Weierstraß-Institut

für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Two convergence results for an alternation maximization procedure

Andreas Andresen , Vladimir Spokoiny

submitted: January 9, 2015

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: andreas.andresen@wias-berlin.de
vladimir.spokoiny@wias-berlin.de

No. 2061

Berlin 2015

2010 *Mathematics Subject Classification.* 62F10, 62J12, 62F25, 62H12.

Key words and phrases. alternating procedure, EM-algorithm, M-estimation, profile maximum likelihood, local linear approximation, spread, local concentration.

This work was partially supported by DFG Research Units 1735 “Structural Inference in Statistics: Adaptation and Efficiency“.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Andresen and Spokoiny's (2013) "critical dimension in semiparametric estimation" provide a technique for the finite sample analysis of profile M-estimators. This paper uses very similar ideas to derive two convergence results for the alternating procedure to approximate the maximizer of random functionals such as the realized log likelihood in MLE estimation. We manage to show that the sequence attains the same deviation properties as shown for the profile M-estimator in Andresen and Spokoiny (2013), i.e. a finite sample Wilks and Fisher theorem. Further under slightly stronger smoothness constraints on the random functional we can show nearly linear convergence to the global maximizer if the starting point for the procedure is well chosen.

1 Introduction

This paper presents a convergence result for an alternating maximization procedure to approximate M-estimators. Let $\mathbb{Y} \in \mathcal{Y}$ denote some observed random data, and $I\!P$ denote the data distribution. In the semiparametric profile M-estimation framework the target of analysis is

$$\boldsymbol{\theta}^* = \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon}^* = \Pi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon}} I\!E_{I\!P} \mathcal{L}(\boldsymbol{\upsilon}, \mathbb{Y}), \tag{1.1}$$

where $\mathcal{L} : \Upsilon \times \mathcal{Y} \to I\!R$, $\Pi_{\boldsymbol{\theta}} : \Upsilon \to I\!R^p$ is a projection and where Υ is some high dimensional or even infinite dimensional parameter space. This paper focuses on finite dimensional parameter spaces $\Upsilon \subseteq I\!R^{p^*}$ with $p^* = p + m \in \mathbb{N}$ being the full dimension, as infinite dimensional maximization problem are computationally anyways not feasible. A prominent way of estimating $\boldsymbol{\theta}^*$ is the profile M-estimator (pME)

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \widetilde{\boldsymbol{\upsilon}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{(\boldsymbol{\theta}, \boldsymbol{\eta})} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}).$$

The alternating maximization procedure is used in situations where a direct computation of the full maximum estimator (ME) $\widetilde{\boldsymbol{\upsilon}} \in I\!R^{p^*}$ is not feasible or simply very difficult to implement. Consider for example the task to calculate the pME where with scalar random observations $\mathbb{Y} = (y_i)_{i=1}^n \subset I\!R$, parameter $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in I\!R^p \times I\!R^m$ and a function basis $(\boldsymbol{e}_k) \subset L^2(I\!R)$

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\frac{1}{2} \sum_{i=1}^{n} \Big| y_i - \sum_{k=0}^{m} \boldsymbol{\eta}_k \boldsymbol{e}_k(\boldsymbol{X}_i^\top \boldsymbol{\theta}) \Big|^2.$$

In this case the maximization problem is high dimensional and non-convex (see Section 3 for more details). But for fixed $\boldsymbol{\theta} \in S_1 \subset I\!R^p$ maximization with respect to $\boldsymbol{\eta} \in I\!R^m$ is rather simple while for fixed $\boldsymbol{\eta} \in I\!R^m$ the maximization with respect to $\boldsymbol{\theta} \in I\!R^p$ can be feasible for low $p \in \mathbb{N}$. This motivates the following iterative procedure. Given some (data dependent) functional $\mathcal{L} : I\!R^p \times I\!R^m \to I\!R$ and an initial guess $\widetilde{\boldsymbol{\upsilon}}_0 \in I\!R^{p+m}$ set for $k \in \mathbb{N}$

$$\begin{aligned} \widetilde{\boldsymbol{\upsilon}}_{k,k+1} &\stackrel{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_{k+1}) = \left(\widetilde{\boldsymbol{\theta}}_k, \operatorname*{argmax}_{\boldsymbol{\eta} \in I\!R^m} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\eta}) \right), \\ \widetilde{\boldsymbol{\upsilon}}_{k,k} &\stackrel{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_k) = \left(\operatorname*{argmax}_{\boldsymbol{\theta} \in I\!R^p} \mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_k), \widetilde{\boldsymbol{\eta}}_k \right). \end{aligned} \tag{1.2}$$

The so called älternation maximization procedure"(or minimization) is a widely applied algorithm in many parameter estimation tasks (see [9], [13], [10] or [17]). Some natural questions arise: Does the sequence $(\widetilde{\boldsymbol{\theta}}_k)$ converge to a limit that satisfies the same statistical properties as the profile estimator? And if the answer is yes, after how many steps does the sequence acquire these properties? Under what circumstances does the sequence actually converge to the global maximizer $\widetilde{\boldsymbol{\upsilon}}$? This problem is hard because the behavior of each step of the sequence is determined by the actual finite sample realization of the functional $\mathcal{L}(\cdot, \mathbb{Y})$. To the authors' knowledge no general "convergence"result is available that answers the questions from above except for the treatment of specific models (see again [9], [13], [10] or [17]).

We address this difficulty via employing new finite sample techniques of [2] and [14] which allow to answer the above questions: with growing iteration number $k \in \mathbb{N}$ the estimators $\widetilde{\boldsymbol{\theta}}_k$ attain the same statistical properties as the profile M-estimator and Theorem 2.2 provides a choice of the necessary number of steps $K \in \mathbb{N}$. Under slightly stronger conditions on the structure of the model we can give a convergence result to the global maximizier that does not rely on unimodality. Further we can address the important question under which ratio of full dimension $p^* = p + m \in \mathbb{N}$ to sample size $n \in \mathbb{N}$ the sequence behaves as desired. For instance for smooth $\mathcal{L}$ our results become sharp if $p^*/\sqrt{n}$ is small and convergence to the full maximizer already occurs if p^*/n is small.

The alternation maximization procedure can be understood as a special case of the Expectation Maximization algorithm (EM algorithm) as we will illustrate below. The EM algorithm itself was derived by [5] who generalized particular versions of this approach and presented a variety of problems where its application can be fruitful; for a brief history of the EM algorithm see [11] (Sect. 1.8). We briefly explain the EM algorithm. Take observations $(\mathbb{X}) \sim I\!P_{\boldsymbol{\theta}}$ for some parametric family $(I\!P_{\boldsymbol{\theta}},\, \boldsymbol{\theta} \in \Theta)$. Assume that a parameter $\boldsymbol{\theta} \in \Theta$ is to be estimated as maximizer of the functional $\mathcal{L}_c(\mathbb{X}, \boldsymbol{\theta}) \in I\!R$, but that only $\mathbb{Y} \in \mathcal{Y}$ is observed, where $\mathbb{Y} = f_Y(\mathbb{X})$ is the image of the complete data set $\mathbb{X} \in \mathcal{X}$ under some map $f_Y : \mathcal{X} \to \mathcal{Y}$. Prominent examples for the map f_Y are projections onto some components of $\mathbb{X}$ if both are vectors. The information lost under the map can be regarded as missing data or latent variables. As a direct maximization of the functional is impossible without knowledge of $\mathbb{X}$ the EM algorithm serves as a workaround. It consists of the iteration of tow steps: starting with some initial guess $\widetilde{\boldsymbol{\theta}}_0$ the kth "Expectation step" derives the functional Q via

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k) = I\!E_{\boldsymbol{\theta}_k}[\mathcal{L}_c(\mathbb{X}, \boldsymbol{\theta})|\mathbb{Y}],$$

which means that on the right hand side the conditional expectation is calculated under the distribution $I\!P_{\boldsymbol{\theta}_k}$. The kth "Maximation step" then simply locates the maximizer $\boldsymbol{\theta}_{k+1}$ of Q.

Since the algorithm is very popular in applications a lot of research on its behaviour has been done. We are only dealing with a special case of this procedure so we restrict our selves to citing the well known convergence result by [16]. Wu presents regularity conditions that ensure that $\mathcal{L}(\boldsymbol{\theta}_{k+1}) \geq \mathcal{L}(\boldsymbol{\theta}_k)$ where

$$\mathcal{L}(\boldsymbol{\theta}, \mathbb{Y}) \stackrel{\text{def}}{=} \log \int_{\{\mathbb{X}|\, \mathbb{Y} = f_Y(\mathbb{X})\}} \exp \mathcal{L}_c(\mathbb{X}, \boldsymbol{\theta}) d\mathbb{X},$$

such that $\mathcal{L}(\boldsymbol{\theta}_k) \to \mathcal{L}^*$ for some limit value $\mathcal{L}^* > 0$, that may depend on the starting point $\boldsymbol{\theta}_0$. Additionally Wu gives conditions that guarantee that the sequence $\boldsymbol{\theta}_k$ (possibly a sequence of sets) converges to $C(\mathcal{L}^*) \stackrel{\text{def}}{=} \{\boldsymbol{\theta}|\, \mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}^*\}$. [5] show that the speed of convergence is linear in the case of point valued $\boldsymbol{\theta}_k$ and of some differentiability criterion being met. A limitation of these results is that it is not clear whether $\mathcal{L}^* = \sup \mathcal{L}(\boldsymbol{\theta})$ and thus it is not guaranteed that $C(\mathcal{L}^*)$ is the desired MLE and not just some local maximum. Of course this problem disappears if $\mathcal{L}(\cdot)$ is unimodal and the regularity conditions are met but this assumption may be too restrictive.

In a recent work [3] present a new way of addressing the properties of the EM sequence in a very general i.i.d. setting, based on concavity of $\boldsymbol{\theta} \mapsto I\!E_{\boldsymbol{\theta}^*}[\mathcal{L}_c(\mathbb{X}, \boldsymbol{\theta})]$. They show that if

additional to concavity the functional $\mathcal{L}_c$ is smooth enough (First order stability) and if for a sample $(\boldsymbol{Y}_i)$ with high probability an uniform bound holds of the kind

$$\sup_{\boldsymbol{\theta}\in B_{\mathtt{r}}(\boldsymbol{\theta}^*)} \left| \sum_{i=1}^{n} \operatorname*{argmax}_{\boldsymbol{\theta}^\circ} I\!E_{\boldsymbol{\theta}}[\mathcal{L}_c(\mathbb{X},\boldsymbol{\theta}^\circ)|\boldsymbol{Y}_i] - \operatorname*{argmax}_{\boldsymbol{\theta}^\circ} I\!E_{\boldsymbol{\theta}^*}[I\!E_{\boldsymbol{\theta}}[\mathcal{L}_c(\mathbb{X},\boldsymbol{\theta}^\circ)|\mathbb{Y}]] \right| \leq \epsilon_n, \tag{1.3}$$

that then with high probability and some $\rho < 1$

$$\|\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\| \leq \rho^k \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\| + \mathtt{C}\epsilon_n. \tag{1.4}$$

Unfortunately this does not answer our two questions to full satisfaction. First the bound (1.3) is rather high level and has to be checked for each model, while we seek (and find) properties of the functional - such as smoothness and bounds on the moments of its gradient - that lead to comparably desirable behavior. Further with (1.4) it remains unclear whether for large $k \in \mathbb{N}$ the alternating sequence satisfies a Fisher expansion or whether a Wilks type phenomenon occurs. In particular it remains open which ratio of dimension to sample size ensures good performance of the procedure. Also the actual convergence of $\widetilde{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}^*$ is not implied, as the right hand side in (1.4) is bounded from below by $\mathtt{C}\epsilon_n > 0$.

Remark 1.1. In the context of the alternating procedure the bound (1.3) would read

$$\max_{\boldsymbol{\theta}^\circ\in B_{\mathtt{r}}(\boldsymbol{\theta}^*)} \left| \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^\circ}) - \operatorname*{argmax}_{\boldsymbol{\theta}} I\!E\mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^\circ}) \right| \leq \epsilon_n,$$

which is still difficult to check.

To see that the procedure (1.2) is a special case of the EM algorithm denote in the notation from above $\mathbb{X} = \big(\operatorname{argmax}_{\boldsymbol{\eta}} \mathcal{L}\{(\boldsymbol{\theta},\boldsymbol{\eta}),\mathbb{Y}\},\mathbb{Y}\big)$ - where $\boldsymbol{\theta}$ is the parameter specifying the distribution $I\!P_{\boldsymbol{\theta}}$ - and $f_Y(\mathbb{X}) = \mathbb{Y}$. Then with $\mathcal{L}_c(\boldsymbol{\theta},\mathbb{X}) = \mathcal{L}_c(\boldsymbol{\theta},\boldsymbol{\eta},\mathbb{Y}) \stackrel{\text{def}}{=} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta})$

$$Q(\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}}_{k-1}) = I\!E_{\widetilde{\boldsymbol{\theta}}_{k-1}}[\mathcal{L}_c(\boldsymbol{\theta},\mathbb{X})|\mathbb{Y}] = \mathcal{L}_c\Big(\boldsymbol{\theta}, \operatorname*{argmax}_{\boldsymbol{\eta}} \mathcal{L}\{(\widetilde{\boldsymbol{\theta}}_{k-1},\boldsymbol{\eta}),\mathbb{Y}\},\mathbb{Y}\Big) = \mathcal{L}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k),$$

and thus the resulting sequence is the same as in (1.2). Consequently the convergence results from above apply to our problem if the involved regularity criteria are met. But as noted these results do not tell us if the limit of the sequence $(\widetilde{\boldsymbol{\theta}}_k)$ actually is the profile and the statistical properties of limit points are not clear without too restrictive assumptions on $\mathcal{L}$ and the data.

This work fills this gap for a wide range of settings. Our main result can be summarized as follows: Under a set of regularity conditions on the data and the functional $\mathcal{L}$ points of the sequence $(\widetilde{\boldsymbol{\theta}}_k)$ behave for large iteration number $k \in \mathbb{N}$ like the pME. To be more precise we show in Theorem 2.2 that when the initial guess $\widetilde{\boldsymbol{\upsilon}}_0 \in \Upsilon$ is good enough, then the step estimator sequence $(\widetilde{\boldsymbol{\theta}}_k)$ satisfies with high probability

$$\big\|\breve{D}\big(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\big) - \breve{\boldsymbol{\xi}}\big\|^2 \leq \epsilon(p^* + \rho^k \mathsf{R}_0),$$

$$\left|\max_{\boldsymbol{\eta}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_k,\boldsymbol{\eta}) - \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*,\boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}\|^2/2\right| \leq (p+\mathtt{x})^{1/2}\epsilon(p^* + \rho^k \mathsf{R}_0),$$

where $\rho < 1$ and $\epsilon > 0$ is some small number, for example $\epsilon = \mathtt{C}p^*/\sqrt{n}$ in the smooth i.i.d setting. Further $\mathtt{R}_0 > 0$ is a bound related to the quality of the initial guess. The random variable $\breve{\boldsymbol{\xi}} \in I\!R^p$ and the matrix $\breve{D} \in I\!R^{p\times p}$ are related to the efficient influence function in semiparametric models and its covariance. These are up to $\rho^k \mathtt{R}_0$ the same properties as those proven for the pME in [2] under nearly the same set of conditions. Further in our second main result we manage to show under slightly stronger smoothness conditions that $(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_k)$ approaches the ME $\widetilde{\boldsymbol{\upsilon}}$ with nearly linear convergence speed, i.e. $\|\mathcal{D}((\boldsymbol{\theta}_k, \boldsymbol{\eta}_k) - \widetilde{\boldsymbol{\upsilon}})\| \leq \tau^{k/\log(k)}$ with some $0 < \tau < 1$ and $\mathcal{D}^2 = I\!E\nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)$ (see Theorem 2.4).

In the following we write $\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)}$ in statements that are true for both $\widetilde{\boldsymbol{\upsilon}}_{k,k+1}$ and $\widetilde{\boldsymbol{\upsilon}}_{k,k}$. Also we do not specify whether the elements of the resulting sequence are sets or single points. All statements made about properties of $\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)}$ are to be understood in the sense that they hold for "every point of $\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)}$".

1.1 Idea of the proof

To motivate the approach first consider the toy model

$$\mathbb{Y} = \boldsymbol{\upsilon}^* + \boldsymbol{\varepsilon}, \text{ where } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbb{F}^{-2}_{\boldsymbol{\upsilon}^*}), \quad \mathbb{F}^2_{\boldsymbol{\upsilon}^*} =: \begin{pmatrix} \mathbb{F}^2_{\boldsymbol{\theta}^*} & A \\ A^\top & \mathbb{F}^2_{\boldsymbol{\eta}^*} \end{pmatrix}.$$

In this case we set $\mathcal{L}$ to be the true log likelihood of the observations

$$\mathcal{L}(\boldsymbol{\upsilon}, \mathbb{Y}) = -\|\mathbb{F}(\boldsymbol{\upsilon}^* - \mathbb{Y})\|^2/2.$$

With any starting initial guess $\widetilde{\boldsymbol{\upsilon}}_0 \in I\!R^{p+m}$ we obtain from (1.2) for $k \in \mathbb{N}$ and the usual first order criterion of maximality the following two equations

$$\begin{aligned} \mathbb{F}_{\boldsymbol{\theta}^*}(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) &= I_{\boldsymbol{\theta}^*}\boldsymbol{\varepsilon}_{\boldsymbol{\theta}} + \mathbb{F}^{-1}_{\boldsymbol{\theta}^*} A(\widetilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}^*), \\ \mathbb{F}_{\boldsymbol{\eta}^*}(\widetilde{\boldsymbol{\eta}}_{k+1} - \boldsymbol{\eta}^*) &= I_{\boldsymbol{\eta}^*}\boldsymbol{\varepsilon}_{\boldsymbol{\eta}} + \mathbb{F}^{-1}_{\boldsymbol{\eta}^*} A^\top(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*). \end{aligned}$$

Combining these two equations we derive, assuming $\|\mathbb{F}^{-1}_{\boldsymbol{\theta}^*} A \mathbb{F}^{-2}_{\boldsymbol{\eta}^*} A^\top I^{-1}_{\boldsymbol{\theta}^*}\| =: \|\boldsymbol{M}_0\| = \nu < 1$

$$\begin{aligned} \mathbb{F}_{\boldsymbol{\theta}^*}(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) &= \mathbb{F}^{-1}_{\boldsymbol{\theta}^*}(\mathbb{F}^2_{\boldsymbol{\theta}^*}\boldsymbol{\varepsilon}_{\boldsymbol{\theta}} - A\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}) + \mathbb{F}^{-1}_{\boldsymbol{\theta}^*} A \mathbb{F}^{-1}_{\boldsymbol{\eta}^*} A^\top \mathbb{F}^{-1}_{\boldsymbol{\theta}^*} \mathbb{F}_{\boldsymbol{\theta}^*}(\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*) \\ &= \sum_{l=1}^{k} \boldsymbol{M}_0^{k-l} \mathbb{F}^{-1}_{\boldsymbol{\theta}^*}(\mathbb{F}^2_{\boldsymbol{\theta}^*}\boldsymbol{\varepsilon}_{\boldsymbol{\theta}} - A\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}) \\ &\quad + \boldsymbol{M}_0^k \mathbb{F}_{\boldsymbol{\theta}^*}(\widetilde{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*) \to \mathbb{F}_{\boldsymbol{\theta}^*}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*). \end{aligned}$$

Because the limit $\widehat{\boldsymbol{\theta}}$ is independent of the initial point $\widetilde{\boldsymbol{\upsilon}}_0$ and because the profile $\widetilde{\boldsymbol{\theta}}$ is a fix point of the procedure the unique limit satisfies $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}$. This argument is based on the fact that in this setting the functional is quadratic such that the gradient satisfies

$$\nabla\mathcal{L}(\boldsymbol{\upsilon}) = \mathbb{F}^2_{\boldsymbol{\upsilon}^*}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) + \mathbb{F}^2_{\boldsymbol{\upsilon}^*}\boldsymbol{\varepsilon}.$$

Any smooth function is quadratic around its maximizer which motivates a local linear approximation of the gradient of the functional $\mathcal{L}$ to derive our results with similar arguments. This is done in the proof of Theorem 2.2.

First it is ensured that the whole sequence $(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)})_{k\in\mathbb{N}_0}$ satisfies for some $\mathsf{R}_0 > 0$

$$\{\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)},\, k \in \mathbb{N}_0\} \subset \{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathsf{R}_0\}, \tag{1.5}$$

where $\mathcal{D}^2 \stackrel{\mathrm{def}}{=} \nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)$ (see Theorem 4.3). In the second step we approximate with $\zeta = \mathcal{L} - I\!\!E\mathcal{L}$

$$\mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \nabla\zeta(\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2 + \alpha(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*), \tag{1.6}$$

where $\alpha(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ is defined by (1.6). Similar to the toy case above this allows using the first order criterion of maximality and (1.5) to obtain a bound of the kind

$$\begin{aligned}\|\mathcal{D}(\boldsymbol{\upsilon}_{k,k} - \boldsymbol{\upsilon}^*)\| &\le \mathtt{C}\sum_{l=0}^{k} \rho^l \left(\|\mathcal{D}^{-1}\nabla\zeta(\boldsymbol{\upsilon}^*)\| + |\alpha(\boldsymbol{\upsilon}_{l,l}, \boldsymbol{\upsilon}^*)|\right) \\ &\le \mathtt{C}_1 \left(\|\mathcal{D}^{-1}\nabla\zeta(\boldsymbol{\upsilon}^*)\| + \epsilon(\mathsf{R}_0)\right) + \rho^k \mathsf{R}_0 \stackrel{\mathrm{def}}{=} \mathtt{r}_k.\end{aligned}$$

This is done in Lemma 4.5 using results from [2] to show that $\epsilon(\mathsf{R}_0)$ is small. Finally the same arguments as in [2] allow to obtain our main result using that with high probability for all $k \in \mathbb{N}_0$ $\widetilde{\boldsymbol{\upsilon}}_{k,k} \in \{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}_k\}$. For the convergence result similar arguments are used. The only difference is that instead of (1.6) we use the approximation

$$\mathcal{L}(\boldsymbol{\upsilon}, \widetilde{\boldsymbol{\upsilon}}) = -\|\mathcal{D}(\boldsymbol{\upsilon} - \widetilde{\boldsymbol{\upsilon}})\|^2/2 + \alpha'(\boldsymbol{\upsilon}, \widetilde{\boldsymbol{\upsilon}}),$$

exploiting that $\nabla\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) \equiv 0$, which allows to obtain actual convergence to the ME.

It is worthy to point out two technical challenges of the analysis. First the sketched approach relies on (1.5). As all estimators $(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)})$ are random this means that we need with some small $\beta > 0$

$$I\!\!P\left(\bigcap_{k\in\mathbb{N}_0}\left\{\widetilde{\boldsymbol{\upsilon}}_{k,k}, \widetilde{\boldsymbol{\upsilon}}_{k,k+1} \in \{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathsf{R}_0\}\right\}\right) \ge 1 - \beta.$$

This is not trivial but the result of Theorem 4.3 serves the result thanks to $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)}) \ge \mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0)$. Second the main result 2.2 is formulated to hold for all $k \in \mathbb{N}_0$. This implies the need of a bound of the kind

$$I\!\!P\left(\bigcap_{k\in\mathbb{N}_0}\left\{\left\|\breve{D}^{-1}\{\breve{\nabla}\boldsymbol{\zeta}(\widetilde{\boldsymbol{\upsilon}}_{k,k}) - \breve{\nabla}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\}\right\| \le \epsilon(\mathtt{r}_k)\right\}\right) \ge 1 - \beta,$$

with some small $\epsilon(\mathtt{r}) > 0$ that is decreasing if $\mathtt{r} > 0$ shrinks. Again this is not trivial and not a direct implication of the results of [2] or [14]. We manage to derive this result in the desired way in Theorem 8.2, which is an adapted version of Theorem D.1 of [2] based on Corollary 2.5 of [14] .

2 Main results

2.1 Conditions

This section collects the conditions imposed on the model. We use the same set of assumptions as in [2] and this section closely follows Section 2.1 of that paper.

Let the full dimension of the problem be finite, i.e. $p^* < \infty$. Our conditions involve the symmetric positive definite information matrix $\mathcal{D}^2 \in I\!R^{p^* \times p^*}$ and a central point $\boldsymbol{\upsilon}^\circ \in I\!R^{p^*}$. In typical situations for $p^* < \infty$, one can set $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$ where $\boldsymbol{\upsilon}^*$ is the "true point" from (1.1). The matrix $\mathcal{D}^2$ can be defined as follows:

$$\mathcal{D}^2 = -\nabla^2 I\!E\mathcal{L}(\boldsymbol{\upsilon}^\circ).$$

Here and in what follows we implicitly assume that the log-functional function $\mathcal{L}(\boldsymbol{\upsilon})\colon I\!R^{p^*} \to I\!R$ is sufficiently smooth in $\boldsymbol{\upsilon} \in I\!R^{p^*}$, $\nabla\mathcal{L}(\boldsymbol{\upsilon}) \in I\!R^{p^*}$ stands for the gradient and $\nabla^2 I\!E\mathcal{L}(\boldsymbol{\upsilon}) \in I\!R^{p^* \times p^*}$ for the Hessian of the expectation $I\!E\mathcal{L} : I\!R^{p^*} \to I\!R$ at $\boldsymbol{\upsilon} \in I\!R^{p^*}$. By smooth enough we mean that we can interchange $\nabla I\!E\mathcal{L} = I\!E\nabla\mathcal{L}$ on $\Upsilon_\circ(\mathsf{R}_0)$, where $\Upsilon_\circ(\mathtt{r})$ is defined in (2.1) and $\mathsf{R}_0 > 0$ in (2.4). It is worth mentioning that $\mathcal{D}^2 = \mathcal{V}^2 \stackrel{\text{def}}{=} \mathrm{Cov}(\nabla\mathcal{L}(\boldsymbol{\upsilon}^*))$ if the model $\boldsymbol{Y} \sim I\!P_{\boldsymbol{\upsilon}^*} \in (I\!P_{\boldsymbol{\upsilon}})$ is correctly specified and sufficiently regular; see e.g. [7].

In the context of semiparametric estimation, it is convenient to represent the information matrix in block form:

$$\mathcal{D}^2 = \begin{pmatrix} D^2 & A \\ A^\top & H^2 \end{pmatrix}.$$

First we state an *identifiability condition*.

$(\boldsymbol{\mathcal{I}})$ It holds for some $\rho < 1$

$$\|H^{-1}A^\top D^{-1}\|_\infty \le \sqrt{\rho}.$$

Remark 2.1. The condition $(\mathcal{I})$ allows to introduce the important $p \times p$ efficient information matrix $\breve{D}^2$ which is defined as the inverse of the $\boldsymbol{\theta}$-block of the inverse of the full dimensional matrix $\mathcal{D}^2$. The exact formula is given by

$$\breve{D}^2 \stackrel{\text{def}}{=} D^2 - AH^{-2}A^\top,$$

and $(\mathcal{I})$ ensures that the matrix $\breve{D}^2$ is well posed.

Using the matrix $\mathcal{D}^2$ and the central point $\boldsymbol{\upsilon}^\circ \in I\!R^{p^*}$, we define the local set $\Upsilon_\circ(\mathtt{r}) \subset \Upsilon \subseteq I\!R^{p^*}$ with some $\mathtt{r} \ge 0$:

$$\Upsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Upsilon\colon \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\| \le \mathtt{r}\bigr\}. \tag{2.1}$$

The following two conditions quantify the smoothness properties on $\Upsilon_\circ(\mathtt{r})$ of the expected log-functional $I\!E\mathcal{L}(\boldsymbol{\upsilon})$ and of the stochastic component $\zeta(\boldsymbol{\upsilon}) = \mathcal{L}(\boldsymbol{\upsilon}) - I\!E\mathcal{L}(\boldsymbol{\upsilon})$.

$(\breve{\mathcal{L}})$ For each $\mathtt{r} \le \mathtt{r}_0$, there is a constant $\delta(\mathtt{r})$ such that it holds on the set $\Upsilon_\circ(\mathtt{r})$:

$$\|D^{-1}D^2(\boldsymbol{\upsilon})D^{-1} - I_p\| \le \delta(\mathtt{r}), \quad \|D^{-1}(A(\boldsymbol{\upsilon}) - A)H^{-1}\| \le \delta(\mathtt{r}),$$
$$\left\|D^{-1}AH^{-1}\left(I_m - H^{-1}H^2(\boldsymbol{\upsilon})H^{-1}\right)\right\| \le \delta(\mathtt{r}).$$

Remark 2.2. This condition describes the local smoothness properties of the function $I\!\!E\mathcal{L}(\boldsymbol{\upsilon})$. In particular, it allows to bound the error of local linear approximation of the projected gradient $\breve{\nabla}_{\boldsymbol{\theta}} I\!\!E\mathcal{L}(\boldsymbol{\upsilon})$ which is defined as

$$\breve{\nabla}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} - AH^{-2}\nabla_{\boldsymbol{\eta}}.$$

Under condition $(\breve{\mathcal{L}}_0)$ it follows from the second order Taylor expansion for any $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})$ (see Lemma B.1 of [2])

$$\|\breve{D}^{-1}\left(\breve{\nabla} I\!\!E\mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla} I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)\right) - \breve{D}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \delta(\mathtt{r})\mathtt{r}. \tag{2.2}$$

In the proofs we actually only need the condition (2.2) which in some cases can be weaker than $(\breve{\mathcal{L}}_0)$.

The next condition concerns the regularity of the stochastic component $\zeta(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \mathcal{L}(\boldsymbol{\upsilon}) - I\!\!E\mathcal{L}(\boldsymbol{\upsilon})$. Similarly to [14], we implicitly assume that the stochastic component $\zeta(\boldsymbol{\upsilon})$ is a separable stochastic process.

$(\breve{\mathcal{E}}\mathcal{D}_1)$ For all $0 < \mathtt{r} < \mathtt{r}_0$, there exists a constant $\omega \le 1/2$ such that for all $|\mu| \le \breve{\mathtt{g}}$ and $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})$

$$\sup_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})} \sup_{\|\boldsymbol{\gamma}\| \le 1} \log I\!\!E \exp\left\{\frac{\mu}{\omega}\frac{\boldsymbol{\gamma}^\top \breve{D}^{-1}\{\breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}')\}}{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|}\right\} \le \frac{\breve{\nu}_1^2\mu^2}{2}.$$

The above conditions allow to derive the main result once the accuracy of the sequence is established. We include another condition that allows to control the deviation behavior of $\|\breve{D}^{-1}\breve{\nabla}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\|$. To present this condition define the covariance matrix $\mathcal{V}^2 \in I\!R^{p^* \times p^*}$ and $\breve{V}^2 \in I\!R^{p \times p}$

$$\mathcal{V}^2 \stackrel{\text{def}}{=} \operatorname{Var}\{\nabla\mathcal{L}(\boldsymbol{\upsilon}^\circ)\}, \quad \breve{V}^2 = \operatorname{Cov}(\breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^\circ)).$$

$\mathcal{V}^2 \in I\!R^{p^* \times p^*}$ describes the variability of the process $\mathcal{L}(\boldsymbol{\upsilon})$ around the central point $\boldsymbol{\upsilon}^\circ$.

$(\breve{\mathcal{E}}\mathcal{D}_0)$ There exist constants $\nu_0 > 0$ and $\breve{\mathtt{g}} > 0$ such that for all $|\mu| \le \breve{\mathtt{g}}$

$$\sup_{\boldsymbol{\gamma} \in I\!R^p} \log I\!\!E \exp\left\{\mu\frac{\langle\breve{\nabla}_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^\circ), \boldsymbol{\gamma}\rangle}{\|\breve{V}\boldsymbol{\gamma}\|}\right\} \le \frac{\breve{\nu}_0^2\mu^2}{2}.$$

So far we only presented conditions that allow to treat the properties of $\widetilde{\boldsymbol{\theta}}_k$ on local sets $\Upsilon_\circ(\mathtt{r}_k)$. To show that $\mathtt{r}_k$ is not to large the following, stronger conditions are employed:

$(\mathcal{L}_0)$ For each $\mathtt{r} \leq \mathtt{r}_0$, there is a constant $\delta(\mathtt{r})$ such that it holds on the set $\Upsilon_\circ(\mathtt{r})$:

$$\bigl\| \mathcal{D}^{-1} \bigl\{ \nabla^2 I\!\!E \mathcal{L}(\boldsymbol{\upsilon}) \bigr\} \mathcal{D}^{-1} - I_{p^*} \bigr\| \leq \delta(\mathtt{r}).$$

$(\mathcal{ED}_1)$ There exists a constant $\omega \leq 1/2$, such that for all $|\mu| \leq \mathtt{g}$ and all $0 < \mathtt{r} < \mathtt{r}_0$

$$\sup_{\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})} \sup_{\|\boldsymbol{\gamma}\|=1} \log I\!\!E \exp \left\{ \frac{\mu \boldsymbol{\gamma}^\top \mathcal{D}^{-1} \bigl\{ \nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}') \bigr\}}{\omega \, \| \mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}') \|} \right\} \leq \frac{\nu_1^2 \mu^2}{2}.$$

$(\mathcal{ED}_0)$ There exist constants $\nu_0 > 0$ and $\mathtt{g} > 0$ such that for all $|\mu| \leq \mathtt{g}$

$$\sup_{\boldsymbol{\gamma} \in I\!R^{p^*}} \log I\!\!E \exp \left\{ \mu \frac{\langle \nabla \zeta(\boldsymbol{\upsilon}^\circ), \boldsymbol{\gamma} \rangle}{\| \mathcal{V} \boldsymbol{\gamma} \|} \right\} \leq \frac{\nu_0^2 \mu^2}{2}.$$

It is important to note, that the constants $\breve{\omega}, \breve{\delta}(\mathtt{r}), \breve{\nu}$ and $\omega, \delta(\mathtt{r}), \nu$ in the respective weak and strong version can differ substantially and may depend on the full dimension $p^* \in \mathbb{N}$ in less or more severe ways ($AH^{-2} \nabla_{\boldsymbol{\eta}} \mathcal{L}$ might be quite smooth while $\nabla_{\boldsymbol{\eta}} \mathcal{L}$ could be less regular). This is why we use both sets of conditions where they suit best, although the list of assumptions becomes rather long. If a short list is preferred the following lemma shows, that the stronger conditions imply the weaker ones from above:

Lemma 2.1. *[[2], Lemma 2.1] Assume* $(\mathcal{I})$. *Then* $(\mathcal{ED}_1)$ *implies* $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\mathcal{L}_0)$ *implies* $(\breve{\mathcal{L}}_0)$, *and* $(\mathcal{ED}_0)$ *implies* $(\breve{\mathcal{E}}\mathcal{D}_0)$ *with*

$$\breve{\mathtt{g}} = \frac{\sqrt{1-\rho^2}}{1 + \rho\sqrt{1+\rho^2}} \mathtt{g}, \; \breve{\nu} = \frac{1 + \rho\sqrt{1+\rho^2}}{\sqrt{1-\rho^2}} \nu, \; \breve{\delta}(\mathtt{r}) = \delta(\mathtt{r}), \;\; \textit{and} \; \breve{\omega} = \omega.$$

Finally we present two conditions that allow to ensure that with a high probability the sequence $(\boldsymbol{\upsilon}_{k,k(+1)})$ stays close to $\boldsymbol{\upsilon}^*$ if the initial guess $\widetilde{\boldsymbol{\upsilon}}_0$ lands close to $\boldsymbol{\upsilon}^*$. These conditions have to be satisfied on the whole set $\Upsilon \subseteq I\!R^{p^*}$.

$(\mathcal{L}\mathtt{r})$ For any $\mathtt{r} > \mathtt{r}_0$ there exists a value $\mathtt{b}(\mathtt{r}) > 0$, such that

$$\frac{-I\!\!E \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ)}{\| \mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ) \|^2} \geq \mathtt{b}(\mathtt{r}), \qquad \boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r}).$$

$(\mathcal{E}\mathtt{r})$ For any $\mathtt{r} \geq \mathtt{r}_0$ there exists a constant $\mathtt{g}(\mathtt{r}) > 0$ such that

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \sup_{\mu \leq \mathtt{g}(\mathtt{r})} \sup_{\boldsymbol{\gamma} \in I\!R^{p^*}} \log I\!\!E \exp \left\{ \mu \frac{\langle \nabla \zeta(\boldsymbol{\upsilon}), \boldsymbol{\gamma} \rangle}{\| \mathcal{D} \boldsymbol{\gamma} \|} \right\} \leq \frac{\nu_{\mathtt{r}}^2 \mu^2}{2}.$$

We impose one further merely technical condition:

$(\mathbf{B}_1)$ We assume for all $\mathtt{r} \geq \frac{6\nu_0}{\mathtt{b}} \sqrt{\mathtt{x} + 4p^*}$

$$1 + \sqrt{\mathtt{x} + 4p^*} \leq \frac{3\nu_{\mathtt{r}}^2}{\mathtt{b}} \mathtt{g}(\mathtt{r}).$$

Remark 2.3. Without this the calculation of $\mathrm{R}_0(\mathtt{x})$ in Section 4.1 would become technically more involved, without that further insight would be gained.

Remark 2.4. For a discussion on how restrictive these conditions are we refer the reader to Remark 2.8 and 2.9 of [2].

2.2 Introduction of important objects

In this section we introduce all objects and bounds that are relevant for Theorem 2.2. This section is quite technical but necessary to understand the results.

First consider the $p^* \times p^*$ matrices $\mathcal{D}^2$ and $\mathcal{V}^2$ from Section 2.1, which could be defined similarly to the Fisher information matrix:

$$\mathcal{D}^2 \stackrel{\mathrm{def}}{=} -\nabla^2 I\!\!E \mathcal{L}(\boldsymbol{\upsilon}^*), \;\; \mathcal{V}^2 \stackrel{\mathrm{def}}{=} \mathrm{Cov}(\nabla \mathcal{L}(\boldsymbol{\upsilon}^*)).$$

We represent the information and covariance matrix in block form:

$$\mathcal{D}^2 = \begin{pmatrix} D^2 & A \\ A^\top & H^2 \end{pmatrix}, \;\; \mathcal{V}^2 = \begin{pmatrix} V^2 & E \\ E^\top & Q^2 \end{pmatrix}.$$

A crucial object is the constant $0 \leq \rho$ defined by

$$\| D^{-1} A H^{-1} \|^2 \stackrel{\mathrm{def}}{=} \rho,$$

which we assume to be smaller 1 ($\| \cdot \|$ here and everywhere denotes the spectral norm when its argument is a matrix). It determines the speed of convergence of the alternating procedure (see Theorem 2.2). Define also the local sets

$$\Upsilon_\circ(\mathtt{r}) \stackrel{\mathrm{def}}{=} \bigl\{ \boldsymbol{\upsilon} : (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)^\top \mathcal{D}^2 (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \leq \mathtt{r}^2 \bigr\},$$
$$\widetilde{\Upsilon}_\circ(\mathtt{r}) \stackrel{\mathrm{def}}{=} \bigl\{ \boldsymbol{\upsilon} : (\boldsymbol{\upsilon} - \widetilde{\boldsymbol{\upsilon}})^\top \mathcal{D}^2 (\boldsymbol{\upsilon} - \widetilde{\boldsymbol{\upsilon}}) \leq \mathtt{r}^2 \bigr\},$$

and the radius $\mathtt{r}_0 > 0$ via

$$\mathtt{r}_0(\mathtt{x}) \stackrel{\mathrm{def}}{=} \inf_{\mathtt{r} \geq 0} \left\{ I\!\!P \left(\operatorname*{argmax}_{\substack{\boldsymbol{\upsilon} \in \Upsilon \\ \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} = \boldsymbol{\theta}^*}} \mathcal{L}(\boldsymbol{\upsilon}), \widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}) \right) \geq 1 - \mathrm{e}^{-\mathtt{x}} \right\}. \tag{2.3}$$

Remark 2.5. This radius can be determined using conditions $(\mathcal{L}_{\mathtt{r}})$ and $(\mathcal{E}\mathtt{r})$ of Section 2.1 and Theorem 4.3 which would yield $\mathtt{r}_0(\mathtt{x}) = \mathtt{C}\sqrt{\mathtt{x} + p^*}$.

Further introduce the $p \times p$ matrix $\breve{D}$ and the p-vectors $\breve{\nabla}_{\boldsymbol{\theta}}$ and $\breve{\boldsymbol{\xi}}$ as

$$\breve{D}^2 = D^2 - A H^{-2} A^\top, \, \breve{\nabla}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} - A H^{-2} \nabla_{\boldsymbol{\eta}}, \, \breve{\boldsymbol{\xi}} = \breve{D}^{-1} \breve{\nabla}_{\boldsymbol{\theta}},$$

and the matrices

$$I\!\!B^2 \stackrel{\mathrm{def}}{=} \mathcal{D}^{-1} \mathcal{V}^2 \mathcal{D}^{-1}, \, I\!\!B_{\boldsymbol{\theta}} \stackrel{\mathrm{def}}{=} D^{-1} V^2 D^{-1}, \, I\!\!B_{\boldsymbol{\eta}} \stackrel{\mathrm{def}}{=} H^{-1} Q^2 H^{-1}.$$

Remark 2.6. The random variable $\breve{\boldsymbol{\xi}} \in I\!R^p$ is related to the efficient influence function in semiparametric models. If the model is regular and correctly specified $\breve{D}^2$ is the covariance of the efficient influence function and its inverse the semiparametric Cramer-Rao lower bound for regular estimators. The matrices $I\!B, I\!B_{\boldsymbol{\theta}}, I\!B_{\boldsymbol{\eta}}$ describe the miss specification of the model and are related to the White-statistic.

For our estimations we need the constant

$$\mathfrak{z}(\mathtt{x}) \stackrel{\mathrm{def}}{=} \mathfrak{z}(\mathtt{x}, I\!B) \vee \mathfrak{z}_Q(\mathtt{x}, 4p^*) \approx \sqrt{p^* + \mathtt{x}},$$

where $\mathfrak{z}(\mathtt{x}, \cdot)$ is explained in Section 7 and $\mathfrak{z}_Q(\mathtt{x}, \cdot)$ is defined in Equation (8.2).

Remark 2.7. The constant $\mathfrak{z}(\mathtt{x})$ is only introduced for ease of notation. This makes some bounds less sharp but allows to address all terms that are of order $\sqrt{p^* + \mathtt{x}}$ with one symbol. The constant $\mathfrak{z}(\mathtt{x}, I\!B)$ is comparable to the "$1 - \mathrm{e}^{-\mathtt{x}}$ quantile of the norm of $\mathcal{D}^{-1}\mathcal{V}\mathbb{X}$, where $\mathbb{X} \sim \mathcal{N}(0, Id_{p^*})$, i.e. it is of order of the trace of $I\!B$. The constant $\mathfrak{z}_Q(\mathtt{x}, \mathbb{Q})$ arises as an exponential deviation bound for the supremum of a smooth process over a set with complexity described by $\mathbb{Q}$.

To bound the deviations of the points of the sequence $(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)})$ we need the following radius:

$$\mathsf{R}_0(\mathtt{x}, \mathsf{K}_0) \stackrel{\mathrm{def}}{=} \mathfrak{z}(\mathtt{x}) \vee \frac{6\nu_0}{\mathtt{b}(1-\rho)} \sqrt{\mathtt{x} + 2.4p^* + \frac{\mathtt{b}^2}{9\nu_0^2} \mathsf{K}_0(\mathtt{x})}, \tag{2.4}$$

which will ensure $\{\widetilde{\boldsymbol{\upsilon}}_0, \widetilde{\boldsymbol{\upsilon}}_{0,1}, \ldots\} \subset \Upsilon_\circ(\mathsf{R}_0)$, where $\mathsf{K}_0(\mathtt{x}) > 0$ is defined as

$$\mathsf{K}_0(\mathtt{x}) \stackrel{\mathrm{def}}{=} \inf_{K>0} \left\{ I\!P\left(\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \geq -K \right) \geq \beta(\mathtt{x}) \right\},$$

for some $\beta(\mathtt{x}) \to 0$ as $\mathtt{x} \to \infty$, see condition (A_1) in 2.3. Finally define the *parametric uniform spread* and the *semiparametric uniform spread*

$$\begin{aligned} \diamondsuit_Q(\mathtt{r}, \mathtt{x}) &\stackrel{\mathrm{def}}{=} \left\{ \delta(\mathtt{r})\mathtt{r} + 6\nu_1 \omega(\mathfrak{z}_Q(\mathtt{x}, 4p^*) + 2\mathtt{r}^2) \right\}, \\ \breve{\diamondsuit}_Q(\mathtt{r}, \mathtt{x}) &\stackrel{\mathrm{def}}{=} \frac{8}{(1-\rho^2)^2} \breve{\delta}(\mathtt{r})\mathtt{r} + 6\nu_1 \breve{\omega} \left(\mathfrak{z}_Q(\mathtt{x}, 2p^* + 2p)^2 + 2\mathtt{r}^2 \right). \end{aligned} \tag{2.5}$$

Remark 2.8. This object is central to our analysis as it describes the accuracy of our main result of Theorem 2.2. It is small for not too large $\mathtt{r}$, if $\breve{\omega}, \breve{\delta}$ from conditions $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\breve{\mathcal{L}}_0)$ from Section 2.1 are small (with Lemma 2.1 it suffices that ω, δ from $(\mathcal{E}\mathcal{D}_1)$, $(\mathcal{L}_0)$ are small). $\breve{\diamondsuit}_Q(\mathtt{r}, \mathtt{x})$ is structurally slightly different from $\breve{\diamondsuit}(\mathtt{r}, \mathtt{x})$ in [2] as it is based on Theorem 8.2 and allows a üniform in k "formulation of our main result Theorem 2.2, but for moderate $\mathtt{x} \in I\!R_+$ they are of similar size.

2.3 Dependence on initial guess

Our main theorem is only valid under the conditions from Section 2.1 and under some constraints on the quality of the initial guess $\widetilde{\boldsymbol{\upsilon}}_0 \in I\!R^{p^*}$ which we denote by (A_1), (A_2) and (A_3):

$(\mathbf{A}_1)$ With probability greater $1-\beta_{(\mathbf{A})}(\mathtt{x})$ the initial guess satisfies $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0,\boldsymbol{\upsilon}^*) \geq -\mathtt{K}_0(\mathtt{x})$ for some $\mathtt{K}_0(\mathtt{x}) \geq 0$.

$(\mathbf{A}_2)$ The conditions $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\breve{\mathcal{L}}_0)$, $(\mathcal{E}\mathcal{D}_1)$ and $(\mathcal{L}_0)$ from Section 2.1 hold for all $\mathtt{r} \leq \mathtt{R}_0(\mathtt{x},\mathtt{K}_0)$ where $\mathtt{R}_0$ is defined in (2.4) with $\beta(\mathtt{x}) = \beta_{(\mathbf{A})}(\mathtt{x})$.

$(\mathbf{A}_3)$ There is some $\epsilon > 0$ such that $\delta(\mathtt{r})/\mathtt{r} \vee 12\nu_1\omega \leq \epsilon$ for all $\mathtt{r} \leq \mathtt{R}_0$. Further $\mathtt{K}_0(\mathtt{x}) \in \mathbb{R}$ and $\epsilon > 0$ are small enough to ensure

$$c(\epsilon,\mathfrak{z}(\mathtt{x})) \stackrel{\mathrm{def}}{=} \epsilon 7\mathtt{C}(\rho)\frac{1}{1-\rho}\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right) < 1, \tag{2.6}$$

$$c(\epsilon,\mathtt{R}_0) \stackrel{\mathrm{def}}{=} \epsilon 7\mathtt{C}(\rho)\frac{1}{1-\rho}\mathtt{R}_0 < 1, \tag{2.7}$$

with

$$\mathtt{C}(\rho) \stackrel{\mathrm{def}}{=} 2\sqrt{2}(1+\sqrt{\rho})(1-\sqrt{\rho})^{-1}. \tag{2.8}$$

Remark 2.9. One way of obtaining condition (A_1) is to show that $\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(R_K)$ with probability greater $1-\beta_{(\mathbf{A})}(\mathtt{x})$ for some finite $R_K(\mathtt{x}) \in \mathbb{R}$ and $0 \leq \beta_{(\mathbf{A})}(\mathtt{x}) < 1$. Then (see Section 4.1)

$$\mathtt{K}_0(\mathtt{x}) \stackrel{\mathrm{def}}{=} (1/2 + 12\nu_0\omega)R_K^2 + (\delta(R_K) + \mathfrak{z}(\mathtt{x}))R_K + 6\nu_0\omega\mathfrak{z}(\mathtt{x})^2.$$

Condition (A_1) is specified by conditions (A_2) and (A_3) and is fundamental, as it allows with dominating probability to concentrate the analysis on a local set $\Upsilon_\circ\big(\mathtt{R}_0(\mathtt{x})\big)$ (see Theorem 4.3). Conditions (A_2) and (A_3) impose a bound on $\mathtt{R}_0(\mathtt{x})$ and thus on $\mathtt{K}_0$ from (A_1). These conditions boil down to $\delta(\mathtt{R}_0)+\omega\mathtt{R}_0$ being significantly smaller than 1. Condition (A_3) ensures that the quality of the main result from [2] can be attained, i.e. that $\breve{\Diamond}_Q(\mathtt{r}_k,\mathtt{x}) \approx \breve{\Diamond}(\mathtt{r}_0,\mathtt{x})$ under rather mild conditions on the size $\mathtt{R}_0$, as we only need $\epsilon\mathtt{R}_0$ to be small. A violation of (A_2) would make it impossible to apply Theorem 8.1 the backbone of our proofs.

Remark 2.10. In the case of iid observations with sample size n one often has $\delta(\mathtt{R}_0) + \omega\mathtt{R}_0 \leq \mathtt{C}\mathtt{R}_0(\mathtt{x})/\sqrt{n}$ which suggests at first glance that (A_2) and (A_3) are only a question of the sample size. But note that in case of iid observations the functional satisfies $n \approx -\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0,\boldsymbol{\upsilon}^*)$ such that the conditions (A_2) and (A_3) are not satisfied automatically with sufficiently large sample size. They are true conditions on the quality of the first guess.

2.4 Statistical properties of the alternating sequence

In this Section we present our main theorem in full rigor, i.e. that the limit of the alternating sequence satisfies a finite sample Wilks Theorem and Fisher expansion.

Theorem 2.2. *Assume that the conditions* $(\mathcal{E}\mathcal{D}_0), (\mathcal{E}\mathcal{D}_1)$, $(\mathcal{L}_0)$, $(\mathcal{L}_\mathtt{r})$ *and* $(\mathcal{E}\mathtt{r})$ *of Section 2.1 are met with a constant* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *and where* $\mathcal{V}^2 = \mathrm{Cov}\left(\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)\right)$, $\mathcal{D}^2 =$

$-\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)$ *and where* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$ *. Assume that* $(\breve{\mathcal{E}}\mathcal{D}_1)$ *and* $(\breve{\mathcal{L}}_0)$ *are met. Further assume* (B_1) *and that the initial guess satisfies* (A_1) *and* (A_2) *of Section 2.3. Then it holds with probability greater* $1 - 8\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})}$ *for all* $k \in \mathbb{N}$

$$\left\| \breve{D}\left(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\right) - \breve{\boldsymbol{\xi}} \right\| \leq \breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x}), \tag{2.9}$$

$$\left| 2\breve{L}(\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}\|^2 \right| \leq 8 \left(\|\breve{\boldsymbol{\xi}}\| + \breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x}) \right) \breve{\diamondsuit}_Q(2(1+\rho)\mathtt{r}_k, \mathtt{x}) + \breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x})^2, \tag{2.10}$$

where

$$\mathtt{r}_k \leq 2\sqrt{2}(1-\sqrt{\rho})^{-1} \left\{ (\mathfrak{z}(\mathtt{x}) + \diamondsuit_Q(R_0, \mathtt{x})) + (1+\sqrt{\rho})\rho^k R_0(\mathtt{x}) \right\}.$$

If further condition (A_3) *is satisfied then* (2.9) *and* (2.10) *are met with*

$$\mathtt{r}_k \leq \mathtt{C}(\rho) \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right) + \epsilon \frac{7^2 \mathtt{C}(\rho)^4}{1 - c(\epsilon, \mathfrak{z}(\mathtt{x}))} \left(\frac{1}{1-\rho} \right) \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^2 + \rho^k \left(\mathtt{C}(\rho) R_0 + \epsilon \frac{7^2 \mathtt{C}(\rho)^4}{1 - c(\epsilon, R_0)} \left(\frac{1}{\rho^{-1} - 1} \right) R_0^2 \right).$$

In particular this means that if

$$k \geq \frac{2\log(\mathfrak{z}(\mathtt{x})) - \log\{2R_0(\mathtt{x}, K_0)\}}{\log(\rho)},$$

we have with $\mathfrak{z}(\mathtt{x})^2 \leq \mathtt{C}_{\mathfrak{z}}(p^* + \mathtt{x})$

$$\breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x}) \approx \breve{\diamondsuit}_Q\left(\mathtt{C}\sqrt{p^* + \mathtt{x}}, \mathtt{x} \right).$$

Remark 2.11. Note that the results are very similar to those in [2] for the profile M estimator $\widetilde{\boldsymbol{\theta}}$. This is evident after noting that (ignoring terms of the order $\epsilon\mathfrak{z}(\mathtt{x})$)

$$\mathtt{r}_k \lesssim \mathtt{C}(\rho) \left(\mathfrak{z}(\mathtt{x}) + \rho^k (\mathrm{R}_0 + \mathtt{C}\epsilon \mathrm{R}_0^2) \right),$$

which for large $k \in \mathbb{N}$ means $\mathtt{r}_k \lesssim \mathtt{C}(\rho)\mathfrak{z}(\mathtt{x})$.

Remark 2.12. Concerning the properties of $\breve{\boldsymbol{\xi}} \in I\!R^p$ we repeat remark 2.1 of [2]. In the case of the correct model specification the deviation properties of the quadratic form $\|\breve{\boldsymbol{\xi}}\|^2 = \|\breve{D}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\|^2$ are essentially the same as of a chi-square random variable with p degrees of freedom; see Theorem 7.1 in the appendix. In the case of a possible model misspecification with, the behavior of the quadratic form $\|\breve{\boldsymbol{\xi}}\|^2$ will depend on the characteristics of the matrix $\breve{I\!B} \stackrel{\mathrm{def}}{=} \breve{D}^{-1} \mathrm{Cov}(\breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}^*))\breve{D}^{-1}$; see again Theorem 7.1. Moreover, in the asymptotic setup the vector $\breve{\boldsymbol{\xi}}$ is asymptotically standard normal; see Section 2.2. of [2] for the i.i.d. case.

Remark 2.13. These results allow to derive some important corollaries like concentration and confidence sets (see [14], Section 3.2).

Remark 2.14. In general an exact numerical computation of

$$\theta(\boldsymbol{\eta}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in I\!R^p} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}), \text{ or } \eta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\eta} \in I\!R^m} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}),$$

is not possible. Define $\widehat{\theta}(\boldsymbol{\eta})$ and $\widehat{\eta}(\boldsymbol{\theta})$ as the numerical approximations to $\theta(\boldsymbol{\eta})$ and $\eta(\boldsymbol{\theta})$ and assume that

$$\|D(\widehat{\theta}(\boldsymbol{\eta}) - \theta(\boldsymbol{\eta}))\| \leq \tau, \text{ for all } \boldsymbol{\eta} \in \Upsilon_{\circ,\boldsymbol{\eta}}(\mathrm{R}_0) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathrm{R}_0),\, \Pi_{\boldsymbol{\eta}} \boldsymbol{\upsilon} = \boldsymbol{\eta}\},$$
$$\|H(\widehat{\eta}(\boldsymbol{\theta}) - \eta(\boldsymbol{\theta}))\| \leq \tau, \text{ for all } \boldsymbol{\theta} \in \Upsilon_{\circ,\boldsymbol{\theta}}(\mathrm{R}_0) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathrm{R}_0),\, \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} = \boldsymbol{\theta}\}.$$

Then we can easily modify the proof of Theorem 2.2 via adding $\mathtt{C}(\rho)\tau$ to the error terms and the radii $\mathtt{r}_k$, where $\mathtt{C}(\rho)$ is some rational function of ρ.

Remark 2.15. Note that under condition (A_3) the size of $\mathtt{r}_k$ for $k \to \infty$ does not depend on $\mathrm{R}_0 > 0$. So as long as $\epsilon \mathrm{R}_0$ is small enough the quality of the initial guess no longer affects the statistical properties of the sequence $(\boldsymbol{\theta}_k)$ for large $k \in \mathbb{N}$.

2.5 Convergence to the ME

Even though Theorem 2.2 tells us, that the statistical properties of the alternating sequence resemble those of its target, the profile ME, it is an interesting question if the underlying approach allows to qualify conditions under which the sequence actually attains the maximizer $\widetilde{\boldsymbol{\upsilon}}$. Without further assumptions Theorem 2.2 yields the following Corollary:

Corollary 2.3. *Under the assumptions of Theorem 2.2 it holds with probability greater* $1 - 8\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})}$

$$\|\breve{D}(\widetilde{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}_k)\| \leq \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}) + \breve{\Diamond}(\mathtt{r}_0, \mathtt{x}),$$

where $\mathtt{r}_0 > 0$ *is defined in* (2.3) *and*

$$\breve{\Diamond}(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} \frac{8}{(1-\rho^2)^2} \breve{\delta}(\mathtt{r})\mathtt{r} + 6\nu_1 \breve{\omega}_{\mathfrak{z}_1}(\mathtt{x}, 2p^* + 2p)\mathtt{r}.$$

Remark 2.16. The value $\mathfrak{z}_1(\mathtt{x}, \cdot)$ is defined in (2.11).

Corollary 2.3 is a first step in the direction of an actual convergence result but the gap $\breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}) + \breve{\Diamond}(\mathtt{r}_0, \mathtt{x})$ is not a zero sequence in $k \in \mathbb{N}$. It turns out that it is possible to prove convergence to the ME with the cost of assuming more smoothness of the functional $\mathcal{L}$ and using the right bound for the maximal eigenvalue of the hessian $\nabla^2 \mathcal{L}(\boldsymbol{\upsilon}^*)$.

Consider the following condition, that basically quantifies how "well behaved"the second derivative $\nabla^2(\mathcal{L} - I\!E\mathcal{L})$ is:

$(\mathcal{ED}_2)$ There exists a constant $\omega \leq 1/2$, such that for all $|\mu| \leq \mathtt{g}$ and all $0 < \mathtt{r} < \mathtt{r}_0$

$$\sup_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})} \sup_{\|\gamma_1\|=1} \sup_{\|\gamma_2\|=1} \log I\!E \exp\left\{ \frac{\mu \boldsymbol{\gamma}_1^\top \mathcal{D}^{-1} \{\nabla^2 \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla^2 \boldsymbol{\zeta}(\boldsymbol{\upsilon}')\} \boldsymbol{\gamma}_2}{\omega_2 \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|} \right\} \leq \frac{\nu_2^2 \mu^2}{2}.$$

Define $\mathfrak{z}(\mathtt{x}, \nabla^2 \mathcal{L}(\boldsymbol{\upsilon}^*))$ via

$$I\!P\left\{ \|\mathcal{D}^{-1} \nabla^2 \mathcal{L}(\boldsymbol{\upsilon}^*)\| \geq \mathfrak{z}\left(\mathtt{x}, \nabla^2 \mathcal{L}(\boldsymbol{\upsilon}^*)\right) \right\} \leq \mathrm{e}^{-\mathtt{x}},$$

and $\varkappa(\mathtt{x}, \mathtt{R}_0)$

$$\varkappa(\mathtt{x}, \mathtt{R}_0) \stackrel{\mathrm{def}}{=} \frac{2\sqrt{2}(1+\sqrt{\rho})}{\sqrt{1-\rho}} \left[\delta(\mathtt{R}_0) + 9\omega_2 \nu_2 \|\mathcal{D}^{-1}\| \mathfrak{z}_1(\mathtt{x}, 6p^*) \mathtt{R}_0 + \|\mathcal{D}^{-1}\| \mathfrak{z}\left(\mathtt{x}, \nabla^2 \mathcal{L}(\boldsymbol{\upsilon}^*)\right) \right],$$

where $\mathfrak{z}_1(\mathtt{x}, \cdot)$ satisfies (see Theorem 9.2)

$$\mathfrak{z}_1(\mathtt{x}, \mathbb{Q}) = \begin{cases} \sqrt{2(\mathtt{x} + \mathbb{Q})} & \text{if } \sqrt{2(\mathtt{x} + \mathbb{Q})} \leq \mathtt{g}_0, \\ \mathtt{g}_0^{-1}(\mathtt{x} + \mathbb{Q}) + \mathtt{g}_0/2 & \text{otherwise.} \end{cases} \tag{2.11}$$

Remark 2.17. For the case that $\mathcal{L}(\boldsymbol{\upsilon}) = \sum_{i=1}^n \ell_i(\boldsymbol{\upsilon})$ with a sum of independent marginal functionals $\ell_i : \Upsilon \to I\!R$ we can use Corollary 3.7 of [15] to obtain

$$\mathfrak{z}\left(\mathtt{x}, \nabla^2 \mathcal{L}(\boldsymbol{\upsilon}^*)\right) = \sqrt{2\tau} \nu_3 \sqrt{\mathtt{x} + p^*},$$

if with a sequence of matrices $(\boldsymbol{A}_i) \subset I\!R^{p^* \times p^*}$

$$\log I\!E \exp \lambda \nabla^2 \ell_i(\boldsymbol{\upsilon}^*) \preceq \nu_3^2 \lambda^2 / 2 \boldsymbol{A}_i, \quad \| \sum_{i=1}^n \boldsymbol{A}_i \| \leq \tau.$$

Remark 2.18. In the case of smooth i.i.d models this means that $\varkappa(\mathtt{x}, \mathtt{R}_0) \leq \mathtt{C}(\mathtt{R}_0 + \mathtt{x} + \log(p^*))/\sqrt{n} + \mathtt{C}\mathtt{R}_0 \sqrt{\mathtt{x} + p^*}/n$. This means that $\varkappa(\mathtt{x}, \mathtt{R}_0) = O((\mathtt{x} + \mathtt{R}_0 + \log(p^*))/\sqrt{n})$ if $p^* + \mathtt{x} = o(n)$.

With these definitions we can prove the following Theorem:

Theorem 2.4. *Let the conditions* $(\mathcal{ED}_2)$, $(\mathcal{L}_0)$, $(\mathcal{L}_\mathtt{r})$ *and* $(\mathcal{E}\mathtt{r})$ *be met with a constant* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *and where* $\mathcal{D}^2 = -\nabla^2 I\!E \mathcal{L}(\boldsymbol{\upsilon}^*)$ *and* $\boldsymbol{\upsilon}^* = \boldsymbol{\upsilon}^\circ$. *Further suppose* (B_1) *and that the initial guess satisfies* (A_1) *and* (A_2). *Assume that* $\varkappa(\mathtt{x}, \mathtt{R}_0) < (1-\rho)$. *Then*

$$I\!P\left(\bigcap_{k \in \mathbb{N}} \left\{ \boldsymbol{\upsilon}_{k,k(+1)} \in \widetilde{\Upsilon}_\circ(\mathtt{r}_k^*) \right\} \right) \geq 1 - 3\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})},$$

where

$$\mathtt{r}_k^* \leq \begin{cases} \rho^k 2\sqrt{2} \frac{1}{1 - \varkappa(\mathtt{x}, \mathtt{R}_0) k} \widetilde{\mathtt{R}_0}, & \varkappa(\mathtt{x}, \mathtt{R}_0) k \leq 1, \\ 2 \frac{1-\rho}{\varkappa(\mathtt{x}, \mathtt{R}_0)} \tau(\mathtt{x})^{k/\log(k)} \widetilde{\mathtt{R}_0}, & \textit{otherwise,} \end{cases} \tag{2.12}$$

with $\widetilde{R_0} \stackrel{\text{def}}{=} R_0 + \mathtt{r}_0$ *and*

$$\tau(\mathtt{x}) \stackrel{\text{def}}{=} \left(\frac{\varkappa(\mathtt{x}, R_0)}{1-\rho} \right)^{L(k)} < 1$$

$$L(k) \stackrel{\text{def}}{=} \left\lfloor \frac{\log(1/\rho) - \frac{1}{k}\left(\log(2\sqrt{2}) - \log(\varkappa(\mathtt{x}, R_0)k - 1)\right)}{\left(1 + \frac{1}{\log(k)}\log(1-\rho)\right)} \right\rfloor \in \mathbb{N},$$

where $\lfloor x \rfloor \in \mathbb{N}$ *denotes the largest natural number smaller than* $x > 0$ *.*

Remark 2.19. This means that we obtain nearly linear convergence to the global maximizer $\widetilde{\boldsymbol{\upsilon}}$.

Remark 2.20. As in Remark 2.14 if no exact numerical computation of the stepwise maximizers is possible we can easily modify the proof of Theorem 2.4 via adding $\mathtt{C}(\rho)\tau$ to $\varkappa(\mathtt{x}, \mathrm{R}_0)$, to address that case.

2.6 Critical dimension

In parallel to [2] we want to address the issue of *critical parameter dimensions* when the full dimension p^* grows with the sample size n . We write $p^* = p_n$. The results of Theorem 2.2 are accurate if the spread function $\breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x})$ from (2.5) is small. The critical size of p^* then depends on the exact bounds on $\breve{\delta}(\cdot)$ and $\breve{\omega}$. In the i.i.d setting $\breve{\delta}(\mathtt{r})/\mathtt{r} \asymp \breve{\omega} \asymp 1/\sqrt{n}$ such that $\breve{\Diamond}(\mathtt{r}_k, \mathtt{x}) \asymp p^*/\sqrt{n}$ for large $k \in \mathbb{N}$. In other words, one needs that "p^{*2}/n is small" to obtain an accurate non asymptotic version of the Wilks phenomenon and the Fisher Theorem for the limit of the alternating sequence. This is not surprising because good performance of the ME itself can only be guaranteed if "p^{*2}/n is small", as is shown in [2]. There are examples where the pME only satisfies a Wilks- or Fisher result if "p^{*2}/n is small", such that in any of those settings the alternating sequence started in the global maximizer does not admit an accurate Wilks- or Fisher expansion.

Interesting enough the constrain $\varkappa(\mathtt{x}, \mathrm{R}_0) < (1-\rho)$ of Theorem 2.4 for the convergence of the sequence to the global maximizer means that one needs $p^*/n \ll 1$ in the smooth i.i.d. setting if $\mathrm{R}_0 \leq \mathtt{C}_{\mathrm{R}_0}\sqrt{p^* + \mathtt{x}}$. Further Theorem 2.4 states a lower bound for the speed of convergence that in the smooth i.i.d. setting decreases if p^*/n grows. Unfortunately we were unable to find an example that meets the conditions of Section 2.1 and where no convergence occurs if p^*/n tends to infinity. So whether this dimension effect on the convergence is an artifact of our proofs or indeed a property of the alternating procedure remains an open question.

3 Application to single index model

We illustrate how the results of Theorem 2.2 and Theorem 2.4 can be applied in Single Index modeling. Consider the following model

$$y_i = f(\boldsymbol{X}_i^\top \boldsymbol{\theta}^*) + \varepsilon_i, \qquad i = 1, ..., n,$$

for some $f : I\!R \to I\!R$ and $\boldsymbol{\theta}^* \in S_1^{p,+} \subset I\!R^p$ and with i.i.d errors $\varepsilon_i \in I\!R$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and i.i.d random variables $\boldsymbol{X}_i \in I\!R^p$ with distribution denoted by $I\!P^{\boldsymbol{X}}$. The single-index model is widely applied in statistics. For example in econometric studies it serves as a compromise between too restrictive parametric models and flexible but hardly estimable purely nonparametric models. Usually the statistical inference focuses on estimating the index vector $\boldsymbol{\theta}^*$. A lot of research has already been done in this field. For instance, [4] show the asymptotic efficiency of the general semiparametric maximum-functional estimator for particular examples and in [6] the right choice of bandwidth for the nonparametric estimation of the link function is analyzed.

To ensure identifiability of $\boldsymbol{\theta}^* \in I\!R^p$ we assume that it lies in the half sphere $S_1^{p,+} \stackrel{\mathrm{def}}{=} \{\boldsymbol{\theta} \in I\!R^p : \|\boldsymbol{\theta}\| = 1, \theta_1 > 0\} \subset I\!R^p$. For simplicity we assume that the support of the $\boldsymbol{X}_i \in I\!R^p$ is contained in the ball of radius $s_{\boldsymbol{X}} > 0$. This allows to approximate $f \in \{f : [-s_{\boldsymbol{X}}, s_{\boldsymbol{X}}] \mapsto I\!R\}$ by an orthonormal C^2-Daubechies-wavelet basis, i.e. for a suitable function $\boldsymbol{e}_0 \stackrel{\mathrm{def}}{=} \psi : [-s_{\mathbb{X}}, s_{\mathbb{X}}] \mapsto I\!R$ we set for $k = (2^{j_k} - 1)13 + r_k$ with $j_k \in \mathbb{N}_0$ and $r_k \in \{0, \ldots, (2^{j_k})13 - 1\}$

$$\boldsymbol{e}_k(t) = 2^{j_k/2}\psi\left(2^{j_k}(t - 2r_k s_{\mathbb{X}})\right),\ k \in \mathbb{N}.$$

A candidate to estimate $\boldsymbol{\theta}^*$ is the profile ME

$$\widetilde{\boldsymbol{\theta}}_m \stackrel{\mathrm{def}}{=} I\!\Pi_{\boldsymbol{\theta}} \operatorname*{argmax}_{(\boldsymbol{\theta},\boldsymbol{\eta}) \in \Upsilon_m} \mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}),$$

where

$$\mathcal{L}_m(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\frac{1}{2}\sum_{i=1}^{n}\Big|y_i - \sum_{k=0}^{m}\boldsymbol{\eta}_k \boldsymbol{e}_k(\boldsymbol{X}_i^\top \boldsymbol{\theta})\Big|^2.$$

and where $\Upsilon_m \subset S_1^{p,+} \times B_{\mathtt{r}^\circ}^m \subset I\!R^p \times I\!R^m$ where $B_{\mathtt{r}^\circ}^m \subset I\!R^m$ denotes the centered ball of radius $\mathtt{r}^\circ > 0$ for some $\mathtt{r}^\circ > 0$. [8] analyzed a very similar estimator in a more general setting based on a kernel estimation of $I\!E\big[y \,\big|\, f(\boldsymbol{\theta}^\top \boldsymbol{X})\big]$ instead of using a parametric sieve approximation $\sum_{k=0}^m \boldsymbol{\eta}_k \boldsymbol{e}_k$. He showed $\sqrt{n}$-consistency and asymptotic normality of the proposed estimator.

In this setting a direct computation of $\widetilde{\boldsymbol{\upsilon}}$ becomes involved, as the maximization problem is high dimensional and not convex. But as noted in the introduction the maximiziation with respect to $\boldsymbol{\eta}$ for given $\boldsymbol{\theta}$ is high dimensional but convex and consequently feasible. Further for moderate $p \in \mathbb{N}$ the maximization with respect to $\boldsymbol{\theta}$ for fixed $\boldsymbol{\eta}$ is computationally realistic. So an alternating maximization procedure is applicable. To show that it behaves in a desired way we apply the technique presented above.

For the initial guess $\widetilde{\boldsymbol{\upsilon}}_0 \in \Upsilon$ one can use a simple grid search. For this generate a uniform grid $G_N \stackrel{\mathrm{def}}{=} (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \subset S_1^+$ and define

$$\widetilde{\boldsymbol{\upsilon}}_0 \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\substack{(\boldsymbol{\theta},\boldsymbol{\eta}) \in \Upsilon \\ \boldsymbol{\theta} \in G_N}} \mathcal{L}(\boldsymbol{\upsilon}). \tag{3.1}$$

Note that given the grid the above maximizer is easily obtained. Simply calculate

$$\widetilde{\boldsymbol{\eta}}_{0,k} \stackrel{\mathrm{def}}{=} \operatorname{argmax} \mathcal{L}(\boldsymbol{\theta}_k, \boldsymbol{\eta}) = \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{e}\boldsymbol{e}^\top (\boldsymbol{X}_i^\top \boldsymbol{\theta}_k) \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} y_i \boldsymbol{e}^\top (\boldsymbol{X}_i^\top \boldsymbol{\theta}_k) \in I\!R^m \quad (3.2)$$

where by abuse of notation $\boldsymbol{e} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m) \in I\!R^m$. Now observe that

$$\widetilde{\boldsymbol{\upsilon}}_0 = \underset{k=1,\ldots,N}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{\theta}_k, \widetilde{\boldsymbol{\eta}}_{0,k}).$$

Define $\tau \stackrel{\mathrm{def}}{=} \sup_{\boldsymbol{\theta},\boldsymbol{\theta}^\circ \in G_N} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|$.

To apply the result presented in Theorem 2.2 and Theorem 2.4 we need a list of assumptions denoted by $(\mathcal{A})$. We start with conditions on the regressors $\boldsymbol{X} \in I\!R^p$:

$(\mathbf{Cond}_{\boldsymbol{X}})$ The measure $I\!P^{\boldsymbol{X}}$ is absolutely continuous with respect to the Lebesgue measure. The Lebesgue density $d_{\boldsymbol{X}} : I\!R^p \to I\!R$ of $I\!P^{\boldsymbol{X}}$ is only positive on the ball $B_{s_{\boldsymbol{X}}}(0) \subset I\!R^p$ and Lipschitz continuous on $B_{s_{\boldsymbol{X}}}(0) \subset I\!R^p$ with Lipschitz constant $L_{d_{\boldsymbol{X}}} > 0$. Further we assume that for any $\boldsymbol{\theta} \perp \boldsymbol{\theta}^*$ with $\|\boldsymbol{\theta}\| = 1$ we have $\operatorname{Var}\left(\boldsymbol{X}^\top \boldsymbol{\theta} \middle| \boldsymbol{X}^\top \boldsymbol{\theta}^* \right) > \sigma^2_{\boldsymbol{X}|\boldsymbol{\theta}^*}$ for some constant $\sigma^2_{\boldsymbol{X}|\boldsymbol{\theta}^*} > 0$ that does not depend on $\mathbb{X}^\top \boldsymbol{\theta}^* \in I\!R$. Also the density $d_{\mathbb{X}} : I\!R^p \to I\!R$ of the regressors satisfies $c_{d_{\boldsymbol{X}}} \leq d_{\boldsymbol{X}} \leq C_{d_{\boldsymbol{X}}}$ on $B_{s_{\boldsymbol{X}}}(0) \subset I\!R^p$ for constants $0 < c_{d_{\boldsymbol{X}}} \leq C_{d_{\boldsymbol{X}}} < \infty$.

$(\mathbf{Cond}_f)$ For some $\boldsymbol{\eta}^* \in l^2$

$$f = f_{\boldsymbol{\eta}^*} = \sum_{k=1}^{\infty} \eta_k^* \boldsymbol{e}_k,$$

where with some $\alpha > 2$ and a constant $C_{\|\boldsymbol{\eta}^*\|} > 0$

$$\sum_{l=0}^{\infty} l^{2\alpha} \eta_l^{*2} \leq C^2_{\|\boldsymbol{\eta}^*\|} < \infty.$$

$(\mathbf{Cond}_{\boldsymbol{X}\boldsymbol{\theta}^*})$ It holds true that $I\!P(|f'_{\boldsymbol{\eta}^*}(\boldsymbol{X}^\top \boldsymbol{\theta}^*)| > c_{f'_{\boldsymbol{\eta}^*}}) > c_{I\!Pf'}$ for some $c_{f'_{\boldsymbol{\eta}^*}}, c_{I\!Pf'} > 0$.

$(\mathbf{Cond}_\varepsilon)$ The errors $(\varepsilon_i) \in I\!R$ are i.i.d. with $I\!E[\varepsilon_i] = 0$, $\operatorname{Cov}(\varepsilon_i) = \sigma^2$ and satisfy for all $|\mu| \leq \widetilde{g}$ for some $\widetilde{g} > 0$ and some $\widetilde{\nu}_{\mathtt{r}} > 0$

$$\log I\!E[\exp\{\mu \varepsilon_1\}] \leq \widetilde{\nu}_{\mathtt{r}}^2 \mu^2 / 2.$$

If these conditions denoted by $(\mathcal{A})$ are met we can proof the following results:

Proposition 3.1. *Let $\tau = o(p^{*-3/2})$ and $p^{*5}/n \to 0$. With initial guess given by Equation* (3.1) *and for $\mathtt{x} \leq 2\widetilde{\nu}^2 \widetilde{\mathtt{g}}^2 n$ the alternating sequence satisfies* (2.9) *and* (2.10) *with probability greater $1 - 9\exp\{-\mathtt{x}\}$ and where with some constant $\mathtt{C}_\diamond \in I\!R$*

$$\breve{\Diamond}_Q(\mathtt{r}, \mathtt{x}) \leq \frac{\mathtt{C}_\diamond (p^* + \mathtt{x})^{3/2}}{\sqrt{n}} (\mathtt{r}^2 + p^* + \mathtt{x}).$$

Remark 3.1. The constraint $\tau = o(p^{*-3/2})$ implies that for the calculation of the initial guess the vector $\widetilde{\boldsymbol{\eta}}_{0,l}$ of (3.2) and the functional $\mathcal{L}(\cdot)$ have to be evaluated $N = p^{*3(p-1)/2}$ times.

Proposition 3.2. *Take the initial guess given by Equation* (3.1). *Assume* $(\mathcal{A})$ *but use a three times continuously differentiable wavelet basis. Further assume that* $p^{*4}/n \to 0$ *and* $\tau = o(p^{*-3/2})$ *. Let* $\mathtt{x} > 0$ *be chosen such that*

$$\mathtt{x} \leq \frac{1}{2}\left(\widetilde{\nu}^2 n\widetilde{\mathtt{g}}^2 - \log(p^*)\right) \wedge p^*.$$

Then we get the claim of Theorem 2.4 with $\beta_{(\mathbf{A})} = \mathrm{e}^{-\mathtt{x}}$ *and*

$$\varkappa(\mathtt{x}, \mathsf{R}_0) = O(\tau m^{3/2} + \sqrt{\tau\mathtt{x}}m^{3/2}/n^{1/4}) + O(p^{*2}/\sqrt{n}) \to 0,$$

for moderate choice of $\mathtt{x} > 0$ *.*

For details see [1].

4 Proof of Theorem 2.2

In this section we will proof Theorem 2.2. Before we start with the actual proof we want to explain the agenda. The first step of the proof is to find a desirable set $\Omega(\mathtt{x}) \subset \Omega$ of high probability, on which a linear approximation of the gradient of the functional $\mathcal{L}(\boldsymbol{\upsilon})$ can be carried out with sufficient accuracy. Once this set is found all subsequent analysis concerns events in $\Omega(\mathtt{x}) \subset \Omega$.

For this purpose define for some $K \in \mathbb{N}$ the set

$$\begin{aligned}
\Omega(\mathtt{x}) &= \bigcap_{k=0}^{K}(C_{k,k} \cap C_{k,k+1}) \cap C(\nabla) \cap \{\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \geq -\mathsf{K}_0(\mathtt{x})\}, \text{ where} \qquad (4.1)\\
C_{k,k(+1)} &= \Big\{\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \boldsymbol{\upsilon}^*)\| \leq \mathsf{R}_0(\mathtt{x}),\ \|D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)\| \leq \mathsf{R}_0(\mathtt{x}),\\
&\qquad \|H(\widetilde{\boldsymbol{\eta}}_{k(+1)} - \boldsymbol{\eta}^*)\| \leq \mathsf{R}_0(\mathtt{x})\Big\},\\
C(\nabla) &= \bigcap_{\mathtt{r} \leq \mathsf{R}_0(\mathtt{x})}\left\{\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})}\left\{\frac{1}{6\omega\nu_1}\|\mathcal{Y}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2\right\} \leq \mathfrak{z}_Q(\mathtt{x}, 4p^*)^2\right\}\\
&\qquad \bigcap_{\mathtt{r} \leq 4\mathsf{R}_0(\mathtt{x})}\left\{\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})}\left\{\frac{1}{6\breve{\omega}\breve{\nu}_1}\|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2\right\} \leq \mathfrak{z}_Q(\mathtt{x}, 2p^* + 2p)^2\right\}\\
&\qquad \cap\Big\{\max\{\|\mathcal{D}^{-1}\nabla\mathcal{L}\|, \|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}\|, \|H^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}\|\} \leq \mathfrak{z}(\mathtt{x})\Big\}\\
&\qquad \cap\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0(\mathtt{x}))\}.
\end{aligned}$$

For $\boldsymbol{\zeta}(\boldsymbol{\upsilon}) = \mathcal{L}(\boldsymbol{\upsilon}) - I\!\!E\mathcal{L}(\boldsymbol{\upsilon})$ the semiparametric normalized stochastic gradient gap is defined as

$$\breve{\mathcal{Y}}(\boldsymbol{\upsilon}) = \breve{D}^{-1}\Big(\breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \breve{\nabla}_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\Big).$$

the parametric normalized stochastic gradient gap $\mathcal{Y}(\boldsymbol{\upsilon})$ is defined as

$$\mathcal{Y}(\boldsymbol{\upsilon}) = \mathcal{D}_0^{-1}\Big(\nabla\zeta(\boldsymbol{\upsilon}) - \nabla\zeta(\boldsymbol{\upsilon}^*)\Big),$$

and $\mathtt{r}_0(\mathtt{x}) > 0$ is chosen such that $I\!\!P(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)) \geq 1 - \mathrm{e}^{-\mathtt{x}}$, where

$$\widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon}\in\Upsilon \\ \Pi_{\boldsymbol{\theta}}\boldsymbol{\upsilon}=\boldsymbol{\theta}^*}} \mathcal{L}(\boldsymbol{\upsilon}).$$

Remark 4.1. We intersect the set with the event $\{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)\}$ where we a priory demand $\mathtt{r}_0(\mathtt{x}) > 0$ to be chosen such that $I\!\!P(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)) \geq 1 - \mathrm{e}^{-\mathtt{x}}$. Note that condition $(\mathcal{E}\mathtt{r})$ together with $(\mathcal{L}\mathtt{r})$ allow to set $\sqrt{p^* + \mathtt{x}} \approx \mathtt{r}_0 \leq \mathrm{R}_0$ (see Theorem 4.3).

In Section 4.1 we show that this set is of probability greater $1 - 8\mathrm{e}^{-\mathtt{x}} - \beta_{(\boldsymbol{A})}$. We want to explain the purpose of this set along the architecture of the proof of our main theorem.

$\{\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \geq -\mathsf{K}_0(\mathtt{x})\}$**:** This set ensures, that the first guess satisfies $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \geq -\mathrm{K}_0(\mathtt{x})$, which means that it is close enough to the target $\boldsymbol{\upsilon}^* \in I\!R^{p^*}$. This fact allows us to obtain an a priori bound for the deviation of the sequence $(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)}) \subset \Upsilon$ from $\boldsymbol{\upsilon}^* \in \Upsilon_\circ(\mathrm{R}_0)$ with Theorem 4.3.

$\{\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \boldsymbol{\upsilon}^*) \leq \mathsf{R}_0(\mathtt{x})\}$**:** As just mentioned this event is of high probability due to $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \geq -\mathrm{K}_0(\mathtt{x})$ and Theorem 4.3. This allows to concentrate the analysis on the set $\Upsilon_\circ(\mathrm{R}_0)$ on which Taylor expansions of the functional $\mathcal{L} : I\!R^{p^*} \to I\!R$ become accurate.

$C(\nabla)$**:** This set ensures that on $\Omega(\mathtt{x}) \subset \Omega$ all occurring random quadratic forms and stochastic errors are controlled by $\mathfrak{z}(\mathtt{x}) \in I\!R$. Consequently we can derive in the proof of Lemma 4.5 an a priori bound of the form $\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \boldsymbol{\upsilon}^*)\| \leq \mathtt{r}_k$ for a decreasing sequence of radii $(\mathtt{r}_k) \subset I\!R_+$ satisfying $\limsup_{k\to\infty} \mathtt{r}_k = \mathtt{C}\mathfrak{z}(\mathtt{x})$. Further this set allows to obtain in Lemma 4.7 the bounds for all $k \in \mathbb{N}$.

On $\Omega(\mathtt{x}) \subset \Omega$ we find $\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} \in \Upsilon_\circ(\mathtt{r}_k)$ such that we can follow the arguments of Theorem 2.2 of [2] to obtain the desired result with accuracy measured by $\breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x})$.

4.1 Probability of desirable set

Here we show that the set $\Omega(\mathtt{x})$ actually is of probability greater $1 - 8\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})}$. We prove the following two Lemmas, which together yield the claim.

Lemma 4.1. *The set $C(\nabla)$ satisfies*

$$I\!P(C(\nabla)) \geq 1 - 7\mathrm{e}^{-\mathtt{x}}.$$

Proof. The proof is similar to the proof of Theorem 3.1 in [14]. Denote

$$\mathcal{A} \stackrel{\mathrm{def}}{=} \bigcap_{\mathtt{r} \le \mathsf{R}_0(\mathtt{x})} \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \left\{ \frac{1}{6\omega\nu_1} \|\mathcal{Y}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2 \right\} \le \mathfrak{z}_Q(\mathtt{x}, 4p^*)^2 \right\}$$

$$\mathcal{B} \stackrel{\mathrm{def}}{=} \bigcap_{\mathtt{r} \le 4\mathsf{R}_0(\mathtt{x})} \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \left\{ \frac{1}{6\breve{\omega}\breve{\nu}_1} \|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2 \right\} \le \mathfrak{z}_Q(\mathtt{x}, 2p^* + 2p)^2 \right\}$$

$$\mathcal{C} \stackrel{\mathrm{def}}{=} \Big\{ \max\{ \|\mathcal{D}^{-1}\nabla\mathcal{L}\|, \|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}\|, \|H^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}\| \} \le \mathfrak{z}(\mathtt{x}) \Big\}.$$

We estimate

$$\begin{aligned} I\!P(C(\nabla)) \ge 1 &- I\!P\left(\mathcal{A}^c\right) - I\!P\left(\mathcal{B}^c\right) - I\!P\left(\mathcal{C}^c\right) \\ &- I\!P\left(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \notin \Upsilon_\circ(\mathtt{r}_0)\right) - I\!P\left(\|\breve{D}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\|^2 > \mathfrak{z}(\mathtt{x}, \breve{I\!B}_{\boldsymbol{\theta}}) \right). \end{aligned}$$

We bound using for both terms Theorem 8.2 which is applicable due to $(\mathcal{E}\mathcal{D}_1)$ and $(\breve{\mathcal{E}}\mathcal{D}_1)$:

$$I\!P(\mathcal{A}^c) \le \mathrm{e}^{-\mathtt{x}}, \quad I\!P\left(\mathcal{B}^c\right) \le \mathrm{e}^{-\mathtt{x}}.$$

For the set $\mathcal{C} \subset \Omega$ observe that we can use $(\boldsymbol{\mathcal{I}})$ and Lemma 4.2 to find

$$\|H^{-1}\nabla_{\boldsymbol{\eta}}\| \vee \|D^{-1}\nabla_{\boldsymbol{\theta}}\| \le \|\mathcal{D}^{-1}\nabla\|.$$

This implies that

$$\begin{aligned} &\{\|\mathcal{D}^{-1}\nabla\| \le \mathfrak{z}(\mathtt{x}, I\!B)\} \\ &\quad \subseteq \{\|D^{-1}\nabla_{\boldsymbol{\theta}}\| \vee \|H^{-1}\nabla_{\boldsymbol{\eta}}\| \le \mathfrak{z}(\mathtt{x}, I\!B)\}. \end{aligned}$$

Using the deviation properties of quadratic forms as sketched in Section 7 we find

$$I\!P\left(\|\mathcal{D}^{-1}\nabla\| > \mathfrak{z}(\mathtt{x}, I\!B) \right) \le 2\mathrm{e}^{-\mathtt{x}}, \quad I\!P\big(\|\breve{D}^{-1}\breve{\nabla}\| > \mathfrak{z}(\mathtt{x}, \breve{I\!B}) \big) \le 2\mathrm{e}^{-\mathtt{x}}.$$

By the choice of $\mathfrak{z}(\mathtt{x}) > 0$ and $\mathtt{r}_0 > 0$ this gives the claim. □

We cite Lemma B.2 of [2]:

Lemma 4.2. *Let*

$$\mathcal{D}^2 = \begin{pmatrix} D^2 & A \\ A^\top & H^2 \end{pmatrix} \in I\!R^{(p+p)\times(p+p)}, \quad D \in I\!R^{p\times p},\ H \in I\!R^{m\times m} \text{ invertible,}$$

$$\|D^{-1}AH^{-1}\| < 1.$$

Then for any $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in I\!R^{p+m}$ *we have* $\|H^{-1}\boldsymbol{\eta}\| \vee \|D^{-1}\boldsymbol{\theta}\| \le \|\mathcal{D}^{-1}\boldsymbol{\upsilon}\|$.

The next step is to show that the set $\bigcap_{k=1}^K (C_{k,k} \cap C_{k,k+1})$ has high probability, that is independent of the number of necessary steps. A close look at the proof of Theorem 4.1 of [14] shows that it actually yields the following modified version:

Theorem 4.3 ([14], Theorem 4.1)**.** *Suppose* $(\mathcal{E}\mathtt{r})$ *and* $(\mathcal{L}\mathtt{r})$ *with* $\mathtt{b}(\mathtt{r}) \equiv \mathtt{b}$ *. Further define the following random set*

$$\Upsilon(K) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} \in \Upsilon : \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \geq -K\}.$$

If for a fixed $\mathtt{r}_0$ *and any* $\mathtt{r} \geq \mathtt{r}_0$ *, the following conditions are fulfilled:*

$$1 + \sqrt{\mathtt{x} + 2p^*} \leq 3\nu_{\mathtt{r}}^2 \mathtt{g}(\mathtt{r})/\mathtt{b},$$
$$6\nu_{\mathtt{r}} \sqrt{\mathtt{x} + 2p^* + \frac{\mathtt{b}}{9\nu_{\mathtt{r}}^2} K} \leq \mathtt{r}\mathtt{b},$$

then

$$I\!P(\Upsilon(K) \subseteq \Upsilon_\circ(\mathtt{r}_0)) \geq 1 - \mathrm{e}^{-\mathtt{x}}.$$

Note that with $(\mathcal{I})$

$$\|D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)\| \vee \|H(\widetilde{\boldsymbol{\eta}}_{k(+1)} - \boldsymbol{\eta}^*)\| \leq \frac{1}{1-\rho} \|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \boldsymbol{\upsilon}^*)\|.$$

With assumption (B_1) and

$$\mathsf{R}_0(\mathtt{x}) = \frac{6\nu_0}{\mathtt{b}(1-\rho)} \sqrt{\mathtt{x} + \mathbb{Q} + \frac{\mathtt{b}}{9\nu_0^2} \mathsf{K}_0(\mathtt{x})},$$

this implies the desired result as $\mathcal{L}(\boldsymbol{\upsilon}_{k,k(+1)}, \boldsymbol{\upsilon}^*) \geq \mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*)$ such that with Theorem 4.3

$$\begin{aligned} I\!P\left(\bigcap_{k=0}^{K} (C_{k,k} \cap C_{k,k+1})\right) &\geq I\!P\left(\bigcap_{k=0}^{K} (C_{k,k} \cap C_{k,k+1}) \cap \{\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \geq -\mathsf{K}_0\}\right) \\ &\quad - I\!P(\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \leq -\mathsf{K}_0) \\ &\geq I\!P\left\{\Upsilon(\mathsf{K}_0(\mathtt{x})) \subset \Upsilon_\circ\Big((1-\rho)\mathsf{R}_0(\mathtt{x})\Big)\right\} - \beta_{(\boldsymbol{A})} \\ &\geq 1 - \mathrm{e}^{-\mathtt{x}} - \beta_{(\boldsymbol{A})}. \end{aligned}$$

Remark 4.2. This also shows that the sets of maximizers $(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)})$ are nonempty and well defined since the maximization always takes place on compact sets of the form $\{\boldsymbol{\theta} \in I\!R^p,\ (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Upsilon_\circ(\mathrm{R}_0)\}$ or $\{\boldsymbol{\eta} \in I\!R^m,\ (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Upsilon_\circ(\mathrm{R}_0)\}$.

To address the claim of remark 2.9 we present the following Lemma:

Lemma 4.4. *On the set* $C(\nabla) \cap \{\widetilde{\boldsymbol{\upsilon}}_0 \in \Upsilon_\circ(R_K)\}$ *it holds*

$$\mathcal{L}(\boldsymbol{\upsilon}_0, \boldsymbol{\upsilon}^*) \geq -(1/2 + 12\nu_0\omega)R_K^2 - (\delta(R_K) + \mathfrak{z}(\mathtt{x}))R_K - 6\nu_0\omega\mathfrak{z}(\mathtt{x})^2.$$

Proof. With similar arguments as in the proof of Lemma 4.5 we have on $C(\nabla) \subset \Omega$ that

$$\begin{aligned}\mathcal{L}(\boldsymbol{\upsilon}_0, \boldsymbol{\upsilon}^*) &\geq I\!E[\mathcal{L}(\boldsymbol{\upsilon}_0, \boldsymbol{\upsilon}^*)] - \|\mathcal{D}^{-1}\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\|R_K - |\{\nabla\boldsymbol{\zeta}(\widehat{\boldsymbol{\upsilon}}) - \nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\}(\boldsymbol{\upsilon}_0 - \boldsymbol{\upsilon}^*)| \\ &\geq -\|\mathcal{D}(\boldsymbol{\upsilon}_0 - \boldsymbol{\upsilon}^*)\|^2/2 - \|\mathcal{D}^{-1}\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\|R_K \\ &\quad -\|\mathcal{D}^{-1}\big\{\nabla\mathcal{L}(\widehat{\boldsymbol{\upsilon}}) - \nabla\mathcal{L}(\boldsymbol{\upsilon}^*)\big\}\|R_K - R_K\delta(R_K) \\ &\geq -(1/2 + 12\nu_0\omega)R_K^2 - (\delta(R_K) + \mathfrak{z}(\mathtt{x}))R_K - 6\nu_0\omega\mathfrak{z}(\mathtt{x})^2.\end{aligned}$$

□

4.2 Proof convergence

We derive the a priori bound $\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} \in \Upsilon_\circ(\mathtt{r}_k)$ with an adequately decreasing sequence $(\mathtt{r}_k) \subset I\!R_+$ using the argument of Section 1.1, where $\limsup \mathtt{r}_k \approx \mathfrak{z}(\mathtt{x})$.

Lemma 4.5. *Assume that*

$$\Omega(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{\boldsymbol{\upsilon}_{k,k(+1)} \in \Upsilon_\circ\left(\mathtt{r}_k^{(l)}\right)\right\}.$$

Then under the assumptions of Theorem 2.2 we get on $\Omega(\mathtt{x})$ for all $k \in \mathbb{N}_0$

$$\begin{aligned}\left\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \boldsymbol{\upsilon}^*)\right\| \;&\leq\; 2\sqrt{2}(1-\sqrt{\rho})^{-1}\left(\mathfrak{z}(\mathtt{x}) + (1+\sqrt{\rho})\rho^k R_0(\mathtt{x})\right) \\ &\quad +2\sqrt{2}(1+\sqrt{\rho})\sum_{r=0}^{k-1}\rho^r \diamondsuit_Q\left(\mathtt{r}_r^{(l)}\right) \\ &=:\; \mathtt{r}_k^{(l+1)}.\end{aligned}$$

Proof. 1. We first show that on $\Omega(\mathtt{x})$

$$\begin{aligned}D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) &= D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*) - D^{-1}A(\widetilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}^*) + \boldsymbol{\tau}\left(\mathtt{r}_k^{(l)}\right), \\ H(\widetilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}^*) &= H^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}^*) - H^{-1}A^\top(\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*) + \boldsymbol{\tau}\left(\mathtt{r}_k^{(l)}\right),\end{aligned} \tag{4.2}$$

where

$$\|\boldsymbol{\tau}(\mathtt{r})\| \leq \diamondsuit_Q(\mathtt{r}, \mathtt{x}) = \left\{\delta(\mathtt{r})\mathtt{r} + 6\nu_1\omega(\mathfrak{z}_Q(\mathtt{x}, 4p^*) + 2\mathtt{r}^2)\right\}.$$

The proof is the same in each step for both statements such that we only prove the first one. The arguments presented here are similar to those of Theorem D.1 in [2]. By assumption on $\Omega(\mathtt{x})$ we have $\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} \in \Upsilon_\circ\left(\mathtt{r}_k^{(l)}\right)$. Define with $\zeta = \mathcal{L} - I\!E\mathcal{L}$

$$\alpha(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) := \mathcal{L}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - \left(\nabla\zeta(\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2\right).$$

Note that

$$\begin{aligned}\mathcal{L}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) &= \nabla\zeta(\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*) - \|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*)\|^2/2 + \alpha(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*) \\ &= \nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*) - \|D(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2/2 + (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top A(\boldsymbol{\eta}-\boldsymbol{\eta}^*) \\ &\quad +\nabla_{\boldsymbol{\eta}}\zeta(\boldsymbol{\upsilon}^*)(\boldsymbol{\eta}-\boldsymbol{\eta}^*) - \|H(\boldsymbol{\eta}-\boldsymbol{\eta}^*)\|^2/2 + \alpha(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*).\end{aligned}$$

Setting $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\eta}}_k)=0$ we find

$$D(\widetilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}^*) - \mathcal{D}^{-1}\big(\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^*) - A(\widetilde{\boldsymbol{\eta}}_k-\boldsymbol{\eta}^*)\big) = \mathcal{D}^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{\upsilon}}_{k,k},\boldsymbol{\upsilon}^*).$$

As we assume that $\widetilde{\boldsymbol{\upsilon}}_{k,k}\in\Upsilon_\circ(\mathsf{R}_0)$ it suffices to show that with dominating probability

$$\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\in\Upsilon_\circ(\mathsf{R}_0)}\|\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\| \le \diamondsuit(\mathtt{r}_k^{(l)}),$$

where

$$\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k) \stackrel{\mathrm{def}}{=} D^{-1}\big\{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_{k,k}) - \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*) - D^2\,(\boldsymbol{\theta}-\boldsymbol{\theta}^*) - A(\widetilde{\boldsymbol{\eta}}_k-\boldsymbol{\eta}^*)\big\}.$$

To see this note first that with Lemma 4.2 $\|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\boldsymbol{\upsilon}\| \le \|\mathcal{D}^{-1}\mathcal{D}\boldsymbol{\upsilon}\|$. This gives by condition $(\boldsymbol{\mathcal{L}_0})$, Lemma 4.2 and Taylor expansion

$$\begin{aligned}\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\in\Upsilon_\circ(\mathtt{r})}\|I\!\!E\mathcal{U}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\| &\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})}\|D^{-1}\Pi_{\boldsymbol{\theta}}\Big(\nabla I\!\!E\mathcal{L}(\boldsymbol{\upsilon}) - \nabla I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*) - \mathcal{D}\,(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*)\Big)\| \\ &\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})}\|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\|\|\mathcal{D}^{-1}\nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon})^2\mathcal{D}^{-1} - I_{p^*}\|^{1/2}\mathtt{r} \\ &\le \delta(\mathtt{r})\mathtt{r}.\end{aligned}$$

For the remainder note that again with Lemma 4.2

$$\|D^{-1}\Big(\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}) - \nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^*)\Big)\| \le \|\mathcal{D}^{-1}\Big(\nabla\zeta(\boldsymbol{\upsilon}) - \nabla\zeta(\boldsymbol{\upsilon}^*)\Big)\|.$$

This yields that on $\Omega(\mathtt{x})$

$$\begin{aligned}&\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\in\Upsilon_\circ(\mathtt{r})}\Big\|\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k) - I\!\!E\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\Big\| \le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})}\Big\|D^{-1}\Big(\nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}) - \nabla_{\boldsymbol{\theta}}\zeta(\boldsymbol{\upsilon}^*)\Big)\Big\| \\ &\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})}\left\{\frac{1}{6\nu_1\omega}\|\mathcal{Y}(\boldsymbol{\upsilon})\|\right\}6\nu_1\omega \le 6\nu_1\omega\big\{\mathfrak{z}_Q(\mathtt{x},4p^*) + 2\mathtt{r}^2\big\}.\end{aligned}$$

Using the same argument for $\widetilde{\boldsymbol{\eta}}_k$ gives the claim.

2. We prove the apriori bound for the distance of the k. estimator to the oracle

$$\big\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \boldsymbol{\upsilon}^*)\big\| \le \mathtt{r}_k^{(l+1)}.$$

To see this we first use the inequality

$$\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \boldsymbol{\upsilon}^*)\| \leq \sqrt{2}\|D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)\| + \sqrt{2}\|H(\widetilde{\boldsymbol{\eta}}_{k(+1)} - \boldsymbol{\eta}^*)\|.$$

Now we find with (4.2)

$$\begin{aligned}\|D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)\| &\leq \|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\| + \|D^{-1}A(\widetilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}^*)\| + \|\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big)\| \\ &\leq \|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\| + \|D^{-1}AH^{-1}\|\|H(\widetilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}^*)\| + \|\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big)\|.\end{aligned}$$

Next we use that on $\mathit{\Omega}(\mathtt{x})$

$$\|D^{-1}AH^{-1}\| \leq \sqrt{\rho},\ \|D^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\| \leq \mathfrak{z}(\mathtt{x}),\ \|H^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\| \leq \mathfrak{z}(\mathtt{x}),$$

and

$$\|H(\widetilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}^*)\| \leq \|H^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\| + \|H^{-1}A^\top(\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*)\| + \|\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big)\|,$$

to derive the recursive formula

$$\|D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)\| \leq (1+\sqrt{\rho})\left(\mathfrak{z}(\mathtt{x}) + \|\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big)\|\right) + \rho\|D(\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*)\|.$$

Deriving the analogous formula for $\|H(\widetilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}^*)\|$ and solving the recursion gives the claim. □

Lemma 4.6. *Assume the same as in Theorem 2.2. Then we get*

$$\mathit{\Omega}(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{\boldsymbol{\upsilon}_{k,k(+1)} \in \mathit{\Upsilon}_\circ\left(\mathtt{r}_k^{(1)}\right)\right\},$$

where

$$\mathtt{r}_k^{(1)} \leq 2\sqrt{2}(1-\sqrt{\rho})^{-1}\left\{(\mathfrak{z}(\mathtt{x}) + \diamondsuit_Q(R_0,\mathtt{x})) + (1+\sqrt{\rho})\rho^k R_0(\mathtt{x})\right\}. \tag{4.3}$$

Further assume that $\delta(\mathtt{r})/\mathtt{r} \vee 12\nu_1\omega \leq \epsilon$ *and that* (2.6) *and* (2.7) *are met with* $\mathtt{C}(\rho)$ *defined in* (2.8)*. Then*

$$\mathit{\Omega}(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{\boldsymbol{\upsilon}_{k,k(+1)} \in \mathit{\Upsilon}_\circ(\mathtt{r}_k^*)\right\},$$

where

$$\begin{aligned}\mathtt{r}_k^* \leq\ & \mathtt{C}(\rho)\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right) + \epsilon\frac{7^2\mathtt{C}(\rho)^4}{1-c(\epsilon,\mathfrak{z}(\mathtt{x}))}\left(\frac{1}{1-\rho}\right)\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^2 \\ &+\rho^k\left(\mathtt{C}(\rho)R_0 + \epsilon\frac{7^2\mathtt{C}(\rho)^4}{1-c(\epsilon,R_0)}\left(\frac{1}{\rho^{-1}-1}\right)R_0^2\right).\end{aligned} \tag{4.4}$$

Proof. We proof this claim via induction. On $\Omega(\mathtt{x})$ we have

$$\boldsymbol{v}_{k,k(+1)} \in \Upsilon_\circ(\mathsf{R}_0), \quad \text{set } \mathtt{r}_k^{(0)} \stackrel{\text{def}}{=} \mathsf{R}_0.$$

Now with Lemma 4.5 we find that

$$\Omega(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{ \boldsymbol{v}_{k,k(+1)} \in \Upsilon_\circ(\mathtt{r}_k^{(l)}) \right\} \text{ implies } \Omega(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{ \boldsymbol{v}_{k,k(+1)} \in \Upsilon_\circ(\mathtt{r}_k^{(l+1)}) \right\},$$

where

$$\begin{aligned}\mathtt{r}_k^{(l)} \leq\ & 2\sqrt{2}(1-\sqrt{\rho})^{-1}\left(\mathfrak{z}(\mathtt{x}) + (1+\sqrt{\rho})\rho^k \mathsf{R}_0(\mathtt{x})\right) \\ & +2\sqrt{2}(1+\sqrt{\rho})\sum_{r=0}^{k-1}\rho^r \diamondsuit_Q\left(\mathtt{r}_r^{(l-1)}, \mathtt{x}\right).\end{aligned}$$

Setting $l=1$ this gives

$$\mathtt{r}_k^{(1)} \leq 2\sqrt{2}(1-\sqrt{\rho})^{-1}\left\{(\mathfrak{z}(\mathtt{x}) + \diamondsuit_Q(\mathsf{R}_0, \mathtt{x})) + (1+\sqrt{\rho})\rho^k \mathsf{R}_0(\mathtt{x})\right\},$$

which gives (4.3). For the second claim we show that

$$\Omega(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{ \boldsymbol{v}_{k,k(+1)} \in \Upsilon_\circ\left(\limsup_{l\to\infty} \mathtt{r}_k^{(l)}\right)\right\} \subseteq \bigcap_{k\in\mathbb{N}} \left\{ \boldsymbol{v}_{k,k(+1)} \in \Upsilon_\circ(\mathtt{r}_k^*)\right\}.$$

So we have to show that $\limsup_{l\to\infty} \mathtt{r}_k^{(l)} \leq \mathtt{r}_k^*$ from (4.4). For this we use $\delta(\mathtt{r})/\mathtt{r} \vee 12\nu_1\omega \leq \epsilon$ to estimate further

$$\begin{aligned}\mathtt{r}_k^{(l)} &\leq 2\sqrt{2}(1-\sqrt{\rho})^{-1}\left(\mathfrak{z}(\mathtt{x}) + (1+\sqrt{\rho})\rho^k \mathsf{R}_0(\mathtt{x})\right) \\ &\quad +2\sqrt{2}(1+\sqrt{\rho})\epsilon\sum_{r=0}^{k-1}\rho^r\left(\left(\mathtt{r}_{k-r}^{(l-1)}\right)^2 + \mathfrak{z}(\mathtt{x})^2\right) \\ &\leq 2\sqrt{2}(1-\sqrt{\rho})^{-1}\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2 + (1+\sqrt{\rho})\rho^k \mathsf{R}_0(\mathtt{x})\right) \\ &\quad +2\sqrt{2}(1+\sqrt{\rho})\epsilon\sum_{r=0}^{k-1}\rho^r\left(\mathtt{r}_{k-r}^{(l-1)}\right)^2 \\ &\leq \mathtt{C}(\rho)\left\{\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right) + \rho^k \mathsf{R}_0 + \epsilon\sum_{r=0}^{k-1}\rho^r\left(\mathtt{r}_{k-r}^{(l-1)}\right)^2\right\},\end{aligned}$$

where $\mathtt{C}(\rho) > 0$ is defined in (2.8). We set

$$A_{s,k}^{(l)} \stackrel{\text{def}}{=} \sum_{r_1=0}^{k-1}\rho^{r_1}\left(\sum_{r_2=0}^{k-r_1-1}\rho^{r_2}\left(\dots\sum_{r_s=0}^{k-r_1-\dots-r_{s-1}-1}\rho^{r_s}\left(\mathtt{r}_{k-r_1-\dots-r_s}^{(l-1)}\right)^2\dots\right)^2\right)^2.$$

Claim

$$A_{s,k}^{(l)} \leq 7^{\sum_{t=0}^{s-1} 2^t} \mathtt{C}(\rho)^{2^s} \left\{ \left(\frac{1}{1-\rho} \right)^{\sum_{t=0}^{s-1} 2^t} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^{2^s} \right. \tag{4.5}$$
$$\left. + \rho^k \left(\frac{1}{\rho^{-1}-1} \right)^{\sum_{t=0}^{s-1} 2^t} \mathsf{R}_0^{2^s} \right\}$$
$$+ 7^{\sum_{t=0}^{s-1} 2^t} (\mathtt{C}(\rho)\epsilon)^{2^s} A_{s+1,k}^{(l-1)}.$$

We proof this claim via induction. Clearly

$$A_{1,k}^{(l)} = \sum_{r_1=0}^{k-1} \rho^{r_1} \left(\mathtt{r}_{k-r_1}^{(l-1)} \right)^2 \leq 7\mathtt{C}(\rho)^2 \sum_{r_1=0}^{k-1} \rho^{r_1} \left\{ \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^2 + \rho^{2(k-r_1)} \mathsf{R}_0^2 \right\}$$
$$+ 7\mathtt{C}(\rho)^2 \epsilon^2 \sum_{r_1=0}^{k-1} \rho^{r_1} \left(\sum_{r_2=0}^{k-r_1-r_2-1} \rho^{r_2} \left(\mathtt{r}_{k-r_1-r_2}^{(l-2)} \right)^2 \right)^2$$
$$\leq 7\mathtt{C}(\rho)^2 \left\{ \frac{1}{1-\rho} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^2 + \frac{\rho^k}{\rho^{-1}-1} \mathsf{R}_0^2 \right\}$$
$$+ 7\mathtt{C}(\rho)^2 \epsilon^2 A_{2,k}^{(l-1)}.$$

Further

$$A_{s,k}^{(l)} \stackrel{\text{def}}{=} \sum_{r_1=0}^{k-1} \rho^{r_1} \left(\sum_{r_2=0}^{k-r_1-1} \rho^{r_2} \left(\dots \sum_{r_s=0}^{k-r_1-\dots-r_{s-1}-1} \rho^{r_s} \left(\mathtt{r}_{k-r_1-\dots-r_s}^{(l-1)} \right)^2 \dots \right)^2 \right)^2$$
$$= \sum_{r_1=0}^{k-1} \rho^{r_1} \left(A_{s-1,k-r_1}^{(l)} \right)^2. \tag{4.6}$$

Plugging in (4.5) we get for $s \geq 2$

$$A_{s,k}^{(l)} \leq \sum_{r_1=0}^{k-1} \rho^{r_1} \left(7^{\sum_{t=0}^{s-2} 2^t} \mathtt{C}(\rho)^{2^{s-1}} \left\{ \left(\frac{1}{1-\rho} \right)^{\sum_{t=0}^{s-2} 2^t} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^{2^{s-1}} \right. \right.$$
$$\left. + \rho^k \left(\frac{1}{\rho^{-1}-1} \right)^{\sum_{t=0}^{s-2} 2^t} \mathsf{R}_0^{2^{s-1}} \right\}$$
$$\left. + 7^{\sum_{t=0}^{s-2} 2^t} (\mathtt{C}(\rho)\epsilon)^{2^{s-1}} A_{s,k-r_1}^{(l-1)} \right)^2.$$

Shifting the index this gives

$$A_{s,k}^{(l)} \le 7 \sum_{r_1=0}^{k-1} \rho^{r_1} \left(7^{\sum_{t=1}^{s-1} 2^t} \mathtt{C}(\rho)^{2^s} \left\{ \left(\frac{1}{1-\rho} \right)^{\sum_{t=1}^{s-1} 2^{t-1}} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^{2^s} \right. \right.$$
$$\left. + \rho^k \left(\frac{1}{\rho^{-1}-1} \right)^{\sum_{t=1}^{s-1} 2^t} \mathsf{R}_0^{2^s} \right\}$$
$$\left. + 7^{\sum_{t=1}^{s-1} 2^t} (\mathtt{C}(\rho)\epsilon)^{2^s} (A_{s,k-r_1}^{(l-1)})^2 \right).$$

Direct calculation then leads to

$$A_{s,k}^{(l)} \le 7^{\sum_{t=0}^{s-1} 2^t} \mathtt{C}(\rho)^{2^s} \left\{ \left(\frac{1}{1-\rho} \right)^{\sum_{t=0}^{s-1} 2^t} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^{2^s} \right.$$
$$\left. + \rho^k \left(\frac{1}{\rho^{-1}-1} \right)^{\sum_{t=0}^{s-1} 2^t} \mathsf{R}_0^{2^s} \right\}$$
$$+ 7^{\sum_{t=0}^{s-1} 2^t} (\mathtt{C}(\rho)\epsilon)^{2^s} \sum_{r_1=0}^{k-1} \rho^{r_1} (A_{s,k-r_1}^{(l-1)})^2,$$

which gives (4.5) with (4.6). Similarly we can prove

$$A_{s,k}^{(1)} = \left(\frac{1}{1-\rho} \right)^{2^s-1} \mathsf{R}_0^{2^s}.$$

Abbreviate

$$\lambda_s \stackrel{\text{def}}{=} 7^{2^s-1} \mathtt{C}(\rho)^{2^s}, \quad \beta_s \stackrel{\text{def}}{=} 7^{2^s-1} (\mathtt{C}(\rho)\epsilon)^{2^s},$$
$$\mathfrak{z}_s(\mathtt{x}) \stackrel{\text{def}}{=} \left(\frac{1}{1-\rho} \right)^{2^s-1} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^{2^s}, \quad \mathsf{R}_s \stackrel{\text{def}}{=} \left(\frac{1}{\rho^{-1}-1} \right)^{2^s-1} \mathsf{R}_0^{2^s}.$$

Then

$$\mathtt{r}_k^{(l)} \le \mathtt{C}(\rho) \left\{ \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right) + \rho^k \mathsf{R}_0 + \epsilon A_{1,k}^{(l)} \right\}$$
$$\le \sum_{s=0}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathfrak{z}_s(\mathtt{x}) + \rho^k \sum_{s=0}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathsf{R}_s + \prod_{r=0}^{l-1} \beta_r \mathsf{R}_l. \tag{4.7}$$

We estimate further

$$\sum_{s=0}^{l-1}\lambda_s\prod_{r=0}^{s-1}\beta_r\mathfrak{z}_s(\mathtt{x})-\mathtt{C}(\rho)\left(\mathfrak{z}(\mathtt{x})+\epsilon\mathfrak{z}(\mathtt{x})^2\right)=\sum_{s=1}^{l-1}\lambda_s\prod_{r=0}^{s-1}\beta_r\mathfrak{z}_s(\mathtt{x})$$

$$\leq\sum_{s=1}^{l-1}7^{2^s}\mathtt{C}(\rho)^{2^{s+1}}\epsilon^{2^s-1}\left(\frac{1}{1-\rho}\right)^{2^s-1}\left(\mathfrak{z}(\mathtt{x})+\epsilon\mathfrak{z}(\mathtt{x})^2\right)^{2^s}$$

$$=\epsilon 7^2\mathtt{C}(\rho)^4\left(\frac{1}{1-\rho}\right)\left(\mathfrak{z}(\mathtt{x})+\epsilon\mathfrak{z}(\mathtt{x})^2\right)^2\sum_{s=1}^{l-1}\left(\epsilon 7\mathtt{C}(\rho)\frac{1}{1-\rho}\left(\mathfrak{z}(\mathtt{x})+\epsilon\mathfrak{z}(\mathtt{x})^2\right)\right)^{2^s-1}.$$

Assuming (2.6) this gives

$$\sum_{s=0}^{l-1}\lambda_s\prod_{r=0}^{s-1}\beta_r\mathfrak{z}_s(\mathtt{x})\leq\mathtt{C}(\rho)\left(\mathfrak{z}(\mathtt{x})+\epsilon\mathfrak{z}(\mathtt{x})^2\right)$$
$$+\epsilon\frac{7^2\mathtt{C}(\rho)^4}{1-c(\epsilon,\mathfrak{z}(\mathtt{x}))}\left(\frac{1}{1-\rho}\right)\left(\mathfrak{z}(\mathtt{x})+\epsilon\mathfrak{z}(\mathtt{x})^2\right)^2.$$

With the same argument we find under (2.7) that

$$\rho^k\sum_{s=0}^{l-1}\lambda_s\prod_{r=0}^{s-1}\beta_r\mathsf{R}_s\leq\rho^k\left(\mathtt{C}(\rho)\mathsf{R}_0+\epsilon\frac{7^2\mathtt{C}(\rho)^4}{1-c(\epsilon,\mathsf{R}_0)}\left(\frac{1}{\rho^{-1}-1}\right)\mathsf{R}_0^2\right).$$

Additionally (2.7) implies

$$\prod_{r=0}^{l-1}\beta_r\mathsf{R}_l\leq\left(\epsilon 7\mathtt{C}(\rho)\frac{1}{\rho^{-1}-1}\right)^{2^{l-1}}\mathsf{R}_0^{2^l}\to 0.$$

Plugging these bounds into (4.7) and letting $l\to\infty$ gives the claim. □

4.3 Result after convergence

In the previous section we showed that

$$\Omega(\mathtt{x})\subset\bigcap_{\mathtt{r}\leq 4\mathsf{R}_0(\mathtt{x})}\left\{\sup_{\boldsymbol{v}\in\Upsilon_\circ(\mathtt{r})}\left\{\frac{1}{6\breve{\omega}\breve{\nu}_1}\|\breve{\mathcal{Y}}(\boldsymbol{v})\|-2\mathtt{r}^2\right\}\leq\mathfrak{z}_Q(\mathtt{x},2p^*+2p)^2\right\}$$
$$\cap\bigcap_{k\in\mathbb{N}}\left\{\boldsymbol{v}_{k,k}\in\Upsilon_\circ\left(\mathtt{r}_k^{(\cdot)}\right),\boldsymbol{v}_{k,k+1}\in\Upsilon_\circ\left(\mathtt{r}_k^{(\cdot)}\right)\right\}\cap\{\widetilde{\boldsymbol{v}},\widetilde{\boldsymbol{v}}_{\boldsymbol{\theta}^*}\in\Upsilon_\circ(\mathtt{r}_0)\},$$

where $\mathtt{r}_k^{(\cdot)}$ is defined in (4.4) or (4.3). The claim of Theorem 2.2 follows with the following lemma:

Lemma 4.7. *Assume* $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\breve{\mathcal{L}}_0)$, *and* $(\mathcal{I})$ *with a central point* $\boldsymbol{\upsilon}^\circ = \boldsymbol{\upsilon}^*$ *and* $\mathcal{D}^2 = \nabla^2 I\!\!E\mathcal{L}(\boldsymbol{\upsilon}^*)$. *Then it holds on* $\Omega(\mathtt{x}) \subseteq \Omega$ *that for all* $k \in \mathbb{N}$

$$\bigl\| \breve{D}\bigl(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\bigr) - \breve{\boldsymbol{\xi}} \bigr\| \le \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}), \tag{4.8}$$

$$\begin{aligned} \bigl| 2\breve{L}(\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}\|^2 \bigr| &\le 8 \Bigl(\|\breve{D}^{-1}\breve{\nabla}\| + \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}) \Bigr) \breve{\Diamond}_Q(2(1+\rho)\mathtt{r}_k, \mathtt{x}) \\ &\quad + \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x})^2, \end{aligned} \tag{4.9}$$

where the spread $\breve{\Diamond}(\mathtt{r}, \mathtt{x})$ *is defined in* (2.5) *and where*

$$\mathtt{r}_k \stackrel{\mathrm{def}}{=} \mathtt{r}_k^{(\cdot)} \vee \mathtt{r}_0.$$

Proof. The proof is nearly the same as that of Theorem 2.2 of [2] which is inspired by the proof of Theorem 1 of [12]. So we only sketch it and refer the reader to [2] for the skipped arguments. We define

$$l : I\!R^p \times \Upsilon \to I\!R, \;\; (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) \mapsto \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\eta} + H^{-2}A^\top(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)).$$

Note that

$$\nabla_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) = \breve{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\eta} + H^{-2}A^\top(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)), \quad \widetilde{\boldsymbol{\theta}}_k = \operatorname*{argmax}_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_k),$$

such that $\breve{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_k) = 0$. This gives

$$\bigl\| \breve{D}\bigl(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\bigr) - \breve{\boldsymbol{\xi}} \bigr\| = \bigl\| \breve{D}^{-1}\breve{\nabla}\mathcal{L}(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_k) - \breve{D}^{-1}\breve{\nabla}\mathcal{L}(\boldsymbol{\upsilon}^*) + \breve{D}\bigl(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\bigr) \bigr\|.$$

Now the right hand side can be bounded just as in the proof of Theorem 2.2 of [2]. This gives (4.8).

For (4.9) we can represent:

$$\breve{L}(\widetilde{\boldsymbol{\theta}}_k) - \breve{L}(\boldsymbol{\theta}^*) = l(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_{k+1}) - l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}),$$

where

$$\widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*} \stackrel{\mathrm{def}}{=} \Pi_{\boldsymbol{\eta}} \operatorname*{argmax}_{\substack{\boldsymbol{\upsilon} \in \Upsilon, \\ \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon} = \boldsymbol{\theta}^*}} \mathcal{L}(\boldsymbol{\upsilon}).$$

Due to the definition of $\widetilde{\boldsymbol{\theta}}_k$ and $\widetilde{\boldsymbol{\eta}}_{k+1}$

$$l(\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) - l(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^*}) \le \breve{L}(\widetilde{\boldsymbol{\theta}}_k) - \breve{L}(\boldsymbol{\theta}^*) \le l(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_{k+1}) - l(\boldsymbol{\theta}^*, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_{k+1}).$$

Again the remaining steps are exactly the same as in the proof of Theorem 2.2 of [2].

□

5 Proof of Corollary 2.3

Proof. Note that with the argument of Section 4.1 $I\!P(\Omega'(\mathtt{x})) \geq 1 - 8\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})}$ where with $\Omega(\mathtt{x})$ from (4.1)

$$\Omega'(\mathtt{x}) = \Omega(\mathtt{x}) \cap \{\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}_0)\}.$$

On $\Omega'(\mathtt{x})$ it holds due to Theorem 2.2 and due to Theorem 2.1 of [2]

$$\|\breve{D}(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\| \leq \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}), \;\; \|\breve{D}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\| \leq \breve{\Diamond}(\mathtt{r}_0, \mathtt{x}).$$

Now the claim follows with the triangular inequality. □

6 Proof of Theorem 2.4

We prove this Theorem in a similar manner to the convergence result in Lemma 4.5. Redefine the set $\Omega(\mathtt{x})$

$$\begin{aligned}
\Omega(\mathtt{x}) &\stackrel{\text{def}}{=} \bigcap_{k=0}^{K} (C_{k,k} \cap C_{k,k+1}) \cap C(\nabla) \cap \{\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \geq -\mathsf{K}_0(\mathtt{x})\}, \text{ where} \\
C_{k,k(+1)} &= \Big\{\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \boldsymbol{\upsilon}^*)\| \leq \mathsf{R}_0(\mathtt{x}), \, \|D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)\| \leq \mathsf{R}_0(\mathtt{x}), \\
&\qquad \|H(\widetilde{\boldsymbol{\eta}}_{k(+1)} - \boldsymbol{\eta}^*)\| \leq \mathsf{R}_0(\mathtt{x})\Big\}, \\
C(\nabla) &= \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathsf{R}_0(\mathtt{x}))} \|\mathcal{Y}(\nabla^2)(\boldsymbol{\upsilon})\| \leq 9\nu_2\omega_2\mathfrak{z}_1(\mathtt{x}, 6p^*)\mathsf{R}_0(\mathtt{x}) \right\} \\
&\qquad \cap \{\|\mathcal{D}^{-1}\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\| \leq \mathfrak{z}(\mathtt{x}, \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*))\}.
\end{aligned}$$

where

$$\mathcal{Y}(\nabla^2)(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \mathcal{D}^{-1}\left(\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\right) \in I\!R^{p^* 2}.$$

We see that on $\Omega(\mathtt{x})$

$$\boldsymbol{\upsilon}_{k,k(+1)} \in \widetilde{\Upsilon}_\circ(\mathsf{R}_0) \stackrel{\text{def}}{=} \{\|\mathcal{D}(\boldsymbol{\upsilon} - \widetilde{\boldsymbol{\upsilon}})\| \leq \mathsf{R}_0 + \mathtt{r}_0\} \cap \Upsilon_\circ(\mathsf{R}_0).$$

Lemma 6.1. *Under the conditions of Theorem 2.4*

$$I\!P(\Omega(\mathtt{x})) \geq 1 - 3\mathrm{e}^{-\mathtt{x}} - \beta_{(\mathbf{A})}.$$

Proof. The proof is very similar to the one presented in Section 4.1, so we only give a sketch. By assumption

$$I\!P\left(\|\mathcal{D}^{-1}\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\| \leq \mathfrak{z}(\mathtt{x}, \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*))\right) \geq 1 - \mathrm{e}^{-\mathtt{x}},$$

and due to $(\mathcal{ED}_2)$ with Theorem 9.2

$$I\!P\left(\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathsf{R}_0(\mathtt{x}))}\|\mathcal{Y}(\nabla^2)(\boldsymbol{\upsilon})\|\le 9\nu_2\omega_2\mathfrak{z}_1(\mathtt{x},6p^*)\mathsf{R}_0(\mathtt{x})\right)\ge 1-\mathrm{e}^{-\mathtt{x}}.$$

□

Lemma 6.2. *Assume for some sequence* $(\mathtt{r}_k^{(l)})$ *that*

$$\bigcap_{k\in\mathbb{N}}\left\{\left\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)}-\widetilde{\boldsymbol{\upsilon}})\right\|\le\mathtt{r}_k^{(l)}\right\}\subseteq\Omega(\mathtt{x}).$$

Then we get on $\Omega(\mathtt{x})$

$$\begin{aligned}\left\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)}-\widetilde{\boldsymbol{\upsilon}})\right\| &\le 2\sqrt{2}(1+\sqrt{\rho})\sum_{r=0}^{k-1}\rho^r\|\boldsymbol{\tau}(\mathtt{r}_{k-r}^{(l)})\|+2\sqrt{2}\rho^k(\mathsf{R}_0+\mathtt{r}_0),\\ &=:\mathtt{r}_k^{(l+1)}.\end{aligned}\tag{6.1}$$

where

$$\|\boldsymbol{\tau}(\mathtt{r})\|\le\left[\delta(\mathsf{R}_0)+9\nu_2\omega_2\|\mathcal{D}^{-1}\|\mathfrak{z}_1(\mathtt{x},6p^*)\mathsf{R}_0+\|\mathcal{D}^{-1}\|\mathfrak{z}(\mathtt{x},\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*))\right]\mathtt{r}.$$

Proof. 1. We first show that on $\Omega(\mathtt{x})$

$$\begin{aligned}D(\widetilde{\boldsymbol{\theta}}_k-\widetilde{\boldsymbol{\theta}})&=-D^{-1}A(\widetilde{\boldsymbol{\eta}}_k-\widetilde{\boldsymbol{\eta}})+\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big),\\ H(\widetilde{\boldsymbol{\eta}}_k-\boldsymbol{\eta}^*)&=-H^{-1}A^\top(\widetilde{\boldsymbol{\theta}}_{k-1}-\widetilde{\boldsymbol{\theta}})+\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big),\end{aligned}$$

The proof is very similar to that of Lemma 4.5. Define

$$\alpha(\boldsymbol{\upsilon},\widetilde{\boldsymbol{\upsilon}}):=\mathcal{L}(\boldsymbol{\upsilon},\widetilde{\boldsymbol{\upsilon}})+\|\mathcal{D}(\boldsymbol{\upsilon}-\widetilde{\boldsymbol{\upsilon}})\|^2/2.$$

Note that

$$\begin{aligned}\mathcal{L}(\boldsymbol{\upsilon},\widetilde{\boldsymbol{\upsilon}})&=\nabla\mathcal{L}(\boldsymbol{\upsilon})-\|\mathcal{D}(\boldsymbol{\upsilon}-\widetilde{\boldsymbol{\upsilon}})\|^2/2+\alpha(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^*)\\ &=-\|D(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})\|^2/2+(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top A(\boldsymbol{\eta}-\widetilde{\boldsymbol{\eta}})\\ &\quad-\|H(\boldsymbol{\eta}-\widetilde{\boldsymbol{\eta}})\|^2/2+\alpha(\boldsymbol{\upsilon},\widetilde{\boldsymbol{\upsilon}}).\end{aligned}$$

Setting $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\eta}}_k)=0$ we find

$$D(\widetilde{\boldsymbol{\theta}}_k-\widetilde{\boldsymbol{\theta}})=D^{-1}A(\widetilde{\boldsymbol{\eta}}_k-\widetilde{\boldsymbol{\eta}})+D^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{\upsilon}}_{k,k},\widetilde{\boldsymbol{\upsilon}}).$$

We want to show

$$\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\in\widetilde{\Upsilon}_\circ\left(\mathtt{r}_k^{(l)}\right)\cap\Upsilon_\circ(\mathsf{R}_0)}D^{-1}\nabla_{\boldsymbol{\theta}}\alpha((\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k),\widetilde{\boldsymbol{\upsilon}})\le\|\boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big)\|,$$

where

$$D^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\boldsymbol{\upsilon},\widetilde{\boldsymbol{\upsilon}}) \stackrel{\mathrm{def}}{=} D^{-1}\big\{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}) - D^2\,(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}) - A(\widetilde{\boldsymbol{\eta}}_k - \widetilde{\boldsymbol{\eta}})\big\}.$$

To see this note that by assumption we have $\Omega(\mathtt{x}) \subseteq \{\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}_0)\} \subseteq \{\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathsf{R}_0)\}$. By condition $(\boldsymbol{\mathcal{L}_0})$, Lemma 4.2 and Taylor expansion we have

$$\begin{aligned}
&\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\in\widetilde{\Upsilon}_\circ(\mathtt{r}_k^{(l)})\cap\Upsilon_\circ(\mathsf{R}_0)} \|I\!E\mathcal{U}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\| \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\widetilde{\Upsilon}_\circ(\mathtt{r}_k^{(l)})\cap\Upsilon_\circ(\mathsf{R}_0)} \|D^{-1}\Pi_{\boldsymbol{\theta}}\Big(\nabla I\!E\mathcal{L}(\boldsymbol{\upsilon}) - \nabla I\!E\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}) - \mathcal{D}\,(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*)\Big)\| \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathsf{R}_0)} \|D^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\|\|\mathcal{D}^{-1}\nabla^2 I\!E\mathcal{L}(\boldsymbol{\upsilon})\mathcal{D}^{-1} - I_{p^*}\|\mathtt{r}_k^{(l)} \\
&\quad\le \delta(\mathsf{R}_0)\mathtt{r}_k^{(l)}.
\end{aligned}$$

For the remainder note that with $\zeta = \mathcal{L} - I\!E\mathcal{L}$ on $\Omega(\mathtt{x})$ using Lemma 4.2 we can bound

$$\begin{aligned}
&\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\in\widetilde{\Upsilon}_\circ(\mathtt{r}_k^{(l)})\cap\Upsilon_\circ(\mathsf{R}_0)} \Big\|\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k) - I\!E\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\Big\| \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\widetilde{\Upsilon}_\circ(\mathtt{r}_k^{(l)})\cap\Upsilon_\circ(\mathsf{R}_0)} \Big\|D^{-1}\Big(\nabla_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\widetilde{\boldsymbol{\upsilon}})\Big)\Big\| \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \big\|\mathcal{D}^{-1}\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon})\mathcal{D}^{-1}\big\|\,\mathtt{r}_k^{(l)} \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathsf{R}_0)} \left\{\frac{1}{9\nu_2\omega_2}\|\mathcal{D}^{-1}\left(\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\right)\mathcal{D}^{-1}\|\right\}6\nu_1\omega\mathtt{r}_k^{(l)} \\
&\qquad + \left\{\|\mathcal{D}^{-1}\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\mathcal{D}^{-1}\|\right\}\mathtt{r}_k^{(l)} \\
&\quad\le \left[9\nu_2\omega_2\|\mathcal{D}^{-1}\|_{\mathfrak{z}1}(\mathtt{x},6p^*)\mathsf{R}_0 + \|\mathcal{D}^{-1}\|_{\mathfrak{z}}(\mathtt{x},\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*))\right]\mathtt{r}_k^{(l)}.
\end{aligned}$$

Using the same argument for $\widetilde{\boldsymbol{\eta}}_k$ gives the claim.

Now the claim follows as in the proof of Lemma 4.5. □

Lemma 6.3. *Assume that* $\delta(\mathtt{r})/\mathtt{r} \vee 9\nu_2\omega_2 \vee \|\mathcal{D}^{-1}\| \le \epsilon_2$. *Further assume that* $\varkappa(\mathtt{x},\mathsf{R}_0) < 1-\rho$ *where*

$$\varkappa(\mathtt{x},\mathsf{R}_0) \stackrel{\mathrm{def}}{=} \frac{2\sqrt{2}(1+\sqrt{\rho})}{\sqrt{1-\rho}}\Big(\delta(\mathsf{R}_0) + 9\omega_2\nu_2\|\mathcal{D}^{-1}\|_{\mathfrak{z}1}(\mathtt{x},6p^*)\mathsf{R}_0 + \|\mathcal{D}^{-1}\|_{\mathfrak{z}}\left(\mathtt{x},\nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right)\Big).$$

Then

$$\Omega(\mathtt{x}) \subseteq \bigcap_{k\in\mathbb{N}} \left\{ \boldsymbol{\upsilon}_{k,k(+1)} \in \widetilde{\Upsilon}_\circ(\mathtt{r}_k) \right\},$$

where $(\mathtt{r}_k)_{k\in\mathbb{N}}$ *satisfy the bound* (2.12)*.*

Proof. Define for all $k \in \mathbb{N}_0$ the sequence $\mathtt{r}_k^{(0)} = \mathsf{R}_0$. We estimate

$$\|\boldsymbol{\tau}(\mathtt{r}_k^{(l)})\| \leq \frac{1}{\sqrt{1-\rho}} \left(\delta(\mathsf{R}_0) + 6\nu_1\omega_2\|\mathcal{D}^{-1}\|\mathfrak{z}_1(\mathtt{x}, 6p^*)\mathsf{R}_0 + \|\mathcal{D}^{-1}\|\mathfrak{z}(\mathtt{x}, I\!B(\nabla^2)\right) \mathtt{r}_k^{(l)},$$

such that by definition

$$2\sqrt{2}(1+\sqrt{\rho}) \sum_{r=0}^{k-1} \rho^r \|\boldsymbol{\tau}(\mathtt{r}_{k-r}^{(l)})\| \leq \varkappa(\mathtt{x}, \mathsf{R}_0) \sum_{r=0}^{k-1} \rho^r \mathtt{r}_{k-r}^{(l)}.$$

Plugging in the recursive formula for $\mathtt{r}_k^{(l)}$ from (6.1) and denoting $\widetilde{\mathsf{R}_0} \stackrel{\text{def}}{=} \mathsf{R}_0 + \mathtt{r}_0$ we find

$$\begin{aligned}
\mathtt{r}_k^{(l)} &\leq \varkappa(\mathtt{x}, \mathsf{R}_0) \sum_{r=0}^{k-1} \rho^r \mathtt{r}_{k-r}^{(l-1)} + 2\sqrt{2}\rho^k \widetilde{\mathsf{R}_0} \\
&\leq \varkappa(\mathtt{x}, \mathsf{R}_0) \sum_{r=0}^{k-1} \rho^r \left(\varkappa(\mathtt{x}, \mathsf{R}_0) \sum_{s=0}^{k-r-1} \rho^s \mathtt{r}_{k-r-s}^{(l-2)} + 2\rho^{k-r}\widetilde{\mathsf{R}_0} \right) + 2\sqrt{2}\widetilde{\mathsf{R}_0}\rho^k \\
&\leq \varkappa(\mathtt{x}, \mathsf{R}_0)^2 \sum_{r=0}^{k-1} \rho^r \sum_{s=0}^{k-r-1} \rho^s \mathtt{r}_{k-r-s}^{(l-2)} + 2\sqrt{2}\rho^k \widetilde{\mathsf{R}_0} \left(\varkappa(\mathtt{x}, \mathsf{R}_0)k + 1\right) \\
&\leq \varkappa(\mathtt{x}, \mathsf{R}_0)^2 \sum_{r=0}^{k-1} \rho^r \sum_{s=0}^{k-r-1} \rho^s \left(\varkappa(\mathtt{x}, \mathsf{R}_0) \sum_{t=0}^{k-r-s-1} \rho^t \mathtt{r}_{k-r-s-t}^{(l-3)} + 2\rho^{k-r-s}\widetilde{\mathsf{R}_0} \right) \\
&\quad + 2\sqrt{2}\rho^k \widetilde{\mathsf{R}_0} \left(\varkappa(\mathtt{x}, \mathsf{R}_0)k + 1\right) \\
&\leq \varkappa(\mathtt{x}, \mathsf{R}_0)^3 \sum_{r=0}^{k-1} \rho^r \sum_{s=0}^{k-r-1} \rho^s \mathtt{r}_{k-r-s}^{(l-3)} + 2\sqrt{2}\rho^k \widetilde{\mathsf{R}_0} \left(\varkappa(\mathtt{x}, \mathsf{R}_0)^2 k^2 + \varkappa(\mathtt{x}, \mathsf{R}_0)k + 1\right).
\end{aligned}$$

By induction this gives for $l \in \mathbb{N}$

$$\begin{aligned}
\mathtt{r}_k^{(l)} &\le \varkappa(\mathtt{x},\mathsf{R}_0)^l \sum_{r_1=0}^{k-1} \rho^{r_1} \sum_{r_2=0}^{k-r_1-1} \rho^{r_2} \ldots \sum_{r_l=0}^{k-\sum_{s=1}^{l-1} r_s - 1} \rho^{r_l} \widetilde{\mathsf{R}_0} \\
&\quad + 2\sqrt{2}\rho^k \widetilde{\mathsf{R}_0} \sum_{s=0}^{l-1} \varkappa(\mathtt{x},\mathsf{R}_0)^s k^s \\
&\le \left(\left(\frac{\varkappa(\mathtt{x},\mathsf{R}_0)}{1-\rho} \right)^l + 2\sqrt{2}\rho^k \sum_{s=0}^{l-1} (\varkappa(\mathtt{x},\mathsf{R}_0)k)^s \right) \widetilde{\mathsf{R}_0} \\
&\le \begin{cases} \left(\left(\frac{\varkappa(\mathtt{x},\mathsf{R}_0)}{1-\rho} \right)^l + 2\sqrt{2}\rho^k \frac{1}{1-\varkappa(\mathtt{x},\mathsf{R}_0)k} \right) \widetilde{\mathsf{R}_0}, & \varkappa(\mathtt{x},\mathsf{R}_0)k \le 1, \\ \varkappa(\mathtt{x},\mathsf{R}_0)^l \left(\left(\frac{1}{1-\rho} \right)^l + 2\sqrt{2}\rho^k \frac{k^l}{\varkappa(\mathtt{x},\mathsf{R}_0)k-1} \right) \widetilde{\mathsf{R}_0}, & \text{otherwise.} \end{cases}
\end{aligned}$$

By Lemma 6.2

$$\Omega(\mathtt{x}) \subset \bigcap_{k \in \mathbb{N}_0} \bigcap_{l \in \mathbb{N}} \left\{ \widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} \in \widetilde{\Upsilon}_\circ\big(\mathtt{r}_k^{(l)}\big) \right\}.$$

Set if $\varkappa(\mathtt{x},\mathsf{R}_0)/(1-\rho) < 1$

$$l(k) \stackrel{\text{def}}{=} \begin{cases} \infty, & \varkappa(\mathtt{x},\mathsf{R}_0)k \le 1, \\ \frac{k\log(\rho)+\log(2\sqrt{2})-\log(\varkappa(\mathtt{x},\mathsf{R}_0)k-1)}{-\log(1-\rho)-\log(k)}, & \text{otherwise.} \end{cases}$$

Then with $\mathtt{r}_k^* \stackrel{\text{def}}{=} \mathtt{r}_k^{(\lfloor l(k) \rfloor)}$ we get

$$\Omega(\mathtt{x}) \subset \bigcap_{k \in \mathbb{N}_0} \left\{ \widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} \in \widetilde{\Upsilon}_\circ\big(\mathtt{r}_k^*\big) \right\}, \quad \mathtt{r}_k^* \le \begin{cases} \frac{\rho^k 2\sqrt{2}}{1-\varkappa(\mathtt{x},\mathsf{R}_0)k} \widetilde{\mathsf{R}_0}, & \varkappa(\mathtt{x},\mathsf{R}_0)k \le 1, \\ 2\left(\frac{\varkappa(\mathtt{x},\mathsf{R}_0)}{1-\rho} \right)^{\frac{k}{\log(k)} L(k) - 1} \widetilde{\mathsf{R}_0}, & \text{otherwise,} \end{cases}$$

as claimed. □

7 Deviation bounds for quadratic forms

This section is the same as Section A of [2]. The following general result from [14] helps to control the deviation for quadratic forms of type $\|\mathbb{B}\boldsymbol{\xi}\|^2$ for a given positive matrix $\mathbb{B}$ and a random vector $\boldsymbol{\xi}$. It will be used several times in our proofs. Suppose that

$$\log \mathbb{E} \exp\big(\boldsymbol{\gamma}^\top \boldsymbol{\xi}\big) \le \|\boldsymbol{\gamma}\|^2/2, \qquad \boldsymbol{\gamma} \in \mathbb{R}^p, \, \|\boldsymbol{\gamma}\| \le \mathtt{g}.$$

For a symmetric matrix $\mathbb{B}$, define

$$\mathtt{p} = \operatorname{tr}(\mathbb{B}^2), \qquad \mathtt{v}^2 = 2\operatorname{tr}(\mathbb{B}^4), \qquad \lambda^* \stackrel{\text{def}}{=} \|\mathbb{B}^2\|_\infty \stackrel{\text{def}}{=} \lambda_{\max}(\mathbb{B}^2).$$

For ease of presentation, suppose that $\mathtt{g}^2 \geq 2\mathtt{p}_{I\!B}$. The other case only changes the constants in the inequalities. Note that $\|\boldsymbol{\xi}\|^2 = \boldsymbol{\eta}^\top I\!B\,\boldsymbol{\eta}$. Define $\mu_c = 2/3$ and

$$\mathtt{g}_c \stackrel{\mathrm{def}}{=} \sqrt{\mathtt{g}^2 - \mu_c \mathtt{p}_{I\!B}},$$
$$2(\mathtt{x}_c + 2) \stackrel{\mathrm{def}}{=} (\mathtt{g}^2/\mu_c - \mathtt{p}_{I\!B})/\lambda^* + \log\det\bigl(I_p - \mu_c I\!B/\lambda^*\bigr).$$

Proposition 7.1. *Let* (ED_0) *hold with* $\nu_0 = 1$ *and* $\mathtt{g}^2 \geq 2\mathtt{p}_{I\!B}$ *. Then for each* $\mathtt{x} > 0$

$$I\!P\bigl(\|I\!B\boldsymbol{\xi}\| \geq \mathfrak{z}(\mathtt{x}, I\!B)\bigr) \leq 2\mathrm{e}^{-\mathtt{x}},$$

where $\mathfrak{z}(\mathtt{x}, I\!B)$ *is defined by*

$$\mathfrak{z}^2(I\!B, \mathtt{x}) \stackrel{\mathrm{def}}{=} \begin{cases} \mathtt{p}_{I\!B} + 2\mathtt{v}_{I\!B}(\mathtt{x}+1)^{1/2}, & \mathtt{x}+1 \leq \mathtt{v}_{I\!B}/(18\lambda^*), \\ \mathtt{p}_{I\!B} + 6\lambda^*(\mathtt{x}+1), & \mathtt{v}_{I\!B}/(18\lambda^*) < \mathtt{x}+1 \leq \mathtt{x}_c + 2, \\ \bigl|\mathtt{y}_c + 2\lambda^*(\mathtt{x} - \mathtt{x}_c + 1)/\mathtt{g}_c\bigr|^2, & \mathtt{x} > \mathtt{x}_c + 1, \end{cases}$$

with $\mathtt{y}_c^2 \leq \mathtt{p}_{I\!B} + 6\lambda^*(\mathtt{x}_c + 2)$ *.*

8 A uniform bound for the norm of a random process

We want to derive for a random process $\breve{\mathcal{Y}}(\boldsymbol{\upsilon}) \in I\!R^p$ a bound of the kind

$$I\!P\left(\sup_{\mathtt{r} \leq \mathtt{r}^*} \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \left\{ \frac{1}{\omega}\|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2 \right\} \geq \mathtt{C}\mathfrak{z}_Q(\mathtt{x}, p^*) \right) \leq \mathrm{e}^{-\mathtt{x}}.$$

This is a slightly stronger result than the one derived in Section D of [2] but the ideas employed here are very similar.

We want to apply Corollary 2.5 of the supplement of [14] which we cite here as a Theorem. Note that we slightly generalized the formulation of the theorem, to make it applicable in out setting. The proof remains the same.

Theorem 8.1. *Let* $(U(\mathtt{r}))_{0 \leq \mathtt{r} \leq \mathtt{r}^*} \subset I\!R^p$ *be a sequence of balls around* $\boldsymbol{\upsilon}^*$ *induced by the metric* $d(\cdot,\cdot)$ *. Let a random real valued process* $\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon})$ *fulfill for any* $0 \leq \mathtt{r} \leq \mathtt{r}^*$ *that* $\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}^*) = 0$ *and*

$(\boldsymbol{\mathcal{E}d})$ *For any* $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in U(\mathtt{r})$

$$\log I\!E \exp\left\{ \lambda \frac{\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}) - \mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}^\circ)}{d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ)} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \qquad |\lambda| \leq \mathtt{g}. \tag{8.1}$$

Finally assume that $\sup_{\boldsymbol{\upsilon} \in U(\mathtt{r})}(\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}))$ *increases in* $\mathtt{r}$ *. Then with probability greater* $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sup_{\boldsymbol{\upsilon} \in U(\mathtt{r})} \left\{ \frac{1}{3\nu_1}\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}) - d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)^2 \right\} \leq \mathfrak{z}_Q(\mathtt{x}, p^*)^2,$$

where $\mathfrak{z}_Q(\mathtt{x}, p^*) \stackrel{\mathrm{def}}{=} \mathbb{Q}(U(\mathtt{r}^*))$ *denotes the entropy of the set* $U(\mathtt{r}^*) \subset I\!R^p$ *and where with* $\mathtt{g}_0 = \nu_0 \mathtt{g}$ *and for some* $\mathbb{Q} > 0$

$$\mathfrak{z}_Q(\mathtt{x}, \mathbb{Q})^2 \stackrel{\mathrm{def}}{=} \begin{cases} (1 + \sqrt{\mathtt{x} + \mathbb{Q}})^2 & \text{if } 1 + \sqrt{\mathtt{x} + \mathbb{Q}} \le \mathtt{g}_0, \\ 1 + \{2\mathtt{g}_0^{-1}(\mathtt{x} + \mathbb{Q}) + \mathtt{g}_0\}^2 & \text{otherwise.} \end{cases} \tag{8.2}$$

To use this result let $\breve{\mathtt{y}}(\boldsymbol{\upsilon})$ be a smooth centered random vector process with values in $I\!R^p$ and let $\mathcal{D} : I\!R^{p^*} \to I\!R^{p^*}$ be some linear operator. We aim at bounding the maximum of the norm $\|\breve{\mathtt{y}}(\boldsymbol{\upsilon})\|$ over a vicinity $\Upsilon_\circ(\mathtt{r}) \stackrel{\mathrm{def}}{=} \{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}\}$ of $\boldsymbol{\upsilon}^*$. Suppose that $\breve{\mathtt{y}}(\boldsymbol{\upsilon})$ satisfies for each $0 < \mathtt{r} < \mathtt{r}^*$ and for all pairs $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in \Upsilon_\circ(\mathtt{r}) = \bigl\{\boldsymbol{\upsilon} \in \Upsilon \colon \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}\bigr\} \subset I\!R^{p^*}$

$$\sup_{\|\boldsymbol{u}\| \le 1} \log I\!E \exp\biggl\{ \lambda \frac{\boldsymbol{u}^\top \bigl(\breve{\mathtt{y}}(\boldsymbol{\upsilon}) - \breve{\mathtt{y}}(\boldsymbol{\upsilon}^\circ)\bigr)}{\omega \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|} \biggr\} \le \frac{\nu_0^2 \lambda^2}{2}. \tag{8.3}$$

Remark 8.1. In the setting of Theorem 2.2 we have

$$\breve{\mathtt{y}}(\boldsymbol{\upsilon}) = \breve{D}^{-1} \Bigl(\breve{\nabla} \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \breve{\nabla} \boldsymbol{\zeta}(\boldsymbol{\upsilon}^*) \Bigr),$$

and condition (8.3) becomes $(\mathcal{E}\mathcal{D}_1)$ from 2.1.

Theorem 8.2. *Let a random* p*-vector process* $\breve{\mathtt{y}}(\boldsymbol{\upsilon})$ *fulfill* $\breve{\mathtt{y}}(\boldsymbol{\upsilon}^*) = 0$ *and let condition* (8.3) *be satisfied. Then for each* $0 \le \mathtt{r} \le \mathtt{r}^*$*, on a set of probability greater* $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sup_{\mathtt{r} \le \mathtt{r}^*} \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \biggl\{ \frac{1}{6\omega\nu_1} \|\breve{\mathtt{y}}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2 \biggr\} \le \mathfrak{z}_Q(\mathtt{x}, 2p^* + 2p)^2,$$

with $\mathtt{g}_0 = \nu_0 \mathtt{g}$*.*

Remark 8.2. Note that the entropy of the original set $\Upsilon_\circ(\mathtt{r}) \subset I\!R^{p^*}$ is equal to $2p^*$. So in order to control the norm $\|\breve{\mathtt{y}}(\boldsymbol{\upsilon})\|$ one only pays with the additional sumand $2p$.

Proof. In what follows, we use the representation

$$\|\breve{\mathtt{y}}(\boldsymbol{\upsilon})\| = \omega \sup_{\|\boldsymbol{u}\| \le \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|} \frac{1}{\omega \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|} \boldsymbol{u}^\top \breve{\mathtt{y}}(\boldsymbol{\upsilon}).$$

This implies

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \|\breve{\mathtt{y}}(\boldsymbol{\upsilon})\| = \omega \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \sup_{\|\boldsymbol{u}\| \le \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|} \frac{1}{\omega \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|} \boldsymbol{u}^\top \breve{\mathtt{y}}(\boldsymbol{\upsilon}).$$

Due to Lemma 8.3 the process $\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}, \boldsymbol{u}) \stackrel{\mathrm{def}}{=} \frac{1}{\omega \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|} \boldsymbol{u}^\top \breve{\mathtt{y}}(\boldsymbol{\upsilon})$ satisfies condition $(\mathcal{E}d)$ (see (8.1)) as process on $U(\mathtt{r}^*)$ where

$$U(\mathtt{r}) \stackrel{\mathrm{def}}{=} \Upsilon_\circ(\mathtt{r}) \times B_{\mathtt{r}}(0). \tag{8.4}$$

Further $\sup_{(\boldsymbol{v},\boldsymbol{u})\in U(\mathtt{r})} \mathcal{U}(\mathtt{r},\boldsymbol{v},\boldsymbol{u})$ is increasing in $\mathtt{r}$. This allows to apply Theorem 8.2 to obtain the desired result. Set $d((\boldsymbol{v},\boldsymbol{u}),(\boldsymbol{v}^\circ,\boldsymbol{u}^\circ))^2 = \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|^2 + \|\boldsymbol{u}-\boldsymbol{u}^\circ\|^2$. We get on a set of probability greater $1-\mathrm{e}^{-\mathtt{x}}$

$$\sup_{(\boldsymbol{v},\boldsymbol{u})\in U(\mathtt{r}^*)} \left\{ \frac{1}{6\omega\nu_1\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|} \boldsymbol{u}^\top \breve{\mathtt{y}}(\boldsymbol{v}) - \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|^2 - \|\boldsymbol{u}\|^2 \right\}$$
$$\leq \mathfrak{z}_Q\Big(\mathtt{x}, \mathbb{Q}\big(U(\mathtt{r}^*)\big)\Big).$$

The constant $\mathbb{Q}\big(U(\mathtt{r}^*)\big) > 0$ quantifies the complexity of the set $U(\mathtt{r}^*) \subset I\!R^{p^*} \times I\!R^p$. We point out that for compact $M \subset I\!R^{p^*}$ we have $\mathbb{Q}(M) = 2p^*$ (see Supplement of [14], Lemma 2.10). This gives $\mathbb{Q}\big(U\big) = 2p^* + 2p$. Finally observe that

$$\sup_{\mathtt{r}\leq\mathtt{r}^*} \sup_{\boldsymbol{v}\in\Upsilon_\circ(\mathtt{r})} \left\{ \frac{1}{6\omega\nu_1} \|\breve{\mathtt{y}}(\boldsymbol{v})\| - 2\mathtt{r}^2 \right\}$$
$$\leq \sup_{\mathtt{r}\leq\mathtt{r}^*} \sup_{(\boldsymbol{v},\boldsymbol{u})\in U(\mathtt{r})} \left\{ \frac{1}{6\omega\nu_1\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|} \boldsymbol{u}^\top \breve{\mathtt{y}}(\boldsymbol{v}) - \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|^2 - \|\boldsymbol{u}\|^2 \right\}$$
$$= \sup_{(\boldsymbol{v},\boldsymbol{u})\in U(\mathtt{r}^*)} \left\{ \frac{1}{6\omega\nu_1\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|} \boldsymbol{u}^\top \breve{\mathtt{y}}(\boldsymbol{v}) - \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|^2 - \|\boldsymbol{u}\|^2 \right\}.$$

□

Lemma 8.3. *Suppose that* $\breve{\mathtt{y}}(\boldsymbol{v})$ *satisfies for each* $\|\boldsymbol{u}\| \leq 1$ *and* $|\lambda| \leq \mathtt{g}$ *the inequality* (8.3)*. Then the process* $\mathcal{U}(\boldsymbol{v},\boldsymbol{u}) = \frac{1}{2\omega\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|} \breve{\mathtt{y}}(\boldsymbol{v})^\top \boldsymbol{u}_1$ *satisfies* $(\boldsymbol{\mathcal{E}d})$ *from* (8.1) *with* $|\lambda| \leq \mathtt{g}/2$, $d((\boldsymbol{v},\boldsymbol{u}),(\boldsymbol{v}^\circ,\boldsymbol{u}^\circ))^2 = \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|^2 + \|\boldsymbol{u}-\boldsymbol{u}^\circ\|^2$, $\nu = 2\nu_0$ *and* $U \subset I\!R^{p^*+p}$ *defined in* (8.4)*, i.e. for any* $(\boldsymbol{v},\boldsymbol{u}_1),(\boldsymbol{v}^\circ,\boldsymbol{u}_2) \in U$

$$\log I\!E \exp\left\{ \lambda \frac{\mathcal{U}(\boldsymbol{v},\boldsymbol{u}_1) - \mathcal{U}(\boldsymbol{v}^\circ,\boldsymbol{u}_2)}{d((\boldsymbol{v},\boldsymbol{u}_1),(\boldsymbol{v}^\circ,\boldsymbol{u}_2))} \right\} \leq \frac{\nu_0^2\lambda^2}{2}, \qquad |\lambda| \leq \mathtt{g}/2.$$

Proof. Let $(\boldsymbol{v},\boldsymbol{u}_1),(\boldsymbol{v}^\circ,\boldsymbol{u}_2) \in U$ and w.l.o.g. $\boldsymbol{u}_1 \leq \|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\| \leq \|\mathcal{D}(\boldsymbol{v}^\circ-\boldsymbol{v}^*)\|$. By

the Hölder inequality and (8.3), we find

$$
\begin{aligned}
&\log I\!E \exp\biggl\{ \lambda \frac{\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{u}_1) - \mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{u}_2)}{d((\boldsymbol{\upsilon}, \boldsymbol{u}_1), (\boldsymbol{\upsilon}^\circ, \boldsymbol{u}_2))} \biggr\} \\
&\quad = \log I\!E \exp\biggl\{ \lambda \frac{\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{u}_1) - \mathcal{U}(\boldsymbol{\upsilon}^\circ, \boldsymbol{u}_1) + \mathcal{U}(\boldsymbol{\upsilon}^\circ, \boldsymbol{u}_1) - \mathcal{U}(\boldsymbol{\upsilon}^\circ, \boldsymbol{u}_2)}{d((\boldsymbol{\upsilon}, \boldsymbol{u}_1), (\boldsymbol{\upsilon}^\circ, \boldsymbol{u}_2))} \biggr\} \\
&\quad \leq \frac{1}{2} \log I\!E \exp\biggl\{ 2\lambda \frac{\boldsymbol{u}_1^\top \bigl(\frac{1}{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|} \breve{\mathcal{Y}}(\boldsymbol{\upsilon}) - \frac{1}{\|\mathcal{D}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)\|} \breve{\mathcal{Y}}(\boldsymbol{\upsilon}^\circ) \bigr)}{\omega \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|} \biggr\} \\
&\qquad + \frac{1}{2} \log I\!E \exp\biggl\{ 2\lambda \frac{(\boldsymbol{u}_1^\top - \boldsymbol{u}_2^\top) \breve{\mathcal{Y}}(\boldsymbol{\upsilon}^\circ)}{\omega \|\boldsymbol{u}_1 - \boldsymbol{u}_2\| \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|} \biggr\} \\
&\quad \leq \sup_{\|\boldsymbol{u}\| \leq 1} \frac{1}{2} \log I\!E \exp\biggl\{ 2\lambda \frac{\boldsymbol{u}^\top \bigl(\breve{\mathcal{Y}}(\boldsymbol{\upsilon}) - \breve{\mathcal{Y}}(\boldsymbol{\upsilon}^\circ) \bigr)}{\omega \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|} \biggr\} \\
&\qquad + \sup_{\|\boldsymbol{u}\| \leq 1} \frac{1}{2} \log I\!E \exp\biggl\{ 2\lambda \frac{\boldsymbol{u}^\top \bigl(\breve{\mathcal{Y}}(\boldsymbol{\upsilon}^\circ) - \breve{\mathcal{Y}}(\boldsymbol{\upsilon}^*) \bigr)}{\omega \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|} \biggr\} \\
&\quad \leq \frac{4\nu_0^2 \lambda^2}{2}, \qquad \lambda \leq \mathtt{g}/2.
\end{aligned}
$$

□

9 A bound for the sprectal norm of a random matrix process

We want to derive for a random process $\breve{\mathcal{Y}}(\boldsymbol{\upsilon}) \in I\!R^{p^* \times p^*}$ a bound of the kind

$$
I\!P \biggl(\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \Bigl\{ \|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| \Bigr\} \geq \mathtt{C} \omega_2 \mathfrak{z}_1(\mathtt{x}, p^*) \mathtt{r} \biggr) \leq \mathrm{e}^{-\mathtt{x}}.
$$

We derive such a bound in a very similar manner to Theorem E.1 of [2].

We want to apply Corollary 2.2 of the supplement of [14]. Again we slightly generalized the formulation but the proof remains the same.

Corollary 9.1. *Let $(U(\mathtt{r}))_{0 \leq \mathtt{r} \leq \mathtt{r}^*} \subset I\!R^p$ be a sequence of balls around $\boldsymbol{\upsilon}^*$ induced by the metric $d(\cdot, \cdot)$. Let a random real valued process $\mathcal{U}(\boldsymbol{\upsilon})$ fulfill that $\mathcal{U}(\boldsymbol{\upsilon}^*) = 0$ and*

$(\mathcal{E}\boldsymbol{d})$ *For any $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in U(\mathtt{r})$*

$$
\log I\!E \exp\biggl\{ \lambda \frac{\mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}^\circ)}{d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ)} \biggr\} \leq \frac{\nu_0^2 \lambda^2}{2}, \qquad |\lambda| \leq \mathtt{g}. \tag{9.1}
$$

Then for each $0 \leq \mathtt{r} \leq \mathtt{r}^*$, on a set of probability greater $1 - \mathrm{e}^{-\mathtt{x}}$

$$
\sup_{\boldsymbol{\upsilon} \in U(\mathtt{r})} \mathcal{U}(\boldsymbol{\upsilon}) \leq 3\nu_1 \mathfrak{z}_1(\mathtt{x}, p^*)^2 d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*),
$$

where $\mathfrak{z}_1(\mathtt{x}, p^*) \stackrel{\mathrm{def}}{=} \mathbb{Q}(U(\mathtt{r}^*))$ *denotes the entropy of the set* $U(\mathtt{r}^*) \subset I\!R^p$ *and where with* $\mathtt{g}_0 = \nu_0 \mathtt{g}$ *and for some* $\mathbb{Q} > 0$

$$\mathfrak{z}_1(\mathtt{x}, \mathbb{Q}) \stackrel{\mathrm{def}}{=} \begin{cases} \sqrt{2(\mathtt{x} + \mathbb{Q})} & \text{if } \sqrt{2(\mathtt{x} + \mathbb{Q})} \leq \mathtt{g}_0, \\ \mathtt{g}_0^{-1}(\mathtt{x} + \mathbb{Q}) + \mathtt{g}_0/2 & \text{otherwise.} \end{cases}$$

To use this result let $\mathcal{Y}(\boldsymbol{\upsilon})$ be a smooth centered random process with values in $I\!R^{p^* \times p^*}$ and let $\mathcal{D} : I\!R^{p^*} \to I\!R^{p^*}$ be some linear operator. We aim at bounding the maximum of the spectral norm $\|\mathcal{Y}(\boldsymbol{\upsilon})\|$ over a vicinity $\Upsilon_\circ(\mathtt{r}) \stackrel{\mathrm{def}}{=} \{\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*\|_{\mathcal{Y}} \leq \mathtt{r}\}$ of $\boldsymbol{\upsilon}^*$. Suppose that $\mathcal{Y}(\boldsymbol{\upsilon})$ satisfies $\mathcal{Y}(\boldsymbol{\upsilon}^*) = 0$ and for each $0 < \mathtt{r} < \mathtt{r}^*$ and for all pairs $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ \in \Upsilon_\circ(\mathtt{r}) = \bigl\{\boldsymbol{\upsilon} \in \Upsilon \colon \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*\|_{\mathcal{Y}} \leq \mathtt{r}\bigr\} \subset I\!R^{p^*}$

$$\sup_{\|\boldsymbol{u}_1\| \leq 1} \sup_{\|\boldsymbol{u}_2\| \leq 1} \log I\!E \exp\biggl\{ \lambda \frac{\boldsymbol{u}_1^\top \bigl(\mathcal{Y}(\boldsymbol{\upsilon}) - \mathcal{Y}(\boldsymbol{\upsilon}^\circ)\bigr) \boldsymbol{u}_2}{\omega_2 \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|} \biggr\} \leq \frac{\nu_2^2 \lambda^2}{2}. \tag{9.2}$$

Remark 9.1. In the setting of Theorem 2.4 we have $\|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|_{\mathcal{Y}} = \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|$ and

$$\mathcal{Y}(\boldsymbol{\upsilon}) = \mathcal{D}^{-1} \nabla^2 \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \mathcal{D}^{-1} \nabla^2 \boldsymbol{\zeta}(\boldsymbol{\upsilon}^*),$$

and condition (9.2) becomes $(\mathcal{E}\mathcal{D}_2)$ from 2.1.

Theorem 9.2. *Let a random process* $\mathcal{Y}(\boldsymbol{\upsilon}) \in I\!R^{p^* \times p^*}$ *fulfill* $\mathcal{Y}(\boldsymbol{\upsilon}^*) = 0$ *and let condition* (9.2) *be satisfied. Then for each* $0 \leq \mathtt{r} \leq \mathtt{r}^*$*, on a set of probability greater* $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{\upsilon})\| \leq 9 \omega_2 \nu_2 \mathfrak{z}_1(\mathtt{x}, 6p^*) \mathtt{r},$$

with $\mathtt{g}_0 = \nu_0 \mathtt{g}$*.*

Remark 9.2. Note that the entropy of the original set $\Upsilon_\circ(\mathtt{r}) \subset I\!R^{p^*}$ is multiplied by 3. So in order to control the spectral norm $\|\mathcal{Y}(\boldsymbol{\upsilon})\|$ one only pays with this factor.

Proof. In what follows, we use the representation

$$\|\mathcal{Y}(\boldsymbol{\upsilon})\| = \omega_2 \sup_{\|\boldsymbol{u}_2\| \leq \mathtt{r}} \sup_{\|\boldsymbol{u}_2\| \leq \mathtt{r}} \frac{1}{\omega_2 \mathtt{r}^2} \boldsymbol{u}_1^\top \breve{\mathcal{Y}}(\boldsymbol{\upsilon}) \boldsymbol{u}_2.$$

This implies

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{\upsilon})\| = \omega \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \sup_{\|\boldsymbol{u}_2\| \leq \mathtt{r}} \sup_{\|\boldsymbol{u}_2\| \leq \mathtt{r}} \frac{1}{\omega \mathtt{r}^2} \boldsymbol{u}_1^\top \breve{\mathcal{Y}}(\boldsymbol{\upsilon}) \boldsymbol{u}_2.$$

Due to Lemma 9.3 the process $\mathcal{U}(\boldsymbol{\upsilon}) \stackrel{\mathrm{def}}{=} \frac{1}{\omega \mathtt{r}^2} \boldsymbol{u}_1^\top \mathcal{Y}(\boldsymbol{\upsilon}) \boldsymbol{u}_2$ satisfies condition $(\mathcal{E}d)$ (see (9.1)) as process on

$$U(\mathtt{r}) \stackrel{\mathrm{def}}{=} \Upsilon_\circ(\mathtt{r}) \times B_{\mathtt{r}}(0) \times B_{\mathtt{r}}(0) \subset I\!R^{3p^*}. \tag{9.3}$$

This allows to apply Corollary 9.1 to obtain the desired result. We get on a set of probability greater $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sup_{\boldsymbol{v} \in \Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{v})\| \leq \sup_{(\boldsymbol{v},\boldsymbol{u}_1,\boldsymbol{u}_2) \in U(\mathtt{r})} \left\{ \frac{1}{\mathtt{r}^2} \boldsymbol{u}_1^\top \mathcal{Y}(\boldsymbol{v}) \boldsymbol{u}_2 \right\} \leq 9\omega_2 \nu_2 \mathfrak{z}_1 \Bigl(\mathtt{x}, \mathbb{Q}\bigl(U(\mathtt{r}^*)\bigr) \Bigr) \mathtt{r}.$$

The constant $\mathbb{Q}\bigl(U(\mathtt{r})\bigr) > 0$ quantifies the complexity of the set $U(\mathtt{r}) \subset I\!R^{3p^*}$. We point out that for compact $M \subset I\!R^{3p^*}$ we have $\mathbb{Q}(M) = 6p^*$ (see Supplement of [14], Lemma 2.10). This gives the claim. □

Lemma 9.3. *Suppose that* $\mathcal{Y}(\boldsymbol{v}) \in I\!R^{p^* \times p^*}$ *satisfies* $\mathcal{Y}(\boldsymbol{v}^*) = 0$ *and for each* $\|\boldsymbol{u}_1\| \leq 1$, $\|\boldsymbol{u}_2\| \leq 1$ *and* $|\lambda| \leq \mathtt{g}$ *the inequality* (9.2)*. Then the process*

$$\mathcal{U}(\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{u}_2) = \frac{1}{2\omega_2 \mathtt{r}^2} \boldsymbol{u}_1^\top \mathcal{Y}(\boldsymbol{v})^\top \boldsymbol{u}_2$$

satisfies $(\boldsymbol{\mathcal{E}d})$ *from* (9.1) *with* $U \subset I\!R^{3p^*}$ *defined in* (9.3)*, with* $|\lambda| \leq \mathtt{g}/3$ *and with*

$$d((\boldsymbol{v}, \boldsymbol{u}), (\boldsymbol{v}^\circ, \boldsymbol{u}^\circ))^2 = \|\mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^*)\|^2 + \|\boldsymbol{u}_1 - \boldsymbol{u}_1^\circ\|^2 + \|\boldsymbol{u}_2 - \boldsymbol{u}_2^\circ\|^2,$$

i.e. for any $(\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{u}_2), (\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ) \in U$

$$\log I\!E \exp\left\{ \lambda \frac{\mathcal{U}(\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{u}_2) - \mathcal{U}(\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ)}{d((\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{v}_2), (\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ))} \right\} \leq \frac{9\nu_2^2 \lambda^2}{2}, \qquad |\lambda| \leq \mathtt{g}/3.$$

Proof. Let $(\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{u}_2), (\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ) \in U$. By the Hölder inequality and (9.2), we find

$$
\begin{aligned}
&\log I\!\!E \exp\biggl\{ \lambda \frac{\mathcal{U}(\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{u}_2) - \mathcal{U}(\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ)}{d((\boldsymbol{v}, \boldsymbol{u}_1, \mathfrak{u}_2), (\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ))} \biggr\} \\
&\quad = \log I\!\!E \exp\biggl\{ \lambda \biggl(\frac{\mathcal{U}(\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{u}_2) - \mathcal{U}(\boldsymbol{v}^\circ, \boldsymbol{u}_1, \boldsymbol{u}_2)}{d((\boldsymbol{v}, \boldsymbol{u}_1, \mathfrak{u}_2), (\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ))} + \frac{\mathcal{U}(\boldsymbol{v}^\circ, \boldsymbol{u}_1, \boldsymbol{u}_2) - \mathcal{U}(\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2)}{d((\boldsymbol{v}, \boldsymbol{u}_1, \boldsymbol{v}_2), (\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ))} \\
&\qquad + \frac{\mathcal{U}(\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2) - \mathcal{U}(\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ)}{d((\boldsymbol{v}, \boldsymbol{u}_1, \mathfrak{u}_2), (\boldsymbol{v}^\circ, \boldsymbol{u}_1^\circ, \boldsymbol{u}_2^\circ))} \biggr) \biggr\} \\
&\quad \le \frac{1}{3} \log I\!\!E \exp\biggl\{ 3\lambda \frac{\boldsymbol{u}_1^\top \bigl(\frac{1}{\mathtt{r}^2} \breve{\mathcal{Y}}(\boldsymbol{v}) - \frac{1}{\mathtt{r}^2} \breve{\mathcal{Y}}(\boldsymbol{v}^\circ) \bigr) \boldsymbol{u}_2}{\omega_2 \| \mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^\circ) \|} \biggr\} \\
&\qquad + \frac{1}{3} \log I\!\!E \exp\biggl\{ 3\lambda \frac{(\boldsymbol{u}_1 - \boldsymbol{u}_1^\circ)^\top) \mathcal{Y}(\boldsymbol{v}^\circ) \boldsymbol{u}_2}{\omega_2 \| \boldsymbol{u}_1 - \boldsymbol{u}_2 \| \mathtt{r}^2} \biggr\} \\
&\qquad + \frac{1}{3} \log I\!\!E \exp\biggl\{ 3\lambda \frac{(\boldsymbol{u}_1^\circ)^\top) \mathcal{Y}(\boldsymbol{v}^\circ) (\boldsymbol{u}_2 - \boldsymbol{u}_2^\circ)}{\omega_2 \| \boldsymbol{u}_1 - \boldsymbol{u}_2 \| \mathtt{r}^2} \biggr\} \\
&\quad \le \frac{1}{3} \sup_{\|\boldsymbol{u}_1\| \le 1} \sup_{\|\boldsymbol{u}_2\| \le 1} \log I\!\!E \exp\biggl\{ 3\lambda \frac{\boldsymbol{u}_1^\top \bigl(\mathcal{Y}(\boldsymbol{v}) - \mathcal{Y}(\boldsymbol{v}^\circ) \bigr) \boldsymbol{u}_2}{\omega_2 \| \mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^\circ) \|} \biggr\} \\
&\qquad + \frac{2}{3} \sup_{\|\boldsymbol{u}_1\| \le 1} \sup_{\|\boldsymbol{u}_2\| \le 1} \log I\!\!E \exp\biggl\{ 3\lambda \frac{\boldsymbol{u}_1^\top \bigl(\mathcal{Y}(\boldsymbol{v}^\circ) - \mathcal{Y}(\boldsymbol{v}^*) \bigr) \boldsymbol{u}_2}{\omega_2 \| \mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^*) \|} \biggr\} \\
&\quad \le \frac{9\nu_2^2 \lambda^2}{2}, \qquad \lambda \le \mathtt{g}/3.
\end{aligned}
$$

□

References

[1] Andresen, A. (2014). Finite sample analysis of profile m-estimation in the single index model. *arXiv:1406.4052*.

[2] Andresen, A. and Spokoiny, V. (2014). Critical dimension in profile semiparametric estimation. *arXiv:1303.4640*.

[3] Balakrishnan, S., Wainwright, M. J., and Yu, B. (2014). Satistical guarantees for the em aogorithm: From population to sample-based analysis. *arXiv: 1408.2156*.

[4] Delecroix., M., Haerdle, W., and Hristache, M. (1997). Efficient estimation in single-index regression. Technical report, SFB 373, Humboldt Univ. Berlin.

[5] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society,*, Series B, 39:1–38.

[6] Haerdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.*, 21:157–178.

[7] Ibragimov, I. and Khas'minskij, R. (1981). *Statistical estimation. Asymptotic theory. Transl. from the Russian by Samuel Kotz.* New York - Heidelberg -Berlin: Springer-Verlag .

[8] Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single- index models. *J Econometrics*, 58:71–120.

[9] Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. *STOC*, pages 665–674.

[10] Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from few entries. *IEEE Transactions onInformation Theory*, 56(6):2980–2998.

[11] McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.

[12] Murphy, S. A. and Van der Vaart, A. W. (1999). Observed information in semi-parametric models. *Bernoulli*, 5(3):381–412.

[13] Netrapalli, P., Jain, P., and Sanghavi, S. (2013). Phase retrieval using alternating minimization. *NIPS*, pages 2796–2804.

[14] Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6):2877–2909. arXiv:1111.3029.

[15] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Compution Mathematics*, 12:389–434.

[16] Wu, C. (1983). On the convergence properties of the em algorithm. *Annals of Statistics*, 11:95–103.

[17] Yi, X., Caramanis, C., and Sanghavi, S. (2013). Alternating minimization for mixed linear regression. *arXiv: 1310.3745*.