Journal of Machine Learning Research 17 (2016) 1-53 Submitted 7/15; Revised 10/15; Published 4/16

Convergence of an Alternating Maximization Procedure

Andreas Andresen ANDREAS.ANDRESEN@POSTEO.DE

Weierstrass-Institute,
Mohrenstr. 39,
10117 Berlin, Germany

Vladimir Spokoiny SPOKOINY@WIAS-BERLIN.DE

Weierstrass-Institute and Humboldt University Berlin,
Higher School of Economics, IITP RAN, MIPT Moscow,
Mohrenstr. 39, 10117 Berlin, Germany

Editor: Jie Peng

Abstract

We derive two convergence results for a sequential alternating maximization 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 by Andresen and Spokoiny (2013), that means 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.

Keywords: alternating maximization, alternating minimization, profile maximum likelihood, EM-algorithm, M-estimation, local linear approximation, local concentration, semiparametric

Contents

1. Introduction

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

$$\boldsymbol{\theta}^* = \Pi_{\boldsymbol{\theta}} \boldsymbol{\upsilon}^* = \Pi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\boldsymbol{\upsilon}} \mathbb{E}_{\mathbb{P}} \mathcal{L}(\boldsymbol{\upsilon}, \mathbb{Y}), \tag{1}$$

where $\mathcal{L} : \Upsilon \times \mathcal{Y} \to \mathbb{R}$ is an appropriate functional, $\Pi_{\boldsymbol{\theta}} : \Upsilon \to \mathbb{R}^p$ is a projection and where Υ is some high dimensional or even infinite dimensional parameter space. A prominent way of estimating $\boldsymbol{\theta}^*$ is the profile M-estimator (pME)

$$\widetilde{\boldsymbol{\theta}} \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \widetilde{\boldsymbol{\upsilon}} \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \operatorname*{argmax}_{(\boldsymbol{\theta},\boldsymbol{\eta})} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}). \tag{2}$$

This paper focuses on finite dimensional parameter spaces $\Upsilon \subseteq \mathbb{R}^{p^*}$ with $p^* = p + m \in \mathbb{N}$ being the full dimension, as infinite dimensional maximization problems are computationally not feasible. This is motivated by the sieve M-estimation technique, which projects the estimation problem to a finite dimensional submodel - see Section 1.2 for details.

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

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

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

$$\begin{aligned} \widetilde{\boldsymbol{\upsilon}}_{k,k+1} &\overset{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_{k+1}) = \left(\widetilde{\boldsymbol{\theta}}_k, \operatorname*{argmax}_{\boldsymbol{\eta} \in \mathbb{R}^m} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\eta}) \right), \\ \widetilde{\boldsymbol{\upsilon}}_{k,k} &\overset{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_k) = \left(\operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_k), \widetilde{\boldsymbol{\eta}}_k \right). \end{aligned} \tag{3}$$

The so called "alternating maximization procedure" (or minimization) is a widely applied algorithm in many parameter estimation tasks (see Jain, Netrapalli, and Sanghavi, 2013; Netrapalli, Jain, and Sanghavi, 2013; Keshavan, Montanari, and Oh, 2010; Yi, Caramanis, and Sanghavi, 2013). 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 again (see Jain, Netrapalli, and Sanghavi, 2013; Netrapalli, Jain, and Sanghavi, 2013; Keshavan, Montanari, and Oh, 2010; Yi, Caramanis, and Sanghavi, 2013) or variants of the procedure (Cheng, 2013).

We address this difficulty via employing new finite sample techniques (Andresen and Spokoiny, 2014; Spokoiny, 2012), 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 7 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 alternating maximization procedure can be understood as a special case of the Expectation Maximization algorithm (EM algorithm) as we illustrate in Section 1.1. The EM algorithm itself was derived in a work of Dempster, Laird, and Rubin (1977) where particular versions of this approach are generalized. This paper (Dempster, Laird, and Rubin, 1977) also contains a variety of problems where an application of the EM algorithm can be fruitful; for a brief history of the EM algorithm (see McLachlan and Krishnan, 1997, Section 1.8). We briefly explain the EM algorithm in Section 1.1.

Since the EM algorithm is very popular in applications a lot of research on its behavior 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 Wu Wu (1983), which is still state of the art in most settings. Unfortunately Wu's result - as most convergence results on these iterative procedures - only ensures convergence to some set of local maximizers or fixpoints of the procedure. Only in very special cases like unimodality can actual convergence to the maximizer be ensured.

In a recent work (Balakrishnan, Wainwright, and Yu, 2014) a new way is presented of addressing the properties of the EM sequence in a very general i.i.d. setting, based on concavity of $\mathcal{L}$. They assume that the functional $\mathcal{L}$ is concave and smooth enough (First order stability) and that for a sample $(\boldsymbol{Y}_i)_{i=1,\dots,n}$ with high probability an uniform bound is satisfied of the kind

$$\max_{\boldsymbol{\theta}^\circ \in B_{\mathtt{r}}(\boldsymbol{\theta}^*)} \left| \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^\circ}) - \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}\mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}^\circ}) \right| \leq \epsilon_n. \tag{4}$$

Under these assumptions, with high probability and some $\nu < 1$ they show

$$\|\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\| \leq \nu^k \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\| + \mathtt{C}\epsilon_n. \tag{5}$$

Unfortunately this does not answer our two questions to full satisfaction. First the bound (4) 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 (5) 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 $\widetilde{\boldsymbol{\theta}}_k \to \widetilde{\boldsymbol{\theta}}$ is not addressed.

These results 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.

Another new work (Cheng, 2013) contains the analysis of a slightly altered algorithm in a very general semiparametric asymptotic framework. Instead of alternatingly maximizing the functional $\mathcal{L}$ a kind of gradient decent procedure for the profile likelihood $pl(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta})$ is analyzed, i.e. they define

$$\boldsymbol{\theta}_k \stackrel{\text{def}}{=} \boldsymbol{\theta}_{k-1} + \widehat{\mathrm{D}}(\boldsymbol{\theta}_{k-1})^{-2} \ell(\boldsymbol{\theta}_{k-1}), \tag{6}$$

where $\widehat{\boldsymbol{\eta}}(\cdot)$ is an estimator of $\operatorname{argmax}_{\boldsymbol{\eta}} \mathbb{E}\mathcal{L}(\cdot, \boldsymbol{\eta})$, $\widehat{\mathrm{D}}(\cdot)$ is an estimator of $\nabla^2 \mathbb{E} \max_{\boldsymbol{\eta}} \mathcal{L}(\cdot, \boldsymbol{\eta})$ and $\ell(\cdot)$ is an estimator of $\nabla \max_{\boldsymbol{\eta}} \mathcal{L}(\cdot, \boldsymbol{\eta})$. Under common regularity conditions it is shown, that $\|\boldsymbol{\theta}_k - \widetilde{\boldsymbol{\theta}}\| = o_{\mathbb{P}}(1/\sqrt{n})$ if $k(n) \in \mathbb{N}$ is chosen such that the rate of the initial guess $\boldsymbol{\theta}_0$ - obtained via a stochastic grid search - and the rate of the estimator of the nuisance parameter are addressed. These results resemble very much what is aimed for in this work but it is important to note a series of differences between the results of that work and the present paper. First and most importantly, the treated algorithm in that paper (Cheng, 2013) is similar in virtue to the alternating procedure, but in fact is a different procedure. It is a gradient descend scheme and involves a very careful data driven choice of step sizes when carrying out the estimations necessary in (6) and in that sense differs substantially from the simple and direct alternating maximization. Also the estimating step of the nuisance component is not object of the analysis but assumed to be sufficiently good for the arguments to go through. Finally the results of (Cheng, 2013) are purely asymptotic.

In this work we carry out a finite sample analysis for the alternating maxmimization procedure in (3). Instead of a general semiparametric framework we address sieve profile estimators also called *finite dimensional linear series estimation* (see Chen, 2007; Andresen and Spokoiny, 2014), see Section 1.2 for more details. In this setting the bias of estimation - induced by projection the full model to a finite dimensional sub model - can be treated separately and the model becomes finite dimensional as far as the algorithm is concerned. This allows a very careful and explicit analysis of the behavior of the procedure. In particular the speed of convergence can be linked to characteristics of the information matrix $-\nabla^2 \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)$ - namely to the constant $\nu < 1$ in (20) - and to the the full dimension of the projected parameter space. The resulting number of iterations necessary for efficient estimation can be given in a rather simple and closed form. Finally our results are nonasymptotic which in this context is crucial as a clear comparison of the computational and the estimation error for finite samples is needed for reasonable inference.

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 7 that if the initial guess $\widetilde{\boldsymbol{\upsilon}}_0 \in \Upsilon$ is good enough the step estimator sequence $(\widetilde{\boldsymbol{\theta}}_k)$ satisfies with high probability

$$\left\| \breve{\mathrm{D}}\big(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\big) - \breve{\boldsymbol{\xi}} \right\|^2 \leq \epsilon(p^* + \nu^k \mathrm{R}_0), \tag{7}$$

$$\left| \max_{\boldsymbol{\eta}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}\|^2/2 \right| \leq (p + \mathtt{x})^{1/2} \epsilon(p^* + \nu^k \mathrm{R}_0), \tag{8}$$

where $\nu < 1$ is introduced in (20) and $\epsilon > 0$ is some small number, for example $\epsilon = \mathtt{C}/\sqrt{n}$ in the smooth i.i.d setting. Further $\mathrm{R}_0 > 0$ is a bound related to the quality of the initial guess. Generally it is proportional to the full dimension and in this way the rate with which the full nuisance can be estimated affects the speed of the convergence of the procedure. The random variable $\breve{\boldsymbol{\xi}} \in \mathbb{R}^p$ and the matrix $\breve{\mathrm{D}} \in \mathbb{R}^{p \times p}$ are related to the efficient influence function in semiparametric models and its covariance. These are up to $\nu^k \mathrm{R}_0$ the same properties as those proven for the pME by Andresen and Spokoiny (2014) under nearly the same set of conditions. Up to the finite sample bounds on the right hand sides this means that the estimating points of the procedure admit a Fisher expansion - in other words are asymptotical normal - and a Wilks expansion. Consequently the usual inference procedures based on confidence and concentration sets can be applied to these estimators. 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 $(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) = \operatorname{argmax} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}^*)$ 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 $\mathcal{D}^2 = -\mathbb{E}\nabla^2 \mathcal{L}(\boldsymbol{\upsilon}^*)$ (see Theorem 14).

To clarify we want to mention that the term convergence refers to the behavior of the sequence $(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_k)$ when the number of iterations $k \in \mathbb{N}$ tends to infinity. We show $(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_k) \to (\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}})$. This has to be distinguished from the usual stochastic convergence results of the M-estimator $(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}})$ towards the target $(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$ or the weak convergence to a normal distribution as the sample size increases. Our setup is assuming finite sample size such that even with $k \to \infty$ there remains a gap between $(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_k)$ and $(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$ and between $\widetilde{\boldsymbol{\theta}}_k$ and $\boldsymbol{\theta}^*$ that is related to the parametric and semiparametric Cramer-Rao lower bounds respectively. This is why we can in the finite sample setting only hope to obtain convergence of the alternating procedure to $(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}})$ but not to $(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$. But for a growing sample size (7) implies the weak convergence results also for the estimator $\widetilde{\boldsymbol{\theta}}_k$ when $k(n) \in \mathbb{N}$ is large enough and ϵp^* vanishes (see Section 1.3).

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 Relation to the EM Algorithm

In the introduction we claimed that the alternating procedure analyzed in this work is related to the EM algorithm. In this section we want to elaborate on that.

First we explain the EM algortihm. Consider data $(\mathbb{X}) \sim \mathbb{P}_{\boldsymbol{\theta}}$ for some parametric family $(\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta)$. Assume that a parameter $\boldsymbol{\theta} \in \Theta$ is to be estimated as maximizer of the functional $\mathcal{L}_c(\boldsymbol{\theta}, \mathbb{X}) \in \mathbb{R}$, 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 subspaces of $\mathcal{X}$ if both $\mathcal{Y}, \mathcal{X}$ are vector spaces. 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 two steps: starting with some initial guess $\widetilde{\boldsymbol{\theta}}_0$ the kth "Expectation step" derives the functional Q via

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

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

Now we can present the convergence result of Wu (1983) in more detail. Wu presents regularity conditions that ensure that $\mathcal{L}(\boldsymbol{\theta}_{k+1}|\mathbb{Y}) \geq \mathcal{L}(\boldsymbol{\theta}_k|\mathbb{Y})$ where

$$\mathcal{L}(\boldsymbol{\theta}|\mathbb{Y}) \stackrel{\text{def}}{=} \mathbb{E}\left[\mathcal{L}_c(\boldsymbol{\theta}, \mathbb{X})|\mathbb{Y} = f_Y(\mathbb{X})\right],$$

such that $\mathcal{L}(\boldsymbol{\theta}_k|\mathbb{Y}) \to \mathcal{L}^*$ for some limit value $\mathcal{L}^*(\mathbb{Y}) > 0$, 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 $C(\mathcal{L}^*) \stackrel{\text{def}}{=} \{\boldsymbol{\theta}|\, \mathcal{L}(\boldsymbol{\theta}|\mathbb{Y}) = \mathcal{L}^*(\mathbb{Y})\}$. Dempster, Laird, and Rubin (1977) 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 $\mathcal{L}^*(\mathbb{Y}) = \sup \mathcal{L}(\boldsymbol{\theta}|\mathbb{Y})$ 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 $\mathcal{L}(\cdot|\mathbb{Y})$ is unimodal and the regularity conditions are met but this assumption may be too restrictive.

To see that the procedure (3) is a special case of the EM algorithm we have to find the right triplet $(\mathbb{X}, f_Y, \mathcal{L}_c)$. For this we take $\mathbb{X} = \big(\boldsymbol{Z}, \mathbb{Y}\big)$ with $\boldsymbol{Z} \sim \operatorname{argmax}_{\boldsymbol{\eta}} \mathcal{L}\{(\boldsymbol{\theta}, \boldsymbol{\eta}), \mathbb{Y}\}$ under $\mathbb{P}_{\boldsymbol{\theta}}$. Further we set $f_Y(\mathbb{X}) = \mathbb{Y}$ and $\mathcal{L}_c(\boldsymbol{\theta}, \boldsymbol{X}) \stackrel{\text{def}}{=} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}, \mathbb{Y})$, where $\boldsymbol{X} = (\boldsymbol{\eta}, \mathbb{Y})$. Then we find

$$\begin{aligned}
Q(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}^{(k-1)}) &= \mathbb{E}_{\widetilde{\boldsymbol{\theta}}^{(k-1)}}[\mathcal{L}_c(\boldsymbol{\theta}, \mathbb{X})|\mathbb{Y}] \\
&= \mathbb{E}_{\widetilde{\boldsymbol{\theta}}^{(k-1)}}\Big[\mathcal{L}_c\Big(\boldsymbol{\theta}, \operatorname*{argmax}_{\boldsymbol{\eta}} \mathcal{L}\{(\boldsymbol{\theta}^{(k-1)}, \boldsymbol{\eta}), \mathbb{Y}\}, \mathbb{Y}\Big)\Big|\mathbb{Y}\Big] \\
&= \mathcal{L}_c\Big(\boldsymbol{\theta}, \operatorname*{argmax}_{\boldsymbol{\eta}} \mathcal{L}\{(\widetilde{\boldsymbol{\theta}}^{(k-1)}, \boldsymbol{\eta}), \mathbb{Y}\}, \mathbb{Y}\Big) \\
&= \mathcal{L}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\eta}}^{(k)}, \mathbb{Y}),
\end{aligned}$$

and thus the resulting sequence is the same as in (3).

1.2 Linear Series Estimators

In semiparametric models the profile M estimator $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^p$ from Equation (2) cannot be calculated in practice if the full model is infinite dimensional. There are various ways to circumvent this problem. Next to non parametric estimation and plugin of the nuisance $\boldsymbol{\eta} \in \mathcal{X}$ a prominent approach is the so called sieve technique that motivates the setting in this work.

The sieve approach was systematically introduced by Grenander (see Grenander, 1981, Chapter 8) and consists in choosing a suitable sequence of subsets $(\Upsilon_m)_{m=1}^{\infty} \subset \Upsilon$ such that for each $\boldsymbol{\upsilon} \in \Upsilon$ there exists a sequence $\Pi_m(\boldsymbol{\upsilon}) \subset \Upsilon_m$ with $\|\boldsymbol{\upsilon} - \Pi_m(\boldsymbol{\upsilon})\| \to 0$. Furthermore the sets $\Upsilon_m \subset \Upsilon$ have to be such that $\sup_{\boldsymbol{\upsilon} \in \Upsilon_m} \mathcal{L}(\boldsymbol{\upsilon})$ can be calculated in practice. In the setting of semiparametric M-Estimation we assume $\Upsilon = \Upsilon_{\boldsymbol{\theta}} \times \Upsilon_{\boldsymbol{\eta}} \subseteq \mathbb{R}^p \times \mathcal{X}$ with some infinite dimensional separable Hilbert space $\mathcal{X}$ and countable basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots\} \subset \mathcal{X}$. We set $\Upsilon_m = \Upsilon_{\boldsymbol{\theta}} \times \Pi_m \Upsilon_{\boldsymbol{\eta}}$, where $\Pi_m : \mathcal{X} \to \mathcal{X}_m$ denotes the orthogonal projection onto $\mathcal{X}_m \overset{\text{def}}{=} \operatorname{span}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m\}$. For each $m \in \mathbb{N}$ the sieve profile M-estimator is defined as

$$\widetilde{\boldsymbol{\theta}}_m \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \widetilde{\boldsymbol{\upsilon}}_m \overset{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \operatorname*{argmax}_{\substack{\boldsymbol{\theta} \in \mathbb{R}^p \\ \boldsymbol{\eta} \in \mathbb{R}^m}} \mathcal{L}\left(\boldsymbol{\theta}, \sum_{k=1}^{m} \eta_k \boldsymbol{e}_k\right).$$

This means that for the calculation of the estimator $\widetilde{\boldsymbol{\theta}}_m$ only a finite dimensional setting has to be considered. In our analysis we will focus on the behavior of the alternating maximization procedure in that case.

But of course the projection onto a finite dimensional submodel induces an approximation bias " $\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*$ " where

$$(\boldsymbol{\theta}_m^*, \boldsymbol{\eta}_m^*) \overset{\text{def}}{=} \boldsymbol{\upsilon}_m^* \overset{\text{def}}{=} \operatorname*{argmax}_{\substack{\boldsymbol{\theta} \in \mathbb{R}^p \\ \boldsymbol{\eta} \in \mathbb{R}^m}} \mathbb{E}\mathcal{L}\left(\boldsymbol{\theta}, \sum_{k=1}^{m} \eta_k \boldsymbol{e}_k\right).$$

In (Andresen and Spokoiny, 2014) it is explained in detail how this bias can be treated. Once the bias is controlled this leads for each $m \in \mathbb{N}$ to bounds of the kind

$$\left\| \breve{\mathrm{D}}(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}_m \right\| \leq \breve{\diamondsuit}(\mathtt{x}) + \alpha(m),$$

$$\left| \max_{\boldsymbol{\eta} \in \Pi_m \Upsilon_{\boldsymbol{\eta}}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_m, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta} \in \Pi_m \Upsilon_{\boldsymbol{\eta}}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}_m\|^2 \right| \leq p \breve{\diamondsuit}(\mathtt{x}) + \alpha(m),$$

where $\alpha(m) \geq 0$ quantifies the impact of the bias " $\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_m^*$ ". The choice of $m \in \mathbb{N}$ then has to balance the two terms $\breve{\diamondsuit}(\mathtt{x})$ and $\alpha(m)$, which leads to common optimal choices for the dimension based on the "smoothness" of the nuisance component $\boldsymbol{\eta}^* \in \mathcal{X}$. To ease notation we drop the $\cdot_m$ in the following, as the treatment of the bias can be done separately, (see Andresen and Spokoiny, 2014).

1.3 Finite sample Wilks and Fisher Theorems

Before we present our main results we want to explain what type of results we aim at and how they can be interpreted. Hopefully this will ease the understanding and will make some of the apparently cumbersome notation more intelligible.

Usually in asymptotic treatments of semiparametric M-estimators like $\widetilde{\boldsymbol{\theta}}$ in (2) the aim is to derive statements of the kind

$$\begin{aligned} \sqrt{n}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{d}^{-1}\breve{\boldsymbol{\xi}} &= o_{\mathbb{P}}(1), & (9) \\ \max_{\boldsymbol{\eta}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}\|^2/2 &= o_{\mathbb{P}}(1), & (10) \\ \breve{\boldsymbol{\xi}} &\xrightarrow{w} \mathcal{N}(0, \breve{d}^{-1}\breve{v}^2\breve{d}^{-1}), \end{aligned}$$

where $n \in \mathbb{N}$ denotes the sample size. The random variable $\breve{\boldsymbol{\xi}} \in \mathbb{R}^p$ is called semiparametric score. Below we will briefly explain its derivation along with the explanation of the matrices $\breve{v}^2, \breve{d}^2 \in \mathbb{R}^{p \times p}$. But before, we sketch how (9) and (10) can be used for the construction of asymptotic confidence sets that yield statistical tests. Given the matrices $\breve{v}^2, \breve{d}^2$ the construction works as follows. Let $q_\alpha^2 > 0$ be an $\alpha-$level quantile of a $\chi_p^2(\breve{d}^{-2}\breve{v}^2\breve{d}^{-2})$-distribution. Set

$$\mathcal{E}(q_\alpha) = \left\{\boldsymbol{\theta} : \sqrt{n}\|(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\| \le q_\alpha\right\}; \tag{11}$$

then one can use (9) to show

$$\mathbb{P}\left\{\boldsymbol{\theta}^* \notin \mathcal{E}\left(q_\alpha\right)\right\} = \mathbb{P}\left\{\sqrt{n}\|(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \ge q_\alpha\right\} \to 1 - \alpha.$$

Similarly one can exploit (10).

Now we explain the definition of $\breve{v}^2$ and $\breve{d}^2$. Introduce

$$\begin{aligned} n\breve{v}^{-2}(\boldsymbol{\upsilon}) &\stackrel{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \operatorname{Cov}(\nabla\mathcal{L}(\boldsymbol{\upsilon}))^{-1}\Pi_{\boldsymbol{\theta}}^\top, \\ n\breve{d}^{-2}(\boldsymbol{\upsilon}) &\stackrel{\text{def}}{=} -\Pi_{\boldsymbol{\theta}} \left(\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})\right)^{-1}\Pi_{\boldsymbol{\theta}}^\top, \end{aligned}$$

where $\Pi_{\boldsymbol{\theta}}$ is the orthogonal projection onto the $\boldsymbol{\theta}$-commponent in $\mathbb{R}^p$ and $\Pi_{\boldsymbol{\theta}}^\top$ is its dual operator. Then $\breve{v}^2 = \breve{v}^2(\boldsymbol{\upsilon}^*)$ and $\breve{d}^2 = \breve{d}^2(\boldsymbol{\upsilon}^*)$.

Remark 1 *Note that these two matrices coincide if the functional $\mathcal{L}$ was the complete loglikelihood of the observations and that then $\breve{d}^2$ would equal the covariance of the efficient influence function (see Kosorok, 2005, for more details).*

For the definition of the semiparametric score $\breve{\boldsymbol{\xi}} \in \mathbb{R}^p$ consider

$$nd^2(\boldsymbol{\upsilon}) = \begin{pmatrix} d^2(\boldsymbol{\upsilon}) & a(\boldsymbol{\upsilon}) \\ a^\top(\boldsymbol{\upsilon}) & h^2(\boldsymbol{\upsilon}) \end{pmatrix} \stackrel{\text{def}}{=} -\nabla^2\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^* \times p^*}.$$

Then

$$\begin{aligned}\breve{\boldsymbol{\xi}} &\stackrel{\text{def}}{=} \frac{1}{\sqrt{n}}\breve{d}(\boldsymbol{\upsilon}^*)(1-\mathbb{E}_\epsilon)\Pi_{\boldsymbol{\theta}} d^{-2}(\boldsymbol{\upsilon}^*)\nabla\mathcal{L}(\boldsymbol{\upsilon}^*) \\ &= \frac{1}{\sqrt{n}}(1-\mathbb{E}_\epsilon)\breve{d}^{-1}(\boldsymbol{\upsilon}^*)\left\{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*) - ah^{-2}(\boldsymbol{\upsilon}^*)\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}^*)\right\},\end{aligned}$$

This random variable is related to the efficient influence function in semiparametric estimation and it plays the role that the usual score $\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)$ plays in the setting of parametric M-estimation.

In this work we derive (9) and (10) for $\widetilde{\boldsymbol{\theta}}_k$ instead of $\widetilde{\boldsymbol{\theta}}$ but with finite sample bounds for the terms on the right-hand sides of (9) and (10). To be more precise we derive statements of the following kind. With probability greater than $1-\mathtt{C}\mathrm{e}^{-\mathtt{x}}$

$$\left\|\sqrt{n}(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) - \breve{d}^{-1}\breve{\boldsymbol{\xi}}\right\| \leq \epsilon(p^* + \nu^k \mathrm{R}_0), \tag{12}$$

$$\left|\max_{\boldsymbol{\eta}}\mathcal{L}(\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\eta}) - \max_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}\|^2/2\right| = (p+\mathtt{x})^{1/2}\epsilon(p^* + \nu^k \mathrm{R}_0), \tag{13}$$

with some small value $\epsilon > 0$ as in (7) and (8). Note that for vanishing right hand sides these equations imply (9) and (10) when n tends to infty. Using the scheme in (11) the bounds (12) and (13) allow the construction of (conservative) finite sample "confidence sets". Assume that (approximate) quantiles q_α for $\|\breve{\boldsymbol{\xi}}\|$ are available, i.e. that with some small $\epsilon > 0$ and any $\alpha \in [0,1]$

$$\mathbb{P}(\|\breve{\boldsymbol{\xi}}\| \leq q_\alpha) \in (\alpha - \delta, \alpha + \delta),$$

then with some generic constant $\mathtt{C} > 0$ (see Andresen and Spokoiny, 2014, Remark 2.13)

$$\begin{aligned}\alpha + \delta + \mathtt{C}\mathrm{e}^{-\mathtt{x}} &\leq \mathbb{P}\left\{\boldsymbol{\theta}^* \in \mathcal{E}(q_\alpha + \epsilon(p^* + \nu^k \mathrm{R}_0)\right\}, \\ \mathbb{P}\left\{\boldsymbol{\theta}^* \in \mathcal{E}(q_\alpha - \epsilon(p^* + \nu^k \mathrm{R}_0)\right\} &\leq \alpha - \delta - \mathtt{C}\mathrm{e}^{-\mathtt{x}}.\end{aligned}$$

The important achievement is that one can make approximate confidence statements for the estimators of the alternating procedure and this even in the finite sample case, without ignoring "hopefully small enough" terms. As remarked above such approximate quantiles could be attained via an plug-in-estimation of $\breve{d}^{-2}\breve{v}^2\breve{d}^{-2}$ combined with a Gaussian approximation or a bootstrap.

2. Main Results

This Section contains the thorough presentation of our main convergence results. It involves the introduction of various technicalities and may appear a bit cumbersome on first read. We recommend to carefully read Section 1.3 first to ease understanding.

2.1 Conditions

This section collects the conditions imposed on the model. We use the same set of assumptions as Andresen and Spokoiny (2014) and this section closely follows Section 2.1 of that paper.

Let the full dimension of the paramter space be finite, i.e. $p^* < \infty$. Our conditions involve the symmetric positive definite *information matrix* $\mathcal{D}_0^2 \in \mathbb{R}^{p^* \times p^*}$ and a central point $\boldsymbol{\upsilon}^* \in \mathbb{R}^{p^*}$. To ease presentation in this paper we identify $\boldsymbol{\upsilon}^*$ with the "true point" from (1) and define

$$\mathcal{D}_0^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*),$$

where we assume, that the second derivative exists. In the context of semiparametric estimation, it is convenient to represent the *information matrix* in block form:

$$\mathcal{D}_0^2 = \begin{pmatrix} \mathrm{D}_0^2 & \mathrm{A}_0 \\ \mathrm{A}_0^\top & \mathrm{H}_0^2 \end{pmatrix}.$$

First we state an *identifiability condition*, which basically imposes that $\mathcal{D}^2$ is positive definite. Note that in this work $\|\cdot\|$ allways denotes the spectral norm when its argument is a matrix.

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

$$\|\mathrm{H}_0^{-1}\mathrm{A}_0^\top \mathrm{D}_0^{-1}\| \leq \sqrt{\nu}. \tag{14}$$

The condition ($\mathcal{I}$) allows to introduce the important $p \times p$ efficient information matrix $\breve{\mathrm{D}}_0^2$ which is defined as the inverse of the $\boldsymbol{\theta}$-block of the inverse of the full dimensional matrix $\mathcal{D}_0^2$. The exact formula is given by

$$\breve{\mathrm{D}}_0^2 \stackrel{\text{def}}{=} \mathrm{D}_0^2 - \mathrm{A}_0\mathrm{H}^{-2}\mathrm{A}_0^\top,$$

and ($\mathcal{I}$) ensures that the matrix $\breve{\mathrm{D}}_0^2$ is positive definite such that $\breve{\mathrm{D}}$ is well defined. At the same time $\nu < 1$ ensures that the alternating sequence actually converges. As can be seen in Theorem 7 the speed of convergence is linear in ν.

Using the matrix $\mathcal{D}_0^2$ and the central point $\boldsymbol{\upsilon}^* \in \mathbb{R}^{p^*}$, we define the local set $\Upsilon_\circ(\mathtt{r}) \subset \Upsilon \subseteq \mathbb{R}^{p^*}$ with some $\mathtt{r} \geq 0$:

$$\Upsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Upsilon \colon \|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \leq \mathtt{r}\bigr\}. \tag{15}$$

2.1.1 Smoothness

Usually in the context of regular M-estimation one assumes local quadraticity, i.e. that

$$pl(\boldsymbol{\theta}) = \nabla pl(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta}^* - \boldsymbol{\theta}^*)^\perp \nabla^2 pl(\boldsymbol{\theta}^*)(\boldsymbol{\theta}^* - \boldsymbol{\theta}^*) + o_{\mathbb{P}}(1),$$

for $\boldsymbol{\theta} \in \mathbb{R}^p$ close enough to $\boldsymbol{\theta}^*$. The functional $pl(\cdot)$ in this context denotes the profile-functional and is defined as

$$pl(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}),$$

see for instance (Cheng, 2013).

In our setting we need a precise bound for the accuracy of a quadratic approximation of the expected profile functional $\mathbb{E}pl$. We want to bound the error of a local linear approximation of the projected gradient $\breve{\nabla}_{\boldsymbol{\theta}} \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$ which is defined as

$$\breve{\nabla}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} - \mathrm{A}_0 \mathrm{H}_0^{-2} \nabla_{\boldsymbol{\eta}}.$$

Instead of local quadraticity of the profile functional we impose that for $(\boldsymbol{\theta}, \boldsymbol{\eta}) = \boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$ - with the local set $\Upsilon_\circ(\mathtt{r})$ defined in (15) - with some small $\breve{\epsilon} > 0$

$$\left\| \breve{\mathrm{D}}^{-1} \left(\breve{\nabla} \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) - \breve{\nabla} \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*) \right) - \breve{\mathrm{D}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\| \leq \breve{\epsilon} \mathtt{r}^2.$$

The following condition serves a less high level way of checking that such an approximation holds (see Andresen and Spokoiny, 2014, Lemma B.1)

$(\breve{\mathcal{L}}_0)$ For each $\mathtt{r} \leq \mathtt{r}_0$, there is a constant $\breve{\epsilon} > 0$ such that it holds on the set $\Upsilon_\circ(\mathtt{r})$:

$$\|\mathrm{D}^{-1} \mathrm{D}^2(\boldsymbol{\upsilon}) \mathrm{D}^{-1} - I_p\| \leq \breve{\epsilon}\mathtt{r}, \quad \|\mathrm{D}^{-1}(\mathrm{A}(\boldsymbol{\upsilon}) - \mathrm{A})\mathrm{H}^{-1}\| \leq \breve{\epsilon}\mathtt{r},$$
$$\left\| \mathrm{D}^{-1} \mathrm{A} \mathrm{H}^{-1} \left(I_m - \mathrm{H}^{-1} \mathrm{H}^2(\boldsymbol{\upsilon}) \mathrm{H}^{-1} \right) \right\| \leq \breve{\epsilon}\mathtt{r}.$$

If $\mathbb{E}\mathcal{L}$ is three times continuously differentiable one obtains $\breve{\epsilon} \leq \mathtt{C} \|\mathcal{D}^{-1}\|$. In i.i.d. models one usually has $\|\mathcal{D}^{-1}\| = O(1/\sqrt{n})$.

Remark 2 *Here and in what follows we implicitly assume that the function $\mathcal{L}(\boldsymbol{\upsilon}) \colon \mathbb{R}^{p^*} \to \mathbb{R}$ is sufficiently smooth in $\boldsymbol{\upsilon} \in \mathbb{R}^{p^*}$, $\nabla \mathcal{L}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^*}$ stands for the gradient and $\nabla^2 \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^* \times p^*}$ for the Hessian of the expectation $\mathbb{E}\mathcal{L} : \mathbb{R}^{p^*} \to \mathbb{R}$ at $\boldsymbol{\upsilon} \in \mathbb{R}^{p^*}$. By smooth enough we mean that we can interchange $\nabla \mathbb{E}\mathcal{L} = \mathbb{E}\nabla\mathcal{L}$ on $\Upsilon_\circ(R_0)$, where $\Upsilon_\circ(\mathtt{r})$ is defined in (15) and $R_0 > 0$ in (19). It is worth mentioning that $\mathcal{D}_0^2 = \mathcal{V}_0^2 \stackrel{\text{def}}{=} \mathrm{Cov}(\nabla \mathcal{L}(\boldsymbol{\upsilon}^*))$ if the model $\boldsymbol{Y} \sim \mathbb{P}_{\boldsymbol{\upsilon}^*} \in (\mathbb{P}_{\boldsymbol{\upsilon}})$ is correctly specified and sufficiently regular; see e.g. (Ibragimov and Khas'minskij, 1981).*

2.1.2 Complexity

The usual approach to gain asymptotic control on profile M-estimators would be to assume that the class

$$\{\breve{\nabla} \mathcal{L}(\boldsymbol{\upsilon}), \boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})\},$$

is $\mathbb{P}$-Donsker (see Cheng, 2013).

To pin down the estimator sequence $(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\eta}}_k)_{k \in \mathbb{N}}$ and to obtain finite sample results we use a more specific approach, which is based on a new finite sample approach (Spokoiny,

2012; Andresen and Spokoiny, 2014). First note that we assume that - as far as the alternating procedure is concerned - the model is finite dimensional, which means that $\Upsilon_\circ(\mathtt{r}) \subset \Upsilon \subset \mathbb{R}^{p^*}$ in (15) is automatically compact for finite radius $\mathtt{r} > 0$. If we can ensure the right smoothness and moment conditions on $\check{\nabla}\mathcal{L}(\boldsymbol{\upsilon})$, we automatically obtain that the above class is $\mathbb{P}$-Donsker. But using the new techniques (Spokoiny, 2012; Andresen and Spokoiny, 2014) we manage to obtain slightly stronger bounds that are useful in a finite sample setting.

To understand the next condition consider first the definition of a subgaussian random vector. A random vector $\boldsymbol{X} \in \mathbb{R}^p$ is called subgaussian, if for any $\mu \in \mathbb{R}$ and some $\boldsymbol{\nu} > 0$

$$\sup_{\|\boldsymbol{\gamma}\| \le 1} \log \mathbb{E} \exp\Big\{ \mu \boldsymbol{\gamma}^\top \boldsymbol{X} \Big\} \le \boldsymbol{\nu}^2 \mu^2 / 2. \tag{16}$$

Obviously this is a strong condition. As it turns out in many situations it is sufficient to assume subexponentiality, which simply relaxes subgaussianity to demanding that (16) is met for any $|\mu| \le \mathtt{g}$ with some $\mathtt{g} > 0$. In this way many distributions that would be excluded by assuming subgaussianity can still be treated.

Our next condition combines subexponentiality with a smoothness constraint on the stochastic component of $\check{\nabla}\mathcal{L}(\boldsymbol{\upsilon})$. It assumes that - with some $\check{\epsilon} > 0$ - the random vector

$$\frac{1}{\check{\epsilon}\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')\|} \check{\mathrm{D}}^{-1} \big\{ \check{\nabla}_{\boldsymbol{\theta}} \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \check{\nabla}_{\boldsymbol{\theta}} \boldsymbol{\zeta}(\boldsymbol{\upsilon}') \big\} \in \mathbb{R}^p,$$

is uniformly subexponential for $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})$, with the stochastic component defined as $\zeta(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \mathcal{L}(\boldsymbol{\upsilon}) - \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$. It reads:

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

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

To convey more intuition consider for some pair $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})$ the function

$$\psi_{\boldsymbol{\gamma}}(t) \stackrel{\text{def}}{=} \boldsymbol{\gamma}^\top \check{\mathrm{D}}^{-1} \big\{ \check{\nabla}_{\boldsymbol{\theta}} \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \check{\nabla}_{\boldsymbol{\theta}} \boldsymbol{\zeta}(\boldsymbol{\upsilon} + t(\boldsymbol{\upsilon}' - \boldsymbol{\upsilon})) \big\}.$$

Then $(\check{\mathcal{E}}\mathcal{D}_1)$ is met with $\check{\epsilon} \le \|\mathcal{D}^{-1}\|$, if - uniformly in $\boldsymbol{\gamma}$ - for any pair $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})$ the corresponding function $\psi_{\boldsymbol{\gamma}} : [0,1] \to \mathbb{R}$ is Lipshitz continuous and the Lipshitz constant a subexponential random variable.

2.1.3 Moments

We need another condition that allows to control the deviation behavior of $\|\check{\mathrm{D}}^{-1} \check{\nabla} \boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\|$. To present this condition define the covariance matrix $\mathcal{V}_0^2 \in \mathbb{R}^{p^* \times p^*}$ and $\check{V}^2 \in \mathbb{R}^{p \times p}$

$$\mathcal{V}_0^2 \stackrel{\text{def}}{=} \mathrm{Cov}\big\{ \nabla \mathcal{L}(\boldsymbol{\upsilon}^*) \big\}, \quad \check{V}^2 = \mathrm{Cov}(\check{\nabla}_{\boldsymbol{\theta}} \zeta(\boldsymbol{\upsilon}^*)).$$

We impose subexponential moments on $\check{V}^{-1} \check{\nabla}_{\boldsymbol{\theta}} \zeta(\boldsymbol{\upsilon}^*)$:

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

$$\sup_{\|\gamma\|\le 1} \log \mathbb{E} \exp\Big\{ \mu \boldsymbol{\gamma}^\top \breve{V}^{-1} \breve{\nabla}_{\boldsymbol{\theta}} \zeta(\boldsymbol{\upsilon}^*) \Big\} \le \frac{\breve{\nu}_0^2 \mu^2}{2}.$$

2.1.4 Conditions for the Full Model

So far we only presented conditions that allow to treat the properties of $\widetilde{\boldsymbol{\theta}}_k$ on local sets $\Upsilon_\circ(\mathtt{r}_k)$, for some sequence $(\mathtt{r}_k)_{k\in\mathbb{N}}$. To show that that the sequence of estimators $(\boldsymbol{\upsilon}_k)_{k\in\mathbb{N}}$ satisfies $\boldsymbol{\upsilon}_k \in \Upsilon_\circ(\mathtt{r}_k)$ for an appropriately decreasing sequence $(\mathtt{r}_k)_{k\in\mathbb{N}}$ the following, stronger conditions are employed, which can be interpreted just as the previous ones.

$(\mathcal{L}_0)$ For each $\mathtt{r} \le \mathtt{r}_0$, there is a constant $\epsilon > 0$ such that it holds on the set $\Upsilon_\circ(\mathtt{r})$:

$$\big\| \mathcal{D}_0^{-1} \big\{ \nabla^2 \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}) \big\} \mathcal{D}_0^{-1} - \boldsymbol{I}_{p^*} \big\| \le \epsilon \mathtt{r}.$$

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

$$\sup_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})} \sup_{\|\gamma\|=1} \log \mathbb{E} \exp\left\{ \frac{\mu\, \boldsymbol{\gamma}^\top \mathcal{D}_0^{-1} \big\{ \nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla \boldsymbol{\zeta}(\boldsymbol{\upsilon}') \big\}}{\epsilon \, \| \mathcal{D}_0 (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}') \|} \right\} \le \frac{\nu_1^2 \mu^2}{2}.$$

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

$$\sup_{\|\gamma\|\le 1} \log \mathbb{E} \exp\Big\{ \mu \boldsymbol{\gamma}^\top \mathcal{V}_0^{-1} \nabla \zeta(\boldsymbol{\upsilon}^*) \Big\} \le \frac{\nu_0^2 \mu^2}{2}.$$

It is important to note, that the constants $\breve{\epsilon}, \breve{\nu}$ and ϵ, ν 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 ($\mathcal{A}\mathrm{H}^{-2}\nabla_{\boldsymbol{\eta}}\mathcal{L}$ might be quite smooth while $\nabla_{\boldsymbol{\eta}}\mathcal{L}$ could be less regular).

For the convergence statement in Theorem 14 we additionally need the following condition, that controls the moments and the smoothness of the process $\nabla^2(\mathcal{L} - \mathbb{E}\mathcal{L})$:

$(\mathcal{ED}_2)$ There exists a constant $\epsilon_2 \le 1/2$, such that for all $|\mu| \le \mathtt{g}$ and all $0 < \mathtt{r} < \mathtt{r}_0$

$$\sup_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}' \in \Upsilon_\circ(\mathtt{r})} \sup_{\|\gamma_1\|=1} \sup_{\|\gamma_2\|=1} \log \mathbb{E} \exp\left\{ \frac{\mu\, \boldsymbol{\gamma}_1^\top \mathcal{D}^{-1} \big\{ \nabla^2 \boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla^2 \boldsymbol{\zeta}(\boldsymbol{\upsilon}') \big\} \boldsymbol{\gamma}_2}{\epsilon_2 \, \| \mathcal{D} (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}') \|} \right\} \le \frac{\nu_2^2 \mu^2}{2}.$$

2.1.5 Quadratic Drift Beats Linear Fluctuation

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 $\Upsilon \subseteq \mathbb{R}^{p^*}$.

The first condition imposes that $\mathbb{E}\mathcal{L}(\boldsymbol{\upsilon})$ decreases nearly quadratically as the distance of $\boldsymbol{\upsilon}$ to $\boldsymbol{\upsilon}^*$ grows.

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

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\upsilon}) - \mathcal{L}(\boldsymbol{\upsilon}^*)\right] \le \mathtt{b}\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2.$$

The next condition bounds moments of the full gradient $\nabla\mathcal{L}(\boldsymbol{\upsilon})$ - again via subexponentiality.

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

$$\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \sup_{\mu\le\mathtt{g}(\mathtt{r})} \sup_{\|\boldsymbol{\gamma}\|\le 1} \log \mathbb{E}\exp\left\{\mu\boldsymbol{\gamma}^\top\mathcal{D}^{-1}\nabla\zeta(\boldsymbol{\upsilon})\right\} \le \frac{\nu_{\mathtt{r}}^2\mu^2}{2}.$$

We impose one further merely technical condition:

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

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

Remark 3 *Without this the calculation of $R_0(\mathtt{x})$ in Section 4.3 would become technically more involved, without that further insight would be gained.*

Remark 4 *The condition ($\mathcal{E}\mathtt{r}$) can be substantially relaxed to $\mathtt{b} = \mathtt{b}(\mathtt{r}) > 0$ that decreases to 0 as $\mathtt{r} \to \infty$. We avoid the resulting technicalities and refer the reader to the original publication for the non constant case (see Spokoiny, 2012, Theorem 4.1).*

2.2 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 $\widetilde{\boldsymbol{\upsilon}}_0 \in \mathbb{R}^{p^*}$ which we denote by (A_1), (A_2) and (A_3):

($\mathbf{A}_1$) With probability greater $1-\beta$ the initial guess satisfies $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0) - \mathcal{L}(\boldsymbol{\upsilon}^*) \ge -\mathrm{K}_0$ for some $\mathrm{K}_0 \ge 0$.

($\mathbf{A}_2$) 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 $\mathtt{r} \le \mathrm{R}_0(\mathtt{x})$ where $\mathrm{R}_0(\mathtt{x})$ can be bounded with (see (19))

$$\mathrm{R}_0(\mathtt{x}) \le \mathtt{C}\sqrt{\mathtt{x}+p^*+\mathrm{K}_0}.$$

($\mathbf{A}_3$) $\mathrm{K}_0 \in \mathbb{R}$ and $\epsilon > 0$ are small enough to ensure

$$\epsilon\mathtt{C}(\nu)\mathrm{R}_0 < 1, \tag{17}$$

with

$$\mathtt{C}(\nu) \stackrel{\text{def}}{=} \frac{16\sqrt{2}(1+\sqrt{\nu})}{(1-\nu)(1-\sqrt{\nu})}, \tag{18}$$

where $\nu > 0$ is defined in (14).

Condition (A_1) allows to concentrate the analysis on a local set $\{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{R}_0(\mathtt{x})\} \subset \Upsilon$ with dominating probability (see Theorem 28). Conditions (A_2) and (A_3) ensure that this neighborhood is small enough to imply convergence of the procedure. They impose a bound on $\mathtt{R}_0(\mathtt{x})$ and thus on $\mathtt{K}_0$ from (A_1). These conditions boil down to $\epsilon\sqrt{\mathtt{K}_0}$ being significantly smaller than 1, which is a quantification of the quality of the first guess. There are numerous ways to initiate the procedure. In Section 3 we use a grid search and show that for a sufficiently fine grid these conditions can be met in the treated model.

Remark 5 *One way of obtaining condition* (A_1) *is to show that* $\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*)\| \le R$ *with probability greater* $1-\beta$ *for some finite* $R \in \mathbb{R}$ *and* $0 \le \beta(R) < 1$. *Then one can use - with some constant* $\mathtt{C} > 0$ *-*

$$K_0 \le (1/2 + \epsilon(1 + 12\nu_0))(R + \mathtt{C}\sqrt{\mathtt{x} + p^*})^2,$$

as we show in Lemma 31.

Remark 6 *The precise definition of* $R_0(\mathtt{x}) > 0$ *reads*

$$R_0(\mathtt{x}) \stackrel{\text{def}}{=} \mathfrak{z}(\mathtt{x}) \vee \frac{6\nu_{\mathtt{r}}}{\mathtt{b}(1-\nu)} \sqrt{\mathtt{x} + 2.4p^* + \frac{\mathtt{b}^2}{9\nu_{\mathtt{r}}^2} K_0}, \tag{19}$$

with the term

$$\mathfrak{z}(\mathtt{x}) \approx \mathtt{C}\sqrt{p^* + \mathtt{x}},$$

which is defined in (30).

2.3 Introduction of Important Objects

In this section we collect the most important objects and bounds that are relevant for Theorem 7. Remember the $p^* \times p^*$ *information matrix* $\mathcal{D}^2$ from Section 2.1, which is defined similarly to the Fisher information matrix:

$$\mathcal{D}^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*) = \begin{pmatrix} \mathrm{D}^2 & \mathcal{A} \\ \mathcal{A}^\top & \mathrm{H}^2 \end{pmatrix}.$$

A crucial object is the constant $0 \le \nu$ defined by

$$\|\mathrm{D}^{-1}\mathcal{A}\mathrm{H}^{-1}\|^2 \stackrel{\text{def}}{=} \nu, \tag{20}$$

which we assume with condition $(\mathcal{I})$ to be smaller 1. It determines the speed of convergence of the alternating procedure (see Theorem 7).

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

$$\begin{aligned} \breve{\mathrm{D}}^2 &= \mathrm{D}^2 - \mathcal{A}\mathrm{H}^{-2}\mathcal{A}^\top, \\ \breve{\boldsymbol{\xi}} &= \breve{\mathrm{D}}^{-1}\breve{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*), \quad \breve{\nabla}_{\boldsymbol{\theta}}\mathcal{L} = \nabla_{\boldsymbol{\theta}}\mathcal{L} - \mathcal{A}\mathrm{H}^{-2}\nabla_{\boldsymbol{\eta}}\mathcal{L}. \end{aligned}$$

The random variable $\breve{\boldsymbol{\xi}} \in \mathbb{R}^p$ is related to the efficient influence function in semiparametric models. If the model is regular and correctly specified $\breve{\mathrm{D}}^2$ is the covariance of the efficient influence function and its inverse the semiparametric Cramer-Rao lower bound for regular estimators.

We define the *semiparametric uniform spread*

$$\breve{\diamondsuit}_Q(\mathtt{r},\mathtt{x}) \stackrel{\mathrm{def}}{=} \breve{\epsilon}\left\{\frac{16}{(1-\nu^2)^2}\mathtt{r}^2 + \mathfrak{z}_Q(\mathtt{x},2p^*+2p)^2\right\}(1+6\breve{\nu}_1^2). \tag{21}$$

This object is central for our analysis as it describes the accuracy of our main results. It is small if $\breve{\epsilon}(\mathtt{r}^2+\mathtt{x}+p^*)$ is small, since $\mathfrak{z}_Q(\mathtt{x},p^*) \approx \sqrt{p^*+\mathtt{x}}$ (see its definition in Equation (42)).

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 7 *Assume that the conditions* $(\mathcal{L}_\mathtt{r})$, $(\mathcal{E}\mathtt{r})$ *and* (B_1) *of Section 2.1 are met. Further assume* (A_1), (A_2) *and* (A_3) *of Section 2.2. Then it holds with probability greater* $1-8\mathrm{e}^{-\mathtt{x}}-\beta$ *for all* $k \in \mathbb{N}$

$$\big\|\breve{D}(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}}\big\| \le \breve{\diamondsuit}_Q(\mathtt{r}_k,\mathtt{x}), \tag{22}$$

$$\big|\max_{\boldsymbol{\eta}} \mathcal{L}(\widetilde{\boldsymbol{\theta}}_k,\boldsymbol{\eta}) - \max_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}^*,\boldsymbol{\eta}) - \|\breve{\boldsymbol{\xi}}\|^2/2\big| \le 5\left(\|\breve{\boldsymbol{\xi}}\| + \breve{\diamondsuit}_Q(\mathtt{r}_k,\mathtt{x})\right)\breve{\diamondsuit}_Q(\mathtt{r}_k,\mathtt{x}), \tag{23}$$

where

$$\mathtt{r}_k \le \mathtt{C}\left(\sqrt{p^*+\mathtt{x}} + \nu^k R_0\right),$$

with a constant $\mathtt{C}$ *that depends on* $\nu<1$ *and* $1-\mathtt{C}(\nu)\epsilon R_0 > 0$. *In particular this means that if*

$$k \ge \frac{\log(p^*+\mathtt{x}) - \log\{R_0\}}{\log(\nu)},$$

we have

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

Remark 8 *Note that with linear convergence speed this leads to statements about* $\widetilde{\boldsymbol{\theta}}_k$ *that are very similar to those in (Andresen and Spokoiny, 2014) for the profile M estimator* $\widetilde{\boldsymbol{\theta}}$.

Remark 9 *Concerning the properties of* $\breve{\boldsymbol{\xi}} \in \mathbb{R}^p$ *we repeat remark 2.1 of (Andresen and Spokoiny, 2014). In case of 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 those of a chi-square random variable with* p *degrees of freedom; see Theorem 39 in the appendix. In the case of a possible*

model misspecification the behavior of the quadratic form $\|\breve{\boldsymbol{\xi}}\|^2$ *will depend on the characteristics of the matrix* $\breve{I\!B} \stackrel{\text{def}}{=} \operatorname{Cov}(\breve{\boldsymbol{\xi}})$ *; see again Theorem 39. Moreover, in the asymptotic setup the vector* $\breve{\boldsymbol{\xi}}$ *is asymptotically standard normal; see Section 2.2. of (Andresen and Spokoiny, 2014) for the i.i.d. case.*

Remark 10 *These results allow to derive some important corollaries like concentration and confidence sets (see Section 1.3).*

Remark 11 *In general an exact numerical computation of*

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

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 - with the local set* $\Upsilon_\circ(\mathtt{r})$ *defined in* (15) *-*

$$\|D(\widehat{\theta}(\boldsymbol{\eta}) - \theta(\boldsymbol{\eta}))\| \le \tau, \quad \text{for all} \quad \boldsymbol{\eta} \in \Upsilon_{\circ,\boldsymbol{\eta}}(R_0) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} \in \Upsilon_\circ(R_0),\, \Pi_{\boldsymbol{\eta}}\boldsymbol{\upsilon} = \boldsymbol{\eta}\},$$

$$\|H(\widehat{\eta}(\boldsymbol{\theta}) - \eta(\boldsymbol{\theta}))\| \le \tau, \quad \text{for all} \quad \boldsymbol{\theta} \in \Upsilon_{\circ,\boldsymbol{\theta}}(R_0) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} \in \Upsilon_\circ(R_0),\, \Pi_{\boldsymbol{\theta}}\boldsymbol{\upsilon} = \boldsymbol{\theta}\}.$$

Then we can easily modify the proof of Theorem 7 via adding $\mathtt{C}(\nu)\tau$ *to the error terms and the radii* $\mathtt{r}_k$ *, where* $\mathtt{C}(\nu)$ *is some rational function of* ν *.*

2.5 Convergence to the ME

Even though Theorem 7 tells us that the statistical properties of the alternating sequence resemble those of its target - the profile ME $\widetilde{\boldsymbol{\theta}}$ - it is an interesting question if the underlying approach allows to qualify conditions under which the sequence actually attains the maximizer $\widetilde{\boldsymbol{\upsilon}}$.

Define the radius $\mathtt{r}_0(\mathtt{x}) > 0$ to be the smallest radius $\mathtt{r} > 0$ such that $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)) \ge 1 - \mathrm{e}^{-\mathtt{x}}$, where

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

Remark 12 *This radius can be determined using conditions* $(\mathcal{L}_{\mathtt{r}})$ *and* $(\mathcal{E}\mathtt{r})$ *of Section 2.1 and Theorem 28 which would yield* $\mathtt{r}_0(\mathtt{x}) = O(\sqrt{\mathtt{x} + p^*})$ *.*

Without further assumptions Theorem 7 yields the following Corollary:

Corollary 13 *Under the assumptions of Theorem 7 it holds with probability greater* $1 - 8\mathrm{e}^{-\mathtt{x}} - \beta$

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

Corollary 13 is a first step in the direction of an actual convergence result but the gap $\breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}) + \breve{\Diamond}_Q(\mathtt{r}_0, \mathtt{x})$ is not a zero sequence in $k \in \mathbb{N}$. It turns out that it is possible to

prove convergence to the ME at the cost of assuming more smoothness of the functional $\mathcal{L}$ and using the right bound for the maximal eigenvalue of the Hessian $\nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)$.

Define $\mathfrak{z}_2(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*))$ via

$$\mathbb{P}\Big(\|\mathcal{D}^{-1}\nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\| \geq \mathfrak{z}_2\left(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right)\Big) \leq \mathrm{e}^{-\mathtt{x}},$$

and $\kappa(\mathtt{x}, \mathrm{R}_0)$ as

$$\kappa(\mathtt{x}, \mathrm{R}_0) \stackrel{\mathrm{def}}{=} \frac{2\sqrt{2}(1+\sqrt{\nu})}{\sqrt{1-\nu}}\left\{\epsilon\mathrm{R}_0 + 9\epsilon_2\nu_2\|\mathcal{D}^{-1}\|\mathfrak{z}_1(\mathtt{x}, 6p^*)\mathrm{R}_0 + \|\mathcal{D}^{-1}\|\mathfrak{z}_2\left(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right)\right\},$$

where $\mathfrak{z}_1(\mathtt{x}, \cdot) \approx \sqrt{\mathtt{x} + p^*}$, see (46). With these definitions we can prove the following Theorem:

Theorem 14 *Let the conditions* $(\mathcal{L}_{\mathtt{r}})$, $(\mathcal{E}\mathtt{r})$ *and* (B_1) *be met. Further suppose* (A_1) *and* (A_2), *with* $(\mathcal{E}\mathcal{D}_2)$ *instead of* $(\mathcal{E}\mathcal{D}_1)$. *Assume that* $\kappa(\mathtt{x}, R_0) < (1-\nu)$. *Then*

$$\mathbb{P}\left(\bigcap_{k\in\mathbb{N}}\left\{\|\mathcal{D}(\boldsymbol{\upsilon}_{k,k(+1)} - \widetilde{\boldsymbol{\upsilon}})\| \leq \mathtt{r}_k^*\right\}\right) \geq 1 - 3\mathrm{e}^{-\mathtt{x}} - \beta,$$

where

$$\mathtt{r}_k^* \leq \begin{cases} \nu^k \frac{4\sqrt{2}}{1-\kappa(\mathtt{x},R_0)k}R_0, & \kappa(\mathtt{x}, R_0)k \leq 1, \\ \nu^{\frac{k}{\log(k)}\log\left(\frac{1-\nu}{\kappa(\mathtt{x},R_0)}\right)}c_k R_0, & \textit{otherwise,} \end{cases} \tag{24}$$

with some sequence $(c_k) \in \mathbb{N}$, *where* $0 < c_k \to 2$.

Remark 15 *This means that we obtain nearly linear convergence to the global maximizer* $\widetilde{\boldsymbol{\upsilon}}$.

Remark 16 *As in Remark 11 if no exact numerical computation of the stepwise maximizers is possible we can easily modify the proof of Theorem 14 via adding* $\mathtt{C}(\nu)\tau$ *to* $\kappa(\mathtt{x}, R_0)$ *to address that case.*

Remark 17 *For the case that* $\mathcal{L}(\boldsymbol{\upsilon}) = \sum_{i=1}^n \ell_i(\boldsymbol{\upsilon})$ *with a sum of independent marginal functionals* $\ell_i : \Upsilon \to \mathbb{R}$ *we can use Corollary 3.7 of (Tropp, 2012) to obtain*

$$\mathfrak{z}_2\left(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right) = \sqrt{2\tau}\nu\sqrt{\mathtt{x} + \log(p^*)},$$

if for some sequence of matrices $(\boldsymbol{A}_i) \subset \mathbb{R}^{p^*\times p^*}$

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

In the case of smooth i.i.d models this means that

$$\kappa(\mathtt{x}, R_0) \leq \frac{\mathtt{C}}{\sqrt{n}}(\mathtt{x} + R_0 + \log(p^*)),$$

if $p^* + \mathtt{x} = o(n)$.

Remark 18 *It may happen that* $\kappa(\mathtt{x}, R_0)/(1-\nu)$ *is very close to or even larger than* 1. *But a close look at the proof of Theorem 14 reveals that this can be improved using Lemma 33. For this purpose bound* $\mathtt{r}_k^* \le \mathtt{C}^*(\mathfrak{z}(\mathtt{x}) + \nu^k R_0)$ *with* $\mathtt{r}_k^*$ *defined in* (33) *and with some constant* $\mathtt{C}^* > 0$. *Then the result of Theorem 14 is true with* $\kappa(\mathtt{x}, \mathtt{C}^*\sqrt{p^* + \mathtt{x}})$ *instead of* $\kappa(\mathtt{x}, R_0)$ *and with probability greater* $1 - 10\mathrm{e}^{-\mathtt{x}}$. *See Remark 38 for more details.*

2.6 Critical Dimension

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 7 are accurate if the spread function $\breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x})$ from (21) is small. The critical size of p_n then depends on the exact bounds on $\epsilon, \breve{\epsilon}$. In the i.i.d setting we have $\epsilon \asymp \breve{\epsilon} \asymp 1/\sqrt{n}$ such that $\breve{\Diamond}(\mathtt{r}_k, \mathtt{x}) \asymp p_n/\sqrt{n}$ for large $k \in \mathbb{N}$. In other words, one needs that "${p_n}^2/n$ 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 "${p_n}^2/n$ is small", as is shown by Andresen and Spokoiny (2014). There are examples where the pME only satisfies a Wilks- or Fisher result if "${p_n}^2/n$ 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 $\kappa(\mathtt{x}, \mathrm{R}_0) < (1-\nu)$ of Theorem 14 for the convergence of the sequence to the global maximizer means that one needs $p_n/n \ll 1$ in the smooth i.i.d. setting if $\mathrm{R}_0 \le \mathtt{C}_{\mathrm{R}_0}\sqrt{p_n + \mathtt{x}}$. Further Theorem 14 states a lower bound for the speed of convergence that in the smooth i.i.d. setting decreases if p_n/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/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 7 and Theorem 14 can be applied in Single Index modeling. This section is based on (Andresen, 2015). See that paper for a more detailed presentation.

Consider the following model

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

for some $f : \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{\theta}^* \in S_1^{p,+} \subset \mathbb{R}^p$ and with i.i.d errors $\varepsilon_i \in \mathbb{R}$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and i.i.d random variables $\boldsymbol{X}_i \in \mathbb{R}^p$ with distribution denoted by $\mathbb{P}^{\boldsymbol{X}}$. 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 non-parametric 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, (Delecroix. et al., 1997) show the asymptotic efficiency of the general semiparametric maximum-functional estimator for particular examples and in (Haerdle et al., 1993) the right choice of band-

width for the nonparametric estimation of the link function is analyzed. We want to use this model to illustrate our theoretical results.

To ensure identifiability of $\boldsymbol{\theta}^* \in \mathbb{R}^p$ we assume that it lies in the half sphere $S_1^{p,+} \stackrel{\text{def}}{=} \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\| = 1, \theta_1 > 0\} \subset \mathbb{R}^p$. For simplicity we assume that the support of the $\boldsymbol{X}_i \in \mathbb{R}^p$ is contained in the ball of radius $s > 0$. This allows to approximate $f \in \{f : [-s, s] \mapsto \mathbb{R}\}$ by an orthonormal C^3-Daubechies-wavelet basis $(\boldsymbol{e}_k)_{k\in\mathbb{N}}$ on the interval.

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

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

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 $\Upsilon_m = S_1^{p,+} \times \mathbb{R}^m$. Again we will suppress the sub index $\cdot_m$ in the following.

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 maximization 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 take a uniform grid $G_N \stackrel{\text{def}}{=} (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \subset S_1^+$ and define

$$\widetilde{\boldsymbol{\upsilon}}_0 \stackrel{\text{def}}{=} \operatorname*{argmax}_{\substack{(\boldsymbol{\theta},\boldsymbol{\eta})\in\Upsilon \\ \boldsymbol{\theta}\in G_N}} \mathcal{L}(\boldsymbol{\upsilon}). \tag{25}$$

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

$$\widetilde{\boldsymbol{\eta}}_{0,k} \stackrel{\text{def}}{=} \operatorname{argmax} \mathcal{L}(\boldsymbol{\theta}_k, \boldsymbol{\eta}) = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{e}\boldsymbol{e}^\top(\boldsymbol{X}_i^\top \boldsymbol{\theta}_k)\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{e}^\top(\boldsymbol{X}_i^\top \boldsymbol{\theta}_k) \in \mathbb{R}^m, \tag{26}$$

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

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

Define the mesh size of the grid $\tau \stackrel{\text{def}}{=} \sup_{\boldsymbol{\theta},\boldsymbol{\theta}^\circ\in G_N} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\|$.

To apply the result presented in Theorem 7 and Theorem 14 we need a list of assumptions denoted by $(\mathcal{A})$.

($\mathbf{Cond}_{\boldsymbol{X}}$) The random variables $(\boldsymbol{X}_i)_{i=1,\ldots,n} \subset \mathbb{R}^p$ are i.i.d and bounded with distribution denoted by $\mathbb{P}^{\boldsymbol{X}}$ and independent of $(\varepsilon_i)_{i=1,\ldots,n} \subset \mathbb{R}$.

The measure $\mathbb{P}^{\boldsymbol{X}}$ is absolutely continuous with respect to the Lebesgue measure. The

Lebesgue density $p_{\boldsymbol{X}}$ of $\mathbb{P}^{\boldsymbol{X}}$ is Lipschitz continuous and positive on $B_s(0) \subset \mathbb{R}^p$. For any pair $\boldsymbol{\theta} \in S_1^{+,p}$ with $\boldsymbol{\theta} \perp \boldsymbol{\theta}^*$ we have almost surely

$$\operatorname{Var}\left(\boldsymbol{X}^\top \boldsymbol{\theta} \middle| \boldsymbol{X}^\top \boldsymbol{\theta}^*\right) > 0.$$

($\mathbf{Cond}_f$) For some $\boldsymbol{\eta}^* \in B_{\mathtt{r}^\circ}(0) \subset l^2 \stackrel{\text{def}}{=} \{(u_k)_{k\in\mathbb{N}} : \sum_{k=1}^\infty u_k^2 < \infty\}$

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

where $\|f'\|_\infty < \infty$ and $\|f''\|_\infty < \infty$ and where with some $\alpha > 2$

$$\sum_{k=0}^\infty k^{2\alpha} \eta_k^{*2} < \infty.$$

On some interval $[t_0 - h, t_0 + h] \subseteq [-s + s]$ with $h > 0$ it holds true that

$$|f'(t)| > 0.$$

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

$$\log \mathbb{E}[\exp\{\mu\varepsilon_1\}] \leq \widetilde{\nu}^2 \mu^2/2.$$

Remark 19 *Note that our assumptions in terms of moments and smoothness are quite common in this model. For instance (Haerdle et al., 1993) assume that the density $p_{\boldsymbol{X}}$ of the regressors $(\boldsymbol{X}_i)$ is twice continuously differentiable, that f has two bounded derivatives and that the errors (ε_i) are centered with bounded polynomial moments of arbitrary degree.*

Remark 20 $\operatorname{Var}\left(\boldsymbol{X}^\top \boldsymbol{\theta}^\circ \middle| \boldsymbol{X}^\top \boldsymbol{\theta}^*\right) = 0$ *would mean that* $\boldsymbol{X}^\top \boldsymbol{\theta}^\circ = a(\boldsymbol{X}^\top \boldsymbol{\theta}^*)$ *for some measurable function* $a : \mathbb{R} \to \mathbb{R}$. *But then we would have for any* $(\alpha, \beta) \in \mathbb{R}^2$ *with* $\alpha^2 + \beta^2 = 1$ *that*

$$f(\boldsymbol{X}^\top(\alpha\boldsymbol{\theta}^* + \beta\boldsymbol{\theta}^\circ)) = f(\alpha \boldsymbol{X}^\top \boldsymbol{\theta}^* + \beta a(\boldsymbol{X}^\top \boldsymbol{\theta}^*)) \stackrel{\text{def}}{=} f^\circ_{\alpha,\beta}(\boldsymbol{X}^\top \boldsymbol{\theta}^*),$$

such that the problem would no longer be identifiable. Also $|f'(t)| > 0$ *on some interval is necessary for identifyability of* $\boldsymbol{\theta}^*$.

Proposition 21 *Let* $\tau = o(p^{*-3/2})$ *and* $p^{*5}/n \to 0$. *With initial guess given by Equation* (25) *and for not too large* $\mathtt{x} > 0$ *the alternating sequence satisfies* (22) *and* (23) *with probability greater* $1 - 9\exp\{-\mathtt{x}\}$ *and where with some constant* $\mathtt{C} \in \mathbb{R}$

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

Proposition 22 *Take the initial guess given by Equation* (25). *Assume* $(\mathcal{A})$. *Further assume that* $p^{*4}/n \to 0$ *and* $\tau = o(p^{*-3/2})$. *Then we get the claim of Theorem 14 with* $\beta = \mathrm{e}^{-\mathtt{x}}$ *and*

$$\kappa(\mathtt{x}, R_0) = O(\tau p^{*3/2} + \sqrt{\tau \mathtt{x}} p^{*3/2}/n^{1/4}) + O(p^{*2}/\sqrt{n}) \to 0,$$

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

Remark 23 *The constraint* $\tau = o(p^{*-3/2})$ *implies that for the calculation of the initial guess the vector* $\widetilde{\boldsymbol{\eta}}_{0,l}$ *of* (26) *and the functional* $\mathcal{L}(\cdot)$ *have to be evaluated* $N = p^{*3(p-1)/2}$ *times.*

For details and proofs see (Andresen, 2015).

4. Proof of Theorem 7

In this section we present the proof of Theorem 7. As the proof is quite technical and complex we want to first explain the basic ideas of the proof. In a second section we will outline more clearly the steps of the proof. Finally we carry out each of these steps which combine to yield the proof.

4.1 Idea of the Proof

To ease the understanding of what follows in the subsequent sections we want to illustrate the central ideas with a simple model. Consider for some positive definite matrix $\mathcal{D} \in \mathbb{R}^{p^* \times p^*}$ and some vector $\boldsymbol{\upsilon}^* = (\boldsymbol{\theta}^*, \boldsymbol{\eta}^*) \in \mathbb{R}^{p+m} = \mathbb{R}^{p^*}$ the model

$$\mathbb{Y} = \boldsymbol{\upsilon}^* + \boldsymbol{\varepsilon} \in \mathbb{R}^{p^*}, \text{ where } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathcal{D}^{-2}), \quad \mathcal{D}^2 = \begin{pmatrix} \mathrm{D}^2 & A \\ A^\top & \mathrm{H}^2 \end{pmatrix} \in \mathbb{R}^{p^* \times p^*}.$$

Set $\mathcal{L}$ to be the true log likelihood of the observations, i.e.

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

With any starting initial guess $\widetilde{\boldsymbol{\upsilon}}_0 \in \mathbb{R}^{p+m}$ we obtain from (3) for $k \in \mathbb{N}$ and the usual first order criterion of maximality the following two equations

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

Combining these two equations we derive, assuming $\|\mathrm{D}^{-1} A \mathrm{H}^{-2} A^\top \mathrm{D}^{-1}\| \overset{\text{def}}{=} \|\boldsymbol{M}_0\| = \nu < 1$

$$\begin{aligned} \mathrm{D}(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) &= \mathrm{D}^{-1}(\mathrm{D}^2 \boldsymbol{\varepsilon}_{\boldsymbol{\theta}} - A\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}) + \mathrm{D}^{-1} A \mathrm{H}^{-1} A^\top \mathrm{D}^{-1} \mathrm{D}(\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*) \\ &= \sum_{l=1}^{k} \boldsymbol{M}_0^{k-l} \mathrm{D}^{-1}(\mathrm{D}^2 \boldsymbol{\varepsilon}_{\boldsymbol{\theta}} - A\boldsymbol{\varepsilon}_{\boldsymbol{\eta}}) + \boldsymbol{M}_0^k \mathrm{D}(\widetilde{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*) \\ &\to \mathbf{X} \overset{\text{def}}{=} \mathrm{D}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*). \end{aligned}$$

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

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

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 7 .

First it is ensured that the whole sequence $(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)})_{k\in\mathbb{N}_0}$ satisfies for some $\mathrm{R}_0(\mathtt{x}) > 0$ and with probability greater that $1 - \mathrm{e}^{-\mathtt{x}}$

$$\{\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)},\, k \in \mathbb{N}_0\} \subset \{\|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathrm{R}_0(\mathtt{x})\}, \tag{27}$$

where $\mathcal{D}^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}\mathcal{L}(\boldsymbol{\upsilon}^*)$ (see Theorem 28). In the second step we approximate with $\zeta = \mathcal{L} - \mathbb{E}\mathcal{L}$

$$\mathcal{L}(\boldsymbol{\upsilon}) - \mathcal{L}(\boldsymbol{\upsilon}^*) = \nabla\zeta(\boldsymbol{\upsilon}^*)(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2/2 + \alpha(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*), \tag{28}$$

where $\alpha(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*)$ is defined by (28). Similar to the toy case above this allows using the first order criterion of maximality and (27) to obtain a bound of the kind

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

This is done in Lemma 32 using results from (Andresen and Spokoiny, 2014) to show that $\epsilon(\mathrm{R}_0)$ is small. Finally the same arguments as in (Andresen and Spokoiny, 2014) 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 (28) we use the approximation

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

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

4.2 A Desirable Set

In this section we will explain the agenda of the proof. 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 - with the local set $\Upsilon_\circ(\mathtt{r})$ defined in (15) - for some $K \in \mathbb{N}$ the set

$$\begin{aligned}
\Omega(\mathtt{x}) &= \bigcap_{k=0}^{K} (C_{k,k} \cap C_{k,k+1}) \cap C(\nabla) \cap \{\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0) - \mathcal{L}(\boldsymbol{\upsilon}^*) \geq -\mathrm{K}_0\}, \text{ where} \qquad (29)\\
C_{k,k(+1)} &= \Big\{ \|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \boldsymbol{\upsilon}^*)\| \leq \mathrm{R}_0(\mathtt{x}),\ \|\mathrm{D}(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)\| \leq \mathrm{R}_0(\mathtt{x}),\\
&\qquad \|\mathrm{H}(\widetilde{\boldsymbol{\eta}}_{k(+1)} - \boldsymbol{\eta}^*)\| \leq \mathrm{R}_0(\mathtt{x}) \Big\},\\
C(\nabla) &= \bigcap_{\mathtt{r} \leq \mathrm{R}_0(\mathtt{x})} \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \left\{ \frac{1}{6\epsilon\nu_1} \|\mathcal{Y}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2 \right\} \leq \mathfrak{z}_Q(\mathtt{x}, 4p^*)^2 \right\}\\
&\qquad \bigcap_{\mathtt{r} \leq 4\mathrm{R}_0(\mathtt{x})} \left\{ \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \left\{ \frac{1}{6\check{\epsilon}\check{\nu}_1} \|\check{\mathcal{Y}}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2 \right\} \leq \mathfrak{z}_Q(\mathtt{x}, 2p^* + 2p)^2 \right\}\\
&\qquad \cap \Big\{ \max\{\|\mathcal{D}^{-1}\nabla\mathcal{L}\|, \|\mathrm{D}^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}\|, \|\mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}\|\} \leq \mathfrak{z}(\mathtt{x}) \Big\}\\
&\qquad \cap \{\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0(\mathtt{x}))\}.
\end{aligned}$$

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

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

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

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

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

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

The constant $\mathfrak{z}(\mathtt{x})$ in the definition of $C(\nabla)$ 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. It is defined as

$$\mathfrak{z}(\mathtt{x}) \stackrel{\text{def}}{=} \mathfrak{z}(\mathtt{x}, \mathcal{D}^{-1}\mathcal{V}^2\mathcal{D}^{-1}) \vee \mathfrak{z}_Q(\mathtt{x}, 4p^*) \approx \sqrt{p^* + \mathtt{x}}, \qquad (30)$$

where $\mathcal{V}^2 \in \mathbb{R}^{p^* \times p^*}$ denotes the *covariance matrix* from Section 2.1

$$\mathcal{V}^2 \stackrel{\text{def}}{=} \mathrm{Cov}(\nabla\mathcal{L}(\boldsymbol{\upsilon}^*)).$$

Remark 24 *$\mathfrak{z}(\mathtt{x}, \cdot)$ is explained in more detail in Section A and $\mathfrak{z}_Q(\mathtt{x}, \cdot)$ is defined in Equation (42). The constant $\mathfrak{z}(\mathtt{x}, I\!B)$ is comparable to the "$1 - \mathrm{e}^{-\mathtt{x}}$"-quantile of the norm*

of $\boldsymbol{X}$, *where* $\boldsymbol{X} \sim \mathcal{N}(0, I\!\!B)$, *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* $\mathbb{Q}$.

Remark 25 *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* $\mathbb{P}(\widetilde{\boldsymbol{\upsilon}}, \widetilde{\boldsymbol{\upsilon}}_{\boldsymbol{\theta}^*} \in \Upsilon_\circ(\mathtt{r}_0)) \geq 1 - \mathrm{e}^{-\mathtt{x}}$. *Note that condition* $(\mathcal{E}\mathtt{r})$ *together with* $(\mathcal{L}\mathtt{r})$ *allow to set* $\sqrt{p^* + \mathtt{x}} \approx \mathtt{r}_0 \leq R_0$ *(see Theorem 28).*

In Section 4.3 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.

$\{\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \geq -\mathbf{K}_0\}$**:** This set ensures, that the first guess satisfies $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \geq -\mathrm{K}_0$, which means that it is close enough to the target $\boldsymbol{\upsilon}^* \in \mathbb{R}^{p^*}$. 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 $\boldsymbol{\upsilon}^* \in \Upsilon$ with Theorem 28.

$\{\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \boldsymbol{\upsilon}^*) \leq \mathbf{R}_0(\mathtt{x})\}$**:** As just mentioned this event is of high probability due to $\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0, \boldsymbol{\upsilon}^*) \geq -\mathrm{K}_0$ and Theorem 28. This allows to concentrate the analysis on the set $\Upsilon_\circ(\mathrm{R}_0)$ on which Taylor expansions of the functional $\mathcal{L} : \mathbb{R}^{p^*} \to \mathbb{R}$ 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 $\mathfrak{z}(\mathtt{x}) \in \mathbb{R}$. Consequently we can derive in the proof of Lemma 32 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 $(\mathtt{r}_k) \subset \mathbb{R}_+$ satisfying $\limsup_{k\to\infty} \mathtt{r}_k = \mathtt{C}\mathfrak{z}(\mathtt{x})$. Further this set allows to obtain in Lemma 34 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 (Andresen and Spokoiny, 2014) to obtain the desired result with accuracy measured by $\breve{\diamondsuit}_Q(\mathtt{r}_k, \mathtt{x})$.

The sketch in figure 4.2 illustrates the behavior of the first steps of the procedure. The axes correspond to the $\boldsymbol{\theta}$- or $\boldsymbol{\eta}$-subspaces respectively. The two ellipsoids with center $\boldsymbol{\upsilon}^*$ and solid frame represent the local sets $\Upsilon_\circ(R_K) \subset \Upsilon_\circ(\mathrm{R}_0)$, with $R_K > 0$ from Remark 5. We see that the initial guess $\widetilde{\boldsymbol{\upsilon}}_0$ lies in $\Upsilon_\circ(R)$. The elements $(\boldsymbol{\upsilon}_{k,k(+1)})$ of the alternating sequence all land inside of the respective $\Upsilon_\circ(\mathtt{r}_k)$, which are represented by shrinking ellipsoids centered in $\boldsymbol{\upsilon}^*$ with doted frames. Note that not the set $\Upsilon_\circ(R_K)$ but $\Upsilon_\circ(\mathrm{R}_0)$ contains all points of the sequence.

4.3 Probability of Desirable Set

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

Lemma 26 *The set* $C(\nabla)$ *satisfies*

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

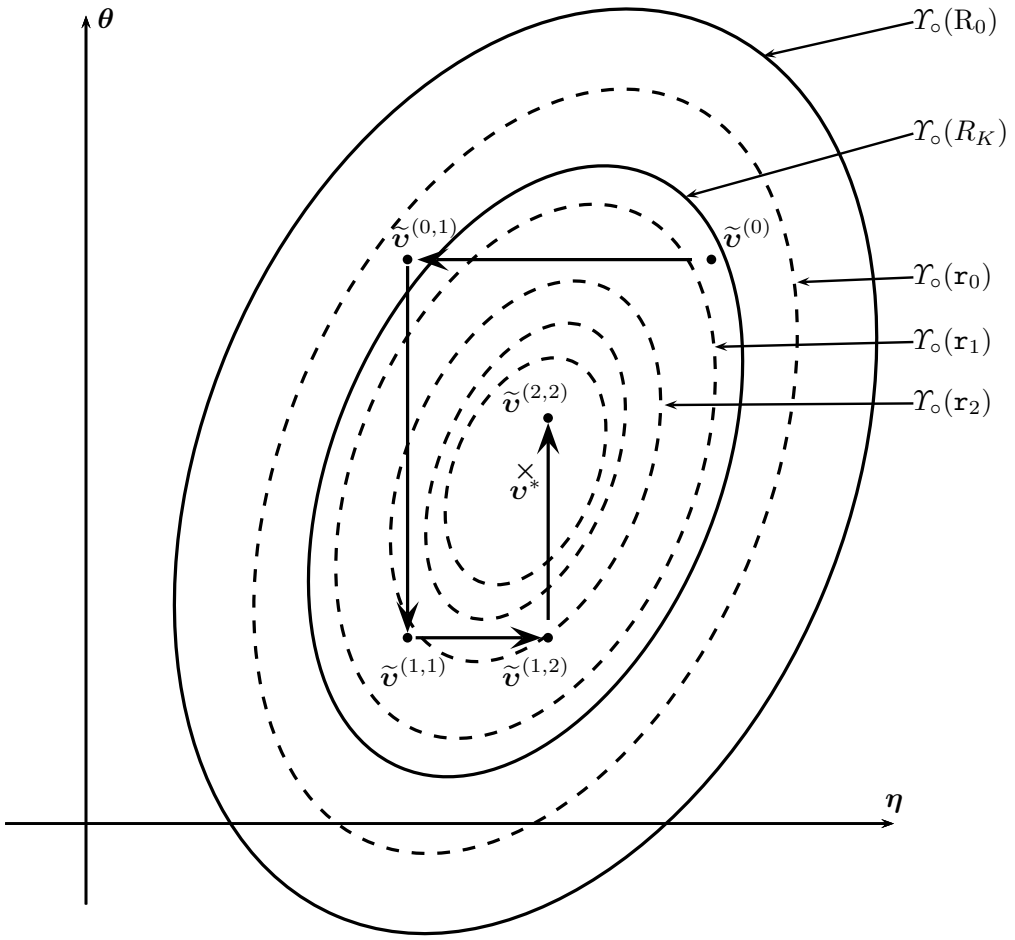

Figure 1: The behavior of the procedure for the first 4 steps of the alternating algorithm.

Proof The proof is similar to the proof of Theorem 3.1 in (Spokoiny, 2012). Denote

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

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

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

We estimate

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

We bound using for both terms Theorem 42 which is applicable due to $(\mathcal{ED}_1)$ and $(\breve{\mathcal{E}}\mathcal{D}_1)$:

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

For the set $\mathcal{C} \subset \Omega$ observe that we can use $(\boldsymbol{\mathcal{I}})$ and Lemma 27 to find

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

This implies that

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

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

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

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

We cite Lemma B.2 of (Andresen and Spokoiny, 2014):

Lemma 27 *Let*

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

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

Then for any $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^{p+m}$ *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 (Spokoiny, 2012) shows that it actually yields the following modified version:

Theorem 28 ((Spokoiny, 2012), 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

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

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

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

With assumption (B_1) and

$$\mathrm{R}_0(\mathtt{x})=\frac{6\nu_{\mathtt{r}}}{\mathtt{b}(1-\nu)}\sqrt{\mathtt{x}+\mathbb{Q}+\frac{\mathtt{b}}{9\nu_{\mathtt{r}}^2}\mathrm{K}_0},$$

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 28

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

Remark 29 *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* $\{\boldsymbol{\theta}\in\mathbb{R}^p,(\boldsymbol{\theta},\boldsymbol{\eta})\in\Upsilon_\circ(R_0)\}$ *or* $\{\boldsymbol{\eta}\in\mathbb{R}^m,(\boldsymbol{\theta},\boldsymbol{\eta})\in\Upsilon_\circ(R_0)\}$.

Remark 30 *To address the claim of Remark 5 we present the following lemma:*

Lemma 31 *On the set* $C(\nabla) \cap \{\widetilde{\boldsymbol{\upsilon}}_0 \in \Upsilon_\circ(R_K)\}$ *it holds*

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

Proof *With similar arguments as in the proof of Lemma 32 we have on* $C(\nabla) \subset \Omega$ *that*

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\upsilon}_0) - \mathcal{L}(\boldsymbol{\upsilon}^*) \\
&\quad \geq \mathbb{E}[\mathcal{L}(\boldsymbol{\upsilon}_0) - \mathcal{L}(\boldsymbol{\upsilon}^*)] - \|\mathcal{D}^{-1}\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\|R - |\{\nabla\boldsymbol{\zeta}(\widehat{\boldsymbol{\upsilon}}) - \nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\}(\boldsymbol{\upsilon}_0 - \boldsymbol{\upsilon}^*)| \\
&\quad \geq -\|\mathcal{D}(\boldsymbol{\upsilon}_0 - \boldsymbol{\upsilon}^*)\|^2/2 - \|\mathcal{D}^{-1}\nabla\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\|R \\
&\qquad -\|\mathcal{D}^{-1}\big\{\nabla\mathcal{L}(\widehat{\boldsymbol{\upsilon}}) - \nabla\mathcal{L}(\boldsymbol{\upsilon}^*)\big\}\|R - \epsilon R_K^2 \\
&\quad \geq -R^2/2 - \mathfrak{z}(\mathtt{x})R - \epsilon(1 + 12\nu_0)\left(R^2 + \mathfrak{z}(\mathtt{x})^2\right).
\end{aligned}$$

■

4.4 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 $(\mathtt{r}_k) \subset \mathbb{R}_+$ using the argument of Section 4.1, where $\limsup \mathtt{r}_k \approx \mathfrak{z}(\mathtt{x})$. For this purpose we define the *parametric uniform spread*

$$\diamondsuit_Q(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} \epsilon\left\{2\mathtt{r}^2 + \mathfrak{z}_Q(\mathtt{x}, 4p^*)^2\right\}(1 + 6\nu_1^2). \tag{31}$$

Lemma 32 *Assume that for some sequence* $(\mathtt{r}_k^{(l)})_{k\in\mathbb{N}}$

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

Then under the assumptions of Theorem 7 we get on $\Omega(\mathtt{x})$ *for all* $k \in \mathbb{N}_0$

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

Proof

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

$$\begin{aligned}
\mathrm{D}(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) &= \mathrm{D}^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\upsilon}^*) - \mathrm{D}^{-1}\mathcal{A}(\widetilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}^*) + \boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big), \\
\mathrm{H}(\widetilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}^*) &= \mathrm{H}^{-1}\nabla_{\boldsymbol{\eta}}\mathcal{L}(\boldsymbol{\upsilon}^*) - \mathrm{H}^{-1}\mathcal{A}^\top(\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*) + \boldsymbol{\tau}\big(\mathtt{r}_k^{(l)}\big),
\end{aligned} \tag{32}$$

where - with $\diamondsuit_Q(\mathtt{r},\mathtt{x})$ defined in (31) -

$$\|\boldsymbol{\tau}(\mathtt{r})\| \leq \diamondsuit_Q(\mathtt{r},\mathtt{x}).$$

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 (Andresen and Spokoiny, 2014). By assumption on $\Omega(\mathtt{x})$ we have $\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} \in \Upsilon_\circ\big(\mathtt{r}_k^{(l)}\big)$. Define with $\zeta = \mathcal{L} - \mathbb{E}\mathcal{L}$

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

Note that

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

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

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

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

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

where

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

To see this note first that with Lemma 27 $\|\mathrm{D}^{-1}\Pi_{\boldsymbol{\theta}}\mathcal{D}\boldsymbol{\upsilon}\| \leq \|\mathcal{D}^{-1}\mathcal{D}\boldsymbol{\upsilon}\|$. This gives by condition $(\boldsymbol{\mathcal{L}_0})$, Lemma 27 and Taylor expansion

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

For the remainder note that again with Lemma 27

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

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

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

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

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

To see this we first use the inequality

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

Now we find with (32)

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

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

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

and

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

to derive the recursive formula

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

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

Lemma 33 *Assume the same as in Theorem 7 . Further assume that* (17) *is met with* $\mathtt{C}(\nu)$ *defined in* (18)*. Then*

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

where - with $\mathtt{C}(\nu) \geq 0$ *defined in* (18) -

$$\begin{aligned} \mathtt{r}_k^* \leq & \left(\mathtt{C}(\nu) + \frac{4\epsilon \mathtt{C}(\nu)^4 \mathfrak{z}(\mathtt{x})}{1 - \epsilon \mathtt{C}(\nu) \mathfrak{z}(\mathtt{x})} \right) 2\mathfrak{z}(\mathtt{x}) \\ & + \nu^k \left(\mathtt{C}(\nu) + \frac{4\epsilon \mathtt{C}(\nu)^4 R_0}{1 - \epsilon \mathtt{C}(\nu) R_0} \right) R_0. \end{aligned} \tag{33}$$

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

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

Now with Lemma 32 we find that if

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

that then

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

where

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

Setting $l = 1$ this gives

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

We show that

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

So we have to show that $\limsup_{l\to\infty} \mathtt{r}_k^{(l)} \le \mathtt{r}_k^*$ in (33). For this we estimate further

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

where $\mathtt{C}(\nu) > 0$ is defined in (18). We set

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

Claim

$$\begin{aligned}
A_{s,k}^{(l)} &\le 4^{\sum_{t=0}^{s-1} 2^t} \mathtt{C}(\nu)^{2^s} \left\{\left(\frac{1}{1-\nu}\right)^{\sum_{t=0}^{s-1} 2^t} \left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^{2^s} \right. \\
&\qquad \left. +\nu^k \left(\frac{1}{\nu^{-1}-1}\right)^{\sum_{t=0}^{s-1} 2^t} \mathrm{R}_0^{2^s}\right\} \\
&\quad +4^{\sum_{t=0}^{s-1} 2^t} (\mathtt{C}(\nu)\epsilon)^{2^s} A_{s+1,k}^{(l-1)}.
\end{aligned} \tag{34}$$

We proof this claim via induction. Clearly

$$\begin{aligned}
A_{1,k}^{(l)} &= \sum_{r_1=0}^{k-1} \nu^{r_1} \left(\mathtt{r}_{k-r_1}^{(l-1)}\right)^2 \le 4\mathtt{C}(\nu)^2 \sum_{r_1=0}^{k-1} \nu^{r_1} \left\{\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^2 + \nu^{2(k-r_1)} \mathrm{R}_0^2\right\} \\
&\quad +4\mathtt{C}(\nu)^2\epsilon^2 \sum_{r_1=0}^{k-1} \nu^{r_1} \left(\sum_{r_2=0}^{k-r_1-r_2-1} \nu^{r_2} \left(\mathtt{r}_{k-r_1-r_2}^{(l-2)}\right)^2\right)^2 \\
&\le 4\mathtt{C}(\nu)^2 \left\{\frac{1}{1-\nu}\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^2 + \frac{\nu^k}{\nu^{-1}-1} \mathrm{R}_0^2\right\} \\
&\quad +4\mathtt{C}(\nu)^2\epsilon^2 A_{2,k}^{(l-1)}.
\end{aligned}$$

Furthermore

$$A_{s,k}^{(l)} \stackrel{\text{def}}{=} \sum_{r_1=0}^{k-1} \nu^{r_1} \left(\sum_{r_2=0}^{k-r_1-1} \nu^{r_2} \left(\dots \sum_{r_s=0}^{k-r_1-\dots-r_{s-1}-1} \nu^{r_s} \left(\mathtt{r}_{k-r_1-\dots-r_s}^{(l-1)} \right)^2 \dots \right) \right)^2$$
$$= \sum_{r_1=0}^{k-1} \nu^{r_1} \left(A_{s-1,k-r_1}^{(l)} \right)^2. \tag{35}$$

Plugging in (34) we get for $s \geq 2$

$$A_{s,k}^{(l)} \leq \sum_{r_1=0}^{k-1} \nu^{r_1} \left(4^{\sum_{t=0}^{s-2} 2^t} \mathtt{C}(\nu)^{2^{s-1}} \left\{ \left(\frac{1}{1-\nu} \right)^{\sum_{t=0}^{s-2} 2^t} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^{2^{s-1}} \right.\right.$$
$$\left.\left. + \nu^k \left(\frac{1}{\nu^{-1}-1} \right)^{\sum_{t=0}^{s-2} 2^t} \mathrm{R}_0^{2^{s-1}} \right\} \right.$$
$$\left. + 4^{\sum_{t=0}^{s-2} 2^t} (\mathtt{C}(\nu)\epsilon)^{2^{s-1}} A_{s,k-r_1}^{(l-1)} \right)^2.$$

Shifting the index this gives

$$A_{s,k}^{(l)} \leq 4 \sum_{r_1=0}^{k-1} \nu^{r_1} \left(4^{\sum_{t=1}^{s-1} 2^t} \mathtt{C}(\nu)^{2^s} \left\{ \left(\frac{1}{1-\nu} \right)^{\sum_{t=1}^{s-1} 2^{t-1}} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^{2^s} \right.\right.$$
$$\left.\left. + \nu^k \left(\frac{1}{\nu^{-1}-1} \right)^{\sum_{t=1}^{s-1} 2^t} \mathrm{R}_0^{2^s} \right\} \right.$$
$$\left. + 4^{\sum_{t=1}^{s-1} 2^t} (\mathtt{C}(\nu)\epsilon)^{2^s} (A_{s,k-r_1}^{(l-1)})^2 \right).$$

Direct calculation then leads to

$$A_{s,k}^{(l)} \leq 4^{\sum_{t=0}^{s-1} 2^t} \mathtt{C}(\nu)^{2^s} \left\{ \left(\frac{1}{1-\nu} \right)^{\sum_{t=0}^{s-1} 2^t} \left(\mathfrak{z}(\mathtt{x}) + \epsilon \mathfrak{z}(\mathtt{x})^2 \right)^{2^s} \right.$$
$$\left. + \nu^k \left(\frac{1}{\nu^{-1}-1} \right)^{\sum_{t=0}^{s-1} 2^t} \mathrm{R}_0^{2^s} \right\}$$
$$+ 4^{\sum_{t=0}^{s-1} 2^t} (\mathtt{C}(\nu)\epsilon)^{2^s} \sum_{r_1=0}^{k-1} \nu^{r_1} (A_{s,k-r_1}^{(l-1)})^2,$$

which gives (34) with (35). Similarly we can prove

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

Abbreviate

$$\lambda_s \stackrel{\text{def}}{=} 4^{2^s-1}\mathtt{C}(\nu)^{2^s}, \quad \beta_s \stackrel{\text{def}}{=} 4^{2^s-1}(\mathtt{C}(\nu)\epsilon)^{2^s},$$
$$\mathfrak{z}_s(\mathtt{x}) \stackrel{\text{def}}{=} \left(\frac{1}{1-\nu}\right)^{2^s-1} \left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^{2^s}, \quad \mathrm{R}_s \stackrel{\text{def}}{=} \left(\frac{1}{\nu^{-1}-1}\right)^{2^s-1} \mathrm{R}_0^{2^s}.$$

Then

$$\begin{aligned}
\mathtt{r}_k^{(l)} &\le \mathtt{C}(\nu)\left\{\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right) + \nu^k \mathrm{R}_0 + \epsilon A_{1,k}^{(l)}\right\} \\
&\le \sum_{s=0}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathfrak{z}_s(\mathtt{x}) + \nu^k \sum_{s=0}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathrm{R}_s + \prod_{r=0}^{l-1} \beta_r \mathrm{R}_l. \qquad (36)
\end{aligned}$$

We estimate further

$$\begin{aligned}
&\sum_{s=0}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathfrak{z}_s(\mathtt{x}) - \mathtt{C}(\nu)\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right) = \sum_{s=1}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathfrak{z}_s(\mathtt{x}) \\
&\quad \le \sum_{s=1}^{l-1} 4^{2^s}\mathtt{C}(\nu)^{2^{s+1}}\epsilon^{2^s-1}\left(\frac{1}{1-\nu}\right)^{2^s-1}\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^{2^s} \\
&\quad = \epsilon 4^2 \mathtt{C}(\nu)^4 \left(\frac{1}{1-\nu}\right)\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)^2 \\
&\qquad \sum_{s=1}^{l-1} \left(\epsilon 4\mathtt{C}(\nu)\frac{1}{1-\nu}\left(\mathfrak{z}(\mathtt{x}) + \epsilon\mathfrak{z}(\mathtt{x})^2\right)\right)^{2^s-1}.
\end{aligned}$$

Assuming (17) and the definition of $\mathrm{R}_0 > \mathfrak{z}(\mathtt{x})$ this gives

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

With the same argument we find under (17) that

$$\nu^k \sum_{s=0}^{l-1} \lambda_s \prod_{r=0}^{s-1} \beta_r \mathrm{R}_s \le \nu^k \left(\mathtt{C}(\nu) + \frac{4\epsilon\mathtt{C}(\nu)^4\mathrm{R}_0}{1-\epsilon\mathtt{C}(\nu)\mathrm{R}_0}\right)\mathrm{R}_0.$$

Additionally (17) implies

$$\prod_{r=0}^{l-1} \beta_r \mathrm{R}_r \le \left(\epsilon 4\mathtt{C}(\nu)\frac{1}{\nu^{-1}-1}\right)^{2^{l-1}} \mathrm{R}_0^{2^l} \to 0.$$

Plugging these bounds into (36) and letting $l \to \infty$ gives the claim. ■

4.5 Result after Convergence

In the previous section we showed that

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

where $\mathtt{r}_k^*$ is defined in (33). The claim of Theorem 7 follows with the following lemma:

Lemma 34 *Assume* $(\breve{\mathcal{E}}\mathcal{D}_1)$, $(\breve{\mathcal{L}}_0)$, *and* $(\mathcal{I})$. *Then it holds on* $\Omega(\mathtt{x}) \subseteq \epsilon$ *that for all* $k \in \mathbb{N}$

$$\left\| \breve{D}\big(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\big) - \breve{\boldsymbol{\xi}} \right\| \le \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}), \tag{37}$$

$$\left| 2\breve{L}(\widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}^*) - \|\breve{\boldsymbol{\xi}}\|^2 \right| \le 5 \left(\|\breve{D}^{-1}\breve{\nabla}\| + \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}) \right) \breve{\Diamond}_Q(\mathtt{r}_k, \mathtt{x}), \tag{38}$$

where the spread $\breve{\Diamond}(\mathtt{r}, \mathtt{x})$ *is defined in* (21) *and where*

$$\mathtt{r}_k \stackrel{\mathrm{def}}{=} \mathtt{r}_k^* \vee \mathtt{r}_0.$$

Proof The proof is nearly the same as that of Theorem 2.2 of (Andresen and Spokoiny, 2014) which is inspired by the proof of Theorem 1 of (Murphy and van der Vaart, 2000). So we only sketch it and refer the reader to (Andresen and Spokoiny, 2014) for the skipped arguments. We define

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

Note that

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

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

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

Now the right hand side can be bounded just as in the proof of Theorem 2.2 of (Andresen and Spokoiny, 2014). This gives (37).

For (38) 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

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

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

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

Again the remaining steps are exactly the same as in the proof of Theorem 2.2 of (Andresen and Spokoiny, 2014).

■

5. Proof of Corollary 13

Proof Note that with the argument of Section 4.3 $\mathbb{P}(\epsilon'(\mathtt{x})) \geq 1 - 8\mathrm{e}^{-\mathtt{x}} - \beta$ where with $\Omega(\mathtt{x})$ from (29)

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

On $\epsilon'(\mathtt{x})$ it holds due to Theorem 7 and due to Theorem 2.1 of (Andresen and Spokoiny, 2014)

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

Now the claim follows with the triangular inequality and noting that $\breve{\Diamond}(\mathtt{r}_0, \mathtt{x}) \leq \breve{\Diamond}_Q(\mathtt{r}_0, \mathtt{x})$.
■

6. Proof of Theorem 14

We prove this Theorem in a similar manner to the convergence result in Lemma 32. Redefine the set $\Omega(\mathtt{x})$

$$\Omega(\mathtt{x}) \stackrel{\text{def}}{=} \bigcap_{k=0}^{K} (C_{k,k} \cap C_{k,k+1}) \cap C(\nabla) \cap \{\mathcal{L}(\widetilde{\boldsymbol{\upsilon}}_0) - \mathcal{L}(\boldsymbol{\upsilon}^*) \geq -\mathrm{K}_0\}, \text{ where} \tag{39}$$

$$\begin{aligned} C_{k,k(+1)} &= \Big\{\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \boldsymbol{\upsilon}^*)\| \leq \mathrm{R}_0(\mathtt{x}),\ \|\mathrm{D}(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)\| \leq \mathrm{R}_0(\mathtt{x}), \\ &\qquad \|\mathrm{H}(\widetilde{\boldsymbol{\eta}}_{k(+1)} - \boldsymbol{\eta}^*)\| \leq \mathrm{R}_0(\mathtt{x})\Big\}, \\ C(\nabla) &= \left\{\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathrm{R}_0(\mathtt{x}))} \|\mathcal{Y}(\nabla^2)(\boldsymbol{\upsilon})\| \leq 9\nu_2\epsilon_2\mathfrak{z}_1(\mathtt{x}, 6p^*)\mathrm{R}_0(\mathtt{x})\right\} \\ &\qquad \cap\{\|\mathcal{D}^{-1}\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\| \leq \mathfrak{z}(\mathtt{x}, \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*))\}. \end{aligned}$$

where

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

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

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

Lemma 35 *Under the conditions of Theorem 14*

$$\mathbb{P}(\Omega(\mathtt{x})) \ge 1 - 3\mathrm{e}^{-\mathtt{x}} - \beta.$$

Proof The proof is very similar to the one presented in Section 4.3, so we only give a sketch. By assumption

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

and due to $(\mathcal{ED}_2)$ with Theorem 47

$$\mathbb{P}\left(\sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathrm{R}_0(\mathtt{x}))} \|\mathcal{Y}(\nabla^2)(\boldsymbol{\upsilon})\| \le 9\nu_2\epsilon_2\mathfrak{z}_1(\mathtt{x}, 6p^*)\mathrm{R}_0(\mathtt{x})\right) \ge 1 - \mathrm{e}^{-\mathtt{x}}.$$

■

Lemma 36 *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 $\Omega(\mathtt{x})$

$$\begin{aligned}\left\|\mathcal{D}(\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} - \widetilde{\boldsymbol{\upsilon}})\right\| &\le 2\sqrt{2}(1+\sqrt{\nu})\sum_{r=0}^{k-1}\nu^r\|\boldsymbol{\tau}(\mathtt{r}_{k-r}^{(l)})\| + 2\sqrt{2}\nu^k(R_0 + \mathtt{r}_0),\\ &=: \mathtt{r}_k^{(l+1)}. \end{aligned} \tag{40}$$

where

$$\|\boldsymbol{\tau}(\mathtt{r})\| \le \left[\epsilon R_0 + 9\nu_2\epsilon_2\|\mathcal{D}^{-1}\|\mathfrak{z}_1(\mathtt{x}, 6p^*)R_0 + \|\mathcal{D}^{-1}\|\mathfrak{z}(\mathtt{x}, \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*))\right]\mathtt{r}.$$

Proof

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

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

The proof is very similar to that of Lemma 32. Define

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

Note that

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

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

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

We want to show

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

where

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

To see this note that by assumption we have $\Omega(\mathtt{x}) \subseteq \{\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathtt{r}_0)\} \subseteq \{\widetilde{\boldsymbol{\upsilon}} \in \Upsilon_\circ(\mathrm{R}_0)\}$. By condition $(\boldsymbol{\mathcal{L}_0})$, Lemma 27 and Taylor expansion we have

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

For the remainder note that with $\zeta = \mathcal{L} - \mathbb{E}\mathcal{L}$ on $\Omega(\mathtt{x})$ using Lemma 27 we can bound

$$\begin{aligned}
&\sup_{(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\in\widetilde{\Upsilon}_\circ(\mathtt{r}_k^{(l)})\cap\Upsilon_\circ(\mathrm{R}_0)} \Big\|\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k) - \mathbb{E}\mathcal{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\widetilde{\boldsymbol{\eta}}_k)\Big\| \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\widetilde{\Upsilon}_\circ(\mathtt{r}_k^{(l)})\cap\Upsilon_\circ(\mathrm{R}_0)} \Big\|\mathrm{D}^{-1}\Big(\nabla_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\widetilde{\boldsymbol{\upsilon}})\Big)\Big\| \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \big\|\mathcal{D}^{-1}\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon})\mathcal{D}^{-1}\big\|\,\mathtt{r}_k^{(l)} \\
&\quad\le \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathrm{R}_0)} \left\{\frac{1}{9\nu_2\epsilon_2}\|\mathcal{D}^{-1}\left(\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}) - \nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\right)\mathcal{D}^{-1}\|\right\} 6\nu_1\epsilon\mathtt{r}_k^{(l)} \\
&\qquad + \left\{\|\mathcal{D}^{-1}\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*)\mathcal{D}^{-1}\|\right\}\mathtt{r}_k^{(l)} \\
&\quad\le \left[9\nu_2\epsilon_2\|\mathcal{D}^{-1}\|_{\mathfrak{z}1}(\mathtt{x},6p^*)\mathrm{R}_0 + \|\mathcal{D}^{-1}\|_{\mathfrak{z}}(\mathtt{x},\nabla^2\boldsymbol{\zeta}(\boldsymbol{\upsilon}^*))\right]\mathtt{r}_k^{(l)}.
\end{aligned}$$

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

Now the claim follows as in the proof of Lemma 32. ∎

Lemma 37 *Assume that* $\kappa(\mathtt{x},R_0) < 1-\nu$ *where*

$$\begin{aligned}
\kappa(\mathtt{x},R_0) \stackrel{\mathrm{def}}{=} \frac{2\sqrt{2}(1+\sqrt{\nu})}{\sqrt{1-\nu}}\Big(&\epsilon R_0 + 9\epsilon_2\nu_2\|\mathcal{D}^{-1}\|_{\mathfrak{z}1}(\mathtt{x},6p^*)R_0 \\
&+ \|\mathcal{D}^{-1}\|_{\mathfrak{z}2}\left(\mathtt{x},\nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right)\Big).
\end{aligned}$$

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* (24).

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

$$\|\boldsymbol{\tau}(\mathtt{r}_k^{(l)})\| \le \frac{1}{\sqrt{1-\nu}}\left(\epsilon\mathrm{R}_0 + 6\nu_1\epsilon_2\|\mathcal{D}^{-1}\|_{\mathfrak{z}1}(\mathtt{x},6p^*)\mathrm{R}_0 + \|\mathcal{D}^{-1}\|_{\mathfrak{z}}(\mathtt{x},\mathit{I\!B}(\nabla^2)\right)\mathtt{r}_k^{(l)},$$

such that by definition

$$2\sqrt{2}(1+\sqrt{\nu})\sum_{r=0}^{k-1}\nu^r\|\boldsymbol{\tau}(\mathtt{r}_{k-r}^{(l)})\| \le \kappa(\mathtt{x},\mathrm{R}_0)\sum_{r=0}^{k-1}\nu^r\mathtt{r}_{k-r}^{(l)}.$$

Plugging in the recursive formula for $\mathtt{r}_k^{(l)}$ from (40) and denoting $\widetilde{\mathrm{R}}_0 \stackrel{\text{def}}{=} \mathrm{R}_0 + \mathtt{r}_0$ we find

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

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

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

By Lemma 36

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

Set if $\kappa(\mathtt{x},\mathrm{R}_0)/(1-\nu) < 1$

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

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

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

$$\mathtt{r}_k^* \leq \begin{cases} \frac{\nu^k 2\sqrt{2}}{1-\kappa(\mathtt{x},\mathrm{R}_0)k} \widetilde{\mathrm{R}}_0, & \kappa(\mathtt{x},\mathrm{R}_0)k \leq 1, \\ 2\left(\frac{\kappa(\mathtt{x},\mathrm{R}_0)}{1-\nu}\right)^{\frac{k}{\log(k)}L(k)-1} \widetilde{\mathrm{R}}_0, & \text{otherwise.} \end{cases}$$

The sequence $L(k) > 0$ is defined as

$$L(k) \stackrel{\text{def}}{=} \left\lfloor \frac{\log(1/\nu) - \frac{1}{k}\left(\log(2\sqrt{2}) - \log(\kappa(\mathtt{x},\mathrm{R}_0)k - 1)\right)}{1 + \frac{1}{\log(k)}\log(1-\nu)} \right\rfloor \in \mathbb{N},$$

where $\lfloor x \rfloor \in \mathbb{N}_0$ denotes the largest natural number smaller than $x > 0$. To ensure that $L(k) > 0$ we assume that $k\log(1/\nu) - \log(2\sqrt{2}) > k$. Further as $\kappa(\mathtt{x},\mathrm{R}_0) < (1-\nu)$ and $L(k)$ is only relevant once $\kappa(\mathtt{x},\mathrm{R}_0)k > 1$ it follows that

$$0 < 1 + \frac{1}{\log(k)}\log(1-\nu) < 1.$$

Then

$$L(k) \geq \log(1/\nu) - \frac{1}{k}\Big(\log(2\sqrt{2}) - \log(\kappa(\mathtt{x},\mathrm{R}_0)k - 1)\Big) > 1.$$

Consequently

$$\begin{aligned} \left(\frac{\kappa(\mathtt{x},\mathrm{R}_0)}{1-\nu}\right)^{\frac{k}{\log(k)}L(k)} &\leq \nu^{\frac{k}{\log(k)}\log\left(\frac{1-\nu}{\kappa(\mathtt{x},\mathrm{R}_0)}\right)} \left(\frac{\kappa(\mathtt{x},\mathrm{R}_0)}{1-\nu}\right)^{-\frac{1}{\log(k)}\left(\log(2\sqrt{2}) - \log(\kappa(\mathtt{x},\mathrm{R}_0)k-1)\right)} \\ &\stackrel{\text{def}}{=} \nu^{\frac{k}{\log(k)}\log\left(\frac{1-\nu}{\kappa(\mathtt{x},\mathrm{R}_0)}\right)} c_k, \end{aligned}$$

where $c_k \to \frac{\kappa(\mathtt{x},\mathrm{R}_0)}{1-\nu}$. Finally note that $\widetilde{\mathrm{R}}_0 \leq 2\mathrm{R}_0$ and the proof is complete. ■

Remark 38 *As pointed out in Remark 18 the above result can be improved. Redefine* $\Omega(\mathtt{x})$ *as the intersection of the two sets in* (29) *and* (39). *Then* $\mathbb{P}(\Omega(\mathtt{x})) \geq 1 - 10\mathrm{e}^{-\mathtt{x}}$. *Also redefine*

$$\kappa(\mathtt{x},\mathtt{r}) \stackrel{\text{def}}{=} \frac{2\sqrt{2}(1+\sqrt{\nu})}{\sqrt{1-\nu}} \Big(\epsilon\mathtt{r} + 3\epsilon_2 \|\mathcal{D}^{-1}\|_{\mathfrak{z}1}(\mathtt{x}, 6p^*)\mathtt{r} + \|\mathcal{D}^{-1}\|_{\mathfrak{z}2}\left(\mathtt{x}, \nabla^2\mathcal{L}(\boldsymbol{\upsilon}^*)\right) \Big).$$

By the arguments of the proof of Theorem 7 we find with $\mathtt{r}_k^*$ *defined in* (33)

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

Using this in Lemma 36 instead of $\cap_{k\in\mathbb{N}}\{\boldsymbol{\upsilon}_{k,k(+1)} \in \Upsilon_\circ(R_0)\}$ *we can bound*

$$\begin{aligned}\|\boldsymbol{\tau}(\mathtt{r}_k^{(l)})\| \leq \frac{1}{\sqrt{1-\nu}}\Big(&\epsilon\mathtt{r}_k^* + 6\nu_1\epsilon_2\|\mathcal{D}^{-1}\|_{\mathfrak{z}1}(\mathtt{x},6p^*)\mathtt{r}_k^* \\ &+\|\mathcal{D}^{-1}\|_{\mathfrak{z}}(\mathtt{x},I\!B(\nabla^2)\Big)\mathtt{r}_k^{(l)}.\end{aligned}$$

Consequently, representing $\mathtt{r}_k^* = \mathtt{C}\left(\mathfrak{z}(\mathtt{x}) + \nu^k R_0\right)$ *we find*

$$\begin{aligned}\mathtt{r}_k^{(l)} &\leq \kappa(\mathtt{x},\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))\sum_{r=0}^{k-1}\nu^r\mathtt{r}_{k-r}^{(l-1)} + 2\sqrt{2}\nu^k(R_0+\mathtt{r}_0) \\ &\quad +\mathtt{C}\epsilon(1+\|\mathcal{D}^{-1}\|_{\mathfrak{z}1}(\mathtt{x},6p^*))\sum_{r=0}^{k-1}\nu^k R_0\mathtt{r}_{k-r}^{(l-1)} \\ &\leq \kappa(\mathtt{x},\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))\sum_{r=0}^{k-1}\nu^r\mathtt{r}_{k-r}^{(l-1)} + \mathtt{C}_1\epsilon R_0 k\nu^k(R_0+\mathtt{r}_0),\end{aligned}$$

where $\mathtt{C}_1 \leq 2\sqrt{2} + \mathtt{C}(1+\|\mathcal{D}^{-1}\|_{\mathfrak{z}1}(\mathtt{x},6p^*))$. *With the same arguments as in the proof of Lemma 37 we infer*

$$\begin{aligned}&\mathtt{r}_k^{(l)}/(R_0+\mathtt{r}_0) \\ &\quad\leq \begin{cases}\left(\left(\frac{\kappa(\mathtt{x},\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))}{1-\nu}\right)^l + k\nu^k\frac{\mathtt{C}_1\epsilon R_0}{1-\kappa(\mathtt{x},\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))k}\right), & \kappa(\mathtt{x},\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))k \leq 1, \\ \kappa(\mathtt{x},\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))^l\left(\left(\frac{1}{1-\nu}\right)^l + \nu^k\frac{k^{l+1}\mathtt{C}_1\epsilon R_0}{\kappa(\mathtt{x},\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))k-1}\right), & \text{otherwise.}\end{cases}\end{aligned}$$

Set

$$l(k) \stackrel{\text{def}}{=} \begin{cases}\infty, & \kappa(\mathtt{x},\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))k \leq 1, \\ \frac{k\log(\nu)+\log(\mathtt{C}_1\epsilon R_0)-\log(\kappa(\mathtt{x},\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))k-1)-\log(k)}{-\log(1-\nu)-\log(k)}, & \text{otherwise.}\end{cases}$$

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

$$\begin{aligned}\Omega(\mathtt{x}) &\subset \bigcap_{k\in\mathbb{N}_0}\left\{\widetilde{\boldsymbol{\upsilon}}_{k,k(+1)} \in \widetilde{\Upsilon}_\circ\big(\mathtt{r}_k^*\big)\right\}, \\ \mathtt{r}_k^* &\leq \begin{cases}\frac{\nu^k 2\sqrt{2}}{1-\kappa(\mathtt{x}\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))k}(R_0+\mathtt{r}_0), & \kappa(\mathtt{x},\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))k \leq 1, \\ 2\left(\frac{\kappa(\mathtt{x},\mathtt{C}_{\mathfrak{z}}(\mathtt{x}))}{1-\nu}\right)^{\frac{k}{\log(k)}L(k)-1}(R_0+\mathtt{r}_0), & \text{otherwise.}\end{cases}\end{aligned}$$

This gives the claim.

Acknowledgments

The first author was supported by Research Units 1735 "Structural Inference in Statistics: Adaptation and Efficiency". The research was partly supported by the Russian Science Foundation grant (project 14-50-00150). Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 649 "Economic Risk" is gratefully acknowledged.

Appendix A. Deviation Bounds for Quadratic Forms

This section is the same as Section A of Andresen and Spokoiny (2014). The following general result from Spokoiny (2012) helps to control the deviation for quadratic forms of type $\|I\!B\boldsymbol{\xi}\|^2$ for a given positive matrix $I\!B$ and a random vector $\boldsymbol{\xi}$. It will be used several times in our proofs. Suppose that

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

For a symmetric matrix $I\!B$, define

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

For ease of presentation, suppose that $\mathtt{g}^2 \ge 2\mathtt{p}_{I\!B}$. 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

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

Proposition 39 *Let (ED_0) hold with $\nu_0 = 1$ and $\mathtt{g}^2 \ge 2\mathtt{p}_{I\!B}$. Then for each $\mathtt{x} > 0$*

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

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

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

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

Appendix B. A Uniform Bound for the Norm of a Random Process

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

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

This is a slightly stronger result than the one derived in Section D of (Andresen and Spokoiny, 2014) but the ideas employed here are very similar.

We want to apply Corollary 2.5 of the supplement of Spokoiny (2012) 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 40 *Let $(U(\mathtt{r}))_{0 \le \mathtt{r} \le \mathtt{r}^*} \subset \mathbb{R}^p$ 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 $0 \le \mathtt{r} \le \mathtt{r}^*$ that $\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}^*) = 0$ and*

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

$$\log \mathbb{E} \exp\biggl\{ \lambda \frac{\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}) - \mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}^{\circ})}{d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^{\circ})} \biggr\} \leq \frac{\nu_0^2 \lambda^2}{2}, \qquad |\lambda| \leq \mathtt{g}. \tag{41}$$

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

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

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

$$\mathfrak{z}_Q(\mathtt{x}, \mathbb{Q})^2 \stackrel{\text{def}}{=} \begin{cases} (1 + \sqrt{\mathtt{x} + \mathbb{Q}})^2 & \text{if } 1 + \sqrt{\mathtt{x} + \mathbb{Q}} \leq \mathtt{g}_0, \\ 1 + \{2\mathtt{g}_0^{-1}(\mathtt{x} + \mathbb{Q}) + \mathtt{g}_0\}^2 & \text{otherwise.} \end{cases} \tag{42}$$

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

$$\sup_{\|\boldsymbol{u}\| \leq 1} \log \mathbb{E} \exp\biggl\{ \lambda \frac{\boldsymbol{u}^{\top} \bigl(\breve{\mathtt{y}}(\boldsymbol{\upsilon}) - \breve{\mathtt{y}}(\boldsymbol{\upsilon}^{\circ}) \bigr)}{\epsilon \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{\circ})\|} \biggr\} \leq \frac{\nu_0^2 \lambda^2}{2}. \tag{43}$$

Remark 41 *In the setting of Theorem 7 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 (43) *becomes* $(\mathcal{E}\mathcal{D}_1)$ *from 2.1.*

Theorem 42 *Let a random* p*-vector process* $\breve{\mathtt{y}}(\boldsymbol{\upsilon})$ *fulfill* $\breve{\mathtt{y}}(\boldsymbol{\upsilon}^*) = 0$ *and let condition* (43) *be satisfied. Then for each* $0 \leq \mathtt{r} \leq \mathtt{r}^*$, *on a set of probability greater* $1 - \mathrm{e}^{-\mathtt{x}}$

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

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

Remark 43 *Note that the entropy of the original set* $\Upsilon_{\circ}(\mathtt{r}) \subset \mathbb{R}^{p^*}$ *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

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

This implies

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

Due to Lemma 44 the process $\mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}, \boldsymbol{u}) \stackrel{\text{def}}{=} \frac{1}{\epsilon \|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*)\|} \boldsymbol{u}^\top \breve{\mathcal{Y}}(\boldsymbol{\upsilon})$ satisfies condition $(\mathcal{E}d)$ (see (41)) as process on $U(\mathtt{r}^*)$ where

$$U(\mathtt{r}) \stackrel{\text{def}}{=} \Upsilon_\circ(\mathtt{r}) \times B_{\mathtt{r}}(0). \tag{44}$$

Further $\sup_{(\boldsymbol{\upsilon},\boldsymbol{u}) \in U(\mathtt{r})} \mathcal{U}(\mathtt{r}, \boldsymbol{\upsilon}, \boldsymbol{u})$ is increasing in $\mathtt{r}$. This allows to apply Theorem 42 to obtain the desired result. Set $d((\boldsymbol{\upsilon}, \boldsymbol{u}), (\boldsymbol{\upsilon}^\circ, \boldsymbol{u}^\circ))^2 = \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 + \|\boldsymbol{u} - \boldsymbol{u}^\circ\|^2$. We get on a set of probability greater $1 - \mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned} &\sup_{(\boldsymbol{\upsilon},\boldsymbol{u}) \in U(\mathtt{r}^*)} \left\{ \frac{1}{6\epsilon\nu_1 \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|} \boldsymbol{u}^\top \breve{\mathcal{Y}}(\boldsymbol{\upsilon}) - \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 - \|\boldsymbol{u}\|^2 \right\} \\ &\quad \le \mathfrak{z}_Q\Big(\mathtt{x}, \mathbb{Q}\big(U(\mathtt{r}^*)\big)\Big). \end{aligned}$$

The constant $\mathbb{Q}\big(U(\mathtt{r}^*)\big) > 0$ quantifies the complexity of the set $U(\mathtt{r}^*) \subset \mathbb{R}^{p^*} \times \mathbb{R}^p$. We point out that for compact $M \subset \mathbb{R}^{p^*}$ we have $\mathbb{Q}(M) = 2p^*$ (see Supplement of Spokoiny (2012), Lemma 2.10). This gives $\mathbb{Q}\big(U\big) = 2p^* + 2p$. Finally observe that

$$\begin{aligned} &\sup_{\mathtt{r} \le \mathtt{r}^*} \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \left\{ \frac{1}{6\epsilon\nu_1} \|\breve{\mathcal{Y}}(\boldsymbol{\upsilon})\| - 2\mathtt{r}^2 \right\} \\ &\quad \le \sup_{\mathtt{r} \le \mathtt{r}^*} \sup_{(\boldsymbol{\upsilon},\boldsymbol{u}) \in U(\mathtt{r})} \left\{ \frac{1}{6\epsilon\nu_1 \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|} \boldsymbol{u}^\top \breve{\mathcal{Y}}(\boldsymbol{\upsilon}) - \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 - \|\boldsymbol{u}\|^2 \right\} \\ &\quad = \sup_{(\boldsymbol{\upsilon},\boldsymbol{u}) \in U(\mathtt{r}^*)} \left\{ \frac{1}{6\epsilon\nu_1 \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|} \boldsymbol{u}^\top \breve{\mathcal{Y}}(\boldsymbol{\upsilon}) - \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 - \|\boldsymbol{u}\|^2 \right\}. \end{aligned}$$

■

Lemma 44 *Suppose that $\breve{\mathcal{Y}}(\boldsymbol{\upsilon})$ satisfies for each $\|\boldsymbol{u}\| \le 1$ and $|\lambda| \le \mathtt{g}$ the inequality* (43)*. Then the process $\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{u}) = \frac{1}{2\epsilon \|\mathcal{D}(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^*)\|} \breve{\mathcal{Y}}(\boldsymbol{\upsilon})^\top \boldsymbol{u}_1$ satisfies $(\boldsymbol{\mathcal{E}d})$ from* (41) *with $|\lambda| \le \mathtt{g}/2$, $d((\boldsymbol{\upsilon}, \boldsymbol{u}), (\boldsymbol{\upsilon}^\circ, \boldsymbol{u}^\circ))^2 = \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 + \|\boldsymbol{u} - \boldsymbol{u}^\circ\|^2$, $\nu = 2\nu_0$ and $U \subset \mathbb{R}^{p^*+p}$ defined in* (44)*, i.e. for any $(\boldsymbol{\upsilon}, \boldsymbol{u}_1), (\boldsymbol{\upsilon}^\circ, \boldsymbol{u}_2) \in U$*

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

Proof Let $(\boldsymbol{v},\boldsymbol{u}_1),(\boldsymbol{v}^\circ,\boldsymbol{u}_2)\in U$ and w.l.o.g. $\boldsymbol{u}_1\le\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|\le\|\mathcal{D}(\boldsymbol{v}^\circ-\boldsymbol{v}^*)\|$. By the Hölder inequality and (43), we find

$$\begin{aligned}
&\log\mathbb{E}\exp\left\{\lambda\frac{\mathcal{U}(\boldsymbol{v},\boldsymbol{u}_1)-\mathcal{U}(\boldsymbol{v},\boldsymbol{u}_2)}{d((\boldsymbol{v},\boldsymbol{u}_1),(\boldsymbol{v}^\circ,\boldsymbol{u}_2))}\right\}\\
&\quad=\log\mathbb{E}\exp\left\{\lambda\frac{\mathcal{U}(\boldsymbol{v},\boldsymbol{u}_1)-\mathcal{U}(\boldsymbol{v}^\circ,\boldsymbol{u}_1)+\mathcal{U}(\boldsymbol{v}^\circ,\boldsymbol{u}_1)-\mathcal{U}(\boldsymbol{v}^\circ,\boldsymbol{u}_2)}{d((\boldsymbol{v},\boldsymbol{u}_1),(\boldsymbol{v}^\circ,\boldsymbol{u}_2))}\right\}\\
&\quad\le\frac{1}{2}\log\mathbb{E}\exp\left\{2\lambda\frac{\boldsymbol{u}_1^\top\left(\frac{1}{\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|}\breve{\mathcal{Y}}(\boldsymbol{v})-\frac{1}{\|\mathcal{D}(\boldsymbol{v}^\circ-\boldsymbol{v}^*)\|}\breve{\mathcal{Y}}(\boldsymbol{v}^\circ)\right)}{\epsilon\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^\circ)\|}\right\}\\
&\qquad+\frac{1}{2}\log\mathbb{E}\exp\left\{2\lambda\frac{(\boldsymbol{u}_1^\top-\boldsymbol{u}_2^\top)\breve{\mathcal{Y}}(\boldsymbol{v}^\circ)}{\epsilon\|\boldsymbol{u}_1-\boldsymbol{u}_2\|\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|}\right\}\\
&\quad\le\sup_{\|\boldsymbol{u}\|\le1}\frac{1}{2}\log\mathbb{E}\exp\left\{2\lambda\frac{\boldsymbol{u}^\top\left(\breve{\mathcal{Y}}(\boldsymbol{v})-\breve{\mathcal{Y}}(\boldsymbol{v}^\circ)\right)}{\epsilon\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^\circ)\|}\right\}\\
&\qquad+\sup_{\|\boldsymbol{u}\|\le1}\frac{1}{2}\log\mathbb{E}\exp\left\{2\lambda\frac{\boldsymbol{u}^\top\left(\breve{\mathcal{Y}}(\boldsymbol{v}^\circ)-\breve{\mathcal{Y}}(\boldsymbol{v}^*)\right)}{\epsilon\|\mathcal{D}(\boldsymbol{v}-\boldsymbol{v}^*)\|}\right\}\\
&\quad\le\frac{4\nu_0^2\lambda^2}{2},\qquad\lambda\le\mathtt{g}/2.
\end{aligned}$$

■

Appendix C. A Bound for the Spectral Norm of a Random Matrix Process

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

$$\mathbb{P}\left(\sup_{\boldsymbol{v}\in\Upsilon_\circ(\mathtt{r})}\left\{\|\breve{\mathcal{Y}}(\boldsymbol{v})\|\right\}\ge\mathtt{C}\epsilon_2\mathfrak{z}_1(\mathtt{x},p^*)\mathtt{r}\right)\le\mathrm{e}^{-\mathtt{x}}.$$

We derive such a bound in a very similar manner to Theorem E.1 of Andresen and Spokoiny (2014).

We want to apply Corollary 2.2 of the supplement of Spokoiny (2012). Again we slightly generalized the formulation but the proof remains the same.

Corollary 45 *Let $(U(\mathtt{r}))_{0\le\mathtt{r}\le\mathtt{r}^*}\subset\mathbb{R}^p$ be a sequence of balls around $\boldsymbol{v}^*$ induced by the metric $d(\cdot,\cdot)$. Let a random real valued process $\mathcal{U}(\boldsymbol{v})$ fulfill that $\mathcal{U}(\boldsymbol{v}^*)=0$ and*

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

$$\log\mathbb{E}\exp\left\{\lambda\frac{\mathcal{U}(\boldsymbol{v})-\mathcal{U}(\boldsymbol{v}^\circ)}{d(\boldsymbol{v},\boldsymbol{v}^\circ)}\right\}\le\frac{\nu_0^2\lambda^2}{2},\qquad|\lambda|\le\mathtt{g}.\tag{45}$$

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

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

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

$$\mathfrak{z}_1(\mathtt{x}, \mathbb{Q}) \stackrel{\text{def}}{=} \begin{cases} \sqrt{2(\mathtt{x} + \mathbb{Q})} & \text{if } \sqrt{2(\mathtt{x} + \mathbb{Q})} \le \mathtt{g}_0, \\ \mathtt{g}_0^{-1}(\mathtt{x} + \mathbb{Q}) + \mathtt{g}_0/2 & \text{otherwise.} \end{cases} \tag{46}$$

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

$$\sup_{\|\boldsymbol{u}_1\| \le 1} \sup_{\|\boldsymbol{u}_2\| \le 1} \log \mathbb{E} \exp\biggl\{ \lambda \frac{\boldsymbol{u}_1^\top \bigl(\mathcal{Y}(\boldsymbol{\upsilon}) - \mathcal{Y}(\boldsymbol{\upsilon}^\circ)\bigr) \boldsymbol{u}_2}{\epsilon_2 \|\mathcal{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\|} \biggr\} \le \frac{\nu_2^2 \lambda^2}{2}. \tag{47}$$

Remark 46 *In the setting of Theorem 14 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 (47) *becomes* $(\mathcal{ED}_2)$ *from 2.1.*

Theorem 47 *Let a random process* $\mathcal{Y}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^* \times p^*}$ *fulfill* $\mathcal{Y}(\boldsymbol{\upsilon}^*) = 0$ *and let condition* (47) *be satisfied. Then for each* $0 \le \mathtt{r} \le \mathtt{r}^*$, *on a set of probability greater than* $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{\upsilon})\| \le 9\epsilon_2 \nu_2 \mathfrak{z}_1(\mathtt{x}, 6p^*) \mathtt{r},$$

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

Remark 48 *Note that the entropy of the original set* $\Upsilon_\circ(\mathtt{r}) \subset \mathbb{R}^{p^*}$ *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

$$\|\mathcal{Y}(\boldsymbol{\upsilon})\| = \epsilon_2 \sup_{\|\boldsymbol{u}_1\| \le \mathtt{r}} \sup_{\|\boldsymbol{u}_2\| \le \mathtt{r}} \frac{1}{\epsilon_2 \mathtt{r}^2} \boldsymbol{u}_1^\top \breve{\mathcal{Y}}(\boldsymbol{\upsilon}) \boldsymbol{u}_2.$$

This implies

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

Due to Lemma 49 the process $\mathcal{U}(\boldsymbol{\upsilon}) \stackrel{\mathrm{def}}{=} \frac{1}{\epsilon \mathtt{r}^2} \boldsymbol{u}_1^\top \mathcal{Y}(\boldsymbol{\upsilon}) \boldsymbol{u}_2$ satisfies condition $(\mathcal{E}d)$ (see (45)) as process on

$$U(\mathtt{r}) \stackrel{\mathrm{def}}{=} \Upsilon_\circ(\mathtt{r}) \times B_{\mathtt{r}}(0) \times B_{\mathtt{r}}(0) \subset \mathbb{R}^{3p^*}. \tag{48}$$

This allows to apply Corollary 45 to obtain the desired result. We get on a set of probability greater $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{\upsilon})\| \le \sup_{(\boldsymbol{\upsilon}, \boldsymbol{u}_1, \boldsymbol{u}_2) \in U(\mathtt{r})} \left\{ \frac{1}{\mathtt{r}^2} \boldsymbol{u}_1^\top \mathcal{Y}(\boldsymbol{\upsilon}) \boldsymbol{u}_2 \right\} \le 9 \epsilon_2 \nu_2 \mathfrak{z}_1 \Bigl(\mathtt{x}, \mathbb{Q}\bigl(U(\mathtt{r}^*)\bigr) \Bigr) \mathtt{r}.$$

The constant $\mathbb{Q}\bigl(U(\mathtt{r})\bigr) > 0$ quantifies the complexity of the set $U(\mathtt{r}) \subset \mathbb{R}^{3p^*}$. We point out that for compact $M \subset \mathbb{R}^{3p^*}$ we have $\mathbb{Q}(M) = 6p^*$ (see Supplement of Spokoiny (2012), Lemma 2.10). This gives the claim. ■

Lemma 49 *Suppose that* $\mathcal{Y}(\boldsymbol{\upsilon}) \in \mathbb{R}^{p^* \times p^*}$ *satisfies* $\mathcal{Y}(\boldsymbol{\upsilon}^*) = 0$ *and for each* $\|\boldsymbol{u}_1\| \le 1$, $\|\boldsymbol{u}_2\| \le 1$ *and* $|\lambda| \le \mathtt{g}$ *the inequality* (47). *Then the process*

$$\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{u}_1, \boldsymbol{u}_2) = \frac{1}{2\epsilon_2 \mathtt{r}^2} \boldsymbol{u}_1^\top \mathcal{Y}(\boldsymbol{\upsilon})^\top \boldsymbol{u}_2$$

satisfies $\boldsymbol{(\mathcal{E}d)}$ *from* (45) *with* $U \subset \mathbb{R}^{3p^*}$ *defined in* (48), *with* $|\lambda| \le \mathtt{g}/3$ *and with*

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

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

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

Proof Let $(\upsilon, u_1, u_2), (\upsilon^\circ, u_1^\circ, u_2^\circ) \in U$. By the Hölder inequality and (47), we find

$$
\begin{aligned}
&\log \mathbb{E} \exp\biggl\{ \lambda \frac{\mathcal{U}(\upsilon, u_1, u_2) - \mathcal{U}(\upsilon^\circ, u_1^\circ, u_2^\circ)}{d((\upsilon, u_1, u_2), (\upsilon^\circ, u_1^\circ, u_2^\circ))} \biggr\} \\
&\quad = \log \mathbb{E} \exp\biggl\{ \lambda \biggl(\frac{\mathcal{U}(\upsilon, u_1, u_2) - \mathcal{U}(\upsilon^\circ, u_1, u_2)}{d((\upsilon, u_1, u_2), (\upsilon^\circ, u_1^\circ, u_2^\circ))} + \frac{\mathcal{U}(\upsilon^\circ, u_1, u_2) - \mathcal{U}(\upsilon^\circ, u_1^\circ, u_2)}{d((\upsilon, u_1, \upsilon_2), (\upsilon^\circ, u_1^\circ, u_2^\circ))} \\
&\qquad + \frac{\mathcal{U}(\upsilon^\circ, u_1^\circ, u_2) - \mathcal{U}(\upsilon^\circ, u_1^\circ, u_2^\circ)}{d((\upsilon, u_1, u_2), (\upsilon^\circ, u_1^\circ, u_2^\circ))} \biggr) \biggr\} \\
&\quad \le \frac{1}{3} \log \mathbb{E} \exp\biggl\{ 3\lambda \frac{u_1^\top \bigl(\frac{1}{\mathtt{r}^2} \breve{\mathcal{Y}}(\upsilon) - \frac{1}{\mathtt{r}^2} \breve{\mathcal{Y}}(\upsilon^\circ) \bigr) u_2}{\epsilon_2 \| \mathcal{D}(\upsilon - \upsilon^\circ) \|} \biggr\} \\
&\qquad + \frac{1}{3} \log \mathbb{E} \exp\biggl\{ 3\lambda \frac{(u_1 - u_1^\circ)^\top) \mathcal{Y}(\upsilon^\circ) u_2}{\epsilon_2 \| u_1 - u_2 \| \mathtt{r}^2} \biggr\} \\
&\qquad + \frac{1}{3} \log \mathbb{E} \exp\biggl\{ 3\lambda \frac{(u_1^\circ)^\top) \mathcal{Y}(\upsilon^\circ) (u_2 - u_2^\circ)}{\epsilon_2 \| u_1 - u_2 \| \mathtt{r}^2} \biggr\} \\
&\quad \le \frac{1}{3} \sup_{\|u_1\| \le 1} \sup_{\|u_2\| \le 1} \log \mathbb{E} \exp\biggl\{ 3\lambda \frac{u_1^\top \bigl(\mathcal{Y}(\upsilon) - \mathcal{Y}(\upsilon^\circ) \bigr) u_2}{\epsilon_2 \| \mathcal{D}(\upsilon - \upsilon^\circ) \|} \biggr\} \\
&\qquad + \frac{2}{3} \sup_{\|u_1\| \le 1} \sup_{\|u_2\| \le 1} \log \mathbb{E} \exp\biggl\{ 3\lambda \frac{u_1^\top \bigl(\mathcal{Y}(\upsilon^\circ) - \mathcal{Y}(\upsilon^*) \bigr) u_2}{\epsilon_2 \| \mathcal{D}(\upsilon - \upsilon^*) \|} \biggr\} \\
&\quad \le \frac{9 \nu_2^2 \lambda^2}{2}, \qquad \lambda \le \mathtt{g}/3.
\end{aligned}
$$

■

References

A. Andresen. Finite sample analysis of profile m-estimation in the single index model. *Electronic Journal of Statistics*, 9(2):2528–2641, 2015.

A. Andresen and V. Spokoiny. Critical dimension in profile semiparametric estimation. *Electron. J. Statist.*, 8(2):3077–3125, 2014. ISSN 1935-7524. doi: 10.1214/14-EJS982.

S. Balakrishnan, M. J. Wainwright, and B. Yu. Satistical guarantees for the em algorithm: From population to sample-based analysis. *arXiv: 1408.2156*, 2014.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:55495632, 2007.

G. Cheng. How many iterations are sufficient for efficient semiparametric estimation? *Scandinavian Journal of Statistics,*, 40(3):592–618, 2013.

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

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

U. Grenander. *Abstract Inference*. Whiley, New York, 1981.

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

I.A. Ibragimov and R.Z. Khas'minskij. *"Statistical estimation. Asymptotic theory. Transl. from the Russian by Samuel Kotz."*. "New York - Heidelberg -Berlin: Springer-Verlag", 1981.

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

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

M.R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer in Statistics, 2005.

G.J. McLachlan and T Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

S. A. Murphy and A. W. van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

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

Vladimir Spokoiny. Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6): 2877–2909, 2012.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

C.F.J Wu. On the convergence properties of the em algorithm. *Annals of Statistics*, 11: 95–103, 1983.

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