisfy a number of regularity conditions. We remark that this assumption

allows for a fair amount of dependence and heterogeneity in the unit-specific observations.[4] Assumption A.3(iii) states that the unit-specific estimator lies in the interior of the parameter space almost surely. If the problem is linear or defined by a convex smooth objective function and continuous covariates, the parameter space can be taken to be $\mathbb{R}^p$, and the condition holds automatically. Assumption A.3(iv) is standard in the M-estimation literature, it requires the gradient of the objective function evaluated at $\boldsymbol{\theta}_i$ to satisfy a CLT. Assumption A.3(v) is a moment condition on the gradient of the objective function. In an i.i.d. setting such an assumption translates into a moment condition on the individual gradients. More generally, this would be implied by appropriate moment and dependence assumption on the individual gradients. Assumption A.3(vi) is also standard in the M-estimation literature; it requires the Hessian to satisfy a uniform law of large numbers. Assumption A.3(vii) effectively requires that the sample Hessian is nonsingular in a small enough neighborhood of $\boldsymbol{\theta}_i$. In a scalar problem, (vii) restricts the possible range of the second derivative as $\boldsymbol{\theta}$ ranges over a shrinking interval around $\boldsymbol{\theta}_i$. In addition, (vii) places an assumption on the moments of deviation from the population limit Hessian. In case of linear regression, the sample and population Hessians do not depend on the slope parameters and (vii) is an assumption on moments of covariates. Assumption A.3($viii$) restricts the spectrum of the matrices $\boldsymbol{\Sigma}_i$ and $\boldsymbol{H}_i$. In particular, it implies a uniform restriction on the asymptotic variance $\boldsymbol{V}_i$ of the individual estimators. Assumption A.3(ix) states that there exists a sequence of nonsingular estimators $\{\hat{\boldsymbol{V}}_i\}$ for the asymptotic variance-covariance matrix of the individual estimator. We remark that Assumptions A.3(iii) and (vii) state that the sequence of unit-specific estimation problems satisfies approprite uniformity conditions. Such conditions allow us to distill the key arguments relevant to our averaging theory and, in a sense, should be intrepreted as a simplifying approximation. In general, (iii) and (vii) would hold with probability approaching one for each unit. In this case all our results would still hold, though under

[4] A classic reference on M-estimation for dependent and heterogeneous data is the book by Pötscher and Prucha (1997).

appropriate rate conditions on (N, T) and trimming to ensure certain well-behavedness of individual estimators. We further note that assumptions (iii) and (vii) might hold in practice in certain special cases regardless (such as linear or nonlinear models with a convex and smooth objective function and continuous covariates).

A.4 (Unit-specific Bias)**.** *There exists a constant C_{Bias}, which does not depend on i, such that $\left\|\mathbb{E}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\right\|_1 \leq C_{Bias}/T$ for all $T > T_0$.*

Assumption A.4 requires that the bias of individual estimators for their own parameters is bounded uniformly in i. The order of the bias is consistent with the results obtained by Rilstone, Srivastava, and Ullah (1996) and Bao and Ullah (2007). The higher order terms can be subsumed into the T^{-1} term for a sufficiently large C_{Bias}.[5] Assumption A.4 is satisfied for linear models under assumption A.3. For nonlinear models it is sufficient that for all s and i it holds that $T\,\mathbb{E}(\|\nabla^s m(\boldsymbol{\theta}_i, z_{it})\|^2) \leq C_s < \infty$ (Rilstone et al., 1996; Bao and Ullah, 2007; Yang, 2015).

A.5 (Focus Parameter)**.** *The focus function $\mu : \Theta \to \mathbb{R}$ is twice-differentiable. There exists a constant $C_{\nabla\mu}$ such that for all $\boldsymbol{\theta} \in \Theta$ it holds that $\|\nabla\mu(\boldsymbol{\theta})\| < C_{\nabla\mu}$. There exists a constant $C_{\nabla^2\mu}$ such that for all $\boldsymbol{\theta} \in \Theta$ the largest and smallest eigenvalues of the Hessian $\nabla^2\mu(\boldsymbol{\theta})$ are bounded in absolute value by $C_{\nabla^2\mu}$. Let $\boldsymbol{d}_0 = \nabla\mu(\boldsymbol{\theta}_0)$ be the gradient of μ at $\boldsymbol{\theta}_0$. Then $\boldsymbol{d}_0 \neq 0$.*

Assumption A.5 lays out mild smoothness assumptions on μ. For simplicity we assume that μ is a scalar focus parameter. However, all our results can be extended to the case in which μ is a vector focus parameter.

[5] Explicitly including terms of order $-3/2$ and higher does not change the analysis, as long as all the constants do not depend on i

3.2 Asymptotic Properties of the Minimum MSE Unit Averaging Estimator

We begin with an auxiliary lemma that establishes the distribution of the unit-specific estimators in the local asymptotic framework of assumption A.1.

Lemma 1. *Assume that assumptions A.1–A.5 are satisfied. Let the unit-specific estimators* $\hat{\boldsymbol{\theta}}_i$ *for* $i = 1,2,\ldots$ *be defined as in eq. (2). Then*

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1\right) \Rightarrow N(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1, \boldsymbol{V}_i) =: \boldsymbol{Z}_i \ ,$$
$$\sqrt{T}\left(\mu(\hat{\boldsymbol{\theta}}_i) - \mu(\boldsymbol{\theta}_1)\right) \Rightarrow N(\boldsymbol{d}_0'\left(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\right), \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0) =: \Lambda_i$$

holds as $T \to \infty$ *for* $i = 1,2,\ldots$. *Convergence is joint (that is, with respect to the product topology), and all* $\boldsymbol{Z}_i$ *and* Λ_i *are independent across* i.

Lemma 1 characterizes the local asymptotic distribution of $\mu(\hat{\boldsymbol{\theta}}_i)$. In particular, the mean and the variance of the limit distribution provide a local asymptotic approximation to the exact moderate-T bias and variance of $\mu(\hat{\boldsymbol{\theta}}_i)$ as an estimator of $\mu(\boldsymbol{\theta}_1)$. By adding together the square mean and the variance of the limit distribution, we obtain a local asymptotic approximation to the MSE (LA-MSE) of each $\mu(\hat{\boldsymbol{\theta}}_i)$.

We now introduce a local asymptotic approximation to the MSE (LA-MSE) of the unit averaging estimator (1). Let $\{\boldsymbol{w}_N\} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots\}$ be a (non-random) sequence where $\boldsymbol{w}_k$ is a k-vector of weights. Suppose that $\boldsymbol{w}_N$ converges to some $\boldsymbol{w} \in \mathbb{R}^\infty$ in the sense defined below. Consider the unit averaging estimator $\hat{\mu}(\boldsymbol{w}_N)$ associated with such a sequence. The following theorem derives $\lim_{N,T\to\infty} T \times MSE\left(\hat{\mu}(\boldsymbol{w}_N)\right)$ where $N, T \to \infty$ jointly. This is a standard notion of risk in a local asymptotic framework (see e.g. Hansen (2016)). A remark on notation is in order before we proceed. In what follows we treat $\boldsymbol{w}_N = (w_{i\,N})$ as an element both in $\mathbb{R}^N$ and in $\mathbb{R}^\infty$ (with coordinates $i > N$ restricted to zero). This duality will not cause any confusion.

Theorem 1. *Let assumptions A.1–A.5 be satisfied. Let* $\{\boldsymbol{w}_N\}$ *be such that* (i) *for each* N, $\boldsymbol{w}_N$ *is measurable with respect to* $\sigma(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N)$, (ii) *for each* N, $w_{i\,N} \geq 0$ *for all* i, $\sum_{i=1}^{N} w_{i\,N} = 1$, $w_{j\,N} = 0$ *for* $j > N$, (iii) $\sup_i |w_{i\,N} - w_i| = o(N^{-1/2})$ *where* $\boldsymbol{w} = (w_i) \in \mathbb{R}^\infty$ *is a vector such that* $w_i \geq 0$ *and* $\sum_{i=1}^{\infty} w_i \leq 1$. *Let* T_0 *be as in assumption A.3.*

Then (i) $\sum_{i=1}^{\infty} w_i \boldsymbol{d}_0' \boldsymbol{\eta}_i$ *and* $\sum_{i=1}^{\infty} w_i^2 \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0$ *exist;* (ii) *for any* N *and* $T > T_0$ *the MSE of the averaging estimator is finite; and* (iii) *as* $N, T \to \infty$ *jointly it holds that*

$$T \times MSE\left(\hat{\mu}(\boldsymbol{w}_N)\right) \to \left(\sum_{i=1}^{\infty} w_i \boldsymbol{d}_0' \boldsymbol{\eta}_i - \boldsymbol{d}_0' \boldsymbol{\eta}_1\right)^2 + \sum_{i=1}^{\infty} w_i^2 \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0.$$

Two remarks on theorem 1 are in order before we proceed. First, the theorem provides a local asymptotic approximation to the MSE of the averaging estimator (LA-MSE). It highlights the bias-variance trade-off associated with the choice of the weights. The two extremes of the trade-off correspond to the individual estimator of unit 1 $\mu(\hat{\boldsymbol{\theta}}_1)$ and the mean group estimator $\hat{\mu}_{MG}$. The individual estimator of unit 1 is obtained by setting $w_{1\,N} = 1$ for all N. It is straightforward to see that $\mu(\hat{\boldsymbol{\theta}}_1)$ is asymptotically unbiased and that its LA-MSE is equal to $\boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0$, which is its asymptotic variance. The mean group estimator is obtained by setting $w_{i\,N} = (N)^{-1} \mathbb{I}_{i \leq N}$ for $i = 1, \ldots, N$ for all N. $\hat{\mu}_{MG}$ has zero asymptotic variance and its LA-MSE is equal to $(\boldsymbol{d}_0' \boldsymbol{\eta}_1)^2$.[6] Second, the theorem requires that the sequence $\{\boldsymbol{w}_N\}$ of weight vectors converge uniformly to some limit $\boldsymbol{w}$.[7] In addition, convergence must occur at a rate faster than $N^{-1/2}$. Notice that this condition is trivially satisfied by the mean group estimator. We also emphasize that the sum of the limit $\boldsymbol{w}$ can be less than one.

In order to study the properties of the two averaging regimes introduced in section 2, we provide a refined version of theorem 1. It imposes a stronger uniform convergence

[6] This corresponds to the estimator of Issler and Lima (2009). In a genuine large-T setting it is feasible to estimate a bias of such of type and correct for it. However, in a moderate-T setting this quantity cannot be consistently estimated.

[7] The assumption is needed to ensure convergence of the bias and the variance series and to prevent a "sliding hump" given by weighting structures like $w_{i\,N} = \mathbb{I}_{i=N}$.

condition on the weights that reflects the assumptions of the large-N regime.

Theorem 2. *Assume that assumptions A.1–A.5 are satisfied. Let* $\{\boldsymbol{w}_N\}$ *be such that* (i) *for each* N, $\boldsymbol{w}_N$ *is measurable w.r.t.* $\sigma(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N)$, (ii) *for each* N, $w_{i\,N} \geq 0$ *for all* i, $\sum_{i=1}^{N} w_{i\,N} = 1$, $w_{j\,N} = 0$ *for* $j > N$, (iii) *for some* $\bar{N} \geq 0$ *it holds that* $\sup_{i>\bar{N}} w_{i\,N} = o(N^{-1/2})$, *and* (iv) $\{w_{i\,N}\}_{i=1}^{\bar{N}} \to \{w_i\}_{i=1}^{\bar{N}}$.

Then as $N, T \to \infty$ *jointly it holds that*

$$\sqrt{T}\left(\hat{\mu}(\boldsymbol{w}_N) - \mu(\boldsymbol{\theta}_1)\right) \Rightarrow N\left(\sum_{i=1}^{\bar{N}} w_i \boldsymbol{d}_0' \boldsymbol{\eta}_i - \boldsymbol{d}_0' \boldsymbol{\eta}_1, \sum_{i=1}^{\bar{N}} w_i^2 \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0\right).$$

Theorem 2 shows that the unit averaging estimator is asymptotically normal. The mean and variance are given by the weighted sums of biases and variances, respectively. The theorem can be applied in both the fixed-N and large-N regimes introduced in Section 2. In the fixed-N regime, suppose only the first $\bar{N} < \infty$ units are being averaged (some of them potentially with zero weights). Then we set $w_{i\,N} = 0$ for all N and $i > \bar{N}$, and condition (iii) holds automatically. The condition that $N \to \infty$ becomes superfluous. Conditions (ii)-(iv) reduce to the requirement that $\boldsymbol{w}_N$ converge (pointwise) to a vector of weights $\boldsymbol{w}$, where for $i > \bar{N}$ the weights satisfy $w_{i\,N} = w_i = 0$ and $\sum_{i=1}^{\bar{N}} w_{i\,N} = \sum_{i=1}^{\bar{N}} w_i = 1$. The limit distribution is normal with mean $\sum_{i=1}^{\bar{N}} w_i \boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)$ and variance $\sum_{i=1}^{\bar{N}} w_i^2 \boldsymbol{d}_1' \boldsymbol{V}_i \boldsymbol{d}_1$. In the large-$N$ regime, order the units with potentially large weights to be the first $\bar{N}$ units where $\bar{N}$ is chosen in advance. Theorem 2 shows that when N is large, the units beyond $\bar{N}$ contribute no variance and approximate bias $-\left(1 - \sum_{i=1}^{\bar{N}} w_i\right) \boldsymbol{d}_0' \boldsymbol{\eta}_1$.

Importantly, the limit distribution of the averaging estimator does not depend on the actual shape of the weight vector beyond $\bar{N}$. The limit distribution is controlled by finitely many weights, $\bar{N}$, and the total weight mass placed beyond $\bar{N}$. The weights in $\boldsymbol{w}_N$ beyond $\bar{N}$ can display strong variations in orders of magnitude, with some weights decaying like $N^{-1/2-\varepsilon}$, and some at a faster rate. The total weight mass placed beyond

$\bar{N}$ in the limit is also not restricted, and may approach 1, like in the case of the mean group estimator. As an application of this observation, if condition (iii) holds with $\bar{N} = 0$ $\sqrt{T}\left(\hat{\mu}(\boldsymbol{w}_N) - \mu(\boldsymbol{\theta}_1)\right) \xrightarrow{p} -\boldsymbol{d}_0'\boldsymbol{\eta}_1$ for *any* weight vector sequence satisfying this condition.

Using theorem 2, we can obtain expressions for the population LA-MSE of the unit averaging estimator (1) in the fixed-N and large-N regimes. Consider the fixed-N case in which we average $\bar{N}$ units. Let $\boldsymbol{w}^{\bar{N}}$ be a $\bar{N}$-vector, the limit vector of theorem 2. In the fixed-N regime the population LA-MSE is

$$LA\text{-}MSE_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) = \sum_{i=1}^{\bar{N}}\sum_{j=1}^{\bar{N}} w_i^{\bar{N}}[\boldsymbol{\Psi}_{\bar{N}}]_{i\,j} w_j^{\bar{N}} \ ,$$

where $\boldsymbol{\Psi}_{\bar{N}}$ is an $\bar{N} \times \bar{N}$ matrix with elements $[\boldsymbol{\Psi}_{\bar{N}}]_{i\,i} = \boldsymbol{d}_0'\left((\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)' + \boldsymbol{V}_i\right)\boldsymbol{d}_0$ and $[\boldsymbol{\Psi}_{\bar{N}}]_{i\,j} = \boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)(\boldsymbol{\eta}_j - \boldsymbol{\eta}_1)'\boldsymbol{d}_0$ when $i \neq j$. Now consider the large-N regime. As theorems 1 and 2 show, the LA-MSE is controlled by a $\bar{N}$-vector $\boldsymbol{w}^{\bar{N},\infty}$ such that $w_i^{\bar{N},\infty} \geq 0$ for all i and $\sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \leq 1$. In the large-$N$ regime, the population LA-MSE is

$$\begin{aligned} LA\text{-}MSE_{\infty}(\boldsymbol{w}^{\bar{N},\infty}) &= \sum_{i=1}^{\bar{N}}\sum_{j=1}^{\bar{N}} w_i^{\bar{N},\infty}[\boldsymbol{\Psi}_{\bar{N}}]_{i\,j} w_j^{\bar{N},\infty} \\ &+ \left(\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\right)\boldsymbol{d}_0'\boldsymbol{\eta}_1 - 2\sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)\right)\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\right)\boldsymbol{d}_0'\boldsymbol{\eta}_1 \ . \end{aligned}$$

The quantities $\widehat{LA\text{-}MSE}_{\bar{N}}$ and $\widehat{LA\text{-}MSE}_{\infty}$ used to define the minimum MSE weights introduced in Section 2 are estimators of the population LA-MSE given above. In the rest of the section we focus on the properties of these estimators as well as the optimal weights (4) and (6) associated with them.

We begin by noting that in our framework the population LA-MSE cannot be consistently estimated. Under local heterogeneity the idiosyncratic components $\boldsymbol{\eta}_i$ cannot be consistently estimated (Hjort and Claeskens, 2003). Instead, following Hjort and Claeskens (2003), we form $\widehat{LA\text{-}MSE}_{\bar{N}}$ and $\widehat{LA\text{-}MSE}_{\infty}$ by plugging in asymptotically unbiased estimators for $\boldsymbol{\eta}_i - \boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_1$. Such estimators are provided by $\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}})$ and

$\sqrt{T}(\hat{\boldsymbol{\theta}}_1 - N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i)$, respectively, as the following lemma establishes.

Lemma 2. *Let assumptions A.1-A.5 hold. Then as $N, T \to \infty$ jointly, it holds that*

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1\right) \Rightarrow N\left(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1, \boldsymbol{V}_i + \boldsymbol{V}_1\right) = \boldsymbol{Z}_i - \boldsymbol{Z}_1,$$
$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i\right) \Rightarrow N(\boldsymbol{\eta}_1, \boldsymbol{V}_1) = \boldsymbol{Z}_1 + \boldsymbol{\eta}_1.$$

Convergence is joint for all i.

We remark that the matrix $\hat{\boldsymbol{\Psi}}_{\bar{N}}$ of equations (3) and (5) is a biased estimator of $\boldsymbol{\Psi}_{\bar{N}}$. Such a bias ensures that $\widehat{LA\text{-}MSE}$ is nonnegative for all admissible weight vectors. An asymptotically unbiased estimator instead would have diagonal elements $\tilde{\boldsymbol{\Psi}}_{ii} = \hat{\boldsymbol{d}}_1'\left(T\left(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1\right)\left(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1\right)' - \left(\hat{\boldsymbol{V}}_i + \hat{\boldsymbol{V}}_1\right)\right)\hat{\boldsymbol{d}}_1$ and off-diagonal elements $\tilde{\boldsymbol{\Psi}}_{ij} = \hat{\boldsymbol{d}}_1'\left(T\left(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1\right)\left(\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}_1\right)' - \hat{\boldsymbol{V}}_1\right)\hat{\boldsymbol{d}}_1$ where $\hat{\boldsymbol{d}}_1 = \nabla\mu\left(\hat{\boldsymbol{\theta}}_1\right)$. However, this can easily fail to be positive definite, as it involves a difference of positive definite matrices. This would lead to the undesirable possibility of obtaining negative estimates of the LA-MSE.[8]

The following two theorems establish the properties of our LA-MSE estimators and the associated minimum MSE weights (4) and (6). The theorem also characterizes the asymptotic distribution of the minimum MSE unit averaging estimator $\hat{\mu}(\hat{\boldsymbol{w}}^{\bar{N}})$ in this regime. First, we state a result for the fixed-N regime. Recall that $\Delta^{\bar{N}} = \{\boldsymbol{w} \in \mathbb{R}^{\bar{N}} : \sum_{i=1}^{\bar{N}} w_i = 1, w_i \geq 0, i = 1, \ldots, \bar{N}\}$.

Theorem 3 (Fixed-N Minimum MSE Unit Averaging)**.** *Let assumptions A.1-A.5 hold.*

(i) For any $\boldsymbol{w}^{\bar{N}} \in \Delta^{\bar{N}}$ it holds that $\widehat{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) \Rightarrow \overline{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) := \boldsymbol{w}^{\bar{N}\prime}\overline{\boldsymbol{\Psi}}_{\bar{N}}\boldsymbol{w}^{\bar{N}}$ as $T \to \infty$, where $\overline{\boldsymbol{\Psi}}_{\bar{N}}$ is an $\bar{N} \times \bar{N}$ matrix with elements $[\overline{\boldsymbol{\Psi}}_{\bar{N}}]_{ij} = \boldsymbol{d}_0'((\boldsymbol{Z}_i - \boldsymbol{Z}_1)(\boldsymbol{Z}_i - \boldsymbol{Z}_1)' + \boldsymbol{V}_i)\boldsymbol{d}_0$ when $i = j$ and $\boldsymbol{d}_0'((\boldsymbol{Z}_i - \boldsymbol{Z}_1)(\boldsymbol{Z}_j - \boldsymbol{Z}_1)')\boldsymbol{d}_0$ when $i \neq j$; and $\boldsymbol{Z}_i$ is as in Lemma 1.

[8] An additional argument in favor of focusing on $\hat{\boldsymbol{\Psi}}_{\bar{N}}$ is given by Liu (2015) who examines the behavior of both $\hat{\boldsymbol{\Psi}}_{\bar{N}}$ and $\tilde{\boldsymbol{\Psi}}_{\bar{N}}$ in the context of model averaging. Liu (2015) finds that $\hat{\boldsymbol{\Psi}}_{\bar{N}}$ leads to superior performance of resulting weights.

(ii) As $T \to \infty$, the minimum MSE weights satisfy

$$\hat{\boldsymbol{w}}^{\bar{N}} = \underset{\boldsymbol{w}^{\bar{N}} \in \Delta^{\bar{N}}}{\arg\min}\, \widehat{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) \Rightarrow \overline{\boldsymbol{w}}^{\bar{N}} = \underset{\boldsymbol{w}^{\bar{N}} \in \Delta^{\bar{N}}}{\arg\min}\, \overline{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}).$$

(iii) As $T \to \infty$, for Λ_i of lemma 1, the minimum MSE unit averaging estimator satisfies

$$\sqrt{T}\left(\hat{\mu}(\hat{\boldsymbol{w}}^{\bar{N}}) - \mu(\boldsymbol{\theta}_1)\right) \Rightarrow \sum_{i=1}^{\bar{N}} \overline{w}_i^{\bar{N}} \Lambda_i.$$

A number of remarks on theorem 3 are in order. First, $\overline{LA\text{-}MSE}_{\bar{N}}$ plays the same role to $\widehat{LA\text{-}MSE}_{\bar{N}}$ as $\boldsymbol{Z}_i$ does to $\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)$ in lemma 1. $\overline{LA\text{-}MSE}_{\bar{N}}$ uses a local approximation to express $\widehat{LA\text{-}MSE}_{\bar{N}}$ in terms of the components of the MSE and the approximate distribution of the individual estimators. We can see that $\overline{LA\text{-}MSE}_{\bar{N}}$ is composed of the population LA-MSE, a bias term, and a noise component. In fact, entries of the matrix $\overline{\boldsymbol{\Psi}}_{\bar{N}}$ may be expressed as

$$\begin{aligned}
[\overline{\boldsymbol{\Psi}}_{\bar{N}}]_{ii} &= [\boldsymbol{\Psi}_{\bar{N}}]_{ii} + \boldsymbol{d}_0'(\boldsymbol{V}_1 + \boldsymbol{V}_i)\boldsymbol{d}_0 + \boldsymbol{d}_0'\boldsymbol{e}_{ii}\boldsymbol{d}_0 \ , \\
[\overline{\boldsymbol{\Psi}}_{\bar{N}}]_{ij} &= [\boldsymbol{\Psi}_{\bar{N}}]_{ij} + \boldsymbol{d}_0'\boldsymbol{V}_1\boldsymbol{d}_0 + \boldsymbol{d}_0'\boldsymbol{e}_{ij}\boldsymbol{d}_0, \quad i \neq j \ ,
\end{aligned}$$

where $\boldsymbol{e}_{ij} = (\boldsymbol{Z}_i - \boldsymbol{Z}_1)(\boldsymbol{Z}_j - \boldsymbol{Z}_1)' - \mathbb{E}\left((\boldsymbol{Z}_i - \boldsymbol{Z}_1)(\boldsymbol{Z}_j - \boldsymbol{Z}_1)'\right)$. The noise terms $\boldsymbol{e}_{ij}$ may be interpreted as the result of the fact that in a moderate-T setting there is limited information about the idiosyncratic components $\boldsymbol{\eta}_i$. These terms are mean zero and independent conditional on unit 1. The bias terms guarantee that $\overline{\boldsymbol{\Psi}}_{\bar{N}}$ is positive definite and arise as a consequence of using the biased positive definite estimator $\hat{\boldsymbol{\Psi}}_{\bar{N}}$. The bias can be split into two components. The $\boldsymbol{d}_0'\boldsymbol{V}_1\boldsymbol{d}_0$ is common for all elements of $\overline{\boldsymbol{\Psi}}_{\bar{N}}$ and does not affect the solution of the MSE minimization problem. The second component $\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0$ only affects the diagonal of $\overline{\boldsymbol{\Psi}}_{\bar{N}}$ by inflating the variance associated to each unit by the variance of the corresponding individual estimator. This component does not modify the ordering of the estimators in terms of their variances. If all the estimators were unbiased,

optimal weights would be inversely proportional to individual variances, and this resulting ordering of the weights would be preserved.

Second, the fixed-N minimum MSE unit averaging estimator has a nonstandard asymptotic distribution in the local heterogeneity framework. Assertion 3 of theorem 3 shows that the limit distribution is a randomly weighted sum of independent normal random variables. In the online appendix, we show how to construct confidence intervals based on this result.

Third, minimizing $\widehat{LA\text{-}MSE}_N$ is natural even in a non-local setting where we drop assumption A.1 and allow the amount of information in each time series to grow as $T \to \infty$. For all i such that $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_1$, the bias estimators $\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1)$ will diverge, while all variance terms remain bounded. Then the procedure will place zero weight asymptotically on all such units i and on the tail term. The minimum MSE weights will only use those units that share the same $\boldsymbol{\theta}_1$. In particular, if the distribution of $\boldsymbol{\eta}$ is continuous, asymptotically the procedure will use only information from unit 1, and the difference between the averaging estimator with minimum MSE weights and the individual estimator will converge to zero in probability. In this case the averaging estimator is asymptotically normal with mean zero and variance $\boldsymbol{d}_0'\boldsymbol{V}_1\boldsymbol{d}_0$. Such a result has a parallel in fixed parameter asymptotics for model averaging. As Fang, Yuan, and Tian (2022) show in a recent paper, the weight assigned to the true (or least wrong) model tends to one as sample size increases.

The following result establishes an analogous result for the minimum MSE weights (6) in the large-N regime. Recall that $\tilde{\Delta}^{\bar{N}} = \{\boldsymbol{w} \in \mathbb{R}^{\bar{N}} : w_i \geq 0, \sum_{i=1}^{N} w_i \leq 1\}$.

Theorem 4 (Large-N Minimum MSE Unit Averaging)**.** *Let assumptions A.1-A.5 hold.*

(i) For any $\boldsymbol{w}^{\bar{N},\infty} \in \tilde{\Delta}^{\bar{N}}$ *it holds that* $\widehat{LA\text{-}MSE}_\infty(\boldsymbol{w}^{\bar{N},\infty}) \Rightarrow \overline{LA\text{-}MSE}_\infty(\boldsymbol{w}^{\bar{N},\infty})$ *as* $N, T \to \infty$ *jointly where*

$$\overline{LA\text{-}MSE}_\infty(\boldsymbol{w}^{\bar{N},\infty}) = \boldsymbol{w}^{\bar{N},\infty\prime}\overline{\boldsymbol{\Psi}}_{\bar{N}}\boldsymbol{w}^{\bar{N},\infty} + \left[\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty}\right)\boldsymbol{d}_0'\left(\boldsymbol{\eta}_1 + \boldsymbol{Z}_1\right)\right.$$

$$- 2 \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \boldsymbol{d}_0' \left(\boldsymbol{Z}_i - \boldsymbol{Z}_1 \right) \Bigg] \left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \right) \boldsymbol{d}_0' \left(\boldsymbol{\eta}_1 + \boldsymbol{Z}_1 \right).$$

(ii) As $N, T \to \infty$, the minimum MSE weights satisfy

$$\hat{\boldsymbol{w}}^{\bar{N},\infty} = \underset{\boldsymbol{w} \in \tilde{\Delta}^{\bar{N}}}{\arg\min}\, \widehat{LA\text{-}MSE}_{\infty}(\boldsymbol{w}) \Rightarrow \overline{\boldsymbol{w}}^{\bar{N},\infty} = \underset{\boldsymbol{w} \in \tilde{\Delta}^{\bar{N}}}{\arg\min}\, \overline{LA\text{-}MSE}_{\infty}(\boldsymbol{w}).$$

(iii) Let $\boldsymbol{v}_{N-\bar{N}} = (v_{\bar{N}\,N}, \ldots, v_{N\,N})$ be a $(N - \bar{N})$-vector such that $\sup_i v_{i\,N-\bar{N}} = o(N^{-1/2})$, $v_{i\,N-\bar{N}} \geq 0$, for each N it holds that $\sum_{i=N-\bar{N}}^{N} v_{i\,N-\bar{N}} = 1$. Then as $N, T \to \infty$ jointly

$$\begin{aligned} &\sqrt{T} \left(\sum_{i=1}^{\bar{N}} \hat{w}_i^{\bar{N},\infty} \mu\left(\hat{\boldsymbol{\theta}}_i \right) + \left(1 - \sum_{i=1}^{\bar{N}} \hat{w}_i^{\bar{N},\infty} \right) \sum_{j=N-\bar{N}}^{N} v_{j\,N-\bar{N}} \mu(\hat{\boldsymbol{\theta}}_j) - \mu(\boldsymbol{\theta}_1) \right) \\ &\Rightarrow \sum_{i=1}^{\bar{N}} \overline{w}_i^{\bar{N},\infty} \Lambda_i - \left(1 - \sum_{i=1}^{\bar{N}} \overline{w}_i^{\bar{N},\infty} \right) \boldsymbol{d}_0' \boldsymbol{\eta}_1. \end{aligned} \tag{7}$$

Observe that the estimator in equation (7) is a valid averaging estimator, as the weights sum to unity by construction. The exact way $\boldsymbol{v}_N$ is picked does not matter, as long as the decay condition holds. All admissible choices lead to the same limit. In particular, we may pick equal weights $v_{i\,N} = 1/(N - \bar{N})$, as we do in eq. (6).

4 Simulation Study

We study the finite sample performance of our minimum MSE unit averaging methodology via a simulation exercise. We consider a model similar to the one we use in our empirical application – a linear dynamic heterogeneous panel model defined as

$$y_{it} = \beta_i x_{it} + \lambda_i y_{it-1} + u_{it}\ , \quad \varepsilon_{it} \overset{i.i.d.}{\sim} N(0, \sigma_i^2)\ ,$$

where we assume $\mathbb{E}(\varepsilon_{it}|y_{it-1}, x_{it}) = 0$. The prediction error ε_{it} is cross-sectionally heteroskedastic, with variance σ_i^2 drawn independently from a standard exponential distribution. The exogenous variable x_{it} is independently drawn from from a standard normal distribution. The initial conditions for y_{i0} are independently drawn from a normal distribution with mean zero and variance $(1+\sigma_i^2)/(1-\lambda_i^2)$ to ensure that $\{y_{it}\}_t$ is covariance stationary. The heterogeneous parameter $\boldsymbol{\theta}_i = (\beta_i,\lambda_i)'$ is specified as $\beta_i = 1 + \eta_{i\beta}/\sqrt{T}$ and $\lambda_i = \eta_{i\lambda}/\sqrt{T}$. The sequences of idiosyncratic components $\{\eta_{i,\beta}\}$ and $\{\eta_{i\lambda}\}$ are independently drawn from, respectively, a standard normal distribution and a uniform distribution with support $[-4,4]$. We simulate panels with a cross-sectional dimension N of $10, 25$, and 50 units and a time dimension T of 60 periods. We remark that $T = 60$ matches the time dimension of the estimation sample in the empirical application. As $T = 60$, the support of λ_i is approximately $[-0.5, 0.5]$.

We consider two focus parameters of interest: λ_1 and the forecast of $y_{1\,T+1}$ given by the conditional mean $\mathbb{E}(y_{1\,T+1}|y_{1\,T}, x_{1\,T} = 1) = \lambda_1 y_{1\,T} + \beta_1$.[9][10] The minimum MSE unit averaging estimators corresponding to these focus parameters are used for estimation. We consider weights based on both the fixed-N and large-N regime. In the large-N regime we consider weights based on the choices $\bar{N} = 10$ and $\bar{N} = 20$.[11] The performance of the minimum MSE unit averaging estimator is benchmarked against a number of alternative approaches. We consider the individual estimators of unit 1, the mean group estimator, as well as the unit averaging estimator based on AIC/BIC weights (Buckland et al., 1997) and MMA weights (Hansen, 2007; Wan et al., 2010). Note that AIC and BIC generate the same weights, since each unit has the same number of coefficients.[12] Similarly, the MMA weights reduce to minimal BIC model selection.

Figure 1 summarizes the main results of the simulation study based on 2500 replications.

[9] The value of $y_{1\,T}$ is determined in each sample by the dynamic process governing y.

[10] In the online appendix, we also report results for estimating β_1 and the long-run effect $\beta_1/(1-\lambda_1)$.

[11] When the panel cross-sectional dimension is $N = 10$ all three approximations are identical.

[12] More formally, our weights effectively correspond to conditional AIC weights (Vaida and Blanchard, 2005) where we treat the random slopes as fixed parameters of interest. See also Vaida and Blanchard (2005); Donohue et al. (2011); Zhang et al. (2014)

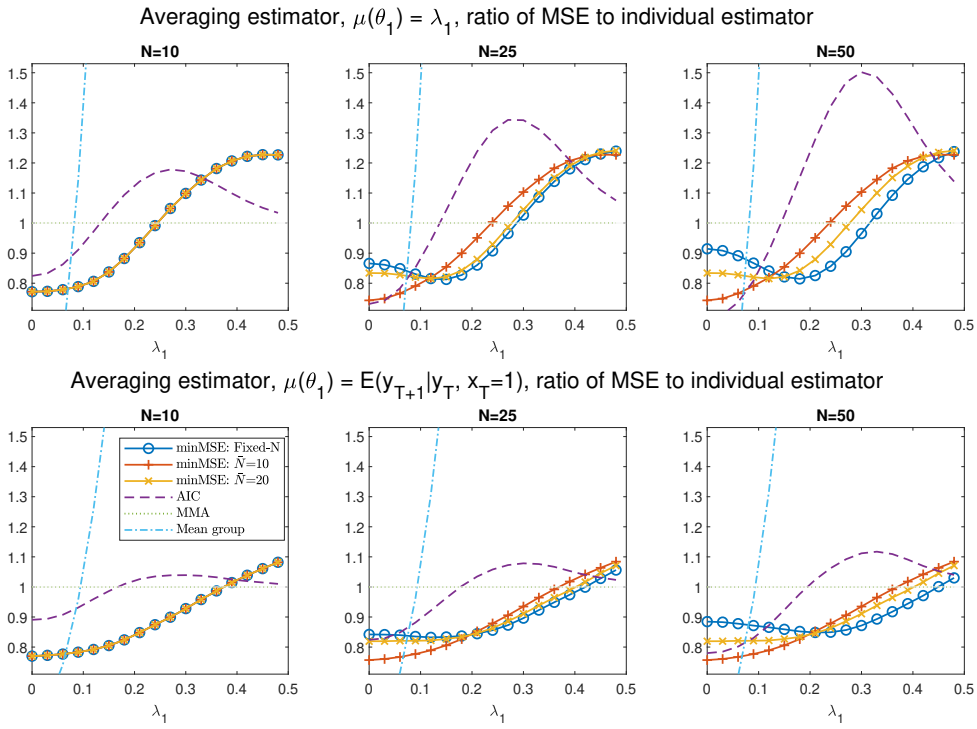

Figure 1: MSE performance of the unit averaging estimator relative to the individual estimator. The focus parameters are the autoregressive parameter λ_1 and conditional mean (forecast). $T = 60$, $N = 10, 25, 50$

We report performance of different averaging strategies in terms of MSE relative to the individual estimator for a range of values of λ_1. As the distribution of λ_1 is symmetric, we report performance for half of the support of λ_1. In each plot, we plot the MSE of different averaging estimators relative to the MSE of the individual estimator of unit 1. A value larger than 1 indicates that the individual estimator is more efficient. The minimum MSE unit averaging estimator performs favorably throughout most of the parameter space. Large-N approximations work better for λ_1 closer to $\mathbb{E}(\lambda_1) = 0$. However, for larger values of λ_1 the flexibility of the fixed-N regime is an advantage, as it can freely choose units most similar to unit 1. None of the averaging estimators dominate the other. As expected, the mean group estimator performs very well close to the mean of λ_1, but bias starts to dominate as λ_1 is increased. AIC weights offers a risk profile somewhat similar to that of the MG estimator for values closer mean of λ_1. As λ_1 becomes larger, AIC moves towards the individual estimator. As noted above, MMA is effectively minimal

BIC model selection in the context of unit averaging. MMA almost always selects unit 1, and correspondingly MMA does not offer improvements over the individual estimator.

5 Empirical Application

We illustrate our methodology with a pseudo-real time nowcasting exercise for quarterly GDP for a panel of European countries. GDP prediction is a natural application of our unit averaging methodology. There is evidence of considerable heterogeneity between countries, yet at the same time pooling the data at least partially improves prediction accuracy (Garcia-Ferrer et al., 1987; Hoogstrate et al., 2000; Marcellino et al., 2003). The design of our application follows standard practices in the nowcasting literature (Marcellino and Schumacher, 2010; Schumacher, 2016). The literature on nowcasting is vast and we do not to cover it here. We refer to Bańbura, Giannone, Modugno, and Reichlin (2013) for a survey.

We use quarterly GDP data from 1995Q1 to 2019Q4 for 12 European countries: the 11 founding Eurozone economies and the UK. We enrich our dataset with a set of 162 monthly GDP predictors for each country. The set of predictors include both real, price, and survey data. Table OA.1 in the online appendix contains the complete list of variables and descriptions.[13] All non-survey data is available from Eurostat whereas the survey data is available from the DG ECFIN. We use final data releases incorporating all revisions, making our study a pseudo-real time one.

Our empirical design takes into account both the delays in publication of monthly data ("ragged-edge problem") and the impact of timing on the information set available ("vintages" of data). First, the predictor variables are typically released with different delays after the end of the corresponding month, which is known as the "ragged-edge" problem (Wallis, 1986).[14] We adopt a stylized release calendar of bimonthly releases

[13]Not all variables are available for all countries at a given time. This only impacts the precision in estimating country-specific factors.

[14]For example, industrial production data is released 6 weeks after the end of the month, while survey

to account for this (table OA.1 in the online appendix lists the release delay for all variables). Second, as the quarter goes by, more data becomes available.[15] Each possible position in time determines a data "vintage". We assume that a month has 4 weeks; in accordance with our release calendar, we nowcast every two weeks starting from the first day of the quarter at -12 weeks (relative to the quarter end) until $+4$ weeks after the end of the quarter (GDP is released at $+6$ weeks). Formally, let t index months. Then $v = -3, -5/2, -2, \ldots, 1/2, +1$ is a fractional value that describes the monthly position (or vintage) relative to the end of the quarter, in increments of two weeks.[16]

We nowcast GDP in quarter $3t$ using all information available at time $3t+v$, separately for each value of $v \in \{-3, -5/2, \ldots, +1\}$. As we have a large number of predictors available at monthly frequency, we opt for factor unrestricted MIDAS (U-MIDAS) (Foroni, Marcellino, and Schumacher, 2015). Given v, for each country we estimate monthly factors f_{it} with $\hat{f}_{it|v}$ for all $t = 1, \ldots, \lfloor T + v \rfloor$ using the full dataset available at $T + v$.[17] The GDP is modeled as

$$y_{i\,3t} = \alpha_{i|v} + \sum_{k=0}^{11} \beta_{i\,k|v} \hat{f}_{i\,\lfloor 3t+v-k \rfloor|v} + \lambda_{i|v} y_{i\,3(t-1)} + \varepsilon_{i\,3t|v},$$

where $y_{i\,3t}$ is GDP of country i in quarter $3t$ and $\varepsilon_{i\,3t|v}$ is the prediction error. The country factors estimates $\hat{f}_{it|v}$ are extracted from the large set of predictor variables using the EM-PCA method (Stock and Watson, 1999). If GDP of quarter $3(t-1)$ is not available at moment v, we use $y_{i\,3(t-2)}$ instead.[18] We use only one factor for prediction following

data is released at the end of the month without delay.

[15] For example, nowcasting Q4 GDP can be done at any moment between October 1 when no data on Q4 is available yet up to the middle of the following February, when GDP data for Q4 is released. The amount of data available increases monotonically between these two dates.

[16] For example, if $v = 0$, nowcasting uses all information that is available at the end of quarter $3t$. If $v = +1$, nowcasting uses all the data available $+4$ weeks after the end of quarter $3t$, the last weekly position we consider. Each step of $-1/2$ corresponds to stepping back 2 weeks until $v = -3$ corresponds to the position of -12 weeks.

[17] For example, suppose we wish to nowcast Q4 GDP. If $v = 1$, we estimate factors up to January of the following year using information available at the end of January. If $v = 1/2$, we estimate factors up to December using all the information available in the middle of January.

[18] Quarterly GDP is released six weeks after the end of the relevant quarter, which corresponds to $v = -3/2$. For $v = -3, -5/2, -2$, we use $y_{i\,3(t-2)}$ in place of $y_{i\,3(t-1)}$.

Marcellino and Schumacher (2010) and we include the lag of GDP following Clements and Galvão (2008). We nowcast GDP for each country using the conditional mean of GDP implied by the U-MIDAS specification. Parameter estimation is carried out using a rolling-windows of sizes 44, 60 and 76 quarters.[19] Factors are also re-estimated every two weeks using the all the data available at each point in time.

We estimate the conditional mean using the fixed-N minimum MSE unit averaging estimator, since cross-sectional dimension is not large and each unit is potentially relevant. The performance of our minimum MSE unit averaging estimator is benchmarked against the individual, mean group, and averaging estimators using AIC and Mallows weights.

In table 1 we provide a summary of forecasting performance results for GDP nowcasting. The table reports the MSE of the individual estimator as well as the MSE relative to the individual estimator for all other strategies. The table reports results for the five largest economies in our sample along with the GDP-weighted mean.[20] We select the vintages that correspond to $-6, 0, +4$ weeks relative to the end of the quarter (corresponding to $v = -3/2, 0, +1$). Full results for all vintages and countries are provided in table OA.7 in the online appendix, and they are similar to the ones reported here.

Our key finding is that averaging with smooth data-dependent criteria – minimum MSE or AIC weights – generally leads to improved forecasting performance, though the degree of improvement varies with the country in question. This is clear from table 1, as the vast majority of entries corresponding to those weights display relative MSE smaller than one, with improvements reaching up to 20%. The average gain in performance is on the scale of about 9% for minimum MSE weights and 5% for AIC weights. We also observe that minimum MSE weights and AIC weights do not uniformly dominate each other.

Figure 2 provides a box plot for relative MSEs for nowcasting GDP for all the

[19] Forecast evaluation begins in 2006Q1 for window size 44, 2010Q1 for T=60 and 2014Q1 for T=76.

[20] Weighing by GDP as in Marcellino et al. (2003) emulates forecasting the Eurozone GDP using individual forecasts.

		−6 weeks			0 weeks			+4 weeks		
	Averaging	44q	60q	76q	44q	60q	76q	44q	60q	76q
Mean	Individual	1.113	0.986	1.167	0.973	1.010	1.196	0.933	0.914	1.124
	minMSE	0.916	0.936	0.907	0.889	0.936	0.910	0.881	0.928	0.901
	AIC	0.933	0.962	0.980	0.908	0.960	0.974	0.878	0.949	0.955
	MMA	1.119	1.178	1.099	1.011	1.036	1.114	1.000	0.921	0.916
	Mean group	1.417	1.570	1.524	1.635	1.696	1.505	1.879	1.704	1.489
DE	Individual	0.661	0.546	0.537	0.509	0.421	0.434	0.565	0.449	0.456
	minMSE	0.793*	0.822	0.815	0.787*	0.818	0.775	0.821*	0.809	0.794
	AIC	0.963*	0.977*	0.989	0.974	0.982*	0.973	0.978	0.989*	0.978
	MMA	1.002	1.257	1.098	0.860	0.970	0.742	0.741	0.830	0.834
	Mean group	0.987	0.937	0.773	1.069	1.157	0.742	0.957	1.153	0.849
FR	Individual	0.194	0.154	0.129	0.143	0.100	0.086	0.155	0.121	0.098
	minMSE	0.988	1.067	1.037	0.971	1.059	1.159	0.916	0.978	1.069
	AIC	0.883*	0.934	0.975	0.833*	0.978	1.049	0.828*	0.935*	0.999
	MMA	1.223	1.343*	1.139	1.082	1.372*	1.659*	1.164	1.048	1.337
	Mean group	2.125*	2.068*	1.348	2.736*	2.942*	2.169*	2.473*	2.652*	2.156*
IT	Individual	0.591	0.253	0.156	0.279	0.178	0.116	0.232	0.131	0.082
	minMSE	0.893*	0.908*	0.852*	0.973	0.974	0.858*	1.046	1.025	0.857*
	AIC	0.945	0.972*	0.980	0.955	0.951*	0.976*	0.947	0.969*	0.975*
	MMA	0.907	0.710*	0.729*	1.323	0.650*	0.704	1.351	0.917	0.688*
	Mean group	0.895	0.901	0.822	1.289	1.042	0.719	1.491*	1.595*	1.239
ES	Individual	0.288	0.198	0.147	0.233	0.121	0.106	0.253	0.114	0.102
	minMSE	0.919	0.909*	0.856*	0.955	0.951	0.927	0.957	0.919	0.889
	AIC	0.958	0.961*	0.974*	0.860	0.940	0.940*	0.813*	0.934	0.933*
	MMA	0.928	0.900	1.000	0.922	0.906	1.002	1.312	0.821	1.000
	Mean group	1.237*	1.427*	1.225	1.011	1.561*	1.352	0.946	1.886*	1.248
UK	Individual	0.281	0.116	0.044	0.254	0.142	0.047	0.244	0.142	0.047
	minMSE	0.928	0.988	0.953	0.840	0.984	0.868	0.743	1.034	0.913
	AIC	0.871*	0.933	0.953	0.876	0.958	0.941	0.726*	0.917*	0.898*
	MMA	1.457	1.375*	1.350	1.240	1.318*	1.580	1.523	1.069	0.854
	Mean group	1.714	2.444*	3.688*	2.530*	2.445*	3.121*	4.330*	2.214*	2.650*

Table 1: Nowcasting MSE. *For individual estimator*: absolute value. *For averaging estimators*: MSE relative to individual estimator. For different estimation window sizes (44, 60, 76); selected weekly horizons relative to quarter end $(-6, 0, +4$ weeks). * – forecasting performance difference significant at 10% in Diebold-Mariano test (Diebold and Mariano, 1995)

countries in the panel for the vintages considered in table 1. The figure illustrates that the favorable performance is robust across countries and not limited to the five biggest economies reported in table 1. Both minimum MSE and AIC weights generally lead to an improvement in performance, as both rarely have relative MSE above one. There is some evidence that the minimum MSE weights have a greater upside, at the price of potentially some more variability in the results, while AIC leads to smaller, but more tightly concentrated improvements. Further, we find that averaging is more attractive for the smallest sample size of $T = 44$, with relative MSE generally approaching one as T increases. This can be clearly seen in figure 2, as the improvement range for AIC and

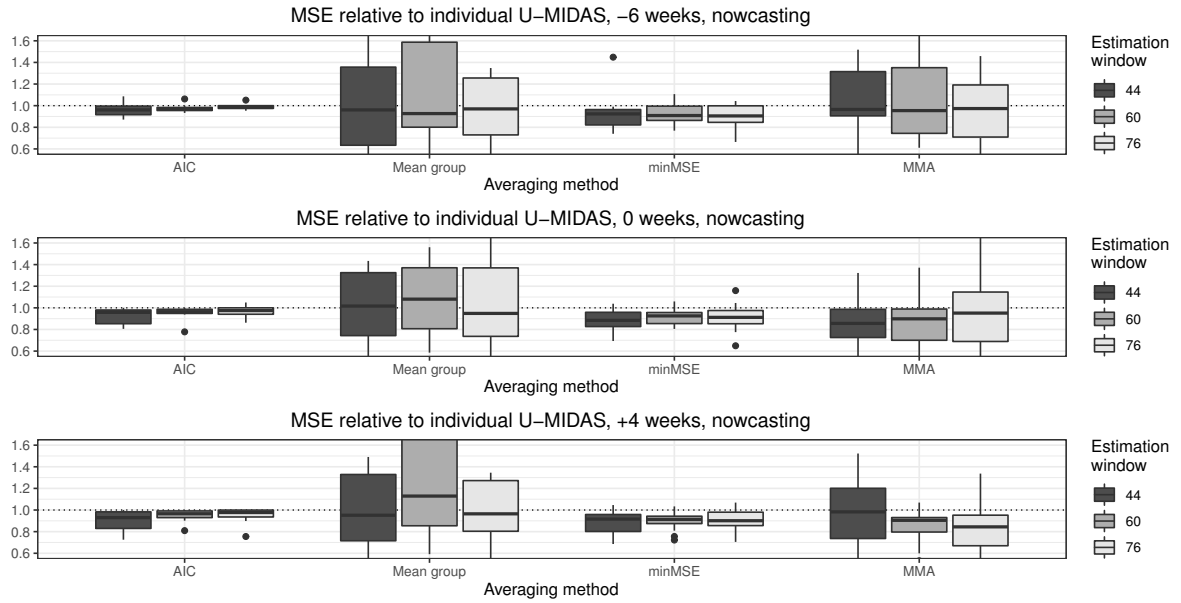

Figure 2: Distribution of relative MSEs across countries. Split by different averaging strategies and estimation window size. Same weekly positions as reported in table 1

minimum MSE weights becomes more concentrated and to closer to one. As previously remarked, as T increases, the minimum MSE estimator converges to the individual estimator; a similar point applies to AIC weights if the log likelihood is not divided by samples size and allowed to diverge as sample size grows.

Averaging methods that are data-independent (mean group) or not smooth (MMA) lead to generally poor results. While capable of offering improvements, the two approaches often lead to significantly worse performance than the individual estimator, regardless of T. This is clear from figure 2, where the distribution of the MSE of both averaging strategies has substantial mass above one.

We remark that the online appendix contains a number of additional robustness checks. First, we consider nowcasting using bridge equations instead of U-MIDAS. Second, we consider one- and two-quarter-ahead GDP forecasting. The evidence emerging from these additional robustness checks on the performance of minimum MSE unit averaging estimator is consistent with the overall evidence reported here.

6 Conclusions

In this work we introduce a unit averaging estimator to recover unit-specific parameters in a general class of panel data models with heterogeneous parameters. The procedure consists in estimating the parameter of a given unit using a weighted average of all the unit-specific parameter estimators in the panel. The weights of the average are determined by minimizing an MSE criterion. The paper studies the properties of the procedures using a local heterogeneity framework that builds upon the literature on frequentist model averaging (Hjort and Claeskens, 2003; Hansen, 2008). An application to GDP nowcasting for a panel of European countries shows that the procedure performs favorably for prediction relative to a number of alternative procedures.

References

Aliprantis, C. D. and Border, K. C. (2006), *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer Berlin Heidelberg, 3rd ed.

Baltagi, B. H. (2013), *Panel data forecasting*, vol. 2, Elsevier B.V.

Baltagi, B. H., Bresson, G., and Pirotte, A. (2008), "To Pool or Not to Pool," in *The Econometrics of Panel Data*, Springer Berlin Heidelberg, chap. 16, pp. 517–546.

Bańbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013), "Now-Casting and the Real-Time Data Flow," in *Handbook of Economic Forecasting, Vol. 2 Part A*, eds. Elliott, G. and Timmermann, A., chap. 4, pp. 195–237.

Bao, Y. and Ullah, A. (2007), "The second-order bias and mean squared error of estimators in time-series models," *Journal of Econometrics*, 140, 650–669.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618.

Claeskens, G. and Hjort, N. L. (2008), *Model Selection and Model Averaging*, Cambridge: Cambridge University Press.

Clements, M. P. and Galvão, A. B. (2008), "Macroeconomic Forecasting with Mixed-Frequency Data: Forecasting Output Growth in the United States," *Journal of Business and Economic Statistics*, 26, 546–554.

Diebold, F. X. and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253.

Donohue, M. C., Overholser, R., Xu, R., and Vaida, F. (2011), "Conditional Akaike Information Under Generalized Linear and Proportional Hazards Mixed Models," *Biometrika*, 98, 685–700.

Elandt, R. C. (1961), "The Folded Normal Distribution: Two Methods of Estimating Parameters from Moments," *Technometrics*, 3, 551–562.

Fang, F., Yuan, C., and Tian, W. (2022), "An Asymptotic Theory for Least Squares Model Averaging with Nested Models," *Econometric Theory*, 1–30.

Foroni, C., Marcellino, M., and Schumacher, C. (2015), "Unrestricted Mixed Data Sampling (MIDAS): MIDAS Regressions with Unrestricted Lag Polynomials," *Journal of the Royal Statistical Society: Series A*, 178, 57–82.

Gao, Y., Zhang, X., Wang, S., and Zou, G. (2016), "Model averaging based on leave-subject-out cross-validation," *Journal of Econometrics*, 192, 139–151.

Garcia-Ferrer, A., Highfield, R. A., Palm, F. C., and Zellner, A. (1987), "Macroeconomic Forecasting Using Pooled International Data," *Journal of Business and Economic Statistics*, 5, 53–67.

Hansen, B. E. (2007), "Least squares model averaging," *Econometrica*, 75, 1175–1189.

— (2008), "Least-squares forecast averaging," *Journal of Econometrics*, 146, 342–350.

— (2016), "Efficient shrinkage in parametric models," *Journal of Econometrics*, 190, 115–132.

Hansen, B. E. and Racine, J. S. (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46.

Hjort, N. L. and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899.

Hoogstrate, A. J., Palm, F. C., and Pfann, G. A. (2000), "Pooling in Dynamic Panel-Data Models: An Application to Forecasting GDP Growth Rates," *Journal of Business and Economic Statistics*, 18, 274–283.

Horn, R. A. and Johnson, C. R. (2012), *Matrix Analysis*, Cambridge University Press, 2nd ed.

Issler, J. V. and Lima, L. R. (2009), "A panel data approach to economic forecasting: The bias-corrected average forecast," *Journal of Econometrics*, 152, 153–164.

Kallenberg, O. (2021), *Foundations of Modern Probability*, Springer Cham, 3rd ed.

Liu, C.-A. (2015), "Distribution Theory of the Least Squares Averaging Estimator," *Journal of Econometrics*, 186, 142–159.

Liu, L., Moon, H. R., and Schorfheide, F. (2020), "Forecasting with Dynamic Pane Data Models," *Econometrica*, 88, 171–201.

Maddala, G. S., Li, H., and Srivastava, V. K. (2001), "A Comparative Study of Different Shrinkage Estimators for Panel Data Models," *Annals of Economics and Finance*, 2, 1–30.

Maddala, G. S., Trost, R. P., Li, H., and Joutz, F. (1997), "Estimation of Short-Run and Long-Run Elasticities of Energy Demand From Panel Data Using Shrinkage Estimators," *Journal of Business and Economic Statistics*, 15, 90–100.

Marcellino, M. and Schumacher, C. (2010), "Factor MIDAS for Nowcasting and Forecasting with Ragged-Edge Data: A Model Comparison for German GDP," *Oxford Bulletin of Economics and Statistics*, 72, 518–550.

Marcellino, M., Stock, J. H., and Watson, M. W. (2003), "Macroeconomic Forecasting in the Euro Area: Country Specific Versus Area-Wide Information," *European Economic Review*, 47, 1–18.

Pesaran, M. H., Shin, Y., and Smith, R. P. (1999), "Pooled Mean Group Estimation of Dynamic Heterogeneous Panels," *Journal of the American Statistical Association*, 94, 621–634.

Pesaran, M. H. and Smith, R. P. (1995), "Estimating long-run relationships from dynamic heterogeneous panels," *Journal of Econometrics*, 6061, 473–477.

Phillips, P. and Moon, H. R. (1999), "Linear regression limit theory for nonstationary panel data," *Econometrica*, 67, 1057–1111.

Pötscher, B. M. and Prucha, I. R. (1997), *Dynamic Nonlinear Econometric Models: Asymptotic Theory*, Springer.

Pruitt, W. E. (1966), "Summability of Independent Random Variables," *Journal of Mathematics and Mechanics*, 15, 769–776.

Rilstone, P., Srivastava, V. K., and Ullah, A. (1996), "The Second-Order Bias and Mean Squared Error of Nonlinear Estimators," *Journal of Econometrics*, 75, 369–395.

Schumacher, C. (2016), "A Comparison of MIDAS and Bridge Equations," *International Journal of Forecasting*, 32, 257–270.

Stock, J. H. and Watson, M. W. (1999), "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series," in *Cointegration, Causality and Forecasting: A Festschrift for Clive W.J. Granger*, eds. Engle, R. F. and White, H., Oxford University Press, pp. 1–44.

Vaida, F. and Blanchard, S. (2005), "Conditional Akaike Information for Mixed-Effects Models," *Biometrika*, 92, 351–370.

Van der Vaart, A. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer.

Wallis, K. F. (1986), "Forecasting with an Econometric Model: The 'Ragged Edge' Problem," *Journal of Forecasting*, 5, 1–13.

Wan, A. T. K., Zhang, X., and Zou, G. (2010), "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283.

Wang, W., Zhang, X., and Paap, R. (2019), "To pool or not to pool: What is a good strategy for parameter estimation and forecasting in panel regressions?" *Journal of Applied Econometrics*, 34, 724–745.

Yang, Z. (2015), "A general method for third-order bias and variance corrections on a nonlinear estimator," *Journal of Econometrics*, 186, 178–200.

Yin, S.-Y., Liu, C.-A., and Lin, C.-C. (2021), "Focused Information Criterion and Model Averaging for Large Panels with a Multifactor Error Structure," *Journal of Business and Economic Statistics*, 39, 54–68.

Zhang, X., Zou, G., and Liang, H. (2014), "Model averaging and weight choice in linear mixed-effects models," *Biometrika*, 101, 205–218.

Proofs of Results in the Main Text

Under assumption A.1 we work conditional on $\{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots\}$. We use $\mathbb{E}[\cdot]$ to denote the expectation operator conditional on $\{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots\}$, whereas $\mathbb{E}_{\boldsymbol{\eta}}[\cdot]$ is the expectation taken with respect the distribution of $\boldsymbol{\eta}$. All results are shown to hold with probability one with respect to the distribution of $\boldsymbol{\eta}$ (denoted $\boldsymbol{\eta}$-a.s.).

A.1 Proof of Lemma 1

Recall that the data vector $\boldsymbol{z}_{it}$ takes values in $\mathcal{Z} \subset \mathbb{R}^d$ and define the data matrix $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}', \ldots, \boldsymbol{z}_{iT}')'$ that takes values in $\mathcal{Z}^T = \prod_{t=1}^T \mathcal{Z}$. Recall that that the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ takes values in $\Theta \subset \mathbb{R}^p$. We denote by $\nabla m(\boldsymbol{\theta}, \boldsymbol{z}_{it})$ the gradient vector of m with respect to $\boldsymbol{\theta}$, by $\nabla^2 m(\boldsymbol{\theta}, \boldsymbol{z}_{it})$ the Hessian matrix of m with respect to $\boldsymbol{\theta}$, by $\nabla_{\theta_k} m(\boldsymbol{\theta}, \boldsymbol{z}_{it})$ the partial derivative of m with respect to θ_k, and by $\nabla^2_{\boldsymbol{\theta}\,\theta_k}$ the gradient vector of $\nabla_{\theta_k} m(\boldsymbol{\theta}, \boldsymbol{z}_{it})$ with respect to $\boldsymbol{\theta}$.

We establish a mean value theorem that does not require compactness of Θ.

Lemma A.1.1. *Suppose assumption A.3 is satisfied. Then for each unit i, any T and any $k = 1, \ldots, p$ there exists a measurable function $\tilde{\boldsymbol{\theta}}_{i\,k}$ from $\mathcal{Z}^T$ to Θ such that the individual estimator $\hat{\boldsymbol{\theta}}_i$ of eq. (2) satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta_k} m(\hat{\boldsymbol{\theta}}_i, \boldsymbol{z}_{it}) = \frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta_k} m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it}) + \left[\frac{1}{T}\sum_{t=1}^{T} \nabla^2_{\boldsymbol{\theta}\,\theta_k} m(\tilde{\boldsymbol{\theta}}_{i\,k}, \boldsymbol{z}_{it})\right]' \left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right) \ ,$$

where $\tilde{\boldsymbol{\theta}}_{i\,k}$ lies on the segment joining $\hat{\boldsymbol{\theta}}_i$ and $\boldsymbol{\theta}_i$.

Further, suppose A.5 is satisfied. Then for each i and any T there exist measurable functions $\bar{\boldsymbol{\theta}}_i$, $\acute{\boldsymbol{\theta}}_i$ and $\check{\boldsymbol{\theta}}_i$ from $\mathcal{Z}^T$ to Θ such that the individual estimator $\hat{\boldsymbol{\theta}}_i$ of eq. (2)

satisfies

$$\mu(\hat{\boldsymbol{\theta}}_i) = \mu(\boldsymbol{\theta}_1) + \nabla\mu(\bar{\boldsymbol{\theta}}_i)'(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1) \ , \tag{A.1.1}$$

$$\mu(\hat{\boldsymbol{\theta}}_i) = \mu(\boldsymbol{\theta}_1) + \boldsymbol{d}_1'(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1) \ , \tag{A.1.2}$$

$$\mu(\hat{\boldsymbol{\theta}}_i) = \mu(\boldsymbol{\theta}_i) + \boldsymbol{d}_i'(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)'\nabla^2\mu(\check{\boldsymbol{\theta}}_i)(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \ , \tag{A.1.3}$$

where $\boldsymbol{d}_1 = \nabla\mu(\boldsymbol{\theta}_1)$*;* $\bar{\boldsymbol{\theta}}_i$ *and* $\acute{\boldsymbol{\theta}}_i$ *lie on the segment joining* $\hat{\boldsymbol{\theta}}_i$ *and* $\boldsymbol{\theta}_1$*; and* $\check{\boldsymbol{\theta}}_i$ *lies on the segment joining* $\hat{\boldsymbol{\theta}}_i$ *and* $\boldsymbol{\theta}_i$*.*

Proof. Fix $k \in \{1, \ldots, p\}$ and define the function $f_i : \mathcal{Z}^T \times [0,1] \to \mathbb{R}$ as

$$\begin{aligned} f_i(\boldsymbol{z}_i, y) = \frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta_k} m(\hat{\boldsymbol{\theta}}_i, \boldsymbol{z}_{it}) - \frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta_k} m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it}) \\ - \left[\frac{1}{T}\sum_{t=1}^{T} \nabla^2_{\boldsymbol{\theta}\,\theta_k} m(y\hat{\boldsymbol{\theta}}_i + (1-y)\boldsymbol{\theta}_i, \boldsymbol{z}_i)\right]' (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \ . \end{aligned}$$

A.3 implies that f_i is well-defined, as for each $y \in [0,1]$ we have that $y\hat{\boldsymbol{\theta}}_i + (1-y)\boldsymbol{\theta}_i \in \Theta$. f_i is a measurable function of $\boldsymbol{z}_i$ for every fixed value $y \in [0,1]$, as $\hat{\boldsymbol{\theta}}_i$ and m are measurable functions of $\boldsymbol{z}_i$ and m is continuously differentiable in $\boldsymbol{\theta}$. f_i is a continuous function of y for every value of $\boldsymbol{z}_i$.

Define the correspondence $\varphi_i : \mathcal{Z}^T \to [0,1]$ as $\varphi_i(\boldsymbol{z}_i) = \{y \in [0,1] : f_i(\boldsymbol{z}_i, y) = 0\}$. The function f_i satisfies the assumptions of corollary 18.8 in Aliprantis and Border (2006), and so φ_i is a measurable correspondence. $\varphi_i(\boldsymbol{z}_i)$ is nonempty for every $\boldsymbol{z}_i$, as by the mean value theorem, for every fixed value of $\boldsymbol{z}_i$ there exists some $\tilde{y} \in [0,1]$ such that

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta_k} m(\hat{\boldsymbol{\theta}}_i, \boldsymbol{z}_{it}) = \frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta_k} m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it}) \\ + \left[\frac{1}{T}\sum_{t=1}^{T} \nabla^2_{\boldsymbol{\theta},\theta_k} m(\tilde{y}\hat{\boldsymbol{\theta}}_i + (1-\tilde{y})\boldsymbol{\theta}_i, \boldsymbol{z}_{it})\right]' \left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right) . \end{aligned}$$

In addition, $\varphi_i(\boldsymbol{z}_i)$ is closed for every $\boldsymbol{z}_i$ as m is twice continuously differentiable in $\boldsymbol{\theta}$ by

assumption A.3. Then by the Kuratowski-Ryll-Nardzewski measurable selection theorem (theorem 18.13 in Aliprantis and Border (2006)), $\varphi_i(\boldsymbol{z}_i)$ admits a measurable selector $\tilde{y}_{ik} = \tilde{y}_{ik}(\boldsymbol{z}_i)$. Finally, define $\tilde{\boldsymbol{\theta}}_{ik} = \tilde{y}_{ik}\hat{\boldsymbol{\theta}}_i + (1 - \tilde{y}_{ik})\boldsymbol{\theta}_i$ and note that $\tilde{\boldsymbol{\theta}}_{ik}$ satisfies the requirements of the lemma. This establishes the first claim of the lemma.

The proof of the second claim of the lemma is analogous. □

The following lemma is needed to prove lemmas 1 and A.2.1.

Lemma A.1.2. *Suppose A.3 is satisfied. Let $\tilde{\boldsymbol{\theta}}_{ij} : \mathcal{Z}^T \to \mathbb{R}^p$ for $j = 1, \ldots, p$ be a sequence of measurable functions that lie on the segment joining $\boldsymbol{\theta}_i$ and $\hat{\boldsymbol{\theta}}_i$ and define*

$$\hat{\boldsymbol{H}}_{iT} = \begin{bmatrix} \left[\frac{1}{T}\sum_{t=1}^{T} \nabla^2_{\boldsymbol{\theta},\theta_1} m(\tilde{\boldsymbol{\theta}}_{i1}, \boldsymbol{z}_{it})\right]' \\ \ldots \\ \left[\frac{1}{T}\sum_{t=1}^{T} \nabla^2_{\boldsymbol{\theta},\theta_p} m(\tilde{\boldsymbol{\theta}}_{ip}, \boldsymbol{z}_{it})\right]' \end{bmatrix}.$$

Then for all $T > T_0$ the matrix $\hat{\boldsymbol{H}}_{iT}$ (i) is a.s. nonsingular and (ii) satisfies

$$\mathbb{E}\left[\left\|\boldsymbol{H}_i^{-1} - \hat{\boldsymbol{H}}_{iT}^{-1}\right\|_\infty^{\frac{2(2+\delta)(1+\delta)}{\delta}}\right] \leq p^{\frac{(2+\delta)(1+\delta)}{\delta}} \underline{\lambda}_{\boldsymbol{H}}^{-\frac{2(2+\delta)(1+\delta)}{\delta}} C_{\nabla^2 m} ,$$

where $\boldsymbol{H}_i = \lim_{T\to\infty} \mathbb{E}\left[T^{-1}\sum_{t=1}^{T} \nabla^2 m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it})\right]$.

Proof. The proof of assertion (i) is based on showing that $\left\|(\boldsymbol{H}_i - \hat{\boldsymbol{H}}_{iT})\boldsymbol{H}_i^{-1}\right\|_\infty < 1$ holds almost surely, which implies that the matrix $\hat{\boldsymbol{H}}_{iT}$ is a.s. nonsingular.[21] Let $\boldsymbol{H}_i^{-1} = (h^{ij})$

[21] This result follows from the standard observation that if $\|\boldsymbol{I} - \boldsymbol{A}\|_\infty < 1$, then $\boldsymbol{A}$ is nonsingular. Write $I = \boldsymbol{H}_i\boldsymbol{H}_i^{-1}$ and $\boldsymbol{A} = \hat{\boldsymbol{H}}_{iT}\boldsymbol{H}_i^{-1}$. Then $\|\boldsymbol{I} - \boldsymbol{A}\|_\infty = \left\|(\boldsymbol{H}_i - \hat{\boldsymbol{H}}_{iT})\boldsymbol{H}_i^{-1}\right\|_\infty < 1$. The matrix $\boldsymbol{A}$ is nonsingular, and $\hat{\boldsymbol{H}}_{iT} = \boldsymbol{A}\boldsymbol{H}_i$ is a product of two nonsingular matrices.

and observe that

$$\hat{\boldsymbol{H}}_{iT}\boldsymbol{H}_i^{-1} = \begin{bmatrix} \sum_{k=1}^{p} \nabla^2_{\theta_k\theta_1} m(\tilde{\boldsymbol{\theta}}_{i1}, \boldsymbol{z}_{it}) h^{k1} & \cdots & \sum_{k=1}^{p} \nabla^2_{\theta_k\theta_1} m(\tilde{\boldsymbol{\theta}}_{i1}, \boldsymbol{z}_{it}) h^{kp} \\ \sum_{k=1}^{p} \nabla^2_{\theta_k\theta_2} m(\tilde{\boldsymbol{\theta}}_{i2}, \boldsymbol{z}_{it}) h^{k1} & \cdots & \sum_{k=1}^{p} \nabla^2_{\theta_k\theta_2} m(\tilde{\boldsymbol{\theta}}_{i2}, \boldsymbol{z}_{it}) h^{kp} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{p} \nabla^2_{\theta_k\theta_p} m(\tilde{\boldsymbol{\theta}}_{ip}, \boldsymbol{z}_{it}) h^{k1} & \cdots & \sum_{k=1}^{p} \nabla^2_{\theta_k\theta_p} m(\tilde{\boldsymbol{\theta}}_{ip}, \boldsymbol{z}_{it}) h^{kp} \end{bmatrix} .$$

Row j of $\hat{\boldsymbol{H}}_{iT}\boldsymbol{H}_i^{-1} - \boldsymbol{I}$ coincides with row j of $\left(T^{-1}\sum_{t=1}^{T} \nabla^2 m\left(\tilde{\boldsymbol{\theta}}_{ij}, \boldsymbol{z}_{it}\right)\right)\boldsymbol{H}^{-1} - \boldsymbol{I}$. Then we have that

$$\begin{aligned} \left\|(\boldsymbol{H}_i - \hat{\boldsymbol{H}}_{iT})\boldsymbol{H}_i^{-1}\right\|_\infty &= \left\|\hat{\boldsymbol{H}}_{iT}\boldsymbol{H}_i^{-1} - \boldsymbol{I}\right\|_\infty \\ &\leq \max_{1\leq j\leq p} \left\| \left(T^{-1}\sum_{t=1}^{T} \nabla^2 m(\tilde{\boldsymbol{\theta}}_{ij}, \boldsymbol{z}_{it})\right)\boldsymbol{H}_i^{-1} - \boldsymbol{I}\right\|_\infty \\ &\leq \sup_{\boldsymbol{\theta}\in[\boldsymbol{\theta}_i, \hat{\boldsymbol{\theta}}_i]} \left\| \left(T^{-1}\sum_{t=1}^{T} \nabla^2 m(\boldsymbol{\theta}, \boldsymbol{z}_{it})\right)\boldsymbol{H}_i^{-1} - \boldsymbol{I}\right\|_\infty \\ &\equiv D_{iT} , \end{aligned} \tag{A.1.4}$$

where the second inequality holds as all $\tilde{\boldsymbol{\theta}}_{ij}$ lie on the segment joining $\boldsymbol{\theta}_i$ and $\hat{\boldsymbol{\theta}}_i$ and where D_{iT} is defined in A.3. A.3 implies $D_{iT} < 1$ a.s. for $T > T_0$, and thus $\left\|(\boldsymbol{H}_i - \hat{\boldsymbol{H}}_{iT})\boldsymbol{H}_i^{-1}\right\|_\infty < 1$ a.s. for $T > T_0$, which implies the first claim.

As $\hat{\boldsymbol{H}}_{iT}$ is invertible for $T > T_0$ we have (Horn and Johnson, 2012, section 5.8)

$$\left\|\boldsymbol{H}_i^{-1} - \hat{\boldsymbol{H}}_{iT}^{-1}\right\|_\infty \leq \left\|\boldsymbol{H}_i^{-1}\right\|_\infty \frac{\left\|\boldsymbol{H}_i^{-1}\hat{\boldsymbol{H}}_{iT} - \boldsymbol{I}\right\|_\infty}{1 - \left\|\boldsymbol{H}_i^{-1}\hat{\boldsymbol{H}}_{iT} - \boldsymbol{I}\right\|_\infty} \leq \left\|\boldsymbol{H}_i^{-1}\right\|_\infty \frac{D_{iT}}{1 - D_{iT}} ,$$

where the last inequality follows from (A.1.4). Taking expectations, we obtain that

$$\begin{aligned} \mathbb{E}\left[\left\|\boldsymbol{H}_i^{-1} - \hat{\boldsymbol{H}}_i^{-1}\right\|_\infty^{\frac{2(2+\delta)(1+\delta)}{\delta}}\right] &\leq \left\|\boldsymbol{H}_i^{-1}\right\|_\infty^{\frac{2(2+\delta)(1+\delta)}{\delta}} \mathbb{E}\left[\left(\frac{D_{iT}}{1 - D_{iT}}\right)^{\frac{2(2+\delta)(1+\delta)}{\delta}}\right] \\ &\leq p^{\frac{(2+\delta)(1+\delta)}{\delta}} \left\|\boldsymbol{H}_i^{-1}\right\|^{\frac{2(2+\delta)(1+\delta)}{\delta}} C_{\nabla^2 m} \end{aligned}$$

$$\leq p^{\frac{(2+\delta)(1+\delta)}{\delta}} \underline{\lambda}_{\boldsymbol{H}}^{-\frac{2(2+\delta)(1+\delta)}{\delta}} C_{\nabla^2 m} \ ,$$

which establishes the second claim. □

Proof of lemma 1. A.3 and Lemma A.1.1 imply that

$$\begin{aligned} 0 &= \frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta_k} m(\hat{\boldsymbol{\theta}}_i, \boldsymbol{z}_{i\,t}) \\ &= \frac{1}{T}\sum_{t=1}^{T} \nabla_{\theta_k} m(\boldsymbol{\theta}_i, \boldsymbol{z}_{i\,t}) + \left[\frac{1}{T}\sum_{t=1}^{T} \nabla^2_{\boldsymbol{\theta},\theta_k} m(\tilde{\boldsymbol{\theta}}_{i\,k}, \boldsymbol{z}_{i\,t})\right]' \left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right) \ , \end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_{i\,k}$ lies on the segment joining $\boldsymbol{\theta}_i$ and $\hat{\boldsymbol{\theta}}_i$. Define the matrix

$$\hat{\boldsymbol{H}}_{i\,T} = \begin{bmatrix} \left[\frac{1}{T}\sum_{t=1}^{T} \nabla^2_{\boldsymbol{\theta},\theta_1} m(\tilde{\boldsymbol{\theta}}_{i\,1}, \boldsymbol{z}_{i\,t})\right]' \\ \cdots \\ \left[\frac{1}{T}\sum_{t=1}^{T} \nabla^2_{\boldsymbol{\theta},\theta_p} m(\tilde{\boldsymbol{\theta}}_{i\,p}, \boldsymbol{z}_{i\,t})\right]' \end{bmatrix} . \tag{A.1.5}$$

As all $\hat{\boldsymbol{\theta}}_{i\,k}$ lie between $\boldsymbol{\theta}_i$ and $\hat{\boldsymbol{\theta}}_i$, by lemma A.1.2 the matrix $\hat{\boldsymbol{H}}_{i\,T}$ is a.s. nonsingular for $T > T_0$. Observe that $\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i = (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1) - (\boldsymbol{\theta}_i - \boldsymbol{\theta}_1)$. Combining the above two observations, we obtain that for $T > T_0$ it holds that

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1\right) = -\hat{\boldsymbol{H}}_{i\,T}^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla m(\boldsymbol{\theta}_i, \boldsymbol{z}_{i\,t}) + (\boldsymbol{\eta}_i - \boldsymbol{\eta}_1).$$

By assumption A.3 and lemma A.1.2, it holds that

$$-\hat{\boldsymbol{H}}_{i\,T}^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \nabla m(\boldsymbol{\theta}_i, \boldsymbol{z}_{i\,t}) \Rightarrow N(0, \boldsymbol{V}_i).$$

The convergence is joint as all units are independent by A.2.

The second assertion follows from the delta method and the observation that $\nabla\mu(\boldsymbol{\theta}_1) = \nabla\mu(\boldsymbol{\theta}_0 + T^{-1/2}\boldsymbol{\eta}_1) \to \nabla\mu(\boldsymbol{\theta}_0) = \boldsymbol{d}_0$ under the continuity assumption of A.5. □

A.2 Proof of Theorem 1

Before presenting the proof of theorem 1 we introduce a number of intermediate results.

Lemma A.2.1. *Suppose A.1 and A.3 are satisfied. Let δ be as in A.3. Then there exist finite constants $C_{\hat{\boldsymbol{\theta}},1}, C_{\hat{\boldsymbol{\theta}},1+\delta/2}, C_{\hat{\boldsymbol{\theta}},2}, C_{\hat{\boldsymbol{\theta}},2+\delta}$, which do not depend on i or T, such that the following moment bounds hold for the individual estimator (2) for all $T > T_0$*

$$\mathbb{E}\left[\left\|\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\right\|^k\right] \leq C_{\hat{\boldsymbol{\theta}},k} \ , \quad k = 1{,}1+\delta/2, 2, 2+\delta,$$
$$\mathbb{E}\left[\left\|\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)\right\|^2\right] \leq C_{\hat{\boldsymbol{\theta}},2} + 2C_{\hat{\boldsymbol{\theta}},1}\left\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\right\| + \left\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\right\|^2 \ .$$

Proof. Let the matrix $\hat{\boldsymbol{H}}_{iT}$ be defined as in eq. (A.1.5). By lemma A.1.2 the matrix $\hat{\boldsymbol{H}}_{iT}$ is non-singular for $T > T_0$. Then, as in the proof of lemma 1, for $T > T_0$ it holds that

$$\begin{aligned}\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right) &= -\hat{\boldsymbol{H}}_{iT}^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it}) \\ &= -\boldsymbol{H}_i^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it}) + \left(\boldsymbol{H}_i^{-1} - \hat{\boldsymbol{H}}_{iT}^{-1}\right)\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it}) \ ,\end{aligned}$$

where $\boldsymbol{H}_i = \lim_{T\to\infty}\mathbb{E}\left(\nabla^2 T^{-1}\sum_{t=1}^{T} m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it})\right)$. We separately bound the $(2+\delta)$-th moment of the norm for the two terms above. For the first term we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{H}_i^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it})\right\|^{2+\delta}\right] \\ &\leq \mathbb{E}\left[\left\|\boldsymbol{H}_i^{-1}\right\|^{2+\delta}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it})\right\|^{2+\delta}\right] \\ &\leq \left\|\boldsymbol{H}_i^{-1}\right\|^{2+\delta}\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i, \boldsymbol{z}_{it})\right\|^{2+\delta}\right] \\ &\leq \underline{\lambda}_{\boldsymbol{H}}^{-2-\delta} C_{\nabla m}^{\frac{2+\delta}{2(1+\delta)}},\end{aligned}$$

where the first inequality follows from $\|Ax\| \leq \|A\|\,\|x\|$, and the last line follows by

assumption A.3 and by Jensen's inequality.

For the second term we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\left(\boldsymbol{H}_i^{-1}-\hat{\boldsymbol{H}}_{iT}^{-1}\right)\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i,\boldsymbol{z}_{it})\right\|^{2+\delta}\right]\\
&\leq p^{\frac{2+\delta}{2}}\,\mathbb{E}\left[\left\|\left(\boldsymbol{H}_i^{-1}-\hat{\boldsymbol{H}}_{iT}^{-1}\right)\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i,\boldsymbol{z}_{it})\right\|_\infty^{2+\delta}\right]\\
&\leq p^{\frac{2+\delta}{2}}\,\mathbb{E}\left[\left\|\boldsymbol{H}_i^{-1}-\hat{\boldsymbol{H}}_{iT}^{-1}\right\|_\infty^{2+\delta}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i,\boldsymbol{z}_{it})\right\|_\infty^{2+\delta}\right]\\
&\leq p^{\frac{2+\delta}{2}}\left(\mathbb{E}\left[\left\|\boldsymbol{H}_i^{-1}-\hat{\boldsymbol{H}}_{iT}^{-1}\right\|_\infty^{\frac{2(2+\delta)(1+\delta)}{\delta}}\right]\right)^{\frac{\delta}{2(1+\delta)}}\left(\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i,\boldsymbol{z}_{it})\right\|_\infty^{2(1+\delta)}\right]\right)^{\frac{1+\delta/2}{1+\delta}}\\
&\leq p^{\frac{2+\delta}{2}}\left(p^{\frac{(2+\delta)(1+\delta)}{\delta}}\underline{\lambda}_{\boldsymbol{H}}^{-\frac{2(2+\delta)(1+\delta)}{\delta}}C_{\nabla^2 m}\right)^{\frac{\delta}{2(1+\delta)}}\left(\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\nabla m(\boldsymbol{\theta}_i,\boldsymbol{z}_{it})\right\|^{2(1+\delta)}\right]\right)^{\frac{1+\delta/2}{1+\delta}}\\
&\leq p^{\frac{2+\delta}{2}}\left(p^{\frac{(2+\delta)(1+\delta)}{\delta}}\underline{\lambda}_{\boldsymbol{H}}^{-\frac{2(2+\delta)(1+\delta)}{\delta}}C_{\nabla^2 m}\right)^{\frac{\delta}{2(1+\delta)}}C_{\nabla\mu}^{\frac{1+\delta/2}{1+\delta}},
\end{aligned}
$$

where the second inequality follows from $\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty$; the third inequality from Hölder's inequality applied with $p = (1+\delta)/(1+\delta/2) > 1$; the fourth inequality from lemma A.1.2, and the last line follows by assumption A.3. Finally, we conclude that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\right\|^{2+\delta}\right]\\
&\leq 2^{1+\delta}\left[\underline{\lambda}_{\boldsymbol{H}}^{-2-\delta}C_{\nabla m}^{\frac{2+\delta}{2(1+\delta)}}+p^{\frac{2+\delta}{2}}\left(p^{\frac{(2+\delta)(1+\delta)}{\delta}}\underline{\lambda}_{\boldsymbol{H}}^{-\frac{2(2+\delta)(1+\delta)}{\delta}}C_{\nabla^2 m}\right)^{\frac{\delta}{2(1+\delta)}}C_{\nabla\mu}^{\frac{1+\delta/2}{1+\delta}}\right]\\
&\equiv C_{\hat{\boldsymbol{\theta}},2+\delta}\ ,
\end{aligned}
$$

where we note that $C_{\hat{\boldsymbol{\theta}},2+\delta}$ does not depend on i or T. By Jensen's inequality we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\right)\right\|^{2}\right] &\leq C_{\hat{\boldsymbol{\theta}},2+\delta}^{\frac{2}{2+\delta}} \equiv C_{\hat{\boldsymbol{\theta}},2},\\
\mathbb{E}\left[\left\|\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\right)\right\|^{1+\delta/2}\right] &\leq C_{\hat{\boldsymbol{\theta}},2+\delta}^{\frac{1}{2}} \equiv C_{\hat{\boldsymbol{\theta}},1+\delta/2},
\end{aligned}
$$

$$\mathbb{E}\left[\left\|\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\right)\right\|\right] \leq C_{\hat{\boldsymbol{\theta}},2+\delta}^{\frac{1}{2+\delta}} \equiv C_{\hat{\boldsymbol{\theta}},1} \ ,$$

which establishes the first part of the claim.

Next we note that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)\right\|^2\right] = \mathbb{E}\left[T(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)'(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)\right] \\
&\leq \mathbb{E}\left[\left\|\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\right\|^2\right] + 2\left|\mathbb{E}\left[T(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)'(\boldsymbol{\theta}_i-\boldsymbol{\theta}_1)\right]\right| + T\left(\boldsymbol{\theta}_i-\boldsymbol{\theta}_1\right)'\left(\boldsymbol{\theta}_i-\boldsymbol{\theta}_1\right) \\
&\leq \mathbb{E}\left[\left\|\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\right\|^2\right] + 2\left\|\boldsymbol{\eta}_i-\boldsymbol{\eta}_1\right\|\mathbb{E}\left[\left\|\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\right\|\right] + \left\|\boldsymbol{\eta}_i-\boldsymbol{\eta}_1\right\|^2 \\
&\leq C_{\hat{\boldsymbol{\theta}},2} + 2C_{\hat{\boldsymbol{\theta}},1}\left\|\boldsymbol{\eta}_i-\boldsymbol{\eta}_1\right\| + \left\|\boldsymbol{\eta}_i-\boldsymbol{\eta}_1\right\|^2 \ ,
\end{aligned}$$

where in the first inequality we add and subtract $\boldsymbol{\theta}_i$ in both parentheses, in the third inequality we apply the Cauchy-Schwarz inequality to the cross term and observe that under A.1 $\sqrt{T}(\boldsymbol{\theta}_i-\boldsymbol{\theta}_1)=\boldsymbol{\eta}_i-\boldsymbol{\eta}_1$. This establishes the second part of the claim. □

Lemma A.2.2. *Suppose A.3 and A.5 are satisfied. Let δ be as in assumption A.3. Then for all i and $T > T_0$ it holds that*

$$\begin{aligned}
\mathbb{E}\left[\left|\mu(\hat{\boldsymbol{\theta}}_i)\right|^{2+\delta}\right] &< \infty \\
\mathbb{E}\left[\left|\sqrt{T}(\mu(\hat{\boldsymbol{\theta}}_i)-\mu(\boldsymbol{\theta}_i))\right|^{2+\delta}\right] &\leq C_{\nabla\mu}^{2+\delta} C_{\hat{\boldsymbol{\theta}},2+\delta}
\end{aligned}$$

Proof. Equation (A.1.1) in lemma A.1.1 implies $\mu(\hat{\boldsymbol{\theta}}_i) = \mu(\boldsymbol{\theta}_i) + \bar{\boldsymbol{d}}_i'(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)$, where $\bar{\boldsymbol{d}}_i = \nabla\mu\left(\bar{\boldsymbol{\theta}}_i\right)$ for $\bar{\boldsymbol{\theta}}_i$ on the segment joining $\boldsymbol{\theta}_i$ and $\hat{\boldsymbol{\theta}}_i$. Raising both sides to the power of $(2+\delta)$ and applying the C_r inequality we obtain that

$$\left|\mu(\hat{\boldsymbol{\theta}}_i)\right|^{2+\delta} \leq 2^{1+\delta}\left[\left|\mu(\boldsymbol{\theta}_i)\right|^{2+\delta} + \left|\bar{\boldsymbol{d}}_i'(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\right|^{2+\delta}\right].$$

By assumption A.5 and the Cauchy-Schwarz inequality it holds that $\left|\bar{\boldsymbol{d}}_i'(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\right|^{2+\delta} \leq$

$\left\|\bar{\boldsymbol{d}}_1\right\|^{2+\delta}\left\|\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\right\|^{2+\delta} \leq C_{\nabla\mu}^{2+\delta}\left\|\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\right\|^{2+\delta}$, hence by lemma A.2.1 it follows that

$$\mathbb{E}\left[\left|\bar{\boldsymbol{d}}_i'(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\right|^{2+\delta}\right] \leq \frac{C_{\nabla\mu}^{2+\delta}C_{\hat{\boldsymbol{\theta}},2+\delta}}{T^{(2+\delta)/2}},$$

where the constants are independent on i. Then both claims of the lemma follow. □

We need an extension of a weighted law of large numbers due to Pruitt (1966).

Lemma A.2.3. *Suppose*

(i) $X_1, X_2, \ldots$ *is a sequence of i.i.d. random variables such that* $\mathbb{E}(X_1) = 0$ *and* $\mathbb{E}|X_1|^{1+1/\gamma} < \infty$ *for some* $\gamma \in (0, 1]$*;*

(ii) $\{\boldsymbol{w}_N\}_N$ *with* $\boldsymbol{w}_N \in \mathbb{R}^\infty$ *is a sequence of weight vectors such that* $w_{i\,N} \geq 0$ *for* $i > 0$, $\sum_{i=1}^N w_{i\,N} \leq 1$, *and* $w_{j\,N} = 0$ *for* $j > N$*;*

(iii) $\boldsymbol{w} \in \mathbb{R}^\infty$ *is a weight vector such that* $w_i \geq 0$ *for* $i > 0$, $\sum_{i=1}^\infty w_i \leq 1$*; and*

(iv) $\{\boldsymbol{w}_N\}$ *and* $\boldsymbol{w}$ *are such that* $\sup_i|w_{i\,N} - w_i| = O(N^{-\gamma})$.

Then it $\sum_{i=1}^\infty w_i X_i$ *exists a.s. and* $\sum_{i=1}^N w_{i\,N}X_i \xrightarrow{a.s.} \sum_{i=1}^\infty w_i X_i$.

Observe that the limit sequence of weights can be defective. If $w_{i\,N} = N^{-1}\,\mathbb{I}_{i\leq N}$ (equal weights), the above result becomes a standard SLLN with a second moment assumption.

Proof. Define $\tilde{\boldsymbol{w}}_N \in \mathbb{R}^\infty$ by $\tilde{w}_{i\,N} = w_{i\,N} - w_i$ for $i \leq N$ and $\tilde{w}_{i\,N} = 0$ for $i > N$. Then

$$\sum_{i=1}^N w_{i\,N}X_i = \sum_{i=1}^N w_i X_i + \sum_{i=1}^N (w_{i\,N} - w_i)X_i = \sum_{i=1}^N w_i X_i + \sum_{i=1}^N \tilde{w}_{i\,N}X_i$$

holds. For any n it holds that $\sum_{i=1}^n \operatorname{Var}(w_i X_i) = \sum_{i=1}^n w_i^2\,\mathbb{E}(X_i^2) = \mathbb{E}(X_i^2)\sum_{i=1}^n w_i^2 \leq \mathbb{E}(X_i^2) < \infty$ since $\gamma \leq 1$. Hence the Kolmogorov two-series theorem (Kallenberg, 2021, lemma 5.16) implies that $\sum_{i=1}^N w_i X_i \xrightarrow{a.s.} \sum_{i=1}^\infty w_i X_i$. The vector $\tilde{\boldsymbol{w}}_N$ satisfies the conditions of theorem 2 of Pruitt (1966) (observe that the condition (1.2) in Pruitt (1966) is not required by the assumption of $\mathbb{E}(X) = 0$ and the remark following (1.3)). Hence the same theorem implies that $\sum_{i=1}^\infty \tilde{w}_{i\,N}X_i \xrightarrow{a.s.} 0$. The claim of the lemma then follows. □

Lemma A.2.4. *Suppose that the assumptions of theorem 1 are satisfied. Then (i)* $\sum_{i=1}^{\infty} w_i \boldsymbol{\eta}_i$ *exists* $\boldsymbol{\eta}$*-a.s. and it holds that*

$$\sum_{i=1}^{N} w_{i\,N}(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) \xrightarrow{a.s.} \sum_{i=1}^{\infty} w_i \boldsymbol{\eta}_i - \boldsymbol{\eta}_1 \ ,$$

and (ii) $\sup_N \sum_{i=1}^{N} w_{i\,N} \left\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\right\|^k < \infty$ *is finite* $\boldsymbol{\eta}$*-a.s. for* $k = 1, 1+\delta/2, 2, 2+\delta$ *for the choice of* δ *in A.3.*

Proof. Notice that $\sum_{i=1}^{N} w_{i\,N}(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) = \sum_{i=1}^{N} w_{i\,N}\boldsymbol{\eta}_i - \boldsymbol{\eta}_1$. By assumption A.1 $\boldsymbol{\eta}_i$ are i.i.d. random vectors with finite third moments and $\sup_i |w_{i\,N} - w_i| = O(N^{-1/2})$. Lemma A.2.3 then implies that $\sum_{i=1}^{\infty} w_i \boldsymbol{\eta}_i$ exists $\boldsymbol{\eta}$-a.s. and that $\sum_{i=1}^{N} w_{i\,N}\boldsymbol{\eta}_i \xrightarrow{a.s.} \sum_{i=1}^{\infty} w_i \boldsymbol{\eta}_i$, which establishes the first claim.

Consider $\left\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\right\|^k$ and note that the triangle and C_r inequalities imply that

$$\left\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\right\|^k \leq (\left\|\boldsymbol{\eta}_i\right\| + \left\|\boldsymbol{\eta}_1\right\|)^k \leq 2^{k-1}(\left\|\boldsymbol{\eta}_k\right\|^k + \left\|\boldsymbol{\eta}_1\right\|^k) \ ,$$

which, in turn, implies

$$\sum_{i=1}^{N} w_{i\,N} \left\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\right\|^k \leq 2^{k-1} \sum_{i=1}^{N} w_{i\,N} \left\|\boldsymbol{\eta}_i\right\|^k + 2^{k-1} \left\|\boldsymbol{\eta}_1\right\|^k \ . \tag{A.2.1}$$

Observe that $\left\|\boldsymbol{\eta}_i\right\|^k$ are i.i.d. random variables with $\mathbb{E}_{\boldsymbol{\eta}}\left[\left\|\boldsymbol{\eta}_i\right\|^{3k}\right] < \infty$ for $k \in [1, 2+\delta]$ by A.1. Then lemma A.2.3 applies with $\gamma = 1/2$, and $\sum_{i=1}^{N} w_{i\,N} \left\|\boldsymbol{\eta}_i\right\|^k$ converges almost surely, which implies that $\sup_N \sum_{i=1}^{N} w_{i\,N} \left\|\boldsymbol{\eta}_i\right\|^k < \infty$ $\boldsymbol{\eta}$-a.s.. Since $\left\|\boldsymbol{\eta}_1\right\|$ is also $\boldsymbol{\eta}$-a.s. finite, together with eq. (A.2.1), this implies the second claim. □

Finally, we present the proof of theorem 1.

Proof of theorem 1. First, from lemma A.2.2 it follows for each N and $T > T_0$

$$\mathbb{E}\left[\hat{\mu}(\boldsymbol{w}_N) - \mu(\boldsymbol{\theta}_1)\right]^2 < \infty \ ,$$

establishing the second assertion of the theorem.

The MSE of the averaging estimator expressed as a sum of squared bias and variance is

$$T\times\mathbb{E}\left[\hat{\mu}(\boldsymbol{w}_N)-\mu(\boldsymbol{\theta}_1)\right]^2=\left(\sum_{i=1}^{N}w_{i\,N}\,\mathbb{E}\left(\sqrt{T}(\mu(\hat{\boldsymbol{\theta}}_i)-\mu(\boldsymbol{\theta}_1))\right)\right)^2+T\,\mathrm{Var}\left(\sum_{i=1}^{N}w_{i\,N}(\mu(\hat{\boldsymbol{\theta}}_i))\right).$$

We examine the bias and the variance separately. We first focus on the bias. By eq. (A.1.2) of lemma A.1.1, we have

$$\mu(\hat{\boldsymbol{\theta}}_i)=\mu(\boldsymbol{\theta}_1)+\boldsymbol{d}_1'\left(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1\right)+\frac{1}{2}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1),\tag{A.2.2}$$

where $\boldsymbol{d}_1=\nabla\mu(\boldsymbol{\theta}_1)$ and $\acute{\boldsymbol{\theta}}_i$ lies on the segment joining $\hat{\boldsymbol{\theta}}_i$ and $\boldsymbol{\theta}_1$. The bias of $\mu(\hat{\boldsymbol{\theta}}_i)$ is

$$\begin{aligned}
&\sqrt{T}\,\mathbb{E}\left(\mu(\hat{\boldsymbol{\theta}}_i)-\mu(\boldsymbol{\theta}_1)\right)\\
&=\mathbb{E}\left[\boldsymbol{d}_1'\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)+\frac{1}{2}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)\right]\\
&=\mathbb{E}\left[\boldsymbol{d}_1'\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)+\frac{1}{2}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)\right]\\
&\quad+\sqrt{T}\boldsymbol{d}_0'(\boldsymbol{\theta}_i-\boldsymbol{\theta}_1)+(\boldsymbol{d}_1-\boldsymbol{d}_0)'\sqrt{T}(\boldsymbol{\theta}_i-\boldsymbol{\theta}_1)\\
&=\mathbb{E}\left[\boldsymbol{d}_1'\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)+\frac{1}{2}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)\right]\\
&\quad+\boldsymbol{d}_0'(\boldsymbol{\eta}_i-\boldsymbol{\eta}_1)+(\boldsymbol{d}_1-\boldsymbol{d}_0)'(\boldsymbol{\eta}_i-\boldsymbol{\eta}_1)\,,
\end{aligned}\tag{A.2.3}$$

where in the first equality we use eq. (A.2.2); in the second equality $\boldsymbol{\theta}_1$ is replaced by $\boldsymbol{\theta}_i$ in the first term using $\boldsymbol{d}_1'\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)-\boldsymbol{d}_1'(\boldsymbol{\eta}_i-\boldsymbol{\eta}_1)=\boldsymbol{d}_1'\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)$; $\boldsymbol{d}_0=\nabla\mu(\boldsymbol{\theta}_0)$; and we use the locality assumption A.1 in the last equality as $\sqrt{T}(\boldsymbol{\theta}_i-\boldsymbol{\theta}_1)=\boldsymbol{\eta}_1-\boldsymbol{\eta}_1$. Define

$$A_{i\,T}\equiv\mathbb{E}\left[\boldsymbol{d}_1'\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\right)\right]+\frac{1}{2}\,\mathbb{E}\left[(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)\right]+(\boldsymbol{d}_1-\boldsymbol{d}_0)'(\boldsymbol{\eta}_i-\boldsymbol{\eta}_1)\,,$$

and note that by eq. (A.2.3), the bias of the averaging estimator can be written as

$$\sum_{i=1}^{N} w_{i\,N}\,\mathbb{E}\left(\sqrt{T}(\mu(\hat{\boldsymbol{\theta}}_i)-\mu(\boldsymbol{\theta}_1))\right)=\sum_{i=1}^{N} w_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i-\boldsymbol{\eta}_1)+\sum_{i=1}^{N} w_{i\,N}A_{i\,T}\ . \quad \text{(A.2.4)}$$

We then proceed by showing that $\left|\sum_{i=1}^{N} w_{i\,N}A_{i\,T}\right| \leq M/\sqrt{T} \to 0$ for some constant $M<\infty$ independent of N.[22] Note that

1. By Hölder's inequality, we obtain $\left|\boldsymbol{d}_1'\,\mathbb{E}\left(\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\right)\right| \leq \|\boldsymbol{d}_1\|_\infty \left\|\sqrt{T}\,\mathbb{E}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\right\|_1$ $\leq C_{\nabla\mu}C_{Bias}T^{-1/2}$, where the last bound follows from assumptions A.4 and A.5;

2. By assumption A.5 the eigenvalues of $\nabla^2\mu$ are bounded in absolute value by $C_{\nabla^2\mu}$. Then $\left|\mathbb{E}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_1)\right| \leq C_{\nabla^2\mu}T^{-1/2}\Big[C_{\hat{\boldsymbol{\theta}},2}+2C_{\hat{\boldsymbol{\theta}},1}\,\|\boldsymbol{\eta}_i-\boldsymbol{\eta}_1\|+$ $\|\boldsymbol{\eta}_i-\boldsymbol{\eta}_1\|^2\Big]$ where the bound is given by lemma A.2.1;

3. By assumption A.5, $\|\boldsymbol{d}_1-\boldsymbol{d}_0\| \equiv \left\|\nabla\mu(\boldsymbol{\theta}_0+T^{-1/2}\boldsymbol{\eta}_1)-\nabla\mu(\boldsymbol{\theta}_0)\right\| \leq C_{\nabla^2\mu}\,\|\boldsymbol{\eta}_1\|\,T^{-1/2}$.

All the $C_\cdot$-constants do not depend in i. Combining the above results, we obtain by the triangle and Cauchy-Scwharz inequalities that

$$|A_{i\,T}| \leq \frac{1}{\sqrt{T}}\left[C_{\nabla\mu}C_{Bias}+C_{\nabla^2\mu}C_{\hat{\boldsymbol{\theta}},2}+C_{\nabla^2\mu}\,\|\boldsymbol{\eta}_i-\boldsymbol{\eta}_1\|^2+C_{\nabla^2\mu}(2C_{\hat{\boldsymbol{\theta}},1}+\|\boldsymbol{\eta}_1\|)\,\|\boldsymbol{\eta}_i-\boldsymbol{\eta}_1\|\right].$$

Define

$$\begin{aligned} M = {} & C_{\nabla\mu}C_{Bias}+C_{\nabla^2\mu}C_{\hat{\boldsymbol{\theta}},2}+C_{\nabla^2\mu}\sup_N\sum_{i=1}^{N} w_{i\,N}\,\|\boldsymbol{\eta}_i-\boldsymbol{\eta}_1\|^2 \\ & +C_{\nabla^2\mu}\left(2C_{\hat{\boldsymbol{\theta}},1}+\|\boldsymbol{\eta}_1\|\right)\sup_N\sum_{i=1}^{N} w_{i\,N}\,\|\boldsymbol{\eta}_i-\boldsymbol{\eta}_1\|\ , \end{aligned}$$

and observe that M does not depend on N or T, and by lemma A.2.4 $M<\infty$ ($\boldsymbol{\eta}$-a.s.).

[22] Recall that all statements are almost surely with respect to the distribution of $\boldsymbol{\eta}$ in line with assumption A.1. M depends on the sequence of $\{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2,\ldots\}$ only, and the sequence is held fixed.

Take the weighted average of A_{iT} to obtain

$$\left|\sum_{i=1}^{N} w_{iN} A_{iT}\right| \leq \sum_{i=1}^{N} w_{iN}|A_{iT}| \leq \frac{M}{\sqrt{T}} \to 0 \text{ as } N, T \to \infty \,. \tag{A.2.5}$$

By lemma A.2.4, $\sum_{i=1}^{N} w_{iN} \boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) \to \sum_{i=1}^{\infty} w_i \boldsymbol{d}_0'\boldsymbol{\eta}_0 - \boldsymbol{d}_0'\boldsymbol{\eta}_i$, where the infinite sum exists. Combining this with eqs. (A.2.4) and (A.2.5), we obtain that the bias converges as $N, T \to \infty$:

$$\sum_{i=1}^{N} w_{iN}\, \mathbb{E}\left(\sqrt{T}\left(\mu(\hat{\boldsymbol{\theta}}_i) - \mu(\boldsymbol{\theta}_1)\right)\right) \to \sum_{i=1}^{\infty} w_i \boldsymbol{d}_0'\boldsymbol{\eta}_0 - \boldsymbol{d}_0'\boldsymbol{\eta}_i, \quad (\boldsymbol{\eta}\text{-a.s.}) \tag{A.2.6}$$

Now turn to the variance series and observe that

$$\begin{aligned}
&T \times \operatorname{Var}\left(\sum_{i=1}^{N} w_{iN}(\mu(\hat{\boldsymbol{\theta}}_i))\right) \\
&= T \sum_{i=1}^{N} w_{iN}^2 \operatorname{Var}\left(\mu(\hat{\boldsymbol{\theta}}_i)\right) \\
&= \sum_{i=1}^{N} w_{iN}^2 \left[\mathbb{E}\left[\sqrt{T}\left(\mu(\hat{\boldsymbol{\theta}}_i) - \mu(\boldsymbol{\theta}_i)\right)\right]^2 - \left[\sqrt{T}\left(\mathbb{E}\left(\mu(\hat{\boldsymbol{\theta}}_i)\right) - \mu(\boldsymbol{\theta}_i)\right)\right]^2\right] \,.
\end{aligned}$$

We tackle the two sums separately. First we show that

$$\sup_N \sum_{i=1}^{N} w_{iN}^2 \left[\sqrt{T}\left(\mu(\boldsymbol{\theta}_i) - \mathbb{E}\left(\mu(\hat{\boldsymbol{\theta}}_i)\right)\right)\right]^2 = O(T^{-1})$$

The argument is similar to that leading up to eq. (A.2.5). By eq. (A.1.3) of lemma A.1.1, we can expand $\mu(\hat{\boldsymbol{\theta}}_i)$ around $\boldsymbol{\theta}_i$ to obtain that

$$\sqrt{T}\left(\mathbb{E}\left(\mu(\hat{\boldsymbol{\theta}}_i)\right) - \mu(\boldsymbol{\theta}_i)\right) = \mathbb{E}\left[\boldsymbol{d}_1'\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)'\nabla^2\mu(\check{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\right] \,,$$

for some $\check{\boldsymbol{\theta}}_i$ on the segment joining $\boldsymbol{\theta}_i$ and $\hat{\boldsymbol{\theta}}_i$. Similarly to the above, we conclude by

lemma A.2.1 and assumption A.4 that

$$\left|\mathbb{E}\left[\boldsymbol{d}_1'\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\right)\right]\right| \leq \frac{C_{\nabla\mu}C_{Bias}}{\sqrt{T}}$$
$$\left|\mathbb{E}\left[(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)'\nabla^2\mu(\check{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\right]\right| \leq \frac{C_{\nabla^2\mu}C_{\hat{\boldsymbol{\theta}},2}}{\sqrt{T}}\ .$$

From this it immediately follows that

$$\sum_{i=1}^{N} w_{iN}^2\left[\sqrt{T}\left(\mathbb{E}\left(\mu(\hat{\boldsymbol{\theta}}_i)\right)-\mu(\boldsymbol{\theta}_i)\right)\right]^2 \leq \frac{1}{T}\left[C_{\nabla\mu}C_{Bias}+C_{\nabla^2\mu}C_{\hat{\boldsymbol{\theta}},2}\right]^2\ , \tag{A.2.7}$$

where the right hand side does not depend on i or N.

Second, we show that

$$\sum_{i=1}^{N} w_{iN}^2\,\mathbb{E}\left[\sqrt{T}\left(\mu(\hat{\boldsymbol{\theta}}_i)-\mu(\boldsymbol{\theta}_i)\right)\right]^2 \to \sum_{i=1}^{\infty} w_i^2\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0.$$

Define $X_{iT}=\mathbb{E}\left[\sqrt{T}(\mu(\hat{\boldsymbol{\theta}}_i)-\mu(\boldsymbol{\theta}_i))\right]^2$. By lemma A.2.2 there exists a constant $C_X<\infty$ that does not depend on i or T such that $X_{iT}\leq C_X$ for $T>T_0$. Then

$$\begin{aligned}
&\sum_{i=1}^{N} w_{iN}^2\,\mathbb{E}\left[\sqrt{T}\left(\mu(\hat{\boldsymbol{\theta}}_i)-\mu(\boldsymbol{\theta}_i)\right)\right]^2\\
&\equiv \sum_{i=1}^{N} w_{iN}^2 X_{iT}\\
&= \sum_{i=1}^{N} w_i^2\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0+\sum_{i=1}^{N}(w_{iN}^2-w_i^2)\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0+\sum_{i=1}^{N}(w_{iN}^2-w_i^2)(X_{iT}-\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0)\\
&\quad+\sum_{i=1}^{N} w_i^2(X_{iT}-\boldsymbol{d}_0'\boldsymbol{V}_0\boldsymbol{d}_0).
\end{aligned}$$

We deal with the four sums separately:

1. By assumption A.3, $\sum_{i=1}^{N} w_i^2\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0 \leq \bar{\lambda}_{\boldsymbol{\Sigma}}\lambda_{\boldsymbol{H}}^2\left\|\boldsymbol{d}_0\right\|^2$, so $\sum_{i=1}^{N} w_i^2\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0$ forms a bounded non-decreasing sequence. Thus $\sum_{i=1}^{N} w_i^2\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0 \to \sum_{i=1}^{\infty} w_i^2\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0$.

2. Consider $\sum_{i=1}^{N}(w_{i\,N}^2 - w_i^2)\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0$

$$\begin{aligned}\left|\sum_{i=1}^{N}(w_{i\,N}^2 - w_i^2)\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0\right| &= \left|\sum_{i=1}^{N}(w_{i\,N} - w_i)(w_{i\,N} + w_i)\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0\right| \\ &\leq \sup_j|w_{j\,N} - w_j|\sum_{i=1}^{N}(w_{i\,N} + w_i)\boldsymbol{d}_0\boldsymbol{V}_i\boldsymbol{d}_0 \\ &\leq 2\bar{\lambda}_{\boldsymbol{\Sigma}}\underline{\lambda}_{\boldsymbol{H}}^2\left\|\boldsymbol{d}_0\right\|^2\sup_j|w_{j\,N} - w_j| \to 0\ ,\end{aligned}$$

where we have used A.3.

3. Similarly we obtain that

$$\begin{aligned}\left|\sum_{i=1}^{N}(w_{i\,N}^2 - w_i^2)(X_{i\,T} - \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0)\right| &= \left|\sum_{i=1}^{N}(w_{i\,N} - w_i)(w_{i\,N} + w_i)(X_{i\,T} - \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0)\right| \\ &\leq \sup_j|w_{j\,N} - w_j|\sum_{i=1}^{N}(w_{i\,N} + w_i)|X_{i\,T} - \boldsymbol{d}_0\boldsymbol{V}_i\boldsymbol{d}_0| \\ &\leq 2\left[\bar{\lambda}_{\boldsymbol{\Sigma}}\underline{\lambda}_{\boldsymbol{H}}^2\left\|\boldsymbol{d}_0\right\|^2 + C_X\right]\sup_j|w_{j\,N} - w_j| \to 0\ .\end{aligned}$$

4. Last, we apply the dominated convergence theorem to show that $\sum_{i=1}^{N} w_i^2(X_{i\,T} - \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0) \to 0$.

   Define $f_{N,T} : \mathbb{N} \to \mathbb{R}$ as $f_{N,T}(i) = w_{i\,N}^2(X_{i\,T} - \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0)$ if $i \leq N$ and $f_{N,T}(i) = 0$ if $i > N$. For each i, $\left\{\sqrt{T}(\mu(\hat{\boldsymbol{\theta}}_i) - \boldsymbol{\theta}_i), T = T_0 + 1, \ldots\right\}$ form a family with uniformly bounded $(2+\delta)$th moments (by lemma A.2.2). By lemma 1 $\sqrt{T}(\mu(\hat{\boldsymbol{\theta}}_i) - \boldsymbol{\theta}_i) \Rightarrow N(0, \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0)$, hence by Vitali's convergence theorem the second moments converge as $X_{i\,T} \to \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0$. This convergence is equivalent to the observation that for each i $f_{N,T}(i)$ converges to zero as $N, T \to \infty$.

   Next, $f_{N,T}$ is dominated: for any i it holds that $|f_{N,T}(i)| \leq w_i^2|X_{i\,T} - \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0| \leq w_i(C_X + \bar{\lambda}_{\boldsymbol{\Sigma}}\underline{\lambda}_{\boldsymbol{H}}^2\left\|\boldsymbol{d}\right\|_0^2)$. The bound is summable: $\sum_{i=1}^{\infty} w_i(C_X + \bar{\lambda}_{\boldsymbol{\Sigma}}\underline{\lambda}_{\boldsymbol{H}}^2\left\|\boldsymbol{d}\right\|_0^2) \leq (C_X + \bar{\lambda}_{\boldsymbol{\Sigma}}\underline{\lambda}_{\boldsymbol{H}}^2\left\|\boldsymbol{d}\right\|_0^2)$, which is independent of N and T.

The dominated convergence theorem applies and so

$$\sum_{i=1}^{N} w_i^2 (X_{iT} - \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0) = \sum_{i=1}^{\infty} f_{N,T}(i) \to \sum_{i=1}^{\infty} 0 = 0 \text{ as } N, T \to \infty.$$

Combining the above arguments, we obtain that as $N, T \to \infty$

$$\sum_{i=1}^{N} w_{iN}^2 \, \mathbb{E}\left[\sqrt{T}\left(\mu(\hat{\boldsymbol{\theta}}_i) - \mu(\boldsymbol{\theta}_i)\right)\right]^2 \to \sum_{i=1}^{\infty} w_i^2 \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0 \ . \tag{A.2.8}$$

Combining together equations (A.2.6), (A.2.7), and (A.2.8) shows that as $N, T \to \infty$

$$T \times \mathbb{E}\left[\hat{\mu}(\boldsymbol{w}_N) - \mu(\boldsymbol{\theta}_1)\right]^2 \to \left(\sum_{i=1}^{\infty} w_i \boldsymbol{d}_0' \boldsymbol{\eta}_i - \boldsymbol{d}_0' \boldsymbol{\eta}_1\right)^2 + \sum_{i=1}^{\infty} w_i^2 \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0 \ .$$

□

A.3 Proof of Theorem 2

Before presenting the proof of theorem 2, we introduce a number of intermediate results.

We first give a straightforward modification of theorem 1 in Phillips and Moon (1999), which allows us to replace sequential convergence (first taking limits as $T \to \infty$, then as $N \to \infty$) by joint convergence ($N, T \to \infty$ jointly).

Lemma A.3.1. *Let Y_{iT} be random variables indexed by $i = 1, \ldots, N$ and $T = 1,2, \ldots,$. Suppose Y_{iT} are independent over i and that*

(i) $Y_{iT} \Rightarrow \Lambda_i$ *as* $T \to \infty$,

(ii) $\sum_{i=1}^{N} w_{iN} \Lambda_i \Rightarrow X$ *as* $N \to \infty$,

(iii) $\limsup_{N,T\to\infty} \sum_{i=1}^{N} w_{iN} |\mathbb{E}(Y_{iT}) - \mathbb{E}(\Lambda_i)| = 0$,

(iv) $\limsup_{N,T\to\infty} \sum_{i=1}^{N} \mathbb{E}|w_{iN} Y_{iT}| < \infty$,

(v) $\limsup_{N\to\infty} \sum_{i=1}^{N} \mathbb{E}\left[w_{iN} |\Lambda_i| \mathbb{I}_{|w_{iN}\Lambda_i| > \varepsilon}\right] = 0$ *for any* $\varepsilon > 0$, *and*

(vi) $\limsup_{N,T\to\infty} \sum_{i=1}^{N} \mathbb{E}\left[w_{iN} |Y_{iT}| \mathbb{I}_{|w_{iN}Y_{iT}| > \varepsilon}\right] = 0$ *for any* $\varepsilon > 0$.

Then as $N, T \to \infty$

$$\sum_{i=1}^{N} w_{i\,N} Y_{i\,T} \Rightarrow X.$$

In particular, if as $N \to \infty$ *it holds that* $\sum_{i=1}^{N} w_{i\,N}\Lambda_i \xrightarrow{p} A$ *for* A *non-random, then as* $N, T \to \infty$ *it holds that* $\sum_{i=1}^{N} w_{i\,N} Y_{i\,T} \xrightarrow{p} A$.

Proof. The proof is close to that of theorem 1 in Phillips and Moon (1999). The key modification consists in replacing $n^{-1}\zeta_{k,n,T}$ (in their notation) by

$$W_{k\,N\,T} = \sum_{1 \le i < k} w_{i\,N} Y_{i\,T} + \sum_{k < i \le N} w_{i\,N}\Lambda_i$$

and every factor $1/n$ by the appropriate weight $w_{i\,N}$. As in their theorem 1, this establishes condition (3.9) of Phillips and Moon (1999): for all bounded continuous f

$$\limsup_{N,T\to\infty} \left| \mathbb{E}\left(f\left(\sum_{i=1}^{N} w_{i\,N} Y_{i\,T} \right) \right) - \mathbb{E}\left(f\left(\sum_{i=1}^{N} w_i \Lambda_i \right) \right) \right| = 0$$

By lemma 6 in Phillips and Moon (1999), this implies the result of the theorem. □

To apply lemma A.3.1, for the remainder of the section define

$$Y_{i\,T} = \sqrt{T}(\mu(\hat{\boldsymbol{\theta}}_i) - \mu(\boldsymbol{\theta}_1)), \tag{A.3.1}$$

and note that $Y_{i\,T} \Rightarrow \Lambda_i$ as $T \to \infty$, where $\Lambda_i \sim N\left(\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1), \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0\right)$ is the random variable that appears on the right hand side in lemma 1. As before, let $\boldsymbol{d}_1 = \nabla\mu(\boldsymbol{\theta}_1)$, $\boldsymbol{d}_0 = \nabla\mu(\boldsymbol{\theta}_0)$.

Lemma A.3.2. *Let* $Y_{i\,T}$ *be defined as in eq.* (A.3.1)*. Under assumptions of theorem 2*

$$\sum_{i=1}^{\bar{N}} w_{i\,N} Y_{i\,T} \Rightarrow \sum_{i=1}^{\bar{N}} w_i \Lambda_i \text{ as } N, T \to \infty.$$

Proof. Note that randomness enters only the T dimension here. As $\{Y_{i\,T}\}_{i=1}^{\bar{N}} \Rightarrow \{\Lambda_i\}_{i=1}^{\bar{N}}$

as $N, T \to \infty$ (N does not matter), and as $N, T \to \infty$ $\{w_{i\,N}\}_{i=1}^{\bar{N}} \to \{w_i\}_{i=1}^{\bar{N}}$ as $N, T \to \infty$. Slutsky's theorem gives the result. □

Recall that under assumption (ii) of theorem 2 it holds that

$$\sup_{i>\bar{N}} w_{i\,N} = o(N^{-\frac{1}{2}}) \ .$$

Lemmas A.3.3-A.3.7 verify conditions (ii)-(vi) of lemma A.3.1 for $\sum_{i=\bar{N}+1}^{N} w_{i\,N} Y_{i\,T}$, $N > \bar{N}$.

Lemma A.3.3. *Let $Y_{i\,T}$ be defined as in eq. (A.3.1). Under assumptions of theorem 2*

$$\sum_{i=\bar{N}+1}^{N} w_{i\,N}\Lambda_i \xrightarrow{p} -\left(1 - \sum_{i=1}^{\bar{N}} w_i\right) \boldsymbol{d}_0'\boldsymbol{\eta}_1 \ \textit{as} \ N \to \infty \ .$$

Proof. By the triangle inequality

$$\begin{aligned}
&\left| \sum_{i=\bar{N}+1}^{N} w_{i\,N}\Lambda_i - \left(1 - \sum_{i=1}^{\bar{N}} w_i\right)(-\boldsymbol{d}_0'\boldsymbol{\eta}_1) \right| \\
&\leq \left| \sum_{i=\bar{N}+1}^{N} w_{i\,N}\Lambda_i - \sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) \right| \\
&\quad + \left| \sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) - \left(1 - \sum_{i=1}^{\bar{N}} w_i\right)(-\boldsymbol{d}_0'\boldsymbol{\eta}_1) \right| . \qquad \text{(A.3.2)}
\end{aligned}$$

We show that both terms on the right hand side converge to zero in probability. First we show that $\left|\sum_{i=\bar{N}+1}^{N} w_{i\,N}\Lambda_i - \sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)\right| \xrightarrow{p} 0$. Consider the variance of $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\Lambda_i$:

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=\bar{N}+1}^{N} w_{i\,N}\Lambda_i\right) &= \sum_{i=\bar{N}+1}^{N} w_{i\,N}^2 \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0 \\
&\leq \left[\sup_{j>\bar{N}} w_{j\,N}\right] \sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0 \\
&\leq \bar{\lambda}_{\boldsymbol{\Sigma}}\underline{\lambda}_{\boldsymbol{H}}^2 \left\|\boldsymbol{d}_0\right\|^2 \left[\sup_{j>\bar{N}} w_{j\,N}\right] ,
\end{aligned}$$

where we used independence of Λ_i, the expressions for variance of Λ_i given in lemma 1, and the bound on variance $\boldsymbol{V}_i = \boldsymbol{H}_i^{-1}\boldsymbol{\Sigma}_i\boldsymbol{H}_i^{-1}$ implied by assumption A.3 on the bounds of eigenvalues of component variance matrices. Since $\mathbb{E}\left(\sum_{i=\bar{N}+1}^{N} w_{i\,N}\Lambda_i\right) = \sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)$, by Chebyshev's inequality and the above bound for variance, for any $\varepsilon > 0$ it holds that

$$P\left(\left|\sum_{i=\bar{N}+1}^{N} w_{i\,N}\Lambda_i - \sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)\right| > \varepsilon\right) \tag{A.3.3}$$
$$\leq \frac{\bar{\lambda}_{\boldsymbol{\Sigma}}\bar{\lambda}_{\boldsymbol{H}}^2 \left\|\boldsymbol{d}_0\right\|^2 \left[\sup_{j>\bar{N}} w_{j\,N}\right]}{\varepsilon} = o(1),$$

by assumption (iii) of theorem 2. Next we show that

$$\left|\sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) - \left(1 - \sum_{i=1}^{\bar{N}} w_i\right)(-\boldsymbol{d}_0'\boldsymbol{\eta}_1)\right| \to 0$$

by considering two cases depending on whether $\sum_{i=1}^{\bar{N}} w_i$ is equal to 1 or not.

Case I: suppose that $\sum_{i=1}^{\bar{N}} w_i \neq 1$. In this case there exist N_0, $\varepsilon_w > 0$ such that for all $N > N_0$ it holds that $\sum_{i=1}^{\bar{N}} w_{i\,N} \leq 1 - \varepsilon_w$. Note that N_0 is necessarily larger than $\bar{N}$. Define $\tilde{w}_{i\,N} = w_{i\,N}/\left(1 - \sum_{i=1}^{\bar{N}} w_{i\,N}\right)$. For $N > N_0$, $(\tilde{w}_{\bar{N}+1\,N}, \tilde{w}_{\bar{N}+2\,N}, \ldots, \tilde{w}_{N-\bar{N}\,N})$ satisfies $\tilde{w}_{i\,N} \geq 0$ and $\sum_{i=\bar{N}+1}^{N} \tilde{w}_{i\,N} = 1$. For all $N > \tilde{N}$ we have that $\tilde{w}_{i\,N} \leq \varepsilon_w^{-1} w_{i\,N}$, which implies that $\sup_{i>\bar{N}} \tilde{w}_{i\,N} \leq \varepsilon_w^{-1} \sup_{j>\bar{N}} w_{i\,N} = o(N^{-1/2})$. By lemma A.2.4 taken with $\gamma = 1/2$, we obtain that $\sum_{i=\bar{N}+1}^{N} \tilde{w}_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) = \sum_{i=\bar{N}+1}^{N} \tilde{w}_{i\,N}\boldsymbol{d}_0'\boldsymbol{\eta}_i - \boldsymbol{d}_0'\boldsymbol{\eta}_1 \to -\boldsymbol{d}_0'\boldsymbol{\eta}_1$ (a.s. with respect to the distribution of $\boldsymbol{\eta}$). The weights $\tilde{w}$ satisfy the hypothesis of lemma A.2.4 with the limit weights equal to the zero sequence as $\sup_{i>\bar{N}} \tilde{w}_{i\,N} = o(N^{-1/2})$. Since $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) = \left(1 - \sum_{i=1}^{\bar{N}} w_{i\,N}\right)\sum_{i=\bar{N}+1}^{N} \tilde{w}_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)$, we obtain that $\left|\sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) - \left(1 - \sum_{i=1}^{\bar{N}} w_i\right)(-\boldsymbol{d}_0'\boldsymbol{\eta}_1)\right| \to 0$. Together with eqs. (A.3.2) and

(A.3.3), this implies that in this case

$$\sum_{i=\bar{N}+1}^{N} w_{i\,N}\Lambda_i \xrightarrow{p} -\left(1-\sum_{i=1}^{\bar{N}} w_i\right)\boldsymbol{d}_0'\boldsymbol{\eta}_1 \ .$$

Case II: suppose that $\sum_{i=1}^{\bar{N}} w_i = 1$. We show that $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) \to 0$ $\boldsymbol{\eta}$-a.s.. First, $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'\boldsymbol{\eta}_1 = \boldsymbol{d}_0'\boldsymbol{\eta}_1 \sum_{i=\bar{N}+1}^{N} w_{i\,N} \to 0$ by the assumption that $\sum_{i=1}^{\bar{N}} w_{i\,N} \to 1$. Second, $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'\boldsymbol{\eta}_i \to \mathbb{E}_{\boldsymbol{\eta}}(\boldsymbol{d}_0'\boldsymbol{\eta}_i) = 0$ by lemma A.2.3, since $\boldsymbol{d}_0'\boldsymbol{\eta}_i$ are i.i.d. variables with finite third moments. As above, this argument and eqs. (A.3.2) and (A.3.3) imply that $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\Lambda_i \xrightarrow{p} 0$.

Combining the two cases yields the assertion. □

Lemma A.3.4. *Let $Y_{i\,T}$ be defined as in eq. (A.3.1). Under assumptions of theorem 2*

$$\limsup_{N,T\to\infty} \sum_{i=\bar{N}+1}^{N} w_{i\,N}|\mathbb{E}(Y_{i\,T}) - \mathbb{E}(\Lambda_i)| = 0 \ \textit{as}\ N, T \to \infty.$$

Proof. First, from lemma A.2.2 it follows that $\mathbb{E}|Y_{i\,T}|$ exists for all i and $T > T_0$. By lemma 1, $\mathbb{E}\,\Lambda_i = \boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)$. By eq. (A.1.2) of lemma A.1.1, we have

$$\mu(\hat{\boldsymbol{\theta}}_i) = \mu(\boldsymbol{\theta}_1) + \boldsymbol{d}_1'\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1\right) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1), \tag{A.3.4}$$

where $\boldsymbol{d}_1 = \nabla\mu(\boldsymbol{\theta}_1)$ and $\acute{\boldsymbol{\theta}}_i$ lies on the segment joining $\hat{\boldsymbol{\theta}}_i$ and $\boldsymbol{\theta}_1$. Then

$$Y_{i\,T} - \mathbb{E}(\Lambda_i) = \boldsymbol{d}_1'\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1\right) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1) - \boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1). \tag{A.3.5}$$

We now establish a bound on $|\mathbb{E}(Y_{i\,T}) - \mathbb{E}(\Lambda_i)|$. Take expectations in eq. (A.3.5):

$$\begin{aligned}
&|\mathbb{E}(Y_{i\,T}) - \mathbb{E}(\Lambda_i)| \\
&= \left|\mathbb{E}\left[\boldsymbol{d}_1'\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1\right) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1) - \boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)\right]\right|
\end{aligned}$$

$$
\begin{aligned}
(*) &\leq \left| \boldsymbol{d}_1' \, \mathbb{E}\left[\sqrt{T} \left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \right) \right] \right| + \left| \mathbb{E}\left[\frac{1}{2} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)' \nabla^2 \mu(\acute{\boldsymbol{\theta}}_i) \sqrt{T} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1) \right] \right| \\
&\quad + \left| (\boldsymbol{d}_1 - \boldsymbol{d}_0)' (\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) \right| \\
(**) &\leq \|\boldsymbol{d}_1\| \frac{C_{Bias}}{\sqrt{T}} + C_{\nabla^2 \mu} \left| \mathbb{E}\left(\sqrt{T} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)' (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1) \right) \right| + \frac{C_{\nabla^2 \mu}}{\sqrt{T}} \|\boldsymbol{\eta}_1\| \, \|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_1\| \\
(***) &\leq \|\boldsymbol{d}_1\| \frac{C_{Bias}}{\sqrt{T}} + \frac{C_{\nabla^2 \mu}}{\sqrt{T}} \mathbb{E} \left\| \sqrt{T} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \right\|^2 + \frac{2 C_{\nabla^2 \mu}}{\sqrt{T}} \mathbb{E} \left\| \sqrt{T} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \right\| \|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\| \\
&\quad + \frac{C_{\nabla^2 \mu}}{\sqrt{T}} \|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\|^2 + \frac{C_{\nabla^2 \mu}}{\sqrt{T}} \|\boldsymbol{\eta}_1\| \, \|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_1\| \\
(****) &\leq C_{\nabla \mu} \frac{C_{Bias}}{\sqrt{T}} + \frac{C_{\nabla^2 \mu} C_{\hat{\boldsymbol{\theta}}, 2}}{\sqrt{T}} + \frac{C_{\nabla^2 \mu}}{\sqrt{T}} \left(2 C_{\hat{\boldsymbol{\theta}}, 1} + \|\boldsymbol{\eta}_1\| \right) \|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\| \\
&\quad + \frac{C_{\nabla^2 \mu}}{\sqrt{T}} \|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\|^2 . \qquad \text{(A.3.6)}
\end{aligned}
$$

where the constants C do not depend on i. Here

(*) $\boldsymbol{\theta}_1$ is replaced by $\boldsymbol{\theta}_i$ in the first term using $\boldsymbol{d}_1' \sqrt{T} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1) - \boldsymbol{d}_1' (\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) = \boldsymbol{d}_1' \sqrt{T} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)$

(**) In the first term we apply Hölder's inequality inside the absolute value as

$$
\left| \boldsymbol{d}_1' \, \mathbb{E}\left[\sqrt{T} \left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \right) \right] \right| \leq \|\boldsymbol{d}_1\|_\infty \left\| \mathbb{E}(\sqrt{T} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)) \right\|_1 \leq \|\boldsymbol{d}_1\|_2 \sqrt{T} \left\| \mathbb{E}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \right\|_1 .
$$

Assumption A.4 bounds $\sqrt{T} \left\| \mathbb{E}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \right\|_1 \leq C_{Bias} / \sqrt{T}$. In the second term apply A.5 to replace the Hessian $\nabla^2 \mu(\acute{\boldsymbol{\theta}}_i)$. In the third term apply assumptions A.1 and A.5: $\nabla \mu$ is a differentiable function with norm of the derivative bounded, which implies that $\|\boldsymbol{d}_1 - \boldsymbol{d}_0\| = \left\| \nabla \mu(\boldsymbol{\theta}_0 + T^{-1/2} \boldsymbol{\eta}_1) - \nabla \mu(\boldsymbol{\theta}_0) \right\| \leq C_{\nabla^2 \mu} \|\boldsymbol{\eta}_1\| / \sqrt{T}$.

(***) Add and subtract $\boldsymbol{\theta}_1$ in both parentheses in the quadratic term, apply the triangle inequality.

(****) Recall that $\boldsymbol{\theta}_i - \boldsymbol{\theta}_1 = (\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) / \sqrt{T}$ by A.1. Expectations of $\left\| \sqrt{T} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \right\|$ are bounded using lemma A.2.1; by A.5 $\|\boldsymbol{d}_1\| \leq C_{\nabla \mu}$

Last, we can consider the sum $\sum_{\bar{N}+1}^{N} w_{i\,N} |\mathbb{E}(Y_{i\,T}) - \mathbb{E}(\Lambda_i)|$, bounded by the corresponding weighted sum of the right hand side of eq. (A.3.6). The first two terms in the bound do

not depend on i, and so

$$\sum_{i=\bar{N}+1}^{N} w_{i\,N} \left[\frac{C_{\nabla\mu} C_{Bias}}{\sqrt{T}} + \frac{C_{\nabla^2\mu} C_{\hat{\boldsymbol{\theta}},2}}{\sqrt{T}} \right] \leq \frac{C_{\nabla\mu} C_{Bias}}{\sqrt{T}} + \frac{C_{\nabla^2\mu} C_{\hat{\boldsymbol{\theta}},2}}{\sqrt{T}} \to 0$$

since $w_{i\,N}$ are part of a weight vector. For the third and the fourth term we make use of the conditions on weight decay and the moments of $\boldsymbol{\eta}_i$. Examine

$$\frac{C_{\nabla^2\mu}}{\sqrt{T}} \left[2C_{\hat{\boldsymbol{\theta}},2} + \|\boldsymbol{\eta}_1\| \right] \sum_{i=\bar{N}+1}^{N} w_{i\,N} \|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\| + \frac{C_{\nabla^2\mu}}{\sqrt{T}} \sum_{i=\bar{N}+1}^{N} w_{i\,N} \|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\|^2 .$$

By lemma A.2.4 $\sup_N \sum_{i=\bar{N}+1}^{N} w_{i\,N} \|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\|^k$, $k = 1, 2$ are finite. Then for some $M < \infty$ the above display is bounded by $M/\sqrt{T}$ and thus converges to zero as well. Combining the last two results together, we obtain that $\sup_N \sum_{i=\bar{N}+1}^{N} w_{i\,N} |\mathbb{E}(Y_{i\,T}) - \mathbb{E}(\Lambda_i)| \to 0$ as $T \to \infty$, giving the result of the lemma. □

Lemma A.3.5. *Let $Y_{i\,T}$ be defined as in eq. (A.3.1). Under assumptions of theorem 2*

$$\limsup_{N,T\to\infty} \sum_{i=\bar{N}+1}^{N} w_{i\,N}\, \mathbb{E}|Y_{i\,T}| < \infty \text{ as } N, T \to \infty.$$

Proof. Existence of $\mathbb{E}|Y_{i\,T}|$ for $T > T_0$ follows from lemma A.2.2. Add and subtract $\mathbb{E}(\Lambda_i)$ under the absolute value in $\mathbb{E}|Y_{i\,T}|$ to get

$$\begin{aligned}
\mathbb{E}|Y_{i\,T}| &\leq |\mathbb{E}(\Lambda_i)| + \mathbb{E}|Y_{i\,T} - \mathbb{E}(\Lambda_i)| \\
&= |\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)| + \mathbb{E}|Y_{i\,T} - \mathbb{E}(\Lambda_i)| \\
&\leq \|\boldsymbol{d}_0\| \, \|\boldsymbol{\eta}_i - \boldsymbol{\eta}_0\| + \mathbb{E}|Y_{i\,T} - \mathbb{E}(\Lambda_i)| \ ,
\end{aligned}$$

where we apply the Cauchy-Schwarz inequality in the last line. Take weighted sums

$$\sum_{i=\bar{N}+1}^{N} w_{i\,N}\,\mathbb{E}|Y_{i\,T}| \leq \|\boldsymbol{d}_0\| \sum_{i=\bar{N}+1}^{N} w_{i\,N}\,\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\| + \sum_{i=\bar{N}+1}^{N} w_{i\,N}\,\mathbb{E}|Y_{i\,T} - \Lambda_i|.$$

We show that both sums are bounded as $N, T \to \infty$. First, as in lemma A.3.4, from lemma A.2.4 it follows that $\sup_N \sum_{i=\bar{N}+1}^{N} w_{i\,N}\,\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\| < \infty$. Now turn to the second sum. Using eq. (A.3.4), we proceed similarly to the proof of lemma A.3.4:

$$\begin{aligned}
&\mathbb{E}|(Y_{i\,T}) - \mathbb{E}(\Lambda_i)| \\
&= \mathbb{E}\left|\left[\boldsymbol{d}_1'\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1\right) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1) - \boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) \pm \boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)\right]\right| \\
&\leq \mathbb{E}\left|\boldsymbol{d}_1'\left[\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right)\right]\right| + \mathbb{E}\left|\left[\frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)\right]\right| + |(\boldsymbol{d}_1 - \boldsymbol{d}_0)'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)| \\
&\leq C_{\nabla\mu}C_{\hat{\boldsymbol{\theta}},1} + \frac{C_{\nabla^2\mu}C_{\hat{\boldsymbol{\theta}},2}}{\sqrt{T}} + \frac{C_{\nabla^2\mu}}{\sqrt{T}}\left[2C_{\hat{\boldsymbol{\theta}},1} + \|\boldsymbol{\eta}_1\|\right]\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\| + \frac{C_{\nabla^2\mu}}{\sqrt{T}}\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\|^2.
\end{aligned}$$

There is one change relative to lemma A.3.4: by the Cauchy-Schwarz inequality and assumption A.5, $\mathbb{E}\left|\boldsymbol{d}_1'\left[\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right)\right]\right| \leq C_{\nabla\mu}\,\mathbb{E}\left\|\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\right\|$, to which we then apply lemma A.2.1. The constant C_{Bias} does not appear in the above bound. Take weighted sums in $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\,\mathbb{E}|Y_{i\,T} - \Lambda_i|$, and use the above bound for each term in the sum. The argument proceeds similarly to lemma A.3.4. The first two terms in the bound satisfy $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\left(C_{\nabla\mu}C_{\hat{\boldsymbol{\theta}},1} + C_{\nabla^2\mu}C_{\hat{\boldsymbol{\theta}},2}/\sqrt{T}\right) \leq C_{\nabla\mu}C_{\hat{\boldsymbol{\theta}},1} + C_{\nabla^2\mu}C_{\hat{\boldsymbol{\theta}},2}/\sqrt{T}$, which is independent of N and convergent in T. Both sums $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\,\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\|$ and $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\,\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\|^2$ are bounded in N regardless of T by lemma A.2.4. We conclude that $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\,\mathbb{E}|Y_{i\,T} - \Lambda_i|$ is bounded in N and T, giving the claim of the lemma. □

Lemma A.3.6. *Let assumptions of theorem 2 hold, and let Λ_i be as in lemma 1. Then for any $\varepsilon > 0$*

$$\limsup_{N\to\infty} \sum_{i=\bar{N}+1}^{N} \mathbb{E}\left[w_{i\,N}|\Lambda_i|\mathbb{I}_{|w_{i\,N}\Lambda_i|>\varepsilon}\right] = 0.$$

Proof. Since $\sup_{i>\bar{N}} w_{i\,N} = o(N^{-1/2})$, there exists some $C_w > 0$ and N_0 such that for all $N > N_0$ it holds that $w_{i\,N} < C_w^{-1} N^{-1/2}$ for all $i > \bar{N}$. Also observe that for $p > 1$ $\mathbb{E}(|X|\,\mathbb{I}_{X>M}) \leq M^{-(p-1)}\,\mathbb{E}(|X|^p)$. Hence for $p > 1$

$$\begin{aligned}\sum_{i=\bar{N}+1}^{N} \mathbb{E}\left[w_{i\,N}|\Lambda_i|\mathbb{I}_{|w_{i\,N}\Lambda_i|>\varepsilon}\right] &\leq \sum_{i=\bar{N}+1}^{N} \mathbb{E}\left[w_{i\,N}|\Lambda_i|\,\mathbb{I}_{|\Lambda_i>C_w N^{1/2}\varepsilon|}\right] \\ &\leq \frac{1}{(C_w\varepsilon N^{1/2})^{p-1}} \sum_{i=\bar{N}+1}^{N} w_{i\,N}\,\mathbb{E}(|\Lambda_i|^p). \qquad \text{(A.3.7)}\end{aligned}$$

Pick $p = 2$. Since $1/(C_w\varepsilon N^{1/2}) \to 0$, it is sufficient to show that $\sum_{i=\bar{N}+1}^{N} w_{i\,N}\,\mathbb{E}(|\Lambda_i|^2)$ is bounded over N.

Since $|\Lambda_i|$ is folded normal, its first two moments are given by (see Elandt (1961)):

$$\begin{aligned}\mathbb{E}|\Lambda_i|^2 &= (\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1))^2 + \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0 - (\mathbb{E}|\Lambda_i|)^2, \\ \mathbb{E}|\Lambda_i| &= \sqrt{\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0}\sqrt{\frac{2}{\pi}}e^{-\frac{(\boldsymbol{d}_0'(\boldsymbol{\eta}_i-\boldsymbol{\eta}_1))^2}{2\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0}} + \boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)\left(1 - 2\Phi\left(-\frac{\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)}{2\sqrt{\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0}}\right)\right).\end{aligned}$$

It is sufficient to establish the boundedness of the weighted sum of each term separately. We proceed in order of appearance in the preceding display.

1. By the Cauchy-Schwarz inequality

$$\sum_{i=\bar{N}+1}^{N} w_{i\,N}(\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1))^2 \leq \|\boldsymbol{d}_0\|^2 \sum_{i=\bar{N}+1}^{N} w_{i\,N}\,\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\|^2.$$

   The sum on the right is bounded over N by lemma A.2.4.

2. By the bound on variance of assumption A.3 it holds that

$$\sum_{i=\bar{N}+1}^{N} w_{i\,N}\boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0 \leq \bar{\lambda}_{\boldsymbol{\Sigma}}\bar{\lambda}_{\boldsymbol{H}}^2\,\|\boldsymbol{d}_0\|^2 \ .$$

3. Consider the first term in $(\mathbb{E}|\Lambda_i|)^2$:

$$\sum_{i=\bar{N}+1}^{N} w_{i\,N} \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0 \frac{2}{\pi} \left[e^{-\frac{(\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1))^2}{2\boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0}} \right]^2 \leq \bar{\lambda}_{\boldsymbol{\Sigma}} \underline{\lambda}_{\boldsymbol{H}}^2 \left\| \boldsymbol{d}_0 \right\|^2 \frac{2}{\pi}.$$

4. Cross-term in $(\mathbb{E}|\Lambda_i|)^2$:

$$\begin{aligned} &\sum_{i=\bar{N}+1}^{N} w_{i\,N} \left| \sqrt{\boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0} \sqrt{\frac{2}{\pi}} e^{-\frac{(\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1))^2}{2\boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0}} \boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) \left(1 - 2\Phi \left(-\frac{\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)}{2\sqrt{\boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0}} \right) \right) \right| \\ \leq &\sqrt{\bar{\lambda}_{\boldsymbol{\Sigma}} \underline{\lambda}_{\boldsymbol{H}}^2} \left\| \boldsymbol{d}_0 \right\|^2 \sqrt{\frac{2}{\pi}} \sum_{i=\bar{N}+1}^{N} w_{i\,N} \left\| \boldsymbol{\eta}_i - \boldsymbol{\eta}_1 \right\|. \end{aligned}$$

   The sum in the last line is bounded over N by lemma A.2.4.

5. Square of the second term:

$$\begin{aligned} &\sum_{i=\bar{N}+1}^{N} w_{i\,N} \left[\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) \right]^2 \left(1 - 2\Phi \left(-\frac{\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)}{2\sqrt{\boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0}} \right) \right)^2 \\ \leq &\sum_{i=\bar{N}+1}^{N} w_{i\,N} \left[\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) \right]^2 \leq \left\| \boldsymbol{d}_0 \right\|^2 \sum_{i=\bar{N}+1}^{N} w_{i\,N} \left\| \boldsymbol{\eta}_i - \boldsymbol{\eta}_1 \right\|^2, \end{aligned}$$

   where the last sum is bounded by lemma A.2.4.

Combining the above arguments, we conclude that $\sup_N \sum_{i=\bar{N}+1}^{N} w_{i\,N} \mathbb{E}(|\Lambda_i|^2) < \infty$. By eq. (A.3.7)

$$\sum_{i=\bar{N}+1}^{N} \mathbb{E}\left[w_{i\,N} |\Lambda_i| \mathbb{I}_{|w_{i\,N}\Lambda_i| > \varepsilon} \right] \leq \frac{1}{C_w \varepsilon N^{1/2}} \sup_N \sum_{i=\bar{N}+1}^{N} w_{i\,N} \mathbb{E}(|\Lambda_i|^2$$

The right hand side tends to 0 as $N \to \infty$. □

Lemma A.3.7. *Let Y_{iT} be defined as in eq. (A.3.1). Under assumptions of theorem 2, for any $\varepsilon > 0$*

$$\limsup_{N,T \to \infty} \sum_{i=\bar{N}+1}^{N} \mathbb{E}\left[w_{i\,N} |Y_{i\,T}| \mathbb{I}_{|w_{i\,N} Y_{i\,T}| > \varepsilon} \right] = 0.$$

Proof. Existence of $\mathbb{E}|Y_{iT}|$ for $T > T_0$ follows from lemma A.2.2. We use the same strategy as in lemma A.3.6. Since $\sup_{i>\bar{N}} w_{iN} = o(N^{-1/2})$, there exists some $C_w > 0$ and N_0 such that for all $N > N_0$ it holds that $w_{iN} < C_w^{-1} N^{-1/2}$ for all $i > \bar{N}$. Then for $p > 1$, if $\mathbb{E}|Y_{iT}|^p$ exists, we obtain that

$$\begin{aligned}
&\sum_{i=\bar{N}+1}^{N} \mathbb{E}\left[w_{iN}|Y_{iT}|\mathbb{I}_{|w_{iN}Y_{iT}|>\varepsilon}\right] \\
&\leq \sum_{i=\bar{N}+1}^{N} \mathbb{E}\left[w_{iN}|Y_{iT}|\mathbb{I}_{|Y_{iT}|>C_w N^{1/2}\varepsilon}\right] \\
&\leq \frac{1}{(C_w \varepsilon N^{1/2})^{p-1}} \sum_{i=\bar{N}+1}^{N} w_{i,N}\, \mathbb{E}\left[|Y_{iT}|^p\right] \\
&\leq \frac{2^{p-1}}{(C_w \varepsilon N^{1/2})^{p-1}} \sum_{i=\bar{N}+1}^{N} w_{iN}\, \mathbb{E}|Y_{iT} - \boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)|^p \\
&\quad + \frac{2^{p-1}}{(C_w \varepsilon N^{1/2})^{p-1}} \sum_{i=\bar{N}+1}^{N} w_{iN}|\boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)|^p. \qquad \text{(A.3.8)}
\end{aligned}$$

It is sufficient to establish convergence of the weighted sums for some $p > 1$, since the leading $N^{(p-1)/2}$ will then drive the expression to zero. Take $p = 1 + \delta'$ where $\delta' = \delta/2$ for δ from assumption A.3.

The second sum in eq. (A.3.8) is bounded over N by lemma A.2.4, as

$$\sum_{\bar{N}+1}^{N} w_{iN}|\boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)|^{1+\delta'} \leq C_{\nabla\mu}^{1+\delta'} \sum_{i=\bar{N}+1}^{N} w_{iN} \|\boldsymbol{\eta}_i - \boldsymbol{\eta}_1\|^{1+\delta'}.$$

Now consider $\sum_{i=\bar{N}+1}^{N} w_{iN}\, \mathbb{E}|Y_{iT} - \boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)|^{1+\delta'}$. We proceed similarly to the proof of lemma A.3.5. First, by lemma A.2.2 $\mathbb{E}|Y_{iT} - \boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)|^{1+\delta'} < \infty$. It remains to show that the weighted sum is bounded over N. Recall from lemma A.3.4 that

$$Y_{iT} - \boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) = \boldsymbol{d}_1'\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)'\nabla^2\mu(\acute{\boldsymbol{\theta}}_i)\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)$$

for $\acute{\boldsymbol{\theta}}_i$ is intermediate between $\hat{\boldsymbol{\theta}}_i$ and $\boldsymbol{\theta}_1$. Then

$$
\begin{aligned}
&|Y_{iT} - \boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)|^{1+\delta'} \\
&\leq 2^{\delta'} \left| \boldsymbol{d}_1' \sqrt{T} \left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \right) \right|^{1+\delta'} + 2^{\delta'} \left| \frac{1}{2} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1)' \nabla^2 \mu(\acute{\boldsymbol{\theta}}_i) \sqrt{T} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1) \right|^{1+\delta'} \\
&\leq 2^{\delta'} \left| \boldsymbol{d}_1' \sqrt{T} \left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \right) \right|^{1+\delta'} + \frac{2^{2\delta'} C_{\nabla^2 \mu}^{1+\delta'}}{T^{(1+\delta')/2}} \left\| \sqrt{T} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_i) \right\|^{2(1+\delta')} \\
&\quad + 2^{1+3\delta'} C_{\nabla^2 \mu}^{1+\delta'} \left| \sqrt{T} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)' \frac{(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)}{\sqrt{T}} \right|^{1+\delta'} + \frac{2^{2\delta'} C_{\nabla^2 \mu}^{1+\delta'}}{T^{(1+\delta')/2}} |(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)' (\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)|^{1+\delta'}.
\end{aligned}
$$

Taking expectations, we obtain

$$
\begin{aligned}
&\mathbb{E}|Y_{iT} - \boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)|^{1+\delta'} \qquad \text{(A.3.9)} \\
&\leq 2^{1+3\delta'} \left[C_{\mu}^{1+\delta'} C_{\hat{\boldsymbol{\theta}}, 1+\delta/2} + \frac{C_{\nabla^2 \mu}^{1+\delta'} C_{\hat{\boldsymbol{\theta}}, 2+\delta}}{T^{(1+\delta')/2}} \right. \\
&\quad \left. + \frac{2 C_{\nabla^2 \mu}^{1+\delta'} C_{\hat{\boldsymbol{\theta}}, 1+\delta/2}}{T^{(1+\delta')/2}} \left\| \boldsymbol{\eta}_i - \boldsymbol{\eta}_1 \right\|^{1+\delta'} + \frac{C_{\nabla^2 \mu}^{1+\delta'}}{T^{(1+\delta')/2}} \left\| \boldsymbol{\eta}_i - \boldsymbol{\eta}_1 \right\|^{2(1+\delta')} \right],
\end{aligned}
$$

where the bounds on $\mathbb{E} \left\| \sqrt{T} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \right\|^k, k = 1 + \delta/2, 2 + \delta$ follow from lemma A.2.1.

Take weighted sums $\sum_{i=\bar{N}+1}^{N} w_{iN} \, \mathbb{E}|Y_{iT} - \boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1)|^{1+\delta'}$. Then for the first two terms it holds that

$$
\sum_{i=\bar{N}+1}^{N} w_{iN} C_{\nabla \mu}^{1+\delta} C_{\hat{\boldsymbol{\theta}}, 1+\delta/2} + \frac{C_{\nabla^2 \mu}^{1+\delta'} C_{\hat{\boldsymbol{\theta}}, 2+\delta}}{T^{(1+\delta')/2}} \leq C_{\nabla \mu}^{1+\delta} C_{\hat{\boldsymbol{\theta}}, 1+\delta/2} + \frac{C_{\nabla^2 \mu}^{1+\delta'} C_{\hat{\boldsymbol{\theta}}, 2+\delta}}{T^{(1+\delta')/2}},
$$

since constants are independent of i. For the third and the fourth term of eq. (A.3.9), it is sufficient to observe that by lemma A.2.4 $\sup_N \sum_{i=\bar{N}+1}^{N} w_{iN} \left\| \boldsymbol{\eta}_i - \boldsymbol{\eta}_1 \right\|^{2(1+\delta')} < \infty$ and $\sup_N \sum_{i=\bar{N}+1}^{N} w_{iN} \left\| \boldsymbol{\eta}_i - \boldsymbol{\eta}_1 \right\|^{1+\delta'} < \infty$.

Hence, both sums in eq. (A.3.8) are bounded uniformly over N. Taking $N \to \infty$ shows the original sum of interest converges to 0. □

Finally, we present the proof of theorem 2.

Proof of theorem 2. Using the fact that $\sum_{i=1}^{N} w_{iN} = 1$ and recalling that $N > \bar{N}$ we write

$$\sqrt{T}\left(\hat{\mu}(\boldsymbol{w}_N) - \boldsymbol{\mu}(\boldsymbol{\theta}_1)\right) = \sum_{i=1}^{N} w_{iN} Y_{iT} = \sum_{i=1}^{\bar{N}} w_{iN} Y_{iT} + \sum_{i=\bar{N}+1}^{N} w_{iN} Y_{iT} \ .$$

The first sum contains the units whose weights are allowed to be asymptotically non-negligible. By lemma A.3.2, as $N, T \to \infty$ jointly, it holds that

$$\sum_{i=1}^{\bar{N}} w_{iN} Y_{iT} \Rightarrow \sum_{i=1}^{\bar{N}} w_i \Lambda_i \sim N\left(\sum_{i=1}^{\bar{N}} w_i \boldsymbol{d}_1'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1), \sum_{i=1}^{\bar{N}} w_i^2 \boldsymbol{d}_1' \boldsymbol{V}_i \boldsymbol{d}_1\right) \ .$$

The second sum contains the units whose weights satisfy $\sup_{i>\bar{N}} w_{iN} = o(N^{-1/2})$. By appealing to lemma A.3.1, we show that $\sum_{i=\bar{N}+1}^{N} w_{iN} Y_{iT} \xrightarrow{p} -\left(1 - \sum_{i=1}^{\bar{N}} w_i\right) \boldsymbol{d}_0' \boldsymbol{\eta}_1$ as $N, T \to \infty$ jointly. We turn to verifying the conditions of lemma A.3.1:

1. *Assumption 1* (large T step): follows from lemma 1 as

$$Y_{iT} \Rightarrow \Lambda_i \sim N\left(\boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1), \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0\right) \ .$$

2. *Assumption 2* (large N step): by lemma A.3.3, $\sum_{i=\bar{N}+1}^{N} w_{iN} \Lambda_i$ converges in probability to $-\left(1 - \sum_{i=1}^{\bar{N}} w_i\right) \boldsymbol{d}_0' \boldsymbol{\eta}_1$

3. *Assumptions 3-6* are verified by lemmas A.3.4-A.3.7, respectively.

Last, by Slutsky's theorem

$$\begin{aligned} &\sum_{i=1}^{\bar{N}} w_{iN} Y_{iT} + \sum_{i=\bar{N}+1}^{N} w_{iN} Y_{iT} \\ &\Rightarrow N\left(\sum_{i=1}^{\bar{N}} w_i \boldsymbol{d}_0'(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1) - \left(1 - \sum_{i=1}^{\bar{N}} w_i\right) \boldsymbol{d}_0' \boldsymbol{\eta}_1, \sum_{i=1}^{\bar{N}} w_i^2 \boldsymbol{d}_0' \boldsymbol{V}_i \boldsymbol{d}_0\right) \ , \end{aligned}$$

which establishes the claim. □

A.4 Proof of Lemma 2

Proof of lemma 2. First assertion: in notation of the proof of lemma 1, for $T > T_0$

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1\right) = \boldsymbol{\eta}_i - \boldsymbol{\eta}_1 + \sqrt{T}\left(\hat{\boldsymbol{H}}_{i\,T}^{-1}\frac{1}{T}\sum_{t=1}^{T}\nabla m(\hat{\boldsymbol{\theta}}_i, \boldsymbol{z}_{i\,t}) - \hat{\boldsymbol{H}}_{1\,T}^{-1}\frac{1}{T}\sum_{t=1}^{T}\nabla m(\hat{\boldsymbol{\theta}}_1, \boldsymbol{z}_{1\,t})\right).$$

By lemma 1, the term in parentheses tends to $\boldsymbol{Z}_i - \boldsymbol{Z}_1 \sim N(\boldsymbol{\eta}_i - \boldsymbol{\eta}_1, \boldsymbol{V}_i + \boldsymbol{V}_1)$, as $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_i$ are independent. Convergence is joint by lemma 1 since $\sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1\right) = \sqrt{T}\left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1\right) - \sqrt{T}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right)$.

Now turn to the second assertion. First, it holds that

$$\sqrt{T}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_1\right) \xrightarrow{p} -\boldsymbol{\eta}_1$$

as $N, T \to \infty$ by theorem 2 with the μ the identity map (which satisfies condition A.5).[23]

Then

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i\right) = \sqrt{T}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right) + \sqrt{T}\left(\boldsymbol{\theta}_1 - \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i\right) \Rightarrow \boldsymbol{Z}_1 + \boldsymbol{\eta}_1 \sim N(\boldsymbol{\eta}_1, \boldsymbol{V}_1),$$

by lemma 1 and Slutsky's theorem.

□

A.5 Proof of Theorems 3 and 4

Proof of theorem 3. Lemma 2 implies that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1) \Rightarrow \boldsymbol{Z}_i - \boldsymbol{Z}_1$$

[23] Formally, we only establish theorem 2 for a scalar parameter μ. To see that it applies to the case of vector $\hat{\boldsymbol{\theta}}$ and $\mu(\boldsymbol{\theta}) = \boldsymbol{\theta}$, it is sufficient to apply the Cramér-Wold device. The Cramér-Wold device succeeds because for each $\boldsymbol{c} \in \mathbb{R}^{\dim \boldsymbol{\theta}}$ $\mu(\boldsymbol{\theta}) = \boldsymbol{c}'\boldsymbol{\theta}$ is a scalar parameter that satisfies assumption A.5. The corresponding gradient is $\boldsymbol{d}_0 = c$. Alternatively, the assertion can be seen by applying lemma A.3.1 directly to the MG estimator, the steps remain unchanged.

jointly for all $i = 1, \ldots, N$. Hence jointly for all i and j it holds that

$$\begin{aligned}
\left[\hat{\boldsymbol{\Psi}}_{\bar{N}}\right]_{ii} &\Rightarrow \boldsymbol{d}_0'((\boldsymbol{Z}_i - \boldsymbol{Z}_1)(\boldsymbol{Z}_i - \boldsymbol{Z}_1)' + \boldsymbol{V}_i)\boldsymbol{d}_0 &&= \left[\overline{\boldsymbol{\Psi}}_{\bar{N}}\right]_{ii}, \\
\left[\hat{\boldsymbol{\Psi}}_{\bar{N}}\right]_{ij} &\Rightarrow \boldsymbol{d}_0'((\boldsymbol{Z}_i - \boldsymbol{Z}_1)(\boldsymbol{Z}_j - \boldsymbol{Z}_1)')\boldsymbol{d}_0 &&= \left[\overline{\boldsymbol{\Psi}}_{\bar{N}}\right]_{ij}, \quad i \neq j.
\end{aligned}$$

Note that $\hat{\boldsymbol{\Psi}}_{\bar{N}}$ is finite-dimensional, and all its elements jointly converge as $T \to \infty$. Then the continuous mapping theorem readily implies that for any $\boldsymbol{w}^{\bar{N}} \in \Delta^{\bar{N}}$

$$\widehat{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) \Rightarrow \overline{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) := \boldsymbol{w}^{\bar{N}'}\overline{\boldsymbol{\Psi}}_{\bar{N}}\boldsymbol{w}^{\bar{N}},$$

which establishes the first claim.

The second claim is an implication of the argmax theorem (theorem 3.2.2 in Van der Vaart and Wellner (1996)). The conditions of that theorem are satisfied since we have that

1. By the first assertion of the theorem, $\widehat{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}}) \Rightarrow \overline{LA\text{-}MSE}_{\bar{N}}(\boldsymbol{w}^{\bar{N}})$ as $T \to \infty$ for every $\boldsymbol{w}^{\bar{N}}$ in the compact set $\Delta^{\bar{N}}$.

2. The limit problem $\arg\min_{\boldsymbol{w}^{\bar{N}} \in \Delta^{\bar{N}}} \boldsymbol{w}^{\bar{N}'}\overline{\boldsymbol{\Psi}}_{\bar{N}}\boldsymbol{w}^{\bar{N}}$ is a problem of minimizing a strictly convex continuous function on a compact convex set $\Delta^{\bar{N}}$, hence it has a unique solution. Strict convexity of the objective function follows since $\overline{\boldsymbol{\Psi}}_{\bar{N}}$ is positive definite. To see that $\overline{\boldsymbol{\Psi}}_{\bar{N}}$ is positive definite, it is sufficient to observe that for any $\boldsymbol{w} \neq 0$ $\boldsymbol{w}'\overline{\boldsymbol{\Psi}}_{\bar{N}}\boldsymbol{w} \geq \min_{i:w_i \neq 0} w_i^2 \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0 > 0$. The inequality follows as $\boldsymbol{w}'\overline{\boldsymbol{\Psi}}_{\bar{N}}\boldsymbol{w}$ is formally the MSE associated with the problem with individual variances given by $\boldsymbol{V}_i$ and biases of the form $(\boldsymbol{Z}_i - \boldsymbol{Z}_1)$. Hence $\boldsymbol{w}'\overline{\boldsymbol{\Psi}}_{\bar{N}}\boldsymbol{w}$= Bias$^2(\boldsymbol{w})$ + Variance$(\boldsymbol{w}) \geq$ Variance$(\boldsymbol{w}) \geq$ the minimal component of variance. Last, $\min_{i:w_i \neq 0} w_i^2 \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0 > 0$ since $\boldsymbol{V}_i$ is positive definite by assumption A.3 and $\boldsymbol{d}_0 \neq 0$.

3. The weights $\hat{\boldsymbol{w}}^{\bar{N}}$ minimize $\widehat{LA\text{-}MSE}_M(\boldsymbol{w}^{\bar{N}})$ over the compact set $\Delta^{\bar{N}}$ for all T.

Then the argmax theorem applies and $\hat{\boldsymbol{w}}^{\bar{N}} \Rightarrow \overline{\boldsymbol{w}}^{\bar{N}} = \arg\min_{\boldsymbol{w}^{\bar{N}} \in \Delta^{\bar{N}}} \boldsymbol{w}^{\bar{N}'}\overline{\boldsymbol{\Psi}}_{\bar{N}}\boldsymbol{w}^{\bar{N}}$ as $T \to \infty$.

The third claim follows from joint convergence of the weights, the estimators being averaged, and the continuous mapping theorem. □

Proof of theorem 4. First assertion: let $\boldsymbol{w}^{\bar{N},\infty} \in \tilde{\Delta}^{\bar{N}}$. Then by lemma 2 and Slutsky's theorem we conclude that as $N, T \to \infty$

$$
\begin{aligned}
&\widehat{LA\text{-}MSE}_{\infty}(\boldsymbol{w}^{\bar{N},\infty}) \\
&= \boldsymbol{w}^{\bar{N},\infty\prime} \hat{\boldsymbol{\Psi}}_{\bar{N}} \boldsymbol{w}^{\bar{N},\infty} + \left[\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \right) \left(\sqrt{T} \hat{\boldsymbol{d}}_1' \left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{\theta}}_i \right) \right) \right. \\
&\quad \left. -2 \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \hat{\boldsymbol{d}}_1' \sqrt{T} \left(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_1 \right) \right] \left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \right) \left(\sqrt{T} \hat{\boldsymbol{d}}_1' \left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{\theta}}_i \right) \right) \\
\Rightarrow &\overline{LA\text{-}MSE}_{\infty}(\boldsymbol{w}^{\bar{N},\infty}) \\
&:= \boldsymbol{w}^{\bar{N},\infty\prime} \overline{\boldsymbol{\Psi}}_{\bar{N}} \boldsymbol{w}^{\bar{N},\infty} + \left[\left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \right) \boldsymbol{d}_0' \left(\boldsymbol{\eta}_1 + \boldsymbol{Z}_1 \right) \right. \\
&\quad \left. -2 \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \boldsymbol{d}_0' \left(\boldsymbol{Z}_i - \boldsymbol{Z}_1 \right) \right] \left(1 - \sum_{i=1}^{\bar{N}} w_i^{\bar{N},\infty} \right) \boldsymbol{d}_0' \left(\boldsymbol{\eta}_1 + \boldsymbol{Z}_1 \right)
\end{aligned}
$$

Second assertion: follows by the same logic as in the fixed-N regime (theorem 3). The objective function $\widehat{LA\text{-}MSE}_{\infty}(\boldsymbol{w}^{\bar{N},\infty})$ can be represented as a quadratic function $\boldsymbol{x}'\hat{\boldsymbol{Q}}\boldsymbol{x}$, where $\boldsymbol{x} \in \Delta^{\bar{N}+1}$ stands in for $\left(\boldsymbol{w}^{\bar{N},\infty}, 1 - \sum_{i=1}^{\bar{N},\infty} w_i \right)$, and

$$
\hat{\boldsymbol{Q}} = \begin{pmatrix} \hat{\boldsymbol{\Psi}}_{\bar{N}} & \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{b}}' & T \left[\hat{\boldsymbol{d}}_1' \left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{\theta}}_i \right) \right]^2 \end{pmatrix} \Rightarrow \overline{\boldsymbol{Q}} = \begin{pmatrix} \overline{\boldsymbol{\Psi}}_{\bar{N}} & \overline{\boldsymbol{b}} \\ \overline{\boldsymbol{b}}' & [\boldsymbol{d}_0'(\boldsymbol{\eta}_1 + \boldsymbol{Z}_1)]^2 \end{pmatrix}
$$

$$
\hat{\boldsymbol{b}} = \begin{pmatrix} -\hat{\boldsymbol{d}}_1' T (\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1) \left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{\theta}}_i \right)' \hat{\boldsymbol{d}}_1 \\ \vdots \\ -\hat{\boldsymbol{d}}_1' T (\hat{\boldsymbol{\theta}}_{\bar{N}} - \hat{\boldsymbol{\theta}}_1) \left(\hat{\boldsymbol{\theta}}_1 - \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{\theta}}_i \right)' \hat{\boldsymbol{d}}_1 \end{pmatrix} \Rightarrow \overline{\boldsymbol{b}} = \begin{pmatrix} \boldsymbol{d}_0' \left(\boldsymbol{Z}_1 - \boldsymbol{Z}_1 \right) \left(\boldsymbol{\eta}_1 + \boldsymbol{Z}_1 \right)' \boldsymbol{d}_0 \\ \vdots \\ \boldsymbol{d}_0' \left(\boldsymbol{Z}_{\bar{N}} - \boldsymbol{Z}_1 \right) \left(\boldsymbol{\eta}_1 + \boldsymbol{Z}_1 \right)' \boldsymbol{d}_0 \end{pmatrix}.
$$

We now verify the condition of the argmax theorem for the problem of minimizing $\boldsymbol{x}'\hat{\boldsymbol{Q}}\boldsymbol{x}$ over $\Delta^{\bar{N}+1}$:

1. By the first assertion of the theorem, for any $\boldsymbol{x}$ in the compact set $\Delta^{\bar{N}+1}$ it holds

that $\boldsymbol{x}'\hat{\boldsymbol{Q}}\boldsymbol{x} \Rightarrow \boldsymbol{x}'\overline{\boldsymbol{Q}}\boldsymbol{x}$ as $N, T \to \infty$ jointly.

2. The limit problem $\arg\min_{\boldsymbol{x}\in\Delta^{\bar{N}+1}} \boldsymbol{x}'\overline{\boldsymbol{Q}}\boldsymbol{x}$ is a problem of minimizing a strictly convex continuous function on a compact convex set $\Delta^{\bar{N}+1}$, hence it has a unique solution. Similarly to the above, strict convexity follows from positive definiteness of $\overline{\boldsymbol{Q}}$. To establish positive definitiness, first let $\boldsymbol{x} \neq 0$ such that at least one of first $\bar{N}$ coordinates are nonzero. For such an $\boldsymbol{x}$ it holds that $\boldsymbol{x}'\overline{\boldsymbol{Q}}\boldsymbol{x} \geq \min_{i=1,\ldots,\bar{N}, x_i\neq 0} x_i^2 \boldsymbol{d}_0'\boldsymbol{V}_i\boldsymbol{d}_0 > 0$ where the inequality follows as in the proof of theorem 3. Alternatively, if the first $\bar{N}$ coordinates of $\boldsymbol{x}$ are zero, then $\boldsymbol{x}'\overline{\boldsymbol{Q}}\boldsymbol{w} = x_{\bar{N}+1}^2 \left(\boldsymbol{d}_0'(\boldsymbol{\eta}_1 + \boldsymbol{Z}_1)\right)^2 > 0$ ($(\boldsymbol{Z}_1)$-a.s.).

3. The vector $\hat{\boldsymbol{x}}^{\bar{N},\infty} = (\hat{\boldsymbol{w}}^{\bar{N},\infty}, 1 - \sum_{i=1}^{\bar{N}} \hat{w}_i^{\bar{N},\infty})$ minimizes $\boldsymbol{x}'\hat{\boldsymbol{Q}}\boldsymbol{x}$ over the compact set $\Delta^{\bar{N}+1}$ for all $N > \bar{N}, T$.

Then the argmax theorem shows that $\hat{\boldsymbol{x}}^{\bar{N},\infty} \Rightarrow \overline{\boldsymbol{x}}^{\bar{N},\infty} := \arg\min_{\boldsymbol{x}\in\Delta^{\bar{N}+1}} \boldsymbol{x}'\overline{\boldsymbol{Q}}\boldsymbol{x}$. Finally, it is sufficient to observe that $\hat{\boldsymbol{w}}^{\bar{N},\infty}$ comprises the first $\bar{N}$-coordinates of $\hat{\boldsymbol{x}}^{\bar{N},\infty}$, and $\overline{\boldsymbol{w}}^{\bar{N},\infty}$ comprises the first $\bar{N}$ coordinates of $\overline{\boldsymbol{x}}^{\bar{N},\infty}$.

The last assertion follows from the joint convergence of $\left(\hat{\boldsymbol{w}}^{\bar{N},\infty}\right)$, $\sqrt{T}(\mu(\hat{\boldsymbol{\theta}}_2 - \mu(\boldsymbol{\theta}_1))), \ldots,$ and $\sqrt{T}(\mu(\hat{\boldsymbol{\theta}}_{\bar{N}}) - \mu(\boldsymbol{\theta}_1)))$ as $N, T \to \infty$, and from the fact that $\sqrt{T}(\sum_{j=\bar{N}+1}^{N} v_{j\,N-\bar{N}}\mu(\hat{\boldsymbol{\theta}}_i) - \mu(\boldsymbol{\theta}_1)) \xrightarrow{p} -\boldsymbol{d}_0'\boldsymbol{\eta}_1$ by theorem 2. □