Finite sample expansions for semiparametric plug-in estimation and inference for BTL model

Vladimir Spokoiny*

Weierstrass Institute and HU Berlin, HSE and IITP RAS,
Mohrenstr. 39, 10117 Berlin, Germany
spokoiny@wias-berlin.de

December 2, 2024

Abstract

This note continues and extends the study from Spokoiny (2024b) about estimation for parametric models with a large parameter dimension.

AMS 2010 Subject Classification: Primary 62F10,62E17. Secondary 62J12
Keywords: penalized MLE, Fisher and Wilks expansions, risk bounds

*Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 1294 "Data assimilation" is gratefully acknowledged.

Contents

1 Introduction

The study in Spokoiny (2023b) offers a new approach to statistical inference and for a special class of *stochastically linear smooth* (SLS) models; see also Spokoiny (2024b). Moreover, Spokoiny (2023b) explains how this appoach can be applied to a high dimensional regression problem. The derived results provide finite sample expansions for the profile maximum likelihood estimator and establish sharp risk bounds. The main limitation of this approach is the so called "critical dimension" condition requiring $p^2 \ll n$, where p is the parameter dimension and n is the sample size. In view of recent results from Katsevich (2023), this condition cannot be avoided in the general setup. Surprisingly, Gao et al. (2023) established very strong inference results for the Bradley-Terry-Luce (BTL) problem of ranking from pairwise comparison under much weaker condition $p \ll n$. One of the aims of this paper is to revisit this problem and to understand the nature of this phenomen. We offer a completely different view on the considered problem and reduce the study to semiparametric plug-in estimation. The general semiparametric setup can be described as follows. Let a model be described by a high dimensional parameter $\boldsymbol{\upsilon}$ consisting of a low dimensional target $\boldsymbol{\theta}$ and a high dimensional nuisance parameter $\boldsymbol{\eta}$. Given a full dimensional log-likelihood function $\mathscr{L}(\boldsymbol{\upsilon}) = \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\eta})$, the full dimensional MLE is defined by a joint optimization

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

The profile MLE $\widetilde{\boldsymbol{\theta}}$ is the $\boldsymbol{\theta}$-component of $\widetilde{\boldsymbol{\upsilon}}$. Unfortunately, the full dimensional optimization is often hard to implement and study. At the same time, partial optimization of $\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\eta})$ w.r.t. $\boldsymbol{\theta}$ for $\boldsymbol{\eta}$ fixed can be a much simpler problem. This suggests a plug-in approach: given a preliminary/pilot estimate $\widehat{\boldsymbol{\eta}}$, define $\widehat{\boldsymbol{\theta}}$ by maximization of $\mathscr{L}(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})$:

$$\widehat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}} \mathscr{L}(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}).$$

Furthermore, one can continue by using $\widehat{\boldsymbol{\theta}}$ as a preliminary estimate of $\boldsymbol{\theta}$ and reestimate $\widehat{\boldsymbol{\eta}}$. This leads to alternated optimization and EM-type methods: starting from $\widehat{\boldsymbol{\theta}}_1$ and $k = 1$, estimate

$$\widehat{\boldsymbol{\eta}}_k = \operatorname*{argmax}_{\boldsymbol{\eta}} \mathscr{L}(\widehat{\boldsymbol{\theta}}_k, \boldsymbol{\eta}), \qquad \widehat{\boldsymbol{\theta}}_{k+1} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathscr{L}(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}_k),$$

and increase k by one; see e.g. Andresen and Spokoiny (2016) and references therein. Analysis of convergence for such procedures requires to evaluate the accuracy of each step. This leads to a semiparametric plug-in study.

Another situation naturally leading to a semiparametric study arises when the MLE $\widetilde{\boldsymbol{\upsilon}}$ of the full dimensional parameter $\boldsymbol{\upsilon} \in \mathbb{R}^p$ is studied componentwise (in sup-norm). For each $j \le p$, we consider the j-th entry υ_j^* of $\boldsymbol{\upsilon}^*$ as the target component while all the remaining components are treated as a nuisance. This approach will be studied in general and then applied to the BTL model. The obtained results state that with a high probability, $\max_{j \le p} \sqrt{\mathbb{N}_j}\, |\widetilde{\upsilon}_j - \upsilon_j^*| \le \mathtt{r}_\infty$ for $\mathtt{r}_\infty \asymp \sqrt{\log p}$ and for $\mathbb{N}_j$ meaning the number of games played by the players j. Moreover, we establish an expansion for every estimate $\widetilde{\upsilon}_j - \upsilon_j^*$ with an explicit error term.

This paper contribution

Estimation and inference for finite samples is a challenging problems, available results are mainly limited to linear models. The approach from Spokoiny (2024b) extends the results to a more general class of SLS models, however, under the critical dimension condition $\mathbb{p}^2 \ll \mathbb{N}$ for the *effective dimension* $\mathbb{p}$ and the *effective sample size* $\mathbb{N}$. This *critical dimension* condition cannot generally be avoided as long as full-dimensional estimation is considered. However, in the case of sup-norm estimation, the problem can be reduced to a special semiparametric setup. This helps to relax the relation between effective dimension and effective sample size; see Section D.4. Finally, the results are applied to the BTL model, the obtained expansions and risk bounds improve the results and conditions from Gao et al. (2023).

The main technical tool of the study is the theory of *perturbed optimization*. The results from Spokoiny (2023b, 2024b) explain how the solution and value of an optimization problem change after a linear, quadratic, or smooth perturbation of the objective function. This paper also considers the case of partial and marginal optimization with a low-dimensional target variable and a possible large-dimensional nuisance variable; Section D.3. This setup requires to evaluate a special *semiparametric bias* caused by using an inexact value of a nuisance variable. It appears that the value of this bias strongly depends on the norm in which smoothness of the objective function w.r.t. the nuisance variable is measured. An important special case is given by the sup-norm; see Section D.4.

The obtained results are applied to the BTL model. We derive the expansions and risk bounds for the MLE $\tilde{\boldsymbol{\upsilon}}$ in the classical ℓ_2-norm and componentwise. The state-of-the-art results from Gao et al. (2023) are restated with more accurate bounds for the error terms in the major expansions. Moreover, the conditions on the model are significantly relaxed. In particular, we allow an arbitrary configuration of the comparison graph, even disconnected. Also, the cases of heterogeneous numbers of comparisons par edge and unbounded skill range are covered. We derive finite-sample bounds with explicit error terms and address the issue of critical dimension.

Organization of the paper

Section 2 presents the results for the BTL models including full dimensional and componentwise expansions. General problem of semiparametric plug-in estimation is discussed in Section 3. Appendix A and Appendix B.1 overview the general SLS theory from Spokoiny (2024b). Appendix C explains how this theory applies to logistic regression. Appendix D collects the results for perturbed optimization including the cases of partial and marginal optimization. Some deviation bounds for Bernoulli vector sums are collected in Appendix E.

2 Estimation in the Bradley-Terry-Luce model

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ stand for a comparison graph, where the vertex set $\mathcal{V} = \{1, 2, \ldots, p\}$ represents the p items of interest. The items j and m are compared if and only if $(jm) \in \mathcal{E}$. One observes independent paired comparisons $Y_{jm}^{(\ell)}$, $\ell = 1, \ldots, N_{jm}$, and $Y_{jm}^{(\ell)} = 1 - Y_{mj}^{(\ell)}$. For modeling and risk analysis, Bradley-Terry-Luce (BTL) model is frequently used; see Bradley and Terry (1952), Luce (1959). The chance of each item winning a paired comparison is determined by the relative scores

$$\mathbb{P}\bigl(\text{item } j \text{ is preferred over item } m\bigr) = \mathbb{P}\bigl(Y_{jm}^{(\ell)} = 1\bigr) = \frac{\mathrm{e}^{\upsilon_j^*}}{\mathrm{e}^{\upsilon_j^*} + \mathrm{e}^{\upsilon_m^*}} = \frac{1}{1 + \mathrm{e}^{\upsilon_m^* - \upsilon_j^*}} \, .$$

The goal is to recover the score vector $\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_p)^\top$ and top$-k$ items. Most of the theoretical results for a BTL model have been established under the following assumptions:

- $\mathcal{G}$ is a random Erdős-Rényi graph with the edge probability α; $\alpha \geq \mathtt{C} p^{-1} \log p$ ensures with overwhelming probability a connected graph;
- the values N_{jm} are all the same; $N_{jm} \equiv L$, for all $(j,m) \in \mathcal{E}$;
- for the ordered sequence $\upsilon^*_{(1)} \geq \upsilon^*_{(2)} \geq \ldots \geq \upsilon^*_{(p)}$, it holds $\upsilon^*_{(k)} - \upsilon^*_{(k+1)} > \Delta$;
- $\upsilon^*_{(1)} - \upsilon^*_{(p)} \leq \mathcal{R}$;

see e.g. Chen et al. (2020); Gao et al. (2023) for a related discussion. Under such assumptions, Chen et al. (2019) and Chen et al. (2020) showed that the conditions

$$\Delta^2 \geq \mathtt{C} \frac{\log p}{p \, \alpha \, L}$$

enables to identify the top- k set with a high probability. Both regularized MLE and a spectral method are rate optimal. We refer to Gao et al. (2023) for an extensive overview and recent results for the BTL model including a non-asymptotic MLE expansion.

Unfortunately, some of the mentioned assumptions could be very restrictive in practical applications. This especially concerns the graph structure and design of the set of comparisons. An assumption of a bounded dynamic range $\mathcal{R}$ is very useful for the theoretical study because it allows to bound the success probability of each game away from zero and one. However, it seems to be questionable for many real-life applications. Our aim is to demonstrate that the general approach of the paper enables us to get accurate results applying under

- arbitrary configuration of the graph $\mathcal{G}$; possibly disconnected;
- heterogeneous numbers N_{jm} of comparisons par edge;
- unbounded range $\upsilon^*_{(1)} - \upsilon^*_{(p)}$.

2.1 Penalized MLE

This section specifies the general results of Section C for logistic regression to the BTL model. For $j < m$, denote $S_{jm} = \sum_{\ell=1}^{N_{jm}} Y_{jm}^{(\ell)}$ and $S_{jm} = 0$ if $N_{jm} = 0$. With $\phi(\upsilon) = \log(1 + \mathrm{e}^{\upsilon})$, the log-likelihood for the parameter vector $\boldsymbol{\upsilon}^*$ reads as follows:

$$L(\boldsymbol{\upsilon}) = \sum_{m=1}^{p} \sum_{j=1}^{m-1} \bigl\{ (\upsilon_j - \upsilon_m) S_{jm} - N_{jm} \phi(\upsilon_j - \upsilon_m) \bigr\}, \tag{2.1}$$

leading to the MLE

$$\widetilde{\boldsymbol{\upsilon}} = \operatorname*{argmax}_{\boldsymbol{\upsilon}} L(\boldsymbol{\upsilon}).$$

The function $\phi(\upsilon) = \log(1+\mathrm{e}^{\upsilon})$ is convex, hence, $L(\boldsymbol{\upsilon})$ is concave. However, representation (2.1) reveals *lack-of-identifiability* problem: $\widetilde{\boldsymbol{\upsilon}}$ is not unique, any shift $\boldsymbol{\upsilon} \to \boldsymbol{\upsilon} + a\boldsymbol{e}$ does not change $L(\boldsymbol{\upsilon})$, $\boldsymbol{e} = (1, \ldots, 1)^\top \in \mathbb{R}^p$. Therefore, the Fisher information matrix $\mathbb{F}(\boldsymbol{\upsilon}) = -\nabla^2 L(\boldsymbol{\upsilon})$ is not positive definite and $L(\boldsymbol{\upsilon})$ is not strongly concave, thus, $\widetilde{\boldsymbol{\upsilon}}$ is not uniquely defined. For a connected graph $\mathcal{G}$, this issue can be resolved by fixing one component of $\boldsymbol{\upsilon}$, e.g. $\upsilon_1 = 0$, or by the condition $\sum_j \upsilon_j = 0$. In general, we need one condition per connected component of the graph $\mathcal{G}$. Alternatively, one can use penalization with a quadratic penalty $\|G\boldsymbol{\upsilon}\|^2/2$. A "non-informative" choice is $G^2 = \lambda \mathbb{I}_p$; cf. Chen et al. (2019). Another option is to replace the constraint $\sum_j \upsilon_j = 0$ by the penalty $\lambda\bigl(\sum_j \upsilon_j\bigr)^2/2$. One more benefit of using a penalization by $\|G\boldsymbol{\upsilon}\|^2/2$ is that it ensures the desired stochastically dominant structure of the Fisher information matrix: each diagonal element is larger than the sum of the off-diagonal for this row. This property will be used for proving a concentration of the penalized MLE

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmax}_{\boldsymbol{\upsilon}} L_G(\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\upsilon}} \bigl(L(\boldsymbol{\upsilon}) - \|G\boldsymbol{\upsilon}\|^2/2\bigr).$$

The structure of the Fisher information matrix $\mathbb{F}(\boldsymbol{\upsilon})$ is given by the next lemma.

Lemma 2.1. *With* $\phi''(\upsilon) = \frac{\mathrm{e}^{\upsilon}}{(1+\mathrm{e}^{\upsilon})^2}$, *the entries* $\mathbb{F}_{jm}(\boldsymbol{\upsilon})$ *of* $\mathbb{F}(\boldsymbol{\upsilon}) = -\nabla^2 \mathbb{E} L(\boldsymbol{\upsilon})$ *satisfy*

$$\begin{aligned} \mathbb{F}_{jj}(\boldsymbol{\upsilon}) &= \sum_{m \neq j} N_{jm} \phi''(\upsilon_j - \upsilon_m), \\ \mathbb{F}_{jm}(\boldsymbol{\upsilon}) &= -N_{jm} \phi''(\upsilon_j - \upsilon_m), \quad j \neq m\,. \end{aligned}$$

This yields $\mathbb{F}_{jj}(\boldsymbol{\upsilon}) = -\sum_{m \neq j} \mathbb{F}_{jm}(\boldsymbol{\upsilon})$. *Moreover, each eigenvalue of* $\mathbb{F}(\boldsymbol{\upsilon})$ *belongs to* $[0, 2\mathbb{F}_{jj}(\boldsymbol{\upsilon})]$ *for some* $j \leq p$.

Proof. The structure of $\mathbb{F}$ follows directly from (2.1). Gershgorin's theorem implies the final statement. □

It has been already mentioned that $\mathbb{F}(\boldsymbol{\upsilon})$ is degenerated. However, adding the penalty term $\|G\boldsymbol{\upsilon}\|^2/2$ with $G^2 > 0$ ensures that $\mathbb{F}_G = \mathbb{F}_G(\boldsymbol{\upsilon}_G^*)$ is positive definite. The smallest eigenvalue of $\mathbb{F}_G$ is the key quantity in all our results, it identifies the accuracy of estimation. Define

$$\mathbb{N}_G \stackrel{\mathrm{def}}{=} \lambda_{\min}(\mathbb{F}_G) = \frac{1}{\|\mathbb{F}_G^{-1}\|}\,.$$

For ease of presentation, we assume later a correctly specified model $\theta^{*}_{jm} = 1/(1 + \mathrm{e}^{\upsilon^{*}_{m} - \upsilon^{*}_{j}})$ yielding $I\!\!E S_{jm} = N_{jm} \theta^{*}_{jm} = N_{jm} \phi'(\upsilon^{*}_{j} - \upsilon^{*}_{m})$. This condition can be relaxed; see Remark 2.1.

Theorem 2.2. *Let $Y^{(\ell)}_{jm} \sim \mathrm{Be}(\theta^{*}_{jm})$ be mutually independent Bernoulli random variables for $(jm) \in \mathcal{G}$ and $\ell \le N_{jm}$. Assume $\theta^{*}_{jm} = (1 + \mathrm{e}^{\upsilon^{*}_{m} - \upsilon^{*}_{j}})^{-1}$. Let, given $\mathtt{x} > 0$,*

$$
I\!\!N_{G}^{-1/2} \bigl(\sqrt{4p} + \sqrt{8\mathtt{x}} \bigr) \le 1/3 . \tag{2.2}
$$

Then $\widetilde{\boldsymbol{\upsilon}}_{G}$ follows Theorem B.2 with τ_{3} given by

$$
\tau_{3} = \frac{\mathrm{e}^{3/4} \sqrt{2}}{\sqrt{I\!\!N_{G}}} \le \frac{3}{\sqrt{I\!\!N_{G}}} . \tag{2.3}
$$

In particular, on a random set $\Omega(\mathtt{x})$ with $I\!\!P\bigl(\Omega(\mathtt{x})\bigr) \ge 1 - 3\mathrm{e}^{-\mathtt{x}}$, it holds

$$
\| I\!\!F_{G}^{1/2} \{ \widetilde{\boldsymbol{\upsilon}}_{G} - \boldsymbol{\upsilon}^{*} - I\!\!F_{G}^{-1} (\nabla \zeta - G^{2} \boldsymbol{\upsilon}^{*}) \} \| \le \frac{9}{4 \sqrt{I\!\!N_{G}}} \bigl(\| I\!\!F_{G}^{-1/2} \nabla \zeta \|^{2} + \mathsf{b}_{G}^{2} \bigr), \tag{2.4}
$$

where $\mathsf{b}_{G} = \| I\!\!F_{G}^{-1/2} G^{2} \boldsymbol{\upsilon}^{*} \|$.

Proof. We apply Proposition C.1. The main condition to be checked is $(\boldsymbol{\Psi}^{\circ})$. Suppose that all the comparisons are polled together and indexed by $i = 1, \ldots, n$, that is, Y_{i} is the result of the game between players $j(i)$ and $m(i)$ in the ith game. The corresponding design vector $\boldsymbol{\Psi}_{i} \in I\!\!R^{p}$ contains one at $j(i)$, minus one at $m(i)$, and zeros elsewhere. Therefore, $\| \boldsymbol{\Psi}_{i} \|^{2} = 2$. Condition $(\boldsymbol{\Psi}^{\circ})$ is fulfilled with $D^{2} \stackrel{\mathrm{def}}{=} I\!\!F_{G}$ and $\delta_{0} = \sqrt{2}\, I\!\!N_{G}^{-1/2}$. Further, $B = 2D\, I\!\!F_{G}^{-1} D = 2 I\!\!I_{p}$ yielding $\mathbb{p}_{D} = 2 \operatorname{tr}(B) = 2p$, $\lambda_{D} = 2\|B\| = 2$, and $\mathtt{r}_{D} \le \sqrt{\mathbb{p}_{D}} + \sqrt{2\mathtt{x}\lambda_{D}} \le \sqrt{2p} + 2\sqrt{\mathtt{x}}$. Therefore, condition (C.8) follows from (2.2). Proposition C.1 enables us to apply the results of Section A. Theorem B.2 with $D = Q = I\!\!F_{G}^{1/2}$ implies (2.4). □

Remark 2.1. Condition $\theta^{*}_{jm} = (1 + \mathrm{e}^{\upsilon^{*}_{m} - \upsilon^{*}_{j}})^{-1}$ can be relaxed to

$$
\theta^{*}_{jm} (1 - \theta^{*}_{jm}) \le \mathtt{C} \frac{\mathrm{e}^{\upsilon^{*}_{j} - \upsilon^{*}_{m}}}{(1 + \mathrm{e}^{\upsilon^{*}_{j} - \upsilon^{*}_{m}})^{2}} , \qquad j \le m,
$$

for some absolute constant $\mathtt{C}$.

2.2 Coordinatewise bounds

The full dimensional expansion (2.4) and the corresponding risk bound for the whole vector $\boldsymbol{\upsilon}$ requires the condition $p \ll I\!\!N_{G}$. This is a serious but presumably unavoidable constraint on the setup. However, Gao et al. (2023) established componentwise results about $\widetilde{\boldsymbol{\upsilon}}$ under a much weaker condition $p \lesssim n/(\log n)^{3/2}$. Here we obtain similar results for the pMLE $\widetilde{\boldsymbol{\upsilon}}_G$ using the componentwise approach from Section D.4. Consider $G^2 = \operatorname{diag}(G_j^2)$ and denote

$$\begin{aligned} \mathbb{D}_j^2 &\stackrel{\text{def}}{=} -\frac{\partial^2}{\partial \upsilon_j \, \partial \upsilon_j} I\!\!E L_G(\boldsymbol{\upsilon}^*) = \sum_{m \neq j} N_{jm} \phi''(\upsilon_j^* - \upsilon_m^*) + G_j^2 \, , \\ \mathbb{D}^2 &= \operatorname{diag}(\mathbb{D}_1^2, \ldots, \mathbb{D}_p^2). \end{aligned} \tag{2.5}$$

This definition differs from the case of full dimensional estimation in Section 2.1 where we used $\mathbb{D}^2 = I\!\!F_G$. In what follows we assume that adding the penalty matrix G^2 ensures that $I\!\!F_G$ is positive definite and for some $\varkappa^2$

$$\mathbb{D}^2 \leq \varkappa^2 \, I\!\!F_G \, . \tag{2.6}$$

Further, define the important quantity ρ_1 by

$$\rho_1^2 = \max_{j=1,\ldots,p} \frac{1}{\mathbb{D}_j^2} \sum_{m \neq j} \frac{1}{\mathbb{D}_m^2} \mathscr{F}_{jm}^2 \, . \tag{2.7}$$

Our results assume that $\rho_1 < 1$. We also assume that $I\!\!N_G = \lambda_{\min}(I\!\!F_G)$ is sufficiently large to ensure $I\!\!N_G^{-1/2} \mathtt{r}_\infty \ll 1$, where

$$\mathtt{r}_\infty \stackrel{\text{def}}{=} \frac{2}{1 - \rho_1} \sqrt{\mathtt{x} + \log p} \, . \tag{2.8}$$

Theorem 2.3. *For* ρ_1 *from* (2.7)*,* $\mathtt{r}_\infty$ *from* (2.8)*, and* $\tau_3 \stackrel{\text{def}}{=} 3 \, \varkappa^{-3/2} I\!\!N_G^{-1/2}$*, let*

$$\rho_1 < 1 \, , \quad \omega \stackrel{\text{def}}{=} \tau_3 \, \mathtt{r}_\infty \leq 1/4 \, , \quad \tau_\infty \sqrt{2(\mathtt{x} + \log p)} \leq 2/5 \, ,$$

where

$$\tau_\infty \stackrel{\text{def}}{=} \Bigl(2.5 + \frac{2(\mathtt{c} + 1)}{(1 - \rho_1)^2} \Bigr) \tau_3 \, , \quad \mathtt{c} \stackrel{\text{def}}{=} \frac{1}{1 - \omega} \left(\rho_1 + \frac{1}{2} + \frac{3(\rho_1 + 1/2)^2}{4(1 - \omega)^2} \right).$$

Define $\varOmega(\mathtt{x}) \stackrel{\text{def}}{=} \bigl\{ \| \mathbb{D}^{-1} \nabla \zeta \|_\infty \leq \sqrt{2(\mathtt{x} + \log p)} \bigr\}$*. Then* $I\!\!P\bigl(\varOmega(\mathtt{x})\bigr) \geq 1 - 4\mathrm{e}^{-\mathtt{x}}$*, and on* $\varOmega(\mathtt{x})$*, it holds*

$$\| \mathbb{D}(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^*) \|_\infty \leq \frac{\sqrt{2}}{1 - \rho_1} \| \mathbb{D}^{-1} \nabla \zeta \|_\infty \leq \mathtt{r}_\infty \, . \tag{2.9}$$

Moreover, on $\Omega(\mathtt{x})$

$$\|\mathbb{D}(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^* - \boldsymbol{F}_G^{-1}\nabla\zeta)\}\|_\infty \le \frac{\tau_\infty}{1-\rho_1}\,\|\mathbb{D}^{-1}\nabla\zeta\|_\infty^2\,, \tag{2.10}$$

$$\|\mathbb{D}(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^* - \mathbb{D}^{-2}\nabla\zeta)\}\|_\infty \le \frac{\tau_\infty}{1-\rho_1}\|\mathbb{D}^{-1}\nabla\zeta\|_\infty^2 + \frac{\rho_1}{1-\rho_1}\,\|\mathbb{D}^{-1}\nabla\zeta\|_\infty\,. \tag{2.11}$$

Proof. It is obvious that the stochastic component $\zeta(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} L(\boldsymbol{\upsilon}) - \boldsymbol{E}L(\boldsymbol{\upsilon})$ is linear in $\boldsymbol{\upsilon}$. Moreover, the gradient $\nabla\zeta = (\nabla_j\zeta)$ admits the closed form representation

$$\nabla_j\zeta = \sum_{m\neq j}\{S_{jm} - \boldsymbol{E}S_{jm}\}.$$

First, we bound in probability the value $\|\mathbb{D}^{-1}\nabla\zeta\|_\infty = \max_{j\le p}|\mathbb{D}_j^{-1}\nabla_j\zeta|$.

Lemma 2.4. *Let* $Y_{jm}^{(\ell)} \sim \mathrm{Be}(\theta_{jm}^*)$ *be mutually independent Bernoulli random variables for* $(jm)\in\mathcal{G}$ *and* $\ell \le N_{jm}$. *If* $\mathtt{x}$ *satisfies* $\sqrt{\mathtt{x}+\log p} \le \frac{2}{3}\min_{j\le p}\mathbb{D}_j$, *then*

$$\boldsymbol{P}\Bigl(\max_{j\le p}|\mathbb{D}_j^{-1}\nabla_j\zeta| > \sqrt{2(\mathtt{x}+\log p)}\Bigr) \le 2\mathrm{e}^{-\mathtt{x}}\,. \tag{2.12}$$

Proof. It holds for each $j\le p$

$$V_j^2 \stackrel{\text{def}}{=} \mathrm{Var}(\nabla_j\zeta) = \mathrm{Var}\sum_{m=1}^{p}S_{jm} = \sum_{m=1}^{p}N_{jm}\,\theta_{jm}^*(1-\theta_{jm}^*) \le \mathbb{D}_j^2\,.$$

Proposition E.1 applied with $w^*=1$ implies by $V_j \le \mathbb{D}_j$

$$\boldsymbol{P}\bigl(\mathbb{D}_j^{-1}|\nabla_j\zeta| > \sqrt{\mathtt{x}}\bigr) \le 2\mathrm{e}^{-\mathtt{x}}, \qquad \sqrt{\mathtt{x}} \le \frac{2}{3}\mathbb{D}_j\,.$$

This bound combined with the Bonferroni device yields the uniform bound (2.12). □

Now we derive the statements of the theorem from Proposition D.27. The required condition $(\mathcal{T}_\infty^*)$ follows from the full dimensional condition $(\mathcal{T}_3^*)$ checked in Section 2.1, however, with $\mathbb{D}^2 = \boldsymbol{F}_G$. In view of (2.6), the use $\mathbb{D}^2$ to (2.5) yields $(\mathcal{T}_3^*)$ with $\varkappa^3\tau_3$ in place of $\tau_3 = 3\,\boldsymbol{N}_G^{-1/2}$; see (2.3). Similarly, $(\mathcal{T}_\infty^*)$ continues to hold with $\tau_{12}=\tau_3$ and $\tau_{21}=\tau_3$; cf. Remark D.5. Now the results (2.9) and (2.10) follow from Proposition D.27. □

3 Semiparametric plug-in estimation

This section explains an approach to semiparametric estimation based on partial optimization which is not limited to the profile MLE. It assumes that a reasonable pilot estimator of the nuisance parameter is available. The aim is to establish some theoretical guarantees on the estimation accuracy which correspond to the dimension of the target parameter only. The idea can be explained as follows. Define a full dimensional truth $\boldsymbol{\upsilon}^* = (\boldsymbol{\theta}^*, \boldsymbol{s}^*)$ by

$$\boldsymbol{\upsilon}^* = (\boldsymbol{\theta}^*, \boldsymbol{s}^*) = \operatorname*{argmax}_{(\boldsymbol{\theta},\boldsymbol{s})} \mathbb{E}\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{s}).$$

Further, consider a family of partial optimization problems with respect to the target parameter $\boldsymbol{\theta}$ for each fixed value of the nuisance parameter $\boldsymbol{s}$:

$$\widetilde{\boldsymbol{\theta}}(\boldsymbol{s}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}} \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{s}),$$

$$\boldsymbol{\theta}^*(\boldsymbol{s}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{s}).$$

We also assume that each partial problem can be easily solved. Further, let a pilot estimate $\widehat{\boldsymbol{s}}$ of the nuisance parameter $\boldsymbol{s}$ be given. This leads to the *plug-in estimator*

$$\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}(\widehat{\boldsymbol{s}}).$$

The profile MLE is a special case of the plug-in method with $\widehat{\boldsymbol{s}} = \widetilde{\boldsymbol{s}}$, where $\widetilde{\boldsymbol{\upsilon}} = (\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{s}})$ is a full dimenstional maximizer of $\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{s})$. However, the results for the profile MLE can be derived from the general results on estimation of the full-dimensional parameter $\boldsymbol{\upsilon}$. Later we do not limit ourselves to any particular choice of $\widehat{\boldsymbol{s}}$. This requires to develop different tools and approaches. The aim is to show that under reasonable conditions on the model and the pilot $\widehat{\boldsymbol{s}}$, this plug-in estimator behaves nearly as the "oracle" estimator $\widetilde{\boldsymbol{\theta}}(\boldsymbol{s}^*)$. A very important requirement for this construction is that the pilot estimator $\widehat{\boldsymbol{s}}$ concentrates on a small vicinity $\mathscr{H}_0$ of the point $\boldsymbol{s}^*$. Further, the results of Section A imply a bound on the difference $\widetilde{\boldsymbol{\theta}}(\boldsymbol{s}) - \boldsymbol{\theta}^*(\boldsymbol{s})$ between the partial estimates $\widetilde{\boldsymbol{\theta}}(\boldsymbol{s})$ and its population counterpart $\boldsymbol{\theta}^*(\boldsymbol{s})$ in terms of the effective target dimension. Moreover, the full dimensional condition (ζ) enables us to state such bounds uniformly over the concentration set $\mathscr{H}_0$ of the pilot $\widehat{\boldsymbol{s}}$. Under a so called "small bias" condition $\boldsymbol{\theta}^*(\boldsymbol{s}) \approx \boldsymbol{\theta}^*$ and $\mathbb{F}(\boldsymbol{s}) \approx \mathbb{F}(\boldsymbol{s}^*)$, behavior of $\widetilde{\boldsymbol{\theta}}(\boldsymbol{s})$ only weakly depends on $\boldsymbol{s} \in \mathscr{H}_0$ yielding the desirable properties of $\widehat{\boldsymbol{\theta}}$.

3.1 Uniform bounds in partial estimation

This section studies *partial estimation* of the target parameter $\boldsymbol{\theta}$ for each $\boldsymbol{s}$ fixed. We intend to establish some bounds that apply uniformly over $\boldsymbol{s} \in \mathscr{H}_0$ for some set $\mathscr{H}_0 \in$

$\mathscr{H}$. Usually this is a concentration set of a pilot estimator $\widehat{\boldsymbol{s}}$. Further, for each $\boldsymbol{s} \in \mathscr{H}_0$, consider $\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{s})$ as a function of $\boldsymbol{\theta}$. We suppose that the function $f_{\boldsymbol{s}}(\boldsymbol{\theta}) = \mathbb{E}\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{s})$ satisfies $(\mathcal{T}^*_{3|\boldsymbol{s}})$. This is a third-order smoothness condition on $f_{\boldsymbol{s}}(\boldsymbol{\theta})$ for $\boldsymbol{s}$ fixed, but with universal constant τ_3, the metric tensor D, and the radius $\mathtt{r}$. An extension to the case when they depend on $\boldsymbol{s}$ is possible, but requires more notations and technical conditions. The matrix

$$
\mathbb{F}(\boldsymbol{s}) = -\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}\mathscr{L}(\boldsymbol{\theta}^*(\boldsymbol{s}), \boldsymbol{s})
$$

describes the information about the target parameter $\boldsymbol{\theta}$ in the partial model with $\boldsymbol{s}$ fixed. Later we need that the matrix $\mathbb{F}(\boldsymbol{s})$ does not vary too much within $\mathscr{H}_0$, that is, $\mathbb{F}(\boldsymbol{s}) \approx \mathbb{F} = \mathbb{F}(\boldsymbol{s}^*)$. More precisely, we assume that for some $\omega \le 1/3$

$$
-\omega D^2 \le \mathbb{F}(\boldsymbol{s}) - \mathbb{F} \le \omega D^2, \quad \forall \boldsymbol{s} \in \mathscr{H}_0. \tag{3.1}
$$

This can be effectively checked under condition $(\mathcal{T}^*_{3,\mathcal{S}})$; see Lemma D.20 and Lemma D.21. Due to assumption (ζ) about linearity of the stochastic term $\zeta(\boldsymbol{v})$, the $\boldsymbol{\theta}$-gradient $\nabla_{\boldsymbol{\theta}}\zeta$ does not depend on $\boldsymbol{s}$. This substantially simplifies our study. We also assume that $\nabla_{\boldsymbol{\theta}}\zeta$ satisfies $(\nabla\zeta)$. Define $B_D = \operatorname{Var}(D\,\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta)$, $\mathbb{p}_D = \operatorname{tr} B_D$, and

$$
\mathtt{r}_D \stackrel{\text{def}}{=} \frac{1}{(1-\omega)^{1/2}}\, z(B_D, \mathtt{x}) = \frac{1}{(1-\omega)^{1/2}}\, \bigl(\sqrt{\operatorname{tr} B_D} + \sqrt{2\mathtt{x}\, \|B_D\|}\bigr). \tag{3.2}
$$

It is important that the radius $\mathtt{r}_D$ corresponds to the dimension of the target parameter $\boldsymbol{\theta}$. Now we apply the general results from Section A to the partial pMLEs $\widetilde{\boldsymbol{\theta}}(\boldsymbol{s})$.

Proposition 3.1. *Assume (3.1) with $\omega \le 1/3$. For any $\boldsymbol{s} \in \mathscr{H}_0$, let $f_{\boldsymbol{s}}(\boldsymbol{\theta}) = \mathbb{E}\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{s})$ satisfy $(\mathcal{T}^*_{3|\boldsymbol{s}})$ at $\boldsymbol{\theta}^*(\boldsymbol{s})$ with $\mathtt{r}$, D, and τ_3 such that*

$$
D^2 \le \varkappa^2\, \mathbb{F}(\boldsymbol{s}), \quad \mathtt{r} \ge \frac{3}{2}\mathtt{r}_D, \quad \tau_3\, \varkappa^2\, \mathtt{r}_D < \frac{4}{9}.
$$

Then on a random set $\Omega(\mathtt{x})$ with $\mathbb{P}(\Omega(\mathtt{x})) \ge 1 - 3\mathrm{e}^{-\mathtt{x}}$, it holds for all $\boldsymbol{s} \in \mathscr{H}_0$

$$
\begin{aligned}
\bigl\|Q\bigl\{\widetilde{\boldsymbol{\theta}}(\boldsymbol{s}) - \boldsymbol{\theta}^*(\boldsymbol{s}) - \mathbb{F}^{-1}(\boldsymbol{s})\,\nabla_{\boldsymbol{\theta}}\zeta\bigr\}\bigr\| &\le \|Q\mathbb{F}^{-1}(\boldsymbol{s})D\|\, \frac{3\tau_3}{4}\, \|D\,\mathbb{F}^{-1}(\boldsymbol{s})\,\nabla_{\boldsymbol{\theta}}\zeta\|^2 \\
&\le \|Q\mathbb{F}^{-1}D\|\, \frac{(3/4)\tau_3}{(1-\omega)^{3/2}}\, \|D\,\mathbb{F}^{-1}\,\nabla_{\boldsymbol{\theta}}\zeta\|^2.
\end{aligned} \tag{3.3}
$$

Also, it holds on $\Omega(\mathtt{x})$

$$
\|Q\{\mathbb{F}^{-1}(\boldsymbol{s})\,\nabla_{\boldsymbol{\theta}}\zeta - \mathbb{F}^{-1}\,\nabla_{\boldsymbol{\theta}}\zeta\}\| \le \|Q\mathbb{F}^{-1}D\|\, \frac{\omega}{1-\omega}\, \|D\,\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta\|. \tag{3.4}
$$

Proof. Condition $(\boldsymbol{\nabla\zeta})$ implies on $\Omega(\mathtt{x})$

$$\|D\,\mathbb{F}^{-1}\,\nabla_{\boldsymbol{\theta}}\zeta\| \leq z(B_D,\mathtt{x})\,.$$

This, (3.2), and (3.1) yield on $\Omega(\mathtt{x})$ for any $\boldsymbol{s}\in\mathscr{H}_0$

$$\bigl\|D\,\mathbb{F}^{-1}(\boldsymbol{s})\,\nabla_{\boldsymbol{\theta}}\zeta\bigr\| \leq \mathtt{r}_D\,.$$

The first bound in (3.3) follows from Theorem A.4 applied to $\widetilde{\boldsymbol{\theta}}(\boldsymbol{s}) = \operatorname{argmax}_{\boldsymbol{\theta}}\{\mathscr{L}(\boldsymbol{\theta},\boldsymbol{s}) - \|G\boldsymbol{\theta}\|^2/2\}$. Further, for each $\boldsymbol{s}\in\mathscr{H}_0$, by (3.1)

$$\bigl\|D^{-1}\mathbb{F}\,\{\mathbb{F}^{-1}(\boldsymbol{s}) - \mathbb{F}^{-1}\}\mathbb{F}(\boldsymbol{s})D^{-1}\bigr\| = \bigl\|D^{-1}\{\mathbb{F} - \mathbb{F}(\boldsymbol{s})\}D^{-1}\bigr\| \leq \omega$$

and

$$\begin{aligned}
&\|Q\{\mathbb{F}^{-1}(\boldsymbol{s})\,\nabla_{\boldsymbol{\theta}}\zeta - \mathbb{F}^{-1}\,\nabla_{\boldsymbol{\theta}}\zeta\}\| \leq \|Q\mathbb{F}^{-1}D\|\;\|D^{-1}\mathbb{F}\{\mathbb{F}^{-1} - \mathbb{F}^{-1}(\boldsymbol{s})\}\nabla_{\boldsymbol{\theta}}\zeta\| \\
&\quad\leq \|Q\mathbb{F}^{-1}D\|\;\omega\;\|D\,\mathbb{F}^{-1}(\boldsymbol{s})\nabla_{\boldsymbol{\theta}}\zeta\|
\end{aligned}$$

With $Q = D^{-1}\mathbb{F}$, this implies

$$(1-\omega)\|D\,\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta\| \leq \|D\,\mathbb{F}^{-1}(\boldsymbol{s})\nabla_{\boldsymbol{\theta}}\zeta\| \leq (1+\omega)\|D\,\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta\|$$

and (3.4) follows as well as the second bound in (3.3). □

A benefit of (3.3) is that the accuracy of estimation corresponds to the dimension of the target component only. Another benefit of (3.3) and (3.4) is that these bounds hold on $\Omega(\mathtt{x})$ for all $\boldsymbol{s}\in\mathscr{H}_0$ simultaneously. This is granted by $(\boldsymbol{\zeta})$, $(\boldsymbol{\nabla\zeta})$, and (3.1). Note, however, that fixing the nuisance parameter $\boldsymbol{s}$ changes the value $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\boldsymbol{s}^*)$ to $\boldsymbol{\theta}^*(\boldsymbol{s})$, and we have to control the variability of this estimate w.r.t. $\boldsymbol{s}\in\mathscr{H}_0$.

3.2 Semiparametric bias under (semi)orthogonality

The results of Proposition 3.1 rely on variability of $\boldsymbol{\theta}^*(\boldsymbol{s}) = \operatorname{argmax}_{\boldsymbol{\theta}} f(\boldsymbol{\theta},\boldsymbol{s})$ w.r.t. the nuisance parameter $\boldsymbol{s}$, where $f(\boldsymbol{\upsilon}) = \mathbb{E}\mathscr{L}(\boldsymbol{\upsilon}) = \mathbb{E}\mathscr{L}(\boldsymbol{\theta},\boldsymbol{s})$. This section studies the *semiparametric bias* $\boldsymbol{\theta}^*(\boldsymbol{s}) - \boldsymbol{\theta}^*$. It appears that local quadratic approximation of the function f in a vicinity of $\boldsymbol{\upsilon}^*$ yields a nearly linear dependence of $\boldsymbol{\theta}^*(\boldsymbol{s})$ in $\boldsymbol{s}$. To see this, represent $\mathscr{F}(\boldsymbol{\upsilon}) = -\nabla^2 f(\boldsymbol{\upsilon})$ in the block form

$$\mathscr{F}(\boldsymbol{\upsilon}) = \begin{pmatrix} \mathscr{F}_{\boldsymbol{\theta\theta}}(\boldsymbol{\upsilon}) & \mathscr{F}_{\boldsymbol{\theta s}}(\boldsymbol{\upsilon}) \\ \mathscr{F}_{\boldsymbol{s\theta}}(\boldsymbol{\upsilon}) & \mathscr{F}_{\boldsymbol{ss}}(\boldsymbol{\upsilon}) \end{pmatrix}. \tag{3.5}$$

We write $I\!\!F(\boldsymbol{\upsilon}) = \mathscr{F}_{\boldsymbol{\theta\theta}}(\boldsymbol{\upsilon})$ and $\mathscr{F} = \mathscr{F}(\boldsymbol{\upsilon}^*)$. Consider first a *quadratic* f. As $\nabla f(\boldsymbol{\upsilon}^*) = 0$, it holds $f(\boldsymbol{\upsilon}) = f(\boldsymbol{\upsilon}^*) - (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)^\top \mathscr{F}\,(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)/2$. For $\boldsymbol{s}$ fixed, the point $\boldsymbol{\theta}^*(\boldsymbol{s})$ satisfies

$$
\begin{aligned}
\boldsymbol{\theta}^*(\boldsymbol{s}) &= \operatorname*{argmax}_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \boldsymbol{s}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top I\!\!F\,(\boldsymbol{\theta} - \boldsymbol{\theta}^*)/2 + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \mathscr{F}_{\boldsymbol{\theta s}}\,(\boldsymbol{s} - \boldsymbol{s}^*)\} \\
&= \boldsymbol{\theta}^* - I\!\!F^{-1}\,\mathscr{F}_{\boldsymbol{\theta s}}\,(\boldsymbol{s} - \boldsymbol{s}^*).
\end{aligned}
$$

This observation is in fact discouraging because the bias $\boldsymbol{\theta}^*(\boldsymbol{s}) - \boldsymbol{\theta}^*$ has the same (in order) magnitude as the nuisance parameter $\boldsymbol{s} - \boldsymbol{s}^*$. If $f(\boldsymbol{\upsilon})$ is not quadratic, the *orthogonality* condition

$$
\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{s}}\, f(\boldsymbol{\theta}, \boldsymbol{s}) \equiv 0, \qquad \forall (\boldsymbol{\theta}, \boldsymbol{s}) \in \mathcal{W},
$$

still ensures a vanishing bias. Indeed, it implies the decomposition $f(\boldsymbol{\theta}, \boldsymbol{s}) = f_1(\boldsymbol{\theta}) + f_2(\boldsymbol{s})$ for some functions f_1 and f_2. As a corollary, the maximizer $\boldsymbol{\theta}^*(\boldsymbol{s})$ and the corresponding negative Hessian $I\!\!F(\boldsymbol{s})$ do not depend on $\boldsymbol{s}$ yielding $\boldsymbol{\theta}^*(\boldsymbol{s}) \equiv \boldsymbol{\theta}^*$ and $I\!\!F(\boldsymbol{s}) \equiv I\!\!F$. This is a very useful property allowing to obtain accurate results about estimation accuracy of the target parameter $\boldsymbol{\theta}$ as if the true value of the nuisance parameter $\boldsymbol{s}^*$ were known. In practice one may apply the plug-in estimator $\widetilde{\boldsymbol{\theta}}(\widehat{\boldsymbol{s}})$, where $\widehat{\boldsymbol{s}}$ is any reasonable estimate of $\boldsymbol{s}^*$.

Unfortunately, the orthogonality condition $\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{s}}\, f(\boldsymbol{\theta}, \boldsymbol{s}) \equiv 0$ is too restrictive and fulfilled only in some special cases. One of them corresponds to the already mentioned additive case $f(\boldsymbol{\theta}, \boldsymbol{s}) = f_1(\boldsymbol{\theta}) + f_2(\boldsymbol{s})$. If $f(\boldsymbol{\theta}, \boldsymbol{s})$ is quadratic, then orthogonality can be achieved by a linear transform of the nuisance parameter $\boldsymbol{s}$. For a general function f, such a linear transform helps to only ensure the *one-point orthogonality* condition $\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{s}}\, f(\boldsymbol{\upsilon}^*) = 0$; see later in this section for more details.

In some situation, $\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{s}}\, f(\boldsymbol{\upsilon}^*) = 0$ implies $\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{s}}\, f(\boldsymbol{\theta}^*, \boldsymbol{s}) = 0$ for all $\boldsymbol{s} \in \mathscr{H}_0$. We refer to this situation as *semi-orthogonality.* A typical example is given by models for which the cross-derivative $\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{s}}\, f(\boldsymbol{\theta}^*, \boldsymbol{s})$ depends on $\boldsymbol{\theta}$ only. In this situation, the semiparametric bias vanishes. More precisely, Lemma D.18 yields the following result.

Proposition 3.2. *Let* $\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{s}}\, f(\boldsymbol{\theta}^*, \boldsymbol{s}) = 0$ *for all* $\boldsymbol{s} \in \mathscr{H}_0$ *and* $f(\boldsymbol{\theta}, \boldsymbol{s})$ *is concave in* $\boldsymbol{\theta}$ *for each* $\boldsymbol{s}$. *Then*

$$
\boldsymbol{\theta}^*(s) \equiv \boldsymbol{\theta}^*, \qquad I\!\!F(s) \stackrel{\text{def}}{=} -\nabla^2_{\boldsymbol{\theta\theta}} f(\boldsymbol{\theta}^*, \boldsymbol{s}) \equiv I\!\!F.
$$

As in the orthogonal case, the condition of semi-orthogonality allows to ignore the semiparametric bias; see Theorem 3.3 later.

3.3 Loss and risk of a plug-in estimator

This section discusses the properties of the plug-in estimator $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}(\widehat{\boldsymbol{s}})$. The major condition on $\widehat{\boldsymbol{s}}$ is that $\widehat{\boldsymbol{s}}$ belongs with high probability to the local set $\mathscr{H}_0$. To simplify our notation, we fix a random set $\Omega(\mathtt{x})$ with $\mathbb{P}(\Omega(\mathtt{x})) \geq 1 - 3\mathrm{e}^{-\mathtt{x}}$ and assume that $\widehat{\boldsymbol{s}} \in \mathscr{H}_0$ and condition (3.1) of Proposition 3.1 hold true on this set.

3.3.1 Orthogonal case

In some special cases like orthogonality or semi-orthogonality considered in Proposition 3.2, it holds $\boldsymbol{\theta}^*(\boldsymbol{s}) \equiv \boldsymbol{\theta}^*$, $\mathbb{F} \equiv \mathbb{F}(\boldsymbol{s})$ yielding $\boldsymbol{\xi}(\boldsymbol{s}) \equiv \boldsymbol{\xi}$. In particular, this condition meets for the special case when $\mathscr{F}_{\boldsymbol{\theta}\boldsymbol{s}}(\boldsymbol{\upsilon})$ and $\mathscr{F}_{\boldsymbol{s}\boldsymbol{s}}(\boldsymbol{\upsilon})$ in (3.5) for $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{s})$ depend on $\boldsymbol{\theta}$ only; see Proposition 3.2. Application of Theorem B.2 to the squared risk of $\widehat{\boldsymbol{\theta}}$ in the partial model with $\boldsymbol{s} = \boldsymbol{s}^*$ yields similarly to Proposition 3.1 the following very strong result.

Theorem 3.3. *Let* $\|H(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)\|_\circ \leq \mathtt{r}_\circ$ *on* $\Omega(\mathtt{x})$, *and let* $\boldsymbol{\theta}^*(\boldsymbol{s}) \equiv \boldsymbol{\theta}^*$, $\mathbb{F}(\boldsymbol{s}) \equiv \mathbb{F}$. *Under the conditions of Proposition 3.1, the plug-in estimator* $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}(\widehat{\boldsymbol{s}})$ *satisfies on* $\Omega(\mathtt{x})$

$$\bigl\|D^{-1}\mathbb{F}\bigl(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta\bigr\| \leq \frac{3\tau_3}{4}\,\|D\,\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta\|^2 .$$

Moreover, for any linear mapping Q, *define* $\mathbb{p}_Q = \mathbb{E}\,\|Q\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta\|^2$ *and*

$$\mathscr{R}_Q \stackrel{\mathrm{def}}{=} \mathbb{E}\bigl\{\|Q\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}\bigr\} \leq \mathbb{p}_Q .$$

With $\mathbb{p}_D$ *and* $\mathtt{r}_D$ *from (3.2), suppose*

$$\alpha_Q \stackrel{\mathrm{def}}{=} \frac{\|Q\,\mathbb{F}^{-1}D\|(3/4)\tau_3\,\mathtt{r}_D\sqrt{\mathbb{p}_D}}{\sqrt{\mathscr{R}_Q}} < 1 .$$

Then

$$(1-\alpha_Q)^2\mathscr{R}_Q \leq \mathbb{E}\bigl\{\|Q\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}\bigr\} \leq (1+\alpha_Q)^2\mathscr{R}_Q .$$

The standardized score $\boldsymbol{\xi} = D^{-1}\nabla_{\boldsymbol{\theta}}\zeta$ does not depend on the pilot $\widehat{\boldsymbol{s}}$. Thus, for the statement of Theorem 3.3, it suffices that $\widehat{\boldsymbol{s}}$ concentrates on $\mathscr{H}_0$, all partial smoothness conditions on $f(\boldsymbol{\theta}, \boldsymbol{s}) = \mathbb{E}\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{s})$ w.r.t. $\boldsymbol{\theta}$ hold uniformly over $\boldsymbol{s} \in \mathscr{H}_0$, and $\boldsymbol{\theta}^*(\boldsymbol{s}) \equiv \boldsymbol{\theta}^*$, $\mathbb{F}(\boldsymbol{s}) \equiv \mathbb{F}$.

3.3.2 Concentration and Fisher expansion for the plug-in estimator

This section presents several results for the semiparametric plug-in estimator $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}(\widehat{\boldsymbol{s}})$ with a given pilot estimator $\widehat{\boldsymbol{s}}$ of the nuisance parameter $\boldsymbol{s}$. The main issue for this study is the semiparametric bias $\boldsymbol{\theta}^*(\boldsymbol{s}) - \boldsymbol{\theta}^*$ caused by using a pilot $\widehat{\boldsymbol{s}}$ in place of the truth $\boldsymbol{s}^*$. This bias can be bounded under smoothness properties of $f(\boldsymbol{v})$ using an accurate expansion of the bias term $\boldsymbol{\theta}^*(\boldsymbol{s}) - \boldsymbol{\theta}^*$ from Section D.3.3 for all $\boldsymbol{s}$ from a vicinity $\mathscr{H}_0$ of $\boldsymbol{s}^*$. The approach allows to incorporate the anisotropic case when conditions on smoothness of $f(\boldsymbol{v}) = f(\boldsymbol{\theta}, \boldsymbol{s})$ are stated in different norm for the target parameter $\boldsymbol{\theta}$ and the nuisance parameter $\boldsymbol{s}$. A typical example is given by using a ℓ_2-norm for $\boldsymbol{\theta}$ and a sup-norm for $\boldsymbol{s}$; see Section D.4. Given a norm $\| \cdot \|_\circ$ in $\mathbb{R}^q$, a metric tensor H on $\boldsymbol{s}$, and a radius $\mathtt{r}_\circ$, consider local sets $\mathscr{H}_0$ of the form

$$
\mathscr{H}_0 = \{\boldsymbol{s} \colon \|H(\boldsymbol{s} - \boldsymbol{s}^*)\|_\circ \le \mathtt{r}_\circ\}, \tag{3.6}
$$

We assume conditions $(\boldsymbol{\mathcal{T}^*_{3|s}})$, $(\boldsymbol{\mathcal{T}^*_{3,\mathcal{S}}})$ from Section D.3. The first one requires that $f(\boldsymbol{\theta}, \boldsymbol{s})$ is smooth in $\boldsymbol{\theta}$ for $\boldsymbol{s}$ fixed, while the second one describes the smoothness properties of $f(\boldsymbol{\theta}^*, \boldsymbol{s})$ w.r.t. $\boldsymbol{s}$. These conditions are weaker than the full dimensional smoothness condition $(\boldsymbol{\mathcal{T}^*_3})$ and only involve partial derivatives in $\boldsymbol{\theta}$ and cross-derivatives of $f(\boldsymbol{v}) = f(\boldsymbol{\theta}, \boldsymbol{s})$; see Section D.3.3 for a detailed discussion. Moreover, in some sense, these conditions are nothing but definitions of the important quantities τ_3, ρ_2, τ_{12}, and τ_{21}. We do not require $\mathscr{F}_{\boldsymbol{\theta s}} = 0$, however, it is implicitly assumed that this operator is close to zero. To quantify this statement, introduce the dual norm of an operator $B \colon \mathbb{R}^q \to \mathbb{R}^p$:

$$
\|B\|_* = \sup_{\boldsymbol{z} \colon \|\boldsymbol{z}\|_\circ \le 1} \|B\boldsymbol{z}\|.
$$

If $p = 1$ and $\| \cdot \|_\circ$ is the sup-norm $\| \cdot \|_\infty$ then $\|B\|_* = \|B\|_1$. Define

$$
\begin{aligned}
\rho_* &\stackrel{\text{def}}{=} \|D^{-1} \mathscr{F}_{\boldsymbol{xs}} H^{-1}\|_*, \\
\rho_2 &\stackrel{\text{def}}{=} \rho_* + \tau_{12}\, \mathtt{r}_\circ / 2.
\end{aligned} \tag{3.7}
$$

Now we apply Proposition D.25 yielding the following *concentration bound* and *semiparametric Fisher expansion.*

Theorem 3.4. *Let* $\widehat{\boldsymbol{s}} \in \mathscr{H}_0$ *on* $\Omega(\mathtt{x})$ *with the local set* $\mathscr{H}_0$ *from* (3.6) *with a norm* $\| \cdot \|_\circ$*, a metric tensor* H *in* $\mathbb{R}^q$*, and a radius* $\mathtt{r}_\circ$*. Let also* $(\boldsymbol{\mathcal{T}^*_{3,\mathcal{S}}})$ *and* $(\boldsymbol{\mathcal{T}^*_{3|s}})$ *with*

$\mathbb{D}_{\boldsymbol{s}} \equiv D$ hold for a metric tensor D and some constants τ_3, τ_{12}, τ_{21}, and $\mathtt{r}$ such that

$$D^2 \le \varkappa^2 \mathbb{F}, \qquad \omega \stackrel{\text{def}}{=} \varkappa^2 \tau_{21} \mathtt{r}_\circ \le 1/4, \qquad \mathtt{r} \ge \frac{3\varkappa^2 \rho_2}{2(1-\omega)} \mathtt{r}_\circ, \qquad \frac{\varkappa^4 \rho_2 \tau_3}{1-\omega} \mathtt{r}_\circ \le \frac{4}{9},$$

with ρ_* and ρ_2 from (3.7). Then it holds on $\Omega(\mathtt{x})$

$$\|D(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \le \frac{3\varkappa \rho_2}{2(1-\omega)} \bigl(\varkappa^2 \rho_2 \, \|H(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)\|_\circ + \|D \, \mathbb{F}^{-1} \nabla_{\boldsymbol{\theta}} \zeta\|\bigr),$$

and, moreover,

$$\begin{aligned} &\|Q\{\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* - \mathbb{F}^{-1} \mathscr{F}_{\boldsymbol{\theta}\boldsymbol{s}}(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*) - \mathbb{F}^{-1} \nabla_{\boldsymbol{\theta}} \zeta\}\| \\ &\quad \le \|Q \, \mathbb{F}^{-1} D\| \Bigl\{(\tau_\circ + \varkappa \tau_{21}) \, \|H(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)\|_\circ^2 + (2\tau_3 + \varkappa^{-1} \tau_{21}/2) \, \|D \, \mathbb{F}^{-1} \nabla_{\boldsymbol{\theta}} \zeta\|^2\Bigr\}, \end{aligned} \tag{3.8}$$

where

$$\tau_\circ \stackrel{\text{def}}{=} \frac{1}{1-\omega} \biggl(\varkappa^2 \rho_* \tau_{21} + \frac{\tau_{12}}{2} + \frac{3\varkappa^4 \rho_2^2 \tau_3}{4(1-\omega)^2}\biggr).$$

3.3.3 Risk of the plug-in estimator

This section describes concentration sets of the semiparametric plug-in estimator $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}(\widehat{\boldsymbol{s}})$ and provides some bounds on the squared localized risk $\mathbb{E}\{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \, \mathbb{1}_{\Omega(\mathtt{x})}\}$ using the Fisher expansion (3.8). Informally it can be written as

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \approx \mathbb{F}^{-1} \mathscr{F}_{\boldsymbol{\theta}\boldsymbol{s}}(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*) + \mathbb{F}^{-1} \nabla \zeta.$$

This decomposition describes two sources of the estimation loss: the *variance* (stochastic) term $\mathbb{F}^{-1} \nabla \zeta$ is due to random errors in observations and the *semiparametric bias* term $\mathbb{F}^{-1} \mathscr{F}_{\boldsymbol{\theta}\boldsymbol{s}}(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)$ is due to the use of the pilot $\widehat{\boldsymbol{s}}$ in place of the truth $\boldsymbol{s}^*$.

Theorem 3.5. *Under conditions of Theorem 3.4, it holds*

$$\begin{aligned} &\Bigl|\mathbb{E}\bigl\{\|Q(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\| \, \mathbb{1}_{\Omega(\mathtt{x})}\bigr\} - \mathbb{E}\Bigl\{\bigl\|Q\bigl\{\mathbb{F}^{-1} \nabla \zeta + \mathbb{F}^{-1} \mathscr{F}_{\boldsymbol{\theta}\boldsymbol{s}}(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)\bigr\}\bigr\| \, \mathbb{1}_{\Omega(\mathtt{x})}\Bigr\}\Bigr| \\ &\le \|Q \, \mathbb{F}^{-1} D\| \Bigl\{(\tau_\circ + \varkappa \tau_{21}) \, \mathbb{p}_H + (2\tau_3 + \varkappa^{-1} \tau_{21}/2) \, \mathbb{p}_D\Bigr\}, \end{aligned} \tag{3.9}$$

where

$$\mathbb{p}_D \stackrel{\text{def}}{=} \mathbb{E} \, \|D \, \mathbb{F}^{-1} \nabla_{\boldsymbol{\theta}} \zeta\|^2, \qquad \mathbb{p}_H \stackrel{\text{def}}{=} \mathbb{E}\bigl\{\|H(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)\|_\circ^2 \, \mathbb{1}_{\Omega(\mathtt{x})}\bigr\} \le \mathtt{r}_\circ^2.$$

Moreover, assume

$$\mathbb{E}\bigl\{\|H(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)\|_\circ^4 \, \mathbb{1}_{\Omega(\mathtt{x})}\bigr\} \le \mathtt{C}_H^2 \, \mathbb{p}_H^2, \qquad \mathbb{E}\bigl\{\|D \, \mathbb{F}^{-1} \nabla_{\boldsymbol{\theta}} \zeta\|^4 \, \mathbb{1}_{\Omega(\mathtt{x})}\bigr\} \le \mathtt{C}_D^2 \, \mathbb{p}_D^2.$$

Then it holds with $\mathscr{R}_Q \stackrel{\text{def}}{=} \mathbb{E}\bigl\{\|Q\mathbb{F}^{-1}\nabla\zeta + Q\mathbb{F}^{-1}\mathscr{F}_{\boldsymbol{\theta}\boldsymbol{s}}(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)\|^2\, \mathbb{1}_{\Omega(\mathtt{x})}\bigr\}$,

$$\begin{aligned}
&\Bigl|\sqrt{\mathbb{E}\,\|Q(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}} - \sqrt{\mathscr{R}_Q}\Bigr| \\
&\qquad \le \|Q\,\mathbb{F}^{-1}D\|\,\Bigl\{(\tau_\circ + \varkappa\,\tau_{21})\,\mathtt{C}_H\,\mathbb{p}_H + (2\tau_3 + \varkappa^{-1}\tau_{21}/2)\mathtt{C}_D\,\mathbb{p}_D\Bigr\}.
\end{aligned}$$

Proof. Bound (3.9) follows by (3.8). Define $\boldsymbol{\varepsilon}_Q = Q\bigl\{\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* - \mathbb{F}^{-1}\mathscr{F}_{\boldsymbol{\theta}\boldsymbol{s}}(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*) - \mathbb{F}^{-1}\,\nabla\zeta\bigr\}$. Then again by (3.8)

$$\begin{aligned}
&\Bigl|\sqrt{\mathbb{E}\,\|Q(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}} - \sqrt{\mathscr{R}_Q}\Bigr| \le \sqrt{\mathbb{E}\,\|\boldsymbol{\varepsilon}_Q\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}} \le \|Q\,\mathbb{F}^{-1}D\| \times \\
&\times\Bigl\{(\tau_\circ + \varkappa\,\tau_{21})\sqrt{\mathbb{E}\|H(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)\|_\circ^4\,\mathbb{1}_{\Omega(\mathtt{x})}} + \Bigl(2\tau_3 + \frac{\tau_{21}}{2\varkappa}\Bigr)\sqrt{\mathbb{E}\|D\,\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta\|^4\,\mathbb{1}_{\Omega(\mathtt{x})}}\Bigr\},
\end{aligned}$$

and the last assertion follows as well. □

3.3.4 Semiparametric adaptivity

This section discusses the issue of semiparametrically adaptive estimation. The question under study is under which conditions the plug-in estimator $\widehat{\boldsymbol{\theta}}$ provides an "oracle" accuracy corresponding to the case of known nuisance parameter $\boldsymbol{s}$. To be more certain, fix $Q = D$. For more transparency, we also assume $\varkappa = 1$, $\tau_3 \asymp n^{-1/2}$ as in **($\boldsymbol{\mathcal{S}_3^*}$)** for the "sample size" n and similarly $\tau_{12} \asymp n^{-1/2}$, $\tau_{21} \asymp n^{-1/2}$. An inspection of the results of Proposition 3.4 and Theorem 3.5 reveals two places to be analyzed. Expansion (3.8) involves the linear term $D\,\mathbb{F}^{-1}\mathscr{F}_{\boldsymbol{\theta}\boldsymbol{s}}(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)$, it should not be larger in magnitude than the term $D\,\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta$. In a high-dimensional situation, the squared norm $\|D\,\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta\|^2$ concentrates around its expectation $\mathbb{p}_D$. By (3.7), it holds on $\Omega(\mathtt{x})$

$$\|D\,\mathbb{F}^{-1}\mathscr{F}_{\boldsymbol{\theta}\boldsymbol{s}}(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)\| \le \rho_*\|H(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)\|_\circ \le \rho_*\,\mathtt{r}_\circ\,,$$

and it is sufficient to check that $\rho_*\,\mathtt{r}_\circ \ll \mathbb{p}_D^{1/2}$. This is the most critical condition, it requires the cross-correlation between the target and the nuisance parameter to be sufficiently small. Similarly, in the remainder of the expansion (3.8), the terms $(\tau_\circ + \tau_{21})\,\|H(\widehat{\boldsymbol{s}} - \boldsymbol{s}^*)\|_\circ^2$ and $(2\tau_3 + \tau_{21}/2)\,\|D\,\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta\|^2$ should not be too large compared to magnitude of $\|D\,\mathbb{F}^{-1}\nabla_{\boldsymbol{\theta}}\zeta\|$. The corresponding relation can be spelled out as

$$n^{-1/2}(\mathbb{p}_D + \mathbb{p}_H) = o(\sqrt{\mathbb{p}_D})\,.$$

3.4 Sup-norm bounds

This section discusses the problem of estimation in sup-norm by a maximum likelihood method

$$\widetilde{\boldsymbol{v}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{v}} \mathscr{L}(\boldsymbol{v}).$$

The aim is to bound the error $\widetilde{\boldsymbol{v}} - \boldsymbol{v}^*$ componentwise. The proposed approach is for each entry $\widetilde{v}_j$ of $\widetilde{\boldsymbol{v}}$ to treat the remaining coordinates as a nuisance parameter and apply the results of Proposition D.27 from Section D.3.

A Properties of the MLE $\widetilde{\boldsymbol{\upsilon}}$ for SLS models

This section collects general results about concentration and expansion of the MLE in the SLS setup which substantially improve the bounds from Spokoiny (2017) and Spokoiny (2023a). We assume to be given a random function $L(\boldsymbol{\upsilon})$, $\boldsymbol{\upsilon} \in \Upsilon \subseteq \mathbb{R}^p$, $p < \infty$. This function can be viewed as log-likelihood or negative loss. Consider in parallel two optimization problems defining the MLE $\widetilde{\boldsymbol{\upsilon}}$ and its population counterpart (the background truth) $\boldsymbol{\upsilon}^*$:

$$\widetilde{\boldsymbol{\upsilon}} = \operatorname*{argmax}_{\boldsymbol{\upsilon}} L(\boldsymbol{\upsilon}), \qquad \boldsymbol{\upsilon}^* = \operatorname*{argmax}_{\boldsymbol{\upsilon}} \mathbb{E} L(\boldsymbol{\upsilon}), \tag{A.1}$$

The corresponding Fisher information matrix $\mathbb{F}(\boldsymbol{\upsilon})$ is given by

$$\mathbb{F}(\boldsymbol{\upsilon}) = -\nabla^2 \mathbb{E} L(\boldsymbol{\upsilon}) .$$

Denote $\mathbb{F} = \mathbb{F}(\boldsymbol{\upsilon}^*)$.

A.1 Basic conditions

Now we present our major conditions. The most important one is about linearity of the stochastic component $\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E} L(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E} L(\boldsymbol{\upsilon})$.

$(\boldsymbol{\zeta})$ *The stochastic component* $\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E} L(\boldsymbol{\upsilon})$ *of the process* $L(\boldsymbol{\upsilon})$ *is linear in* $\boldsymbol{\upsilon}$. *We denote by* $\nabla\zeta \equiv \nabla\zeta(\boldsymbol{\upsilon}) \in \mathbb{R}^p$ *its gradient* .

Below we assume some concentration properties of the stochastic vector $\nabla\zeta$. More precisely, we require that $\nabla\zeta$ obeys the following condition.

$(\boldsymbol{\nabla\zeta})$ *There exists* $V^2 \geq \operatorname{Var}(\nabla\zeta)$ *such that for all considered* $B \in \mathfrak{M}_p$ *and* $\mathtt{x} > 0$

$$\mathbb{P}\bigl(\|B^{1/2} V^{-1} \nabla\zeta\| \geq z(B, \mathtt{x})\bigr) \leq 3 \mathrm{e}^{-\mathtt{x}},$$
$$z^2(B, \mathtt{x}) \stackrel{\text{def}}{=} \operatorname{tr} B + 2\sqrt{\mathtt{x} \operatorname{tr} B^2} + 2\mathtt{x} \|B\| . \tag{A.2}$$

This condition can be effectively checked if the errors in the data exhibit sub-gaussian or sub-exponential behavior; see Spokoiny (2024c), Spokoiny (2024a). The important special case corresponds to $B = \mathbb{F}^{-1/2} V^2 \mathbb{F}^{-1/2}$ and $\mathtt{x} \approx \log n$ leading to the bound

$$\mathbb{P}\bigl(\|\mathbb{F}^{-1/2} \nabla\zeta\| > z(B, \mathtt{x})\bigr) \leq 3/n.$$

The value $\mathbb{p}_G = \operatorname{tr}(\mathbb{F}^{-1} V^2)$ can be called the *effective dimension*; see Spokoiny (2017).

We also assume that the log-likelihood $L(\boldsymbol{v})$ or, equivalently, its deterministic part $\mathbb{E}L(\boldsymbol{v})$ is a concave function. It can be relaxed using localization; see Section **??**.

$(\mathcal{C})$ *The function $\mathbb{E}L(\boldsymbol{v})$ is concave on Υ which is open and convex set in $\mathbb{R}^p$.*

Later we will also need some smoothness conditions on the function $f(\boldsymbol{v}) = \mathbb{E}L(\boldsymbol{v})$ within a local vicinity of the point $\boldsymbol{v}^*$. The notion of locality is given in terms of a metric tensor $D \in \mathfrak{M}_p$. In most of the results later on, one can use $D = \mathbb{F}^{1/2}$. In general, we only assume $D^2 \le \varkappa^2 \mathbb{F}$ for some $\varkappa > 0$. Introduce the error of the second-order Taylor approximation at a point $\boldsymbol{v}$ in a direction $\boldsymbol{u}$ by

$$\begin{aligned} \delta_3(\boldsymbol{v},\boldsymbol{u}) &= f(\boldsymbol{v}+\boldsymbol{u}) - f(\boldsymbol{v}) - \langle \nabla f(\boldsymbol{v}), \boldsymbol{u} \rangle - \frac{1}{2}\langle \nabla^2 f(\boldsymbol{v}), \boldsymbol{u}^{\otimes 2} \rangle, \\ \delta_3'(\boldsymbol{v},\boldsymbol{u}) &= \langle \nabla f(\boldsymbol{v}+\boldsymbol{u}), \boldsymbol{u} \rangle - \langle \nabla f(\boldsymbol{v}), \boldsymbol{u} \rangle - \langle \nabla^2 f(\boldsymbol{v}), \boldsymbol{u}^{\otimes 2} \rangle . \end{aligned}$$

Second order smoothness means a bound of the form

$$\delta_3(\boldsymbol{v},\boldsymbol{u}) \le \omega \|D\boldsymbol{u}\|^2 , \quad \delta_3'(\boldsymbol{v},\boldsymbol{u}) \le \omega' \|D\boldsymbol{u}\|^2 , \qquad \|D\boldsymbol{u}\| \le \mathtt{r} , \tag{A.3}$$

for some radius $\mathtt{r}$ and small constants ω and ω'. These quantities can be effectively bounded under smoothness conditions $(\mathcal{T}_3)$, $(\mathcal{T}_3^*)$, or $(\mathcal{S}_3^*)$ given in Section D. For instance, under $(\mathcal{T}_3)$, by Lemma D.1, it holds for a small constant τ_3

$$\omega' \le \tau_3 \, \mathtt{r} , \qquad \omega \le \tau_3 \, \mathtt{r}/3.$$

Also under $(\mathcal{S}_3^*)$, the same bounds apply with $\tau_3 = \mathsf{c}_3 \, n^{-1/2}$; see Lemma D.2.

The class of models satisfying the conditions (ζ), $(\nabla\zeta)$, and $(\mathcal{C})$ with a smooth function $f(\boldsymbol{v}) = \mathbb{E}L(\boldsymbol{v})$ will be referred to as *stochastically linear smooth* (SLS). This class includes linear regression, generalized linear models (GLM), and log-density models; see Spokoiny and Panov (2021), Ostrovskii and Bach (2021) or Spokoiny (2023a). However, this class is much larger. For instance, nonlinear regression can be adapted to the SLS framework by an extension of the parameter space; see Section **??**.

A.2 Concentration of the MLE $\widetilde{\boldsymbol{v}}$. 2S-expansions

This section discusses properties of the MLE $\widetilde{\boldsymbol{v}} = \operatorname{argmax}_{\boldsymbol{v}} L(\boldsymbol{v})$ under second-order smoothness conditions. Fix $\mathtt{x} > 0$ and define with V^2 from $(\nabla\zeta)$ and $B = \mathbb{F}^{-1/2} V^2 \mathbb{F}^{-1/2}$

$$\mathcal{U} \stackrel{\text{def}}{=} \bigl\{ \boldsymbol{u} \colon \|\mathbb{F}^{1/2}\boldsymbol{u}\| \le \tfrac{4}{3} \mathtt{r}_B \bigr\}, \qquad \mathtt{r}_B \stackrel{\text{def}}{=} z(B,\mathtt{x}), \tag{A.4}$$

where $z(B, \mathtt{x})$ is given by (A.2). By $(\nabla\zeta)$, on a random set $\Omega(\mathtt{x})$ with $\mathbb{P}(\Omega(\mathtt{x})) \geq 1 - 3\mathrm{e}^{-\mathtt{x}}$, it holds $\|\mathbb{F}^{-1/2}\nabla\zeta\| \leq \mathtt{r}$. Further, for the metric tensor D from (A.3), define

$$\omega \stackrel{\text{def}}{=} \sup_{\boldsymbol{u}\in\mathcal{U}} \frac{2|\delta_3(\boldsymbol{v}^*, \boldsymbol{u})|}{\|D\boldsymbol{u}\|^2} , \qquad \omega' \stackrel{\text{def}}{=} \sup_{\boldsymbol{u}\in\mathcal{U}} \frac{|\delta_3'(\boldsymbol{v}^*, \boldsymbol{u})|}{\|D\boldsymbol{u}\|^2} . \tag{A.5}$$

Proposition A.1. *Suppose (ζ), $(\nabla\zeta)$, and $(\mathcal{C})$. Let also $D^2 \leq \varkappa^2 \mathbb{F}$ and $\omega' \varkappa^2 < 1/4$; see (A.5). Then on $\Omega(\mathtt{x})$, it holds*

$$\|\mathbb{F}^{1/2}(\widetilde{\boldsymbol{v}} - \boldsymbol{v}^*)\| \leq \frac{4}{3}\mathtt{r}_B , \qquad \|D(\widetilde{\boldsymbol{v}} - \boldsymbol{v}^*)\| \leq \frac{4\varkappa}{3}\mathtt{r}_B .$$

Proof. Apply Proposition D.4 to $f(\boldsymbol{v}) = \mathbb{E}L(\boldsymbol{v})$, $\nu = 3/4$, and $\boldsymbol{A} = \nabla\zeta$. □

Concentration of $\widetilde{\boldsymbol{v}}$ around $\boldsymbol{v}^*$ can be used to establish a version of the Fisher expansion for the estimation error $\widetilde{\boldsymbol{v}} - \boldsymbol{v}^*$ and the Wilks expansion for the excess $L(\widetilde{\boldsymbol{v}}) - L(\boldsymbol{v}^*)$. The result substantially improves the bounds from Ostrovskii and Bach (2021) for M-estimators and follows by Proposition D.5.

Theorem A.2. *Assume the conditions of Proposition A.1. Then on $\Omega(\mathtt{x})$*

$$2L(\widetilde{\boldsymbol{v}}) - 2L(\boldsymbol{v}^*) - \bigl\|\mathbb{F}^{-1/2}\nabla\zeta\bigr\|^2 \leq \frac{\omega}{1 - \varkappa^2\omega}\bigl\|D\mathbb{F}^{-1}\nabla\zeta\bigr\|^2 ,$$
$$2L(\widetilde{\boldsymbol{v}}) - 2L(\boldsymbol{v}^*) - \bigl\|\mathbb{F}^{-1/2}\nabla\zeta\bigr\|^2 \geq -\frac{\omega}{1 + \varkappa^2\omega}\bigl\|D\mathbb{F}^{-1}\nabla\zeta\bigr\|^2 .$$

Also

$$\bigl\|D\bigl(\widetilde{\boldsymbol{v}} - \boldsymbol{v}^* - \mathbb{F}^{-1}\nabla\zeta\bigr)\bigr\| \leq \frac{\sqrt{3\omega}}{1 - \varkappa^2\omega}\bigl\|D\mathbb{F}^{-1}\nabla\zeta\bigr\| ,$$
$$\bigl\|D\bigl(\widetilde{\boldsymbol{v}} - \boldsymbol{v}^*\bigr)\bigr\| \leq \frac{1 + \sqrt{3\omega}}{1 - \varkappa^2\omega}\bigl\|D\mathbb{F}^{-1}\nabla\zeta\bigr\| .$$

A.3 Expansions and risk bounds under third-order smoothness

The results of Theorem A.2 can be refined under third-order smoothness conditions. Namely, Proposition D.6 yields the following Wilks expansion for the MLE $\widetilde{\boldsymbol{v}}$.

Theorem A.3. *Assume (ζ), $(\nabla\zeta)$, and $(\mathcal{C})$. Let also $(\mathcal{T}_3)$ hold at $\boldsymbol{v}^*$ with a metric tensor D and values $\mathtt{r}$ and τ_3 satisfying*

$$D^2 \leq \varkappa^2\, \mathbb{F}, \quad \mathtt{r} \geq \frac{4\varkappa}{3}\mathtt{r}_B, \quad \tau_3\, \varkappa^3\, \mathtt{r}_B < \frac{1}{4} ,$$

for $\mathtt{r}_B$ from (A.4). Then on $\Omega(\mathtt{x})$, it holds

$$\|\mathbb{F}^{1/2}(\widetilde{\boldsymbol{v}} - \boldsymbol{v}^*)\| \leq \frac{4}{3}\mathtt{r}_B , \qquad \|D(\widetilde{\boldsymbol{v}} - \boldsymbol{v}^*)\| \leq \frac{4\varkappa}{3}\mathtt{r}_B ,$$

and

$$\Bigl| 2L(\widetilde{\boldsymbol{\upsilon}}) - 2L(\boldsymbol{\upsilon}^{*}) - \|\mathbb{F}^{-1/2}\nabla\zeta\|^{2} \Bigr| \le \frac{\tau_3}{2}\, \|D\mathbb{F}^{-1}\nabla\zeta\|^{3} .$$

Under $(\mathcal{T}_3^{*})$, Proposition D.8 yields an advanced Fisher expansion. Define

$$B_D = D\mathbb{F}^{-1}V^2\mathbb{F}^{-1}D,$$

$$\mathbb{p}_D \stackrel{\text{def}}{=} \operatorname{tr} B_D\,, \quad \mathtt{r}_D \stackrel{\text{def}}{=} z(B_D,\mathtt{x}) \le \sqrt{\operatorname{tr} B_D} + \sqrt{2\mathtt{x}\,\|B_D\|}\,; \tag{A.6}$$

cf. (A.2). By $(\nabla\zeta)$, it holds $\mathbb{P}(\|D\,\mathbb{F}^{-1}\nabla\zeta\| > \mathtt{r}_D) \le 3\mathrm{e}^{-\mathtt{x}}$. The result follows by limiting to the set $\Omega(\mathtt{x})$ on which $\|D\,\mathbb{F}^{-1}\nabla\zeta\| \le \mathtt{r}_D$ and by applying Proposition D.8.

Theorem A.4. *Assume* (ζ), $(\nabla\zeta)$, *and* $(\mathcal{C})$. *Let* $(\mathcal{T}_3^{*})$ *hold at* $\boldsymbol{\upsilon}^{*}$ *with a metric tensor* D *and values* $\mathtt{r}$ *and* τ_3 *satisfying*

$$D^2 \le \varkappa^2\,\mathbb{F}, \quad \mathtt{r} \ge \frac{3}{2}\,\mathtt{r}_D\,, \quad \tau_3\,\varkappa^2\,\mathtt{r}_D < \frac{4}{9}\,, \tag{A.7}$$

where $\mathtt{r}_D$ *is from* (A.6). *With* $\Omega(\mathtt{x}) = \{\|D\,\mathbb{F}^{-1}\nabla\zeta\| \le \mathtt{r}_D\}$, *it holds* $\mathbb{P}(\Omega(\mathtt{x})) \ge 1 - 3\mathrm{e}^{-\mathtt{x}}$ *and on* $\Omega(\mathtt{x})$

$$\|D^{-1}\mathbb{F}(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^{*} - \mathbb{F}^{-1}\nabla\zeta)\| \le \frac{3\tau_3}{4}\|D\,\mathbb{F}^{-1}\nabla\zeta\|^{2} . \tag{A.8}$$

Expansion (A.8) yields accurate risk bounds.

Theorem A.5. *Assume* (ζ), $(\nabla\zeta)$, *and* $(\mathcal{C})$. *Let* $f(\boldsymbol{\upsilon}) = \mathbb{E}L(\boldsymbol{\upsilon})$ *satisfy* $(\mathcal{T}_3^{*})$ *at* $\boldsymbol{\upsilon}^{*}$ *with some* D, $\mathtt{r}$, *and* τ_3. *Let also*

$$D^2 \le \varkappa^2\,\mathbb{F}\,, \qquad \mathtt{r} \ge \frac{3}{2}\mathtt{r}_D\,, \qquad \varkappa^2\tau_3\,\mathtt{r}_D < \frac{4}{9}\,;$$

see (A.6). *For any linear mapping* $Q\colon \mathbb{R}^p \to \mathbb{R}^q$, *it holds on* $\Omega(\mathtt{x})$

$$\|Q(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^{*} - \mathbb{F}^{-1}\nabla\zeta)\| \le \|Q\mathbb{F}^{-1}D\|\,\frac{3\tau_3}{4}\,\|D\mathbb{F}^{-1}\nabla\zeta\|^{2} .$$

Also, introduce

$$\mathscr{R}_Q \stackrel{\text{def}}{=} \mathbb{E}\{\|Q\mathbb{F}^{-1}\nabla\zeta\|^{2}\,\mathbb{1}_{\Omega(\mathtt{x})}\} \le \mathbb{p}_Q$$

with $\mathbb{p}_Q \stackrel{\text{def}}{=} \mathbb{E}\|Q\mathbb{F}^{-1}\nabla\zeta\|^{2} = \operatorname{tr}\operatorname{Var}(Q\mathbb{F}^{-1}\nabla\zeta)$. *Then*

$$\mathbb{E}\bigl\{\|Q(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^{*})\|\,\mathbb{1}_{\Omega(\mathtt{x})}\bigr\} \le \mathscr{R}_Q^{1/2} + \|Q\mathbb{F}^{-1}D\|\,\frac{3\tau_3}{4}\,\mathbb{p}_D\,.$$

Further, assume $\mathbb{E}\bigl\{\|D\mathbb{F}^{-1}\nabla\zeta\|^4\,\mathbb{I}_{\Omega(\mathtt{x})}\bigr\} \leq \mathtt{C}_4^2\,\mathbb{p}_D^2$ *and define*

$$\alpha_Q \stackrel{\text{def}}{=} \frac{\|Q\mathbb{F}^{-1}D\|\,(3/4)\tau_3\,\mathtt{C}_4\,\mathbb{p}_D}{\sqrt{\mathscr{R}_Q}}\,. \tag{A.9}$$

If $\alpha_Q < 1$ *then*

$$(1-\alpha_Q)^2\mathscr{R}_Q \leq \mathbb{E}\bigl\{\|Q\,(\widetilde{\boldsymbol{\upsilon}}-\boldsymbol{\upsilon}^*)\|^2\,\mathbb{I}_{\Omega(\mathtt{x})}\bigr\} \leq (1+\alpha_Q)^2\mathscr{R}_Q\,. \tag{A.10}$$

A.4 Effective and critical dimension in ML estimation

This section discusses the important question of the critical parameter dimension still ensuring the validity of the presented results. To be more specific, we only consider the 3S-results of Theorem A.4. Also, assume $\varkappa \equiv 1$. The important constant τ_3 is identified by $(\boldsymbol{\mathcal{S}_3^*})$: $\tau_3 = \mathtt{c}_3/\sqrt{n}$, where the scaling factor n means the sample size. It can be defined as the smallest eigenvalue of the Fisher operator $\mathbb{F}$.

First, we discuss the case $Q = D = \mathbb{F}^{1/2}$. It appears that in this full dimensional situation, all the obtained results apply and are meaningful under the condition $\mathbb{p} \ll n$, where $\mathbb{p} = \operatorname{tr}(B)$ for $B = \mathbb{F}^{-1/2}V^2\mathbb{F}^{-1/2}$ is the *effective dimension* of the problem. Indeed, $\mathtt{r}_D^2 = \mathtt{r}_B^2 \approx \operatorname{tr}(B) = \mathbb{p}$, and condition (A.7) requires $\tau_3\,\mathtt{r}_D \ll 1$ which can be spelled out as $\mathbb{p} \ll n$. Expansion (A.8) means

$$\|\mathbb{F}^{1/2}(\widetilde{\boldsymbol{\upsilon}}-\boldsymbol{\upsilon}^*)\| \leq \|\mathbb{F}^{-1/2}\nabla\zeta\| + \frac{3\tau_3}{4}\|\mathbb{F}^{-1/2}\nabla\zeta\|^2\,,$$

and the second term on the right-hand side of this bound is smaller than the first one under the same condition $\tau_3\,\mathtt{r}_D \ll 1$. Similar observations apply to bound (A.10) of Theorem A.5 which is meaningful only if α_Q in (A.9) is small. As $\mathscr{R}_Q \approx \mathbb{p}_Q = \mathbb{p}$, the condition $\tau_3\,\mathtt{r}_D \ll 1$ implies $\alpha_Q \ll 1$ and hence, the bound (A.10) is sharp. We conclude that the main properties of the MLE $\widetilde{\boldsymbol{\upsilon}}$ are valid under the condition $\mathbb{p} \ll n$ meaning sufficiently many observations per effective number of parameters.

The situation changes drastically if Q is not full-dimensional as e.g. in semiparametric estimation, when Q projects onto a low-dimensional target component. We will see in Section **??** that in this case, (A.9) requires $\mathbb{p}^2 \ll n$.

B Penalization, bias-variance decomposition

This section explains how the results for SLS models can be extended to the penalized maximum likelihood approach.

B.1 Penalization bias

A common approach for improving the performance of MLE is based on regularization or penalization. The objective function $L(\boldsymbol{v})$ is extended by including a penalty term $\operatorname{pen}_G(\boldsymbol{v})$ which is responsible for complexity (roughness) of the parameter $\boldsymbol{v}$. A typical example to keep in mind is $\operatorname{pen}_G(\boldsymbol{v}) = \|G\boldsymbol{v}\|^2/2$ for a penalization matrix G^2. Penalization by $\operatorname{pen}_G(\boldsymbol{v})$ can gradially improve stability and numerical properties of the estimator, however, it leads to a change of the "truth" $\boldsymbol{v}^*$, and hence, to some bias. This section describes the bias caused by a smooth penalty. Define the penalized MLE $\widetilde{\boldsymbol{v}}_G$

$$
\widetilde{\boldsymbol{v}}_G \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{v}} L_G(\boldsymbol{v}) = \operatorname*{argmax}_{\boldsymbol{v}} \{L(\boldsymbol{v}) - \operatorname{pen}_G(\boldsymbol{v})\} .
$$

Compared to (A.1), consider three optimization problems

$$
\widetilde{\boldsymbol{v}}_G = \operatorname*{argmax}_{\boldsymbol{v}} L_G(\boldsymbol{v}), \qquad \boldsymbol{v}_G^* = \operatorname*{argmax}_{\boldsymbol{v}} \mathbb{E} L_G(\boldsymbol{v}), \qquad \boldsymbol{v}^* = \operatorname*{argmax}_{\boldsymbol{v}} \mathbb{E} L(\boldsymbol{v}).
$$

Due to Proposition A.1, the penalized MLE $\widetilde{\boldsymbol{v}}_G$ estimates rather $\boldsymbol{v}_G^*$ than $\boldsymbol{v}^*$. This section describes the bias $\boldsymbol{v}_G^* - \boldsymbol{v}^*$ caused by penalization.

Define the penalized Fisher information $\mathbb{F}_G(\boldsymbol{v}) = -\nabla^2 \mathbb{E} L_G(\boldsymbol{v})$ and introduce $\boldsymbol{M}_G(\boldsymbol{v}) \stackrel{\text{def}}{=} \nabla \operatorname{pen}_G(\boldsymbol{v})$. Set $\mathbb{F}_G = \mathbb{F}_G(\boldsymbol{v}_G^*)$,

$$
\mathbb{F}_G = \mathbb{F}_G(\boldsymbol{v}_G^*), \qquad \boldsymbol{M}_G \stackrel{\text{def}}{=} \nabla \operatorname{pen}_G(\boldsymbol{v}^*), \quad \mathsf{b}_D \stackrel{\text{def}}{=} \|D \mathbb{F}_G^{-1} \boldsymbol{M}_G\| . \tag{B.1}
$$

For a quadratic penalty $\operatorname{pen}_G(\boldsymbol{v}) = \|G\boldsymbol{v}\|^2/2$, this results in

$$
\boldsymbol{M}_G = G^2 \boldsymbol{v}^*, \qquad \mathsf{b}_D \stackrel{\text{def}}{=} \|D \mathbb{F}_G^{-1} G^2 \boldsymbol{v}^*\| .
$$

Proposition D.14 yields the following result.

Proposition B.1. *Let $f_G(\boldsymbol{v}) = \mathbb{E} L_G(\boldsymbol{v})$ satisfy $(\mathcal{T}_3^*)$ at $\boldsymbol{v}_G^*$ with some metric tensor D and values $\mathtt{r}$ and τ_3 such that*

$$
D^2 \le \varkappa^2 \, \mathbb{F}_G , \qquad \mathtt{r} \ge 3\mathsf{b}_D/2 , \qquad \tau_3 \, \varkappa^2 \, \mathsf{b}_D < 4/9,
$$

for b_D from (B.1). Then

$$
\|D^{-1} \mathbb{F}_G(\boldsymbol{v}_G^* - \boldsymbol{v}^* + \mathbb{F}_G^{-1} \boldsymbol{M}_G)\| \le \frac{3\tau_3}{4} \, \mathsf{b}_D^2 . \tag{B.2}
$$

The same bounds apply with $\mathbb{F}_G(\boldsymbol{v}^*)$ in place of $\mathbb{F}_G = \mathbb{F}_G(\boldsymbol{v}_G^*)$.

B.2 Loss and risk of the pMLE. Bias-variance decomposition

Now we combine the previous results about the stochastic term $\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*_G$ and the bias term $\boldsymbol{\upsilon}^*_G - \boldsymbol{\upsilon}^*$ to obtain sharp bounds on the loss and risk of the pMLE $\widetilde{\boldsymbol{\upsilon}}_G$.

Theorem B.2. *Assume* $(\boldsymbol{\zeta})$, $(\boldsymbol{\nabla\zeta})$, *and* $(\boldsymbol{\mathcal{C}})$. *Let* $f_G(\boldsymbol{\upsilon}) = \mathbb{E} L_G(\boldsymbol{\upsilon})$ *satisfy* $(\boldsymbol{\mathcal{T}^*_3})$ *at* $\boldsymbol{\upsilon}^*_G$ *with some* D, $\mathtt{r}$, *and* τ_3. *With* $(\mathtt{r}_D \vee \mathsf{b}_D) \stackrel{\text{def}}{=} \max\{\mathtt{r}_D, \mathsf{b}_D\}$, *assume*

$$D^2 \le \varkappa^2\, \mathbb{F}_G\,, \qquad \mathtt{r} \ge \frac{3}{2}(\mathtt{r}_D \vee \mathsf{b}_D)\,, \qquad \varkappa^2 \tau_3\,(\mathtt{r}_D \vee \mathsf{b}_D) < \frac{4}{9}\,;$$

see (A.6) *and* (B.1). *For any linear mapping* $Q \colon \mathbb{R}^p \to \mathbb{R}^q$, *it holds on* $\Omega(\mathtt{x})$

$$\|Q(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* - \mathbb{F}_G^{-1}\nabla\zeta + \mathbb{F}_G^{-1}\boldsymbol{M}_G)\| \le \|Q\mathbb{F}_G^{-1}D\|\,\frac{3\tau_3}{4}\,\bigl(\|D\mathbb{F}_G^{-1}\nabla\zeta\|^2 + \mathsf{b}_D^2\bigr)\,. \tag{B.3}$$

Also, introduce $\mathbb{p}_Q \stackrel{\text{def}}{=} \mathbb{E}\|Q\mathbb{F}_G^{-1}\nabla\zeta\|^2 = \operatorname{tr}\operatorname{Var}(Q\mathbb{F}_G^{-1}\nabla\zeta)$ *and*

$$\mathscr{R}_Q \stackrel{\text{def}}{=} \mathbb{E}\{\|Q\mathbb{F}_G^{-1}(\nabla\zeta - \boldsymbol{M}_G)\|^2\, \mathbb{1}_{\Omega(\mathtt{x})}\} \le \mathbb{p}_Q + \|Q\mathbb{F}_G^{-1}\boldsymbol{M}_G\|^2\,.$$

Then

$$\mathbb{E}\bigl\{\|Q(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*)\|\, \mathbb{1}_{\Omega(\mathtt{x})}\bigr\} \le \mathscr{R}_Q^{1/2} + \|Q\mathbb{F}_G^{-1}D\|\,\frac{3\tau_3}{4}\bigl(\mathbb{p}_D + \mathsf{b}_D^2\bigr)\,. \tag{B.4}$$

Further, assume $\mathbb{E}\bigl\{\|D\mathbb{F}_G^{-1}\nabla\zeta\|^4\, \mathbb{1}_{\Omega(\mathtt{x})}\bigr\} \le \mathsf{C}_4^2\, \mathbb{p}_D^2$ *and define*

$$\alpha_Q \stackrel{\text{def}}{=} \frac{\|Q\mathbb{F}_G^{-1}D\|\,(3/4)\tau_3\,(\mathsf{C}_4\,\mathbb{p}_D + \mathsf{b}_D^2)}{\sqrt{\mathscr{R}_Q}}\,. \tag{B.5}$$

If $\alpha_Q < 1$ *then*

$$(1-\alpha_Q)^2\mathscr{R}_Q \le \mathbb{E}\bigl\{\|Q\,(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*)\|^2\, \mathbb{1}_{\Omega(\mathtt{x})}\bigr\} \le (1+\alpha_Q)^2\mathscr{R}_Q\,. \tag{B.6}$$

Proof. It holds by (A.8) and (B.2)

$$\begin{aligned} \|Q(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*_G - \mathbb{F}_G^{-1}\nabla\zeta\| &\le \|Q\mathbb{F}_G^{-1}D\|\,\frac{3\tau_3}{4}\,\|D\mathbb{F}_G^{-1}\nabla\zeta\|^2\,, \\ \|Q(\boldsymbol{\upsilon}^*_G - \boldsymbol{\upsilon}^* + \mathbb{F}_G^{-1}\boldsymbol{M}_G)\| &\le \|Q\mathbb{F}_G^{-1}D\|\,\frac{3\tau_3}{4}\,\mathsf{b}_D^2\,, \end{aligned} \tag{B.7}$$

and hence,

$$\|Q(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* - \mathbb{F}_G^{-1}\nabla\zeta + \mathbb{F}_G^{-1}\boldsymbol{M}_G)\| \le \|Q\mathbb{F}_G^{-1}D\|\,\frac{3\tau_3}{4}\,\bigl(\|D\mathbb{F}_G^{-1}\nabla\zeta\|^2 + \mathsf{b}_D^2\bigr)$$

yielding (B.3) and (B.4). Further, define

$$\boldsymbol{\varepsilon}_G \stackrel{\text{def}}{=} Q\bigl\{\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* - \mathbb{F}_G^{-1}(\nabla\zeta - \boldsymbol{M}_G)\bigr\}\,.$$

It holds by (B.7)

$$\mathbb{E}^{1/2}\bigl\{\|\varepsilon_Q\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}\bigr\} \le \|Q\mathbb{F}_G^{-1}D\|\frac{3\tau_3}{4}\,\bigl\{\mathbb{E}^{1/2}\bigl(\|D\mathbb{F}_G^{-1}\nabla\zeta\|^4\,\mathbb{1}_{\Omega(\mathtt{x})}\bigr)+\mathsf{b}_D^2\bigr\} \le \alpha_Q\,\mathscr{R}_Q^{1/2},$$

and therefore,

$$\begin{aligned}\mathbb{E}^{1/2}\bigl\{\|Q\,(\widetilde{\boldsymbol{\upsilon}}_G-\boldsymbol{\upsilon}^*)\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}\bigr\} &= \mathbb{E}^{1/2}\bigl\{\|Q\mathbb{F}_G^{-1}(\nabla\zeta-\boldsymbol{M}_G)+\varepsilon_Q\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}\bigr\}\\ &\le \mathbb{E}^{1/2}\bigl\{\|Q\mathbb{F}_G^{-1}(\nabla\zeta-\boldsymbol{M}_G)\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}\bigr\}+\mathbb{E}^{1/2}\bigl\{\|\varepsilon_Q\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}\bigr\} \le (1+\alpha_Q)\,\mathscr{R}_Q^{1/2}.\end{aligned}$$

This yields (B.6). □

Remark B.1. The condition $D^2 \le \varkappa^2\mathbb{F}_G$ implies $\|Q\mathbb{F}_G^{-1}D\| \le \varkappa^2\|QD^{-1}\|$ which can be used in the remainder for all risk bounds.

Remark B.2. Due to (B.6)

$$\mathbb{E}\bigl\{\|Q\,(\widetilde{\boldsymbol{\upsilon}}_G-\boldsymbol{\upsilon}^*)\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}\bigr\} = \bigl(\mathbb{p}_Q+\|Q\mathbb{F}_G^{-1}\boldsymbol{M}_G\|^2\bigr)\,\bigl\{1+o(1)\bigr\}. \tag{B.8}$$

This relation is usually referred to as "bias-variance decomposition". Our bound is sharp in the sense that for the special case of linear models, (B.8) becomes equality. Under the so-called "small bias" condition $\|Q\mathbb{F}_G^{-1}\boldsymbol{M}_G\|^2 \ll \mathbb{p}_Q$, the impact of the bias induced by penalization is negligible. The relation $\|Q\mathbb{F}_G^{-1}\boldsymbol{M}_G\|^2 \asymp \mathbb{p}_Q$ is called "bias-variance trade-off", it leads to minimax rate of estimation; see Section **??**.

If the constant α_Q from (B.5) satisfies $\alpha_Q \ll 1$ then by (B.6), $\mathbb{E}\bigl\{\|Q\,(\widetilde{\boldsymbol{\upsilon}}_G-\boldsymbol{\upsilon}^*)\|^2\,\mathbb{1}_{\Omega(\mathtt{x})}\bigr\} = (1+o(1))\mathscr{R}_Q$.

C Anisotropic logistic regression

This section considers a popular logistic regression model. It is widely used e.g. in binary classification in machine learning for binary classification or in binary response models in econometrics. The results presented here can be viewed as an extension of Spokoiny and Panov (2021) and Spokoiny (2023a).

Suppose we are given a vector of independent observations/labels $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^\top$ and a set of the corresponding feature vectors $\boldsymbol{\Psi}_i \in \mathbb{R}^p$. Each binary label Y_i is modelled as a Bernoulli random variable with the parameter $\theta_i^* = \mathbb{P}(Y_i = 1)$. Logistic regression reduces this model to linear regression for the canonical parameter $\upsilon_i^* = \log\frac{\theta_i^*}{1-\theta_i^*}$ in the form $\upsilon_i^* = \langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon}^* \rangle$, where $\boldsymbol{\upsilon}$ is the parameter vector in $\mathbb{R}^p$. The corresponding log-likelihood reads

$$L(\boldsymbol{\upsilon}) = \sum_{i=1}^{n} \Big\{ Y_i \langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle - \phi\big(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle \big) \Big\} \tag{C.1}$$

with $\phi(\upsilon) = \log\big(1 + \mathrm{e}^{\upsilon}\big)$. This function $\phi(\cdot)$ is convex with $\phi''(\upsilon) = \frac{\mathrm{e}^{\upsilon}}{(1+\mathrm{e}^{\upsilon})^2}$. We also assume that the $\boldsymbol{\Psi}_i$'s are deterministic, otherwise, we condition on the design. A penalized MLE $\widetilde{\boldsymbol{\upsilon}}_G$ is defined by maximization of the penalized log-likelihood $L_G(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \|G\boldsymbol{\upsilon}\|^2/2$ for the quadratic penalty $\|G\boldsymbol{\upsilon}\|^2/2$:

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} L_G(\boldsymbol{\upsilon}).$$

The truth and the penalized truth are defined via the expected log-likelihood

$$\boldsymbol{\upsilon}^* = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} \mathbb{E} L(\boldsymbol{\upsilon}), \quad \boldsymbol{\upsilon}_G^* = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} \mathbb{E} L_G(\boldsymbol{\upsilon}).$$

The Fisher matrix $\mathbb{F}(\boldsymbol{\upsilon})$ at $\boldsymbol{\upsilon}$ is given by

$$\mathbb{F}(\boldsymbol{\upsilon}) = -\nabla^2 \mathbb{E} L(\boldsymbol{\upsilon}) = \sum_{i=1}^{n} w_i(\boldsymbol{\upsilon}) \, \boldsymbol{\Psi}_i \boldsymbol{\Psi}_i^{\top}, \quad w_i(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \phi''\big(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle \big) . \tag{C.2}$$

We also write

$$\mathbb{F}_G(\boldsymbol{\upsilon}) = \mathbb{F}(\boldsymbol{\upsilon}) + G^2, \qquad \mathbb{F}_G = \mathbb{F}_G(\boldsymbol{\upsilon}_G^*) = \mathbb{F}(\boldsymbol{\upsilon}_G^*) + G^2.$$

Alternatively one can define $\mathbb{F}_G = \mathbb{F}_G(\boldsymbol{\upsilon}^*)$. Later we need a metric tensor D which defines a local vicinity of $\boldsymbol{\upsilon}^*$. We assume $\mathbb{F} \le D^2 \le \mathbb{F}_G$. If $\mathbb{F}$ is well posed, then $D^2 = \mathbb{F}$ is a default choice.

Further, we discuss the stochastic component of $L(\boldsymbol{\upsilon})$

$$\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E} L(\boldsymbol{\upsilon}) = \sum_{i=1}^{n} \big(Y_i - \mathbb{E} Y_i \big) \langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle$$

It is obviously linear in $\boldsymbol{\upsilon}$ with

$$\nabla \zeta = \sum_{i=1}^{n} \big(Y_i - \mathbb{E} Y_i \big) \boldsymbol{\Psi}_i . \tag{C.3}$$

Further, with $Y_i \sim \mathrm{Be}(\theta_i^*)$

$$\operatorname{Var}(\nabla \zeta) = \sum_{i=1}^{n} \operatorname{Var}(Y_i) \, \boldsymbol{\Psi}_i \boldsymbol{\Psi}_i^{\top} = \sum_{i=1}^{n} \theta_i^* (1 - \theta_i^*) \, \boldsymbol{\Psi}_i \boldsymbol{\Psi}_i^{\top} . \tag{C.4}$$

If model (C.1) is correctly specified, that is, $\theta_i^* = \mathrm{e}^{\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon}^* \rangle}/(1+\mathrm{e}^{\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon}^* \rangle})$, then $\operatorname{Var}(\nabla\zeta) = I\!\!F(\boldsymbol{\upsilon}^*)$; see (C.2). Instead, we assume $\operatorname{Var}(\nabla\zeta) \le D^2$. This can be relaxed to $\operatorname{Var}(\nabla\zeta) \le \mathtt{C} D^2$ for some fixed constant $\mathtt{C}$. Now we check the general conditions from Section A. Convexity of $\phi(\cdot)$ yields concavity of $L(\boldsymbol{\upsilon})$ and thus, $(\boldsymbol{\mathcal{C}})$. Condition $(\boldsymbol{\zeta})$ is granted by (C.3). For checking the other conditions, we need some regularity of the design $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_n$. Let $\boldsymbol{\upsilon}^\circ$ be either $\boldsymbol{\upsilon}^*$, or $\boldsymbol{\upsilon}_G^*$. Define an elliptic vicinity $\varUpsilon^\circ$ of $\boldsymbol{\upsilon}^\circ$ as

$$\varUpsilon^\circ = \{\boldsymbol{\upsilon} \colon \|D\,(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)\| \le \mathtt{r}\}. \tag{C.5}$$

$(\boldsymbol{\Psi}^\circ)$ *For some* $\delta_0 > 0$

$$\max_{i \le n} \|D^{-1}\, \boldsymbol{\Psi}_i\| \le \delta_0 . \tag{C.6}$$

It appears that the condition that δ_0 is sufficiently small ensures applicability of all the results from Section A. We also assume for some $\delta \ge 0$

$$\sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle^4\, w_i(\boldsymbol{\upsilon}) \le \delta^2\, \|D\boldsymbol{z}\|^4, \qquad \boldsymbol{z} \in I\!\!R^p . \tag{C.7}$$

Later we show that this condition follows from $(\boldsymbol{\Psi}^\circ)$ for $\delta^2 \le \sqrt{\mathrm{e}}\,\delta_0^2$. However, this is a conservative upper bound, condition (C.7) may apply with much smaller values of δ. As the final accuracy bound depends on δ rather then on δ_0, we keep (C.7) as a separate condition.

The penalty $\|G\boldsymbol{\upsilon}\|^2/2$ in the pMLE $\widetilde{\boldsymbol{\upsilon}}_G$ results in some bias. It can be measured by

$$\mathsf{b}_D \stackrel{\mathrm{def}}{=} \|D I\!\!F_G^{-1} G^2 \boldsymbol{\upsilon}^*\| .$$

Proposition C.1. *Consider the logistic regression model* (C.1) *and let* $\operatorname{Var}(\nabla\zeta) \le D^2$*; see* (C.2) *and* (C.4)*. With* $B = 2(D\, I\!\!F_G^{-1} D)^2$*, define* $\mathtt{r}_D = z(B, \mathtt{x}) \le \sqrt{\mathbb{p}_D} + \sqrt{2\mathtt{x}\,\lambda_D}$ *for* $\mathbb{p}_D = \operatorname{tr}(B)$*,* $\lambda_D = \|B\|$*. Assume* $(\boldsymbol{\Psi}^\circ)$ *with* δ_0 *satisfying*

$$(\mathtt{r}_D \vee \mathsf{b}_D)\,\delta_0 \le 1/3, \qquad 1/(2\delta_0) \ge \sqrt{2\mathtt{x}} + \sqrt{\mathbb{p}_D/\lambda_D} . \tag{C.8}$$

Then (C.7) *holds for* $\delta^2 \le \sqrt{\mathrm{e}}\,\delta_0^2$ *all the results of Theorem A.3 through Theorem B.2 apply with* $\tau_3 \le \sqrt{\mathrm{e}}\,\delta \le \mathrm{e}^{3/4}\delta_0$.

Proof. We systematically check the conditions of Theorem A.3 through Theorem B.2. Conditions $(\boldsymbol{\zeta})$ and $(\boldsymbol{\mathcal{C}})$ are granted by (C.1) with ϕ convex. Next, we check $(\boldsymbol{\nabla\zeta})$ for a matrix V^2 such that $V^2 \ge 2\operatorname{Var}(\nabla\zeta)$ and $V^2 \ge D^2$.

Lemma C.2. *Let* $V^2 \geq 2\,\mathrm{Var}(\nabla\zeta)$ *and* $V^2 \geq D^2$. *For* $B \in \mathfrak{M}_p$, *define*

$$\mathbb{p}_B = \mathrm{tr}\bigl(B\bigr), \qquad \lambda_B = \|B\|, \qquad \mathtt{r}_B = z(B,\mathtt{x}) \leq \sqrt{\mathbb{p}_B} + \sqrt{2\mathtt{x}\lambda_B}\,.$$

With δ_0 *from* (C.6), *let*

$$1/(2\delta_0) > \sqrt{\mathbb{p}_B/\lambda_B}\,.$$

Then $(\boldsymbol{\nabla\zeta})$ *is fulfilled with for this* B *and all* $\mathtt{x}$ *such that* $\sqrt{2\mathtt{x}} \leq 1/(2\delta_0) - \sqrt{\mathbb{p}_B/\lambda_B}$.

Proof. Let $\mathtt{g} = \log(2)/\delta_0$ and $\mathtt{x}_c$ be given by (**??**). By Theorem E.2,

$$\mathbb{P}\bigl(\|B^{1/2}V^{-1}\nabla\zeta\| \geq z(B,\mathtt{x})\bigr) \leq 3\mathrm{e}^{-\mathtt{x}}$$

for all $\mathtt{x} \leq \mathtt{x}_c$, and by (**??**), $\sqrt{\mathtt{x}_c} \geq \mathtt{g}/2 - \sqrt{\mathbb{p}_B/(2\lambda_B)}$. This yields the assertion. □

Lemma C.3. *Assume* $(\boldsymbol{\varPsi}^{\circ})$. *Let* $\varUpsilon^{\circ}$ *be from* (C.5) *with* $\mathtt{r}$ *satisfying* $\delta_0\,\mathtt{r} \leq 1/2$. *Then for any* $\boldsymbol{\upsilon} \in \varUpsilon^{\circ}$

$$\phi''\bigl(\langle\boldsymbol{\varPsi}_i,\boldsymbol{\upsilon}\rangle\bigr) \leq \sqrt{\mathrm{e}}\;\phi''\bigl(\langle\boldsymbol{\varPsi}_i,\boldsymbol{\upsilon}^{\circ}\rangle\bigr), \qquad i = 1,\ldots,n, \tag{C.9}$$

$$\frac{1}{\sqrt{\mathrm{e}}}\,\mathbb{F}(\boldsymbol{\upsilon}^{\circ}) \leq \mathbb{F}(\boldsymbol{\upsilon}) \leq \sqrt{\mathrm{e}}\,\mathbb{F}(\boldsymbol{\upsilon}^{\circ}). \tag{C.10}$$

Also, (C.7) *holds with* $\delta^2 \leq \sqrt{\mathrm{e}}\,\delta_0^2$.

Proof. The function $\phi(\upsilon) = \log(1+\mathrm{e}^{\upsilon})$ satisfies for all $\upsilon \in \mathbb{R}$

$$|\phi^{(k)}(\upsilon)| \leq \phi''(\upsilon), \qquad k = 3,4. \tag{C.11}$$

Indeed, it holds

$$\begin{aligned}
\phi'(\upsilon) &= \frac{\mathrm{e}^{\upsilon}}{1+\mathrm{e}^{\upsilon}}\,,\\
\phi''(\upsilon) &= \frac{\mathrm{e}^{\upsilon}}{(1+\mathrm{e}^{\upsilon})^2}\,,\\
\phi^{(3)}(\upsilon) &= \frac{\mathrm{e}^{\upsilon}}{(1+\mathrm{e}^{\upsilon})^2} - \frac{2\mathrm{e}^{2\upsilon}}{(1+\mathrm{e}^{\upsilon})^3}\,,\\
\phi^{(4)}(\upsilon) &= \frac{\mathrm{e}^{\upsilon}}{(1+\mathrm{e}^{\upsilon})^2} - \frac{6\mathrm{e}^{2\upsilon}}{(1+\mathrm{e}^{\upsilon})^3} + \frac{6\mathrm{e}^{3\upsilon}}{(1+\mathrm{e}^{\upsilon})^4}\,.
\end{aligned}$$

It is straightforward to see that $|\phi^{(k)}(\upsilon)| \leq \phi''(\upsilon)$ for $k = 3,4$ and any υ.

Next, we check local variability of $\phi''(\upsilon)$. Fix $\upsilon^{\circ} < 0$. As the function $\phi''(\upsilon)$ is monotonously increasing in υ, it holds

$$\sup_{|\upsilon-\upsilon^{\circ}|\le b} \frac{\phi''(\upsilon)}{\phi''(\upsilon^{\circ})} = \frac{\phi''(\upsilon^{\circ}+b)}{\phi''(\upsilon^{\circ})} \le \mathrm{e}^{b}. \tag{C.12}$$

Putting together (C.11) and (C.12) leads to a bound on variability of $I\!\!F(\boldsymbol{\upsilon})$ for $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}^{\circ} + \boldsymbol{u}$ and $\|D\boldsymbol{u}\| \le \mathtt{r}$. By definition,

$$I\!\!F(\boldsymbol{\upsilon}) = \sum_{i=1}^{n} \boldsymbol{\Psi}_i \boldsymbol{\Psi}_i^{\top} \phi''\bigl(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle\bigr).$$

Now (C.6) and $\delta_0 \mathtt{r} \le 1/2$ imply $\bigl|\langle \boldsymbol{\Psi}_i, \boldsymbol{u} \rangle\bigr| \le \|D^{-1}\boldsymbol{\Psi}_i\| \, \|D\boldsymbol{u}\| \le \delta_0 \mathtt{r} \le 1/2$ for each $i \le n$, and (C.9), (C.10) follow by (C.12).

By definition $\|I\!\!F^{1/2}(\boldsymbol{\upsilon})\boldsymbol{z}\|^2 = \sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle^2 w_i(\boldsymbol{\upsilon})$ and the use of (C.6) yields

$$\begin{aligned}
\sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle^4 w_i(\boldsymbol{\upsilon}) &= \sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle^2 \, \langle D^{-1}\boldsymbol{\Psi}_i, D\boldsymbol{z} \rangle^2 \, w_i(\boldsymbol{\upsilon}) \\
&\le \delta_0^2 \|D\boldsymbol{z}\|^2 \sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle^2 \, w_i(\boldsymbol{\upsilon}) = \delta_0^2 \|D\boldsymbol{z}\|^2 \, \|I\!\!F^{1/2}(\boldsymbol{\upsilon})\boldsymbol{z}\|^2.
\end{aligned}$$

Therefore, $D^2 = I\!\!F(\boldsymbol{\upsilon}^{\circ}) \ge \mathrm{e}^{-1/2} I\!\!F(\boldsymbol{\upsilon})$ implies $\delta^2 \le \sqrt{\mathrm{e}}\, \delta_0^2$. □

Further we check $(\mathcal{T}_3^*)$, $(\mathcal{T}_4^*)$ at $\boldsymbol{\upsilon}^*$ with $\mathtt{r} = 3\mathtt{r}_D/2$, $\tau_3 = \sqrt{\mathrm{e}}\, \delta$, and $\tau_4 = \sqrt{\mathrm{e}}\, \delta^2$.

Lemma C.4. *Assume* $(\boldsymbol{\Psi}^{\circ})$. *Let* $\mathtt{r}$ *satisfy* $\delta_0 \mathtt{r} \le 1/2$. *Then* $(\mathcal{T}_3^*)$ *and* $(\mathcal{T}_4^*)$ *hold with* $\tau_3 = \sqrt{\mathrm{e}}\, \delta$ *and* $\tau_4 = \sqrt{\mathrm{e}}\, \delta^2$ *for* δ *from* (C.7).

Proof. First we evaluate the derivative $\nabla^k f(\boldsymbol{\upsilon})$ for $f(\boldsymbol{\upsilon}) = I\!\!E L(\boldsymbol{\upsilon})$ over $\boldsymbol{\upsilon} \in \Upsilon^{\circ}$. For any $\boldsymbol{z} \in I\!\!R^p$, it holds

$$\bigl\langle \nabla^k f(\boldsymbol{\upsilon}), \boldsymbol{z}^{\otimes k} \bigr\rangle = - \sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle^k \, \phi^{(k)}\bigl(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle\bigr).$$

With $w_i(\boldsymbol{\upsilon}) = \phi''\bigl(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle\bigr)$, we derive by (C.6), (C.7), (C.11), and (C.10)

$$\begin{aligned}
&\bigl|\bigl\langle \nabla^3 f(\boldsymbol{\upsilon}), \boldsymbol{z}^{\otimes 3} \bigr\rangle\bigr| \le \sum_{i=1}^{n} \bigl|\langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle\bigr|^3 \phi''\bigl(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle\bigr) \le \sqrt{\mathrm{e}} \sum_{i=1}^{n} \bigl|\langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle\bigr|^3 w_i(\boldsymbol{\upsilon}^{\circ}) \\
&\quad \le \sqrt{\mathrm{e}} \biggl(\sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle^2 w_i(\boldsymbol{\upsilon}^{\circ}) \biggr)^{1/2} \biggl(\sum_{i=1}^{n} \langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle^4 w_i(\boldsymbol{\upsilon}^{\circ}) \biggr)^{1/2} \le \sqrt{\mathrm{e}}\, \delta \, \|I\!\!F^{1/2}(\boldsymbol{\upsilon}^{\circ})\boldsymbol{z}\| \, \|D\boldsymbol{z}\|^2
\end{aligned}$$

and $(\mathcal{T}_3^*)$ follows with $\tau_3 = \sqrt{\mathrm{e}}\, \delta$ by $I\!\!F(\boldsymbol{\upsilon}^{\circ}) = D^2$. Similarly $(\mathcal{T}_4^*)$ holds with $\tau_4 = \sqrt{\mathrm{e}}\, \delta^2$. □

All the required conditions from Section A have been checked and the results about the behavior of the pMLE $\widetilde{\boldsymbol{v}}_G$ apply. □

D Local smoothness and linearly perturbed optimization

This section discusses the problem of linearly and quadratically perturbed optimization of a smooth and concave function $f(\boldsymbol{v})$, $\boldsymbol{v} \in I\!\!R^p$.

D.1 Gateaux smoothness and self-concordance

Below we assume the function $f(\boldsymbol{v})$ to be strongly concave with the negative Hessian $\mathbb{F}(\boldsymbol{v}) \stackrel{\text{def}}{=} -\nabla^2 f(\boldsymbol{v}) \in \mathfrak{M}_p$ positive definite. Also, assume $f(\boldsymbol{v})$ three or sometimes even four times Gateaux differentiable in $\boldsymbol{v} \in \varUpsilon$. For any particular direction $\boldsymbol{u} \in I\!\!R^p$, we consider the univariate function $f(\boldsymbol{v} + t\boldsymbol{u})$ and measure its smoothness in t. Local smoothness of f will be described by the relative error of the Taylor expansion of the third or fourth order. Namely, define

$$\begin{aligned} \delta_3(\boldsymbol{v},\boldsymbol{u}) &= f(\boldsymbol{v}+\boldsymbol{u}) - f(\boldsymbol{v}) - \langle \nabla f(\boldsymbol{v}), \boldsymbol{u} \rangle - \frac{1}{2} \langle \nabla^2 f(\boldsymbol{v}), \boldsymbol{u}^{\otimes 2} \rangle, \\ \delta_3'(\boldsymbol{v},\boldsymbol{u}) &= \langle \nabla f(\boldsymbol{v}+\boldsymbol{u}), \boldsymbol{u} \rangle - \langle \nabla f(\boldsymbol{v}), \boldsymbol{u} \rangle - \langle \nabla^2 f(\boldsymbol{v}), \boldsymbol{u}^{\otimes 2} \rangle , \end{aligned}$$

and

$$\delta_4(\boldsymbol{v},\boldsymbol{u}) \stackrel{\text{def}}{=} f(\boldsymbol{v}+\boldsymbol{u}) - f(\boldsymbol{v}) - \langle \nabla f(\boldsymbol{v}), \boldsymbol{u} \rangle - \frac{1}{2} \langle \nabla^2 f(\boldsymbol{v}), \boldsymbol{u}^{\otimes 2} \rangle - \frac{1}{6} \langle \nabla^3 f(\boldsymbol{v}), \boldsymbol{u}^{\otimes 3} \rangle .$$

Now, for each $\boldsymbol{v}$, suppose to be given a positive symmetric operator $\mathbb{D}(\boldsymbol{v}) \in \mathfrak{M}_p$ defining a local metric and a local vicinity around $\boldsymbol{v}$:

$$\mathcal{U}_{\mathtt{r}}(\boldsymbol{v}) = \bigl\{ \boldsymbol{u} \in I\!\!R^p \colon \|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\| \le \mathtt{r} \bigr\}$$

for some radius $\mathtt{r}$.

Local smoothness properties of f at $\boldsymbol{v}$ are given via the quantities

$$\omega(\boldsymbol{v}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{u} \colon \|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\| \le \mathtt{r}} \frac{2|\delta_3(\boldsymbol{v},\boldsymbol{u})|}{\|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\|^2} , \quad \omega'(\boldsymbol{v}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{u} \colon \|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\| \le \mathtt{r}} \frac{|\delta_3'(\boldsymbol{v},\boldsymbol{u})|}{\|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\|^2} . \tag{D.1}$$

The definition yields for any $\boldsymbol{u}$ with $\|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\| \le \mathtt{r}$

$$\bigl| \delta_3(\boldsymbol{v},\boldsymbol{u}) \rangle \bigr| \le \frac{\omega(\boldsymbol{v})}{2} \|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\|^2 , \qquad \bigl| \delta_3'(\boldsymbol{v},\boldsymbol{u}) \bigr| \le \omega'(\boldsymbol{v}) \|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\|^2 . \tag{D.2}$$

The approximation results can be improved provided a third order upper bound on the error of Taylor expansion.

$(\mathcal{T}_3)$ *For some* τ_3

$$\bigl|\delta_3(\boldsymbol{v},\boldsymbol{u})\bigr| \leq \frac{\tau_3}{6}\|\mathbb{D}(\boldsymbol{v})\,\boldsymbol{u}\|^3\,, \quad \bigl|\delta_3'(\boldsymbol{v},\boldsymbol{u})\bigr| \leq \frac{\tau_3}{2}\|\mathbb{D}(\boldsymbol{v})\,\boldsymbol{u}\|^3\,, \quad \boldsymbol{u} \in \mathcal{U}_{\mathtt{r}}(\boldsymbol{v}).$$

$(\mathcal{T}_4)$ *For some* τ_4

$$\bigl|\delta_4(\boldsymbol{v},\boldsymbol{u})\bigr| \leq \frac{\tau_4}{24}\|\mathbb{D}(\boldsymbol{v})\,\boldsymbol{u}\|^4\,, \qquad \boldsymbol{u} \in \mathcal{U}_{\mathtt{r}}(\boldsymbol{v}).$$

We also present a version of $(\mathcal{T}_3)$ resp. $(\mathcal{T}_4)$ in terms of the third (resp. fourth) derivative of f.

$(\mathcal{T}_3^*)$ $f(\boldsymbol{v})$ *is three times differentiable and*

$$\sup_{\boldsymbol{u}\colon \|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\|\leq \mathtt{r}} \sup_{\boldsymbol{z}\in \mathbb{R}^p} \frac{\bigl|\langle \nabla^3 f(\boldsymbol{v}+\boldsymbol{u}), \boldsymbol{z}^{\otimes 3}\rangle\bigr|}{\|\mathbb{D}(\boldsymbol{v})\boldsymbol{z}\|^3} \leq \tau_3\,.$$

$(\mathcal{T}_4^*)$ $f(\boldsymbol{v})$ *is four times differentiable and*

$$\sup_{\boldsymbol{u}\colon \|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\|\leq \mathtt{r}} \sup_{\boldsymbol{z}\in \mathbb{R}^p} \frac{\bigl|\langle \nabla^4 f(\boldsymbol{v}+\boldsymbol{u}), \boldsymbol{z}^{\otimes 4}\rangle\bigr|}{\|\mathbb{D}(\boldsymbol{v})\boldsymbol{z}\|^4} \leq \tau_4\,.$$

By Banach's characterization Banach (1938), $(\mathcal{T}_3^*)$ implies

$$\bigl|\langle \nabla^3 f(\boldsymbol{v}+\boldsymbol{u}), \boldsymbol{z}_1\otimes\boldsymbol{z}_2\otimes\boldsymbol{z}_3\rangle\bigr| \leq \tau_3\|\mathbb{D}(\boldsymbol{v})\boldsymbol{z}_1\|\,\|\mathbb{D}(\boldsymbol{v})\boldsymbol{z}_2\|\,\|\mathbb{D}(\boldsymbol{v})\boldsymbol{z}_3\| \tag{D.3}$$

for any $\boldsymbol{u}$ with $\|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\| \leq \mathtt{r}$ and all $\boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_3 \in \mathbb{R}^p$. Similarly under $(\mathcal{T}_4^*)$

$$\bigl|\langle \nabla^4 f(\boldsymbol{v}+\boldsymbol{u}), \boldsymbol{z}_1\otimes\boldsymbol{z}_2\otimes\boldsymbol{z}_3\otimes\boldsymbol{z}_4\rangle\bigr| \leq \tau_4\prod_{k=1}^{4}\|\mathbb{D}(\boldsymbol{v})\boldsymbol{z}_k\|\,, \quad \boldsymbol{z}_1,\boldsymbol{z}_2,\boldsymbol{z}_3,\boldsymbol{z}_4 \in \mathbb{R}^p\,. \tag{D.4}$$

Lemma D.1. *Under* $(\mathcal{T}_3)$ *or* $(\mathcal{T}_3^*)$, *the values* $\omega(\boldsymbol{v})$ *and* $\omega'(\boldsymbol{v})$ *from* (D.1) *satisfy*

$$\omega(\boldsymbol{v}) \leq \frac{\tau_3\,\mathtt{r}}{3}\,, \qquad \omega'(\boldsymbol{v}) \leq \frac{\tau_3\,\mathtt{r}}{2}\,, \qquad \boldsymbol{v} \in \varUpsilon^{\circ}.$$

Proof. For any $\boldsymbol{u} \in \mathcal{U}_{\mathtt{r}}(\boldsymbol{v})$ with $\|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\| \leq \mathtt{r}$

$$\bigl|\delta_3(\boldsymbol{v},\boldsymbol{u})\bigr| \leq \frac{\tau_3}{6}\,\|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\|^3 \leq \frac{\tau_3\,\mathtt{r}}{6}\,\|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\|^2,$$

and the bound for $\omega(\boldsymbol{v})$ follows. The proof for $\omega'(\boldsymbol{v})$ is similar. □

The values τ_3 and τ_4 are usually very small. Some quantitative bounds are given later in this section under the assumption that the function $f(\boldsymbol{v})$ can be written in the form $-f(\boldsymbol{v}) = nh(\boldsymbol{v})$ for a fixed smooth function $h(\boldsymbol{v})$ with the Hessian $\nabla^2 h(\boldsymbol{v})$. The factor n has meaning of the sample size.

$(\mathcal{S}_3^*)$ $-f(\boldsymbol{v}) = nh(\boldsymbol{v})$ *for* $h(\boldsymbol{v})$ *three times differentiable and*

$$\sup_{\boldsymbol{u}\colon \|\mathbb{m}(\boldsymbol{v})\boldsymbol{u}\| \le \mathtt{r}/\sqrt{n}} \frac{\bigl|\langle \nabla^3 h(\boldsymbol{v}+\boldsymbol{u}), \boldsymbol{u}^{\otimes 3}\rangle\bigr|}{\|\mathbb{m}(\boldsymbol{v})\boldsymbol{u}\|^3} \le \mathsf{c}_3 .$$

$(\mathcal{S}_4^*)$ *the function* $h(\cdot)$ *satisfies* $(\mathcal{S}_3^*)$ *and*

$$\sup_{\boldsymbol{u}\colon \|\mathbb{m}(\boldsymbol{v})\boldsymbol{u}\| \le \mathtt{r}/\sqrt{n}} \frac{\bigl|\langle \nabla^4 h(\boldsymbol{v}+\boldsymbol{u}), \boldsymbol{u}^{\otimes 4}\rangle\bigr|}{\|\mathbb{m}(\boldsymbol{v})\boldsymbol{u}\|^4} \le \mathsf{c}_4 .$$

$(\mathcal{S}_3^*)$ and $(\mathcal{S}_4^*)$ are local versions of the so-called self-concordance condition; see Nesterov (1988) and Ostrovskii and Bach (2021). In fact, they require that each univariate function $h(\boldsymbol{v}+t\boldsymbol{u})$ of $t \in \mathbb{R}$ is self-concordant with some universal constants c_3 and c_4. Under $(\mathcal{S}_3^*)$ and $(\mathcal{S}_4^*)$, with $\mathbb{D}^2(\boldsymbol{v}) = n\,\mathbb{m}^2(\boldsymbol{v})$, the values $\delta_3(\boldsymbol{v},\boldsymbol{u})$, $\delta_4(\boldsymbol{v},\boldsymbol{u})$, and $\omega(\boldsymbol{v})$, $\omega'(\boldsymbol{v})$ can be bounded.

Lemma D.2. *Suppose* $(\mathcal{S}_3^*)$. *Then* $(\mathcal{T}_3)$ *follows with* $\tau_3 = \mathsf{c}_3 n^{-1/2}$. *Moreover, for* $\omega(\boldsymbol{v})$ *and* $\omega'(\boldsymbol{v})$ *from* (D.1), *it holds*

$$\omega(\boldsymbol{v}) \le \frac{\mathsf{c}_3\,\mathtt{r}}{3n^{1/2}}, \qquad \omega'(\boldsymbol{v}) \le \frac{\mathsf{c}_3\,\mathtt{r}}{2n^{1/2}}. \tag{D.5}$$

Also $(\mathcal{T}_4)$ *follows from* $(\mathcal{S}_4^*)$ *with* $\tau_4 = \mathsf{c}_4 n^{-1}$.

Proof. For any $\boldsymbol{u} \in \mathcal{U}_{\mathtt{r}}(\boldsymbol{v})$ and $t \in [0,1]$, by the Taylor expansion of the third order

$$\begin{aligned} |\delta(\boldsymbol{v},\boldsymbol{u})| &\le \frac{1}{6}\bigl|\langle \nabla^3 f(\boldsymbol{v}+t\boldsymbol{u}), \boldsymbol{u}^{\otimes 3}\rangle\bigr| = \frac{n}{6}\bigl|\langle \nabla^3 h(\boldsymbol{v}+t\boldsymbol{u}), \boldsymbol{u}^{\otimes 3}\rangle\bigr| \le \frac{n\,\mathsf{c}_3}{6}\|\mathbb{m}(\boldsymbol{v})\boldsymbol{u}\|^3 \\ &= \frac{n^{-1/2}\,\mathsf{c}_3}{6}\|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\|^3 \le \frac{n^{-1/2}\,\mathsf{c}_3\,\mathtt{r}}{6}\|\mathbb{D}(\boldsymbol{v})\boldsymbol{u}\|^2 . \end{aligned}$$

This implies $(\mathcal{T}_3)$ as well as (D.5); see (D.2). The statement about $(\mathcal{T}_4)$ is similar. □

D.2 Optimization after linear perturbation. A basic lemma

Let $f(\boldsymbol{v})$ be a smooth concave function,

$$\boldsymbol{v}^* = \operatorname*{argmax}_{\boldsymbol{v}} f(\boldsymbol{v}),$$

and $\mathbb{F} = -\nabla^2 f(\boldsymbol{v}^*)$. Later we study the question of how the point of maximum and the value of maximum of f change if we add a linear or quadratic component to f. More precisely, let another function $g(\boldsymbol{v})$ satisfy for some vector $\boldsymbol{A}$

$$g(\boldsymbol{v}) - g(\boldsymbol{v}^*) = \langle \boldsymbol{v}-\boldsymbol{v}^*, \boldsymbol{A}\rangle + f(\boldsymbol{v}) - f(\boldsymbol{v}^*). \tag{D.6}$$

A typical example corresponds to $f(\boldsymbol{v}) = \mathbb{E}L(\boldsymbol{v})$ and $g(\boldsymbol{v}) = L(\boldsymbol{v})$ for a random function $L(\boldsymbol{v})$ with a linear stochastic component $\zeta(\boldsymbol{v}) = L(\boldsymbol{v}) - \mathbb{E}L(\boldsymbol{v})$. Then (D.6) is satisfied with $\boldsymbol{A} = \nabla\zeta$. Define

$$\boldsymbol{v}^{\circ} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{v}} g(\boldsymbol{v}), \qquad g(\boldsymbol{v}^{\circ}) = \max_{\boldsymbol{v}} g(\boldsymbol{v}). \tag{D.7}$$

The aim of the analysis is to evaluate the quantities $\boldsymbol{v}^{\circ} - \boldsymbol{v}^{*}$ and $g(\boldsymbol{v}^{\circ}) - g(\boldsymbol{v}^{*})$. First, we consider the case of a quadratic function f.

Lemma D.3. *Let $f(\boldsymbol{v})$ be quadratic with $\nabla^2 f(\boldsymbol{v}) \equiv -\mathbb{F}$ and $g(\boldsymbol{v})$ satisfy (D.6). Then*

$$\boldsymbol{v}^{\circ} - \boldsymbol{v}^{*} = \mathbb{F}^{-1}\boldsymbol{A}, \qquad g(\boldsymbol{v}^{\circ}) - g(\boldsymbol{v}^{*}) = \frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2.$$

Proof. If $f(\boldsymbol{v})$ is quadratic, then, of course, under (D.6), $g(\boldsymbol{v})$ is quadratic as well with $-\nabla^2 g(\boldsymbol{v}) \equiv \mathbb{F}$. This implies

$$\nabla g(\boldsymbol{v}^{*}) - \nabla g(\boldsymbol{v}^{\circ}) = \mathbb{F}(\boldsymbol{v}^{\circ} - \boldsymbol{v}^{*}).$$

Further, (D.6) and $\nabla f(\boldsymbol{v}^{*}) = 0$ yield $\nabla g(\boldsymbol{v}^{*}) = \boldsymbol{A}$. Together with $\nabla g(\boldsymbol{v}^{\circ}) = 0$, this implies $\boldsymbol{v}^{\circ} - \boldsymbol{v}^{*} = \mathbb{F}^{-1}\boldsymbol{A}$. The Taylor expansion of g at $\boldsymbol{v}^{\circ}$ yields by $\nabla g(\boldsymbol{v}^{\circ}) = 0$

$$g(\boldsymbol{v}^{*}) - g(\boldsymbol{v}^{\circ}) = -\frac{1}{2}\|\mathbb{F}^{1/2}(\boldsymbol{v}^{\circ} - \boldsymbol{v}^{*})\|^2 = -\frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2$$

and the assertion follows. □

The next result describes the concentration properties of $\boldsymbol{v}^{\circ}$ from (D.7) in a local elliptic set of the form

$$\mathcal{A}(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{v}\colon \|\mathbb{F}^{1/2}(\boldsymbol{v} - \boldsymbol{v}^{*})\| \le \mathtt{r}\}, \tag{D.8}$$

where $\mathtt{r}$ is slightly larger than $\|\mathbb{F}^{-1/2}\boldsymbol{A}\|$.

Proposition D.4. *Let $f(\boldsymbol{v})$ be a strongly concave function with $f(\boldsymbol{v}^{*}) = \max_{\boldsymbol{v}} f(\boldsymbol{v})$ and $\mathbb{F} = -\nabla^2 f(\boldsymbol{v}^{*})$. Let further $g(\boldsymbol{v})$ and $f(\boldsymbol{v})$ be related by (D.6) with some vector $\boldsymbol{A}$. Fix $\nu < 1$ and $\mathtt{r}$ such that $\|\mathbb{F}^{-1/2}\boldsymbol{A}\| \le \nu\,\mathtt{r}$. Suppose now that $f(\boldsymbol{v})$ satisfy (D.1) for $\boldsymbol{v} = \boldsymbol{v}^{*}$, $\mathbb{D}(\boldsymbol{v}^{*}) = \mathbb{D} \le \varkappa\,\mathbb{F}^{1/2}$ with some $\varkappa > 0$ and ω' such that*

$$1 - \nu - \omega'\varkappa^2 > 0. \tag{D.9}$$

Then for $\boldsymbol{v}^{\circ}$ from (D.7), it holds

$$\|\mathbb{F}^{1/2}(\boldsymbol{v}^{\circ} - \boldsymbol{v}^{*})\| \le \mathtt{r} \quad \textit{and} \quad \|\mathbb{D}(\boldsymbol{v}^{\circ} - \boldsymbol{v}^{*})\| \le \varkappa\,\mathtt{r}\,.$$

Proof. Rescaling $\mathbb{D}$ by $\varkappa^{-1}$ reduces the proof to $\varkappa = 1$. The bound $\|\mathbb{F}^{-1/2}\boldsymbol{A}\| \leq \nu\,\mathtt{r}$ implies for any $\boldsymbol{u}$

$$\bigl|\langle \boldsymbol{A}, \boldsymbol{u}\rangle\bigr| = \bigl|\langle \mathbb{F}^{-1/2}\boldsymbol{A}, \mathbb{F}^{1/2}\boldsymbol{u}\rangle\bigr| \leq \nu\,\mathtt{r}\|\mathbb{F}^{1/2}\boldsymbol{u}\|\,.$$

Let $\boldsymbol{\upsilon}$ be a point on the boundary of the set $\mathcal{A}(\mathtt{r})$ from (D.8). We also write $\boldsymbol{u} = \boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*$. The idea is to show that the derivative $\frac{d}{dt} g(\boldsymbol{\upsilon}^* + t\boldsymbol{u}) < 0$ is negative for $t > 1$. Then all the extreme points of $g(\boldsymbol{\upsilon})$ are within $\mathcal{A}(\mathtt{r})$. We use the decomposition

$$g(\boldsymbol{\upsilon}^* + t\boldsymbol{u}) - g(\boldsymbol{\upsilon}^*) = \langle \boldsymbol{A}, \boldsymbol{u}\rangle\, t + f(\boldsymbol{\upsilon}^* + t\boldsymbol{u}) - f(\boldsymbol{\upsilon}^*).$$

With $h(t) = f(\boldsymbol{\upsilon}^* + t\boldsymbol{u}) - f(\boldsymbol{\upsilon}^*) + \langle \boldsymbol{A}, \boldsymbol{u}\rangle\, t$, it holds

$$\frac{d}{dt} f(\boldsymbol{\upsilon}^* + t\boldsymbol{u}) = -\langle \boldsymbol{A}, \boldsymbol{u}\rangle + h'(t). \tag{D.10}$$

By definition of $\boldsymbol{\upsilon}^*$, it also holds $h'(0) = \langle \boldsymbol{A}, \boldsymbol{u}\rangle$. The identity $-\nabla^2 f(\boldsymbol{\upsilon}^*) = \mathbb{F}$ yields $-h''(0) = \|\mathbb{F}^{1/2}\boldsymbol{u}\|^2$. Bound (D.2) implies for $|t| \leq 1$

$$\bigl|h'(t) - h'(0) - t h''(0)\bigr| \leq t\,\|\mathbb{D}\boldsymbol{u}\|^2\,\omega'\,.$$

For $t = 1$, we obtain by (D.9)

$$h'(1) \leq -\langle \boldsymbol{A}, \boldsymbol{u}\rangle - \|\mathbb{F}^{1/2}\boldsymbol{u}\|^2 + \|\mathbb{D}\boldsymbol{u}\|^2\,\omega' \leq -\|\mathbb{F}^{1/2}\boldsymbol{u}\|^2(1 - \omega' - \nu) < 0.$$

Moreover, concavity of $h(t)$ imply that $h'(t) - h'(0)$ decreases in t for $t > 1$. Further, summing up the above derivation yields

$$\frac{d}{dt} g(\boldsymbol{\upsilon}^* + t\boldsymbol{u})\Big|_{t=1} \leq -\|\mathbb{F}^{1/2}\boldsymbol{u}\|^2(1 - \nu - \omega') < 0.$$

As $\frac{d}{dt} g(\boldsymbol{\upsilon}^* + t\boldsymbol{u})$ decreases with $t \geq 1$ together with $h'(t)$ due to (D.10), the same applies to all such t. This implies the assertion. □

The result of Proposition D.4 allows to localize the point $\boldsymbol{\upsilon}^\circ = \operatorname{argmax}_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon})$ in the local vicinity $\mathcal{A}(\mathtt{r})$ of $\boldsymbol{\upsilon}^*$. The use of smoothness properties of g or, equivalently, of f, in this vicinity helps to obtain rather sharp expansions for $\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*$ and for $g(\boldsymbol{\upsilon}^\circ) - g(\boldsymbol{\upsilon}^*)$.

Proposition D.5. *Under the conditions of Proposition D.4,*

$$-\frac{\omega}{1 + \varkappa^2\omega}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 \leq 2g(\boldsymbol{\upsilon}^\circ) - 2g(\boldsymbol{\upsilon}^*) - \|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 \leq \frac{\omega}{1 - \varkappa^2\omega}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2\,. \tag{D.11}$$

Also

$$
\begin{aligned}
\|\mathbb{D}(\boldsymbol{v}^{\circ}-\boldsymbol{v}^{*}-\mathbb{F}^{-1}\boldsymbol{A})\| &\leq \frac{\sqrt{3\omega}}{1-\varkappa^{2}\omega}\,\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|\,, \\
\|\mathbb{D}(\boldsymbol{v}^{\circ}-\boldsymbol{v}^{*})\| &\leq \frac{1+\sqrt{3\omega}}{1-\varkappa^{2}\omega}\,\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|\,.
\end{aligned}
\tag{D.12}
$$

Proof. As in the proof of Proposition D.4, rescaling $\mathbb{D}$ by $\varkappa^{-1}$ reduces the statement to $\varkappa=1$. By (D.1), for any $\boldsymbol{v}\in\mathcal{A}(\mathtt{r})$

$$
\Bigl| f(\boldsymbol{v}^{*})-f(\boldsymbol{v})-\frac{1}{2}\|\mathbb{F}^{1/2}(\boldsymbol{v}-\boldsymbol{v}^{*})\|^{2}\Bigr| \leq \frac{\omega}{2}\|\mathbb{D}(\boldsymbol{v}-\boldsymbol{v}^{*})\|^{2}. \tag{D.13}
$$

Further,

$$
\begin{aligned}
& g(\boldsymbol{v})-g(\boldsymbol{v}^{*})-\frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^{2} \\
&\quad = \bigl\langle \boldsymbol{v}-\boldsymbol{v}^{*},\boldsymbol{A}\bigr\rangle + f(\boldsymbol{v})-f(\boldsymbol{v}^{*})-\frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^{2} \\
&\quad = -\frac{1}{2}\bigl\|\mathbb{F}^{1/2}(\boldsymbol{v}-\boldsymbol{v}^{*})-\mathbb{F}^{-1/2}\boldsymbol{A}\bigr\|^{2} + f(\boldsymbol{v})-f(\boldsymbol{v}^{*})+\frac{1}{2}\|\mathbb{F}^{1/2}(\boldsymbol{v}-\boldsymbol{v}^{*})\|^{2}.
\end{aligned}
\tag{D.14}
$$

As $\boldsymbol{v}^{\circ}\in\mathcal{A}(\mathtt{r})$ and it maximizes $g(\boldsymbol{v})$, we derive by (D.13)

$$
\begin{aligned}
g(\boldsymbol{v}^{\circ})-g(\boldsymbol{v}^{*})-\frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^{2} &= \max_{\boldsymbol{v}\in\mathcal{A}(\mathtt{r})}\Bigl\{ g(\boldsymbol{v})-g(\boldsymbol{v}^{*})-\frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^{2}\Bigr\} \\
&\leq \max_{\boldsymbol{v}\in\mathcal{A}(\mathtt{r})}\Bigl\{ -\frac{1}{2}\bigl\|\mathbb{F}^{1/2}(\boldsymbol{v}-\boldsymbol{v}^{*})-\mathbb{F}^{-1/2}\boldsymbol{A}\bigr\|^{2}+\frac{\omega}{2}\|\mathbb{D}(\boldsymbol{v}-\boldsymbol{v}^{*})\|^{2}\Bigr\}.
\end{aligned}
$$

Denote $\boldsymbol{u}=\mathbb{F}^{1/2}(\boldsymbol{v}-\boldsymbol{v}^{*})$, $\boldsymbol{\xi}=\mathbb{F}^{-1/2}\boldsymbol{A}$, and $\mathbb{B}=\mathbb{F}^{-1/2}\,\mathbb{D}^{2}\,\mathbb{F}^{-1/2}$. As $\mathbb{D}^{2}\leq\mathbb{F}$ and $\omega<1$, it holds $\|\mathbb{B}\|\leq 1$ and

$$
\begin{aligned}
& \max_{\boldsymbol{v}\in\mathcal{A}(\mathtt{r})}\bigl\{ -\bigl\|\mathbb{F}^{1/2}(\boldsymbol{v}-\boldsymbol{v}^{*})-\mathbb{F}^{-1/2}\boldsymbol{A}\bigr\|^{2}+\omega\|\mathbb{D}(\boldsymbol{v}-\boldsymbol{v}^{*})\|^{2}\bigr\} \\
&\quad = \max_{\|\boldsymbol{u}\|\leq\mathtt{r}}\bigl\{ -\|\boldsymbol{u}-\boldsymbol{\xi}\|^{2}+\omega\,\boldsymbol{u}^{\top}\mathbb{B}\boldsymbol{u}\bigr\} = \boldsymbol{\xi}^{\top}\bigl\{ (\mathit{I\!I}-\omega\,\mathbb{B})^{-1}-\mathit{I\!I}\bigr\}\boldsymbol{\xi} \leq \frac{\omega}{1-\omega}\boldsymbol{\xi}^{\top}\mathbb{B}\,\boldsymbol{\xi}
\end{aligned}
$$

yielding

$$
g(\boldsymbol{v}^{\circ})-g(\boldsymbol{v}^{*})-\frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^{2} \leq \frac{\omega}{2(1-\omega)}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^{2}.
$$

Similarly

$$
\begin{aligned}
& g(\boldsymbol{\upsilon}^{\circ}) - g(\boldsymbol{\upsilon}^{*}) - \frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 \\
& \quad \geq \max_{\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})} \Bigl\{ -\frac{1}{2}\bigl\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*}) - \mathbb{F}^{-1/2}\boldsymbol{A}\bigr\|^2 - \frac{\omega}{2}\|\mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*})\|^2 \Bigr\} \\
& \quad \geq \frac{1}{2}\boldsymbol{\xi}^{\top}\bigl\{(I\!\!I + \omega\,\mathbb{B})^{-1} - I\!\!I\bigr\}\boldsymbol{\xi} \geq -\frac{\omega}{2(1+\omega)}\,\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2.
\end{aligned}
\tag{D.15}
$$

These bounds imply (D.11).

Now we derive similarly to (D.14) that for $\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})$

$$
g(\boldsymbol{\upsilon}) - g(\boldsymbol{\upsilon}^{*}) \leq \bigl\langle \boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*}, \boldsymbol{A} \bigr\rangle - \frac{1}{2}\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*})\|^2 + \frac{\omega}{2}\|\mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*})\|^2.
$$

A particular choice $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}^{\circ}$ yields

$$
g(\boldsymbol{\upsilon}^{\circ}) - g(\boldsymbol{\upsilon}^{*}) \leq \bigl\langle \boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*}, \boldsymbol{A} \bigr\rangle - \frac{1}{2}\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})\|^2 + \frac{\omega}{2}\|\mathbb{D}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})\|^2.
$$

Combining this inequality with (D.15) allows to bound

$$
\bigl\langle \boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*}, \boldsymbol{A} \bigr\rangle - \frac{1}{2}\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})\|^2 + \frac{\omega}{2}\|\mathbb{D}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})\|^2 \geq \frac{1}{2}\boldsymbol{\xi}^{\top}(I\!\!I + \omega\,\mathbb{B})^{-1}\boldsymbol{\xi}.
$$

With $\boldsymbol{u}^{\circ} = \mathbb{F}^{1/2}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})$, this implies

$$
2\bigl\langle \boldsymbol{u}^{\circ}, \boldsymbol{\xi} \bigr\rangle - \boldsymbol{u}^{\circ\top}(1 - \omega\mathbb{B})\boldsymbol{u}^{\circ} \geq \boldsymbol{\xi}^{\top}(I\!\!I + \omega\,\mathbb{B})^{-1}\boldsymbol{\xi}\,,
$$

and hence,

$$
\begin{aligned}
& \bigl\{\boldsymbol{u}^{\circ} - (I\!\!I - \omega\mathbb{B})^{-1}\boldsymbol{\xi}\bigr\}^{\top}(I\!\!I - \omega\mathbb{B})\bigl\{\boldsymbol{u}^{\circ} - (I\!\!I - \omega\mathbb{B})^{-1}\boldsymbol{\xi}\bigr\} \\
& \quad \leq \boldsymbol{\xi}^{\top}\bigl\{(I\!\!I - \omega\,\mathbb{B})^{-1} - (I\!\!I + \omega\,\mathbb{B})^{-1}\bigr\}\boldsymbol{\xi} \leq \frac{2\omega}{(1+\omega)(1-\omega)}\,\boldsymbol{\xi}^{\top}\mathbb{B}\,\boldsymbol{\xi}\,.
\end{aligned}
$$

Introduce $\|\cdot\|_{\mathbb{z}}$ by $\|\boldsymbol{x}\|_{\mathbb{z}}^2 \stackrel{\mathrm{def}}{=} \boldsymbol{x}^{\top}(I\!\!I - \omega\mathbb{B})\boldsymbol{x}$, $\boldsymbol{x} \in I\!\!R^p$. Then

$$
\|\boldsymbol{u}^{\circ} - (I\!\!I - \omega\mathbb{B})^{-1}\boldsymbol{\xi}\|_{\mathbb{z}}^2 \leq \frac{2\omega}{1-\omega^2}\,\boldsymbol{\xi}^{\top}\mathbb{B}\,\boldsymbol{\xi}\,.
$$

As

$$
\|\boldsymbol{\xi} - (I\!\!I - \omega\mathbb{B})^{-1}\boldsymbol{\xi}\|_{\mathbb{z}}^2 = \omega^2(\mathbb{B}\boldsymbol{\xi})^{\top}(I\!\!I - \omega\mathbb{B})^{-1}\mathbb{B}\boldsymbol{\xi} \leq \frac{\omega^2}{1-\omega}\,\|\mathbb{B}\boldsymbol{\xi}\|^2 \leq \frac{\omega^2}{1-\omega}\,\boldsymbol{\xi}^{\top}\mathbb{B}\,\boldsymbol{\xi}
$$

we conclude for $\omega \leq 1/3$ by the triangle inequality

$$
\|\boldsymbol{u}^{\circ} - \boldsymbol{\xi}\|_{\mathbb{z}} \leq \Biggl(\sqrt{\frac{\omega^2}{1-\omega}} + \sqrt{\frac{2\omega}{1-\omega^2}}\Biggr)\sqrt{\boldsymbol{\xi}^{\top}\mathbb{B}\,\boldsymbol{\xi}} \leq \sqrt{\frac{3\omega}{1-\omega}}\,\sqrt{\boldsymbol{\xi}^{\top}\mathbb{B}\,\boldsymbol{\xi}}\,,
$$

and (D.12) follows by $I\!\!I - \omega\mathbb{B} \geq (1-\omega)I\!\!I$. □

Remark D.1. The roles of the functions f and g are exchangeable. In particular, the results from (D.12) apply with $\mathbb{F} = -\nabla^2 g(\boldsymbol{\upsilon}^\circ) = -\nabla^2 f(\boldsymbol{\upsilon}^\circ)$ provided that (D.1) is fulfilled at $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}^\circ$.

D.2.1 Basic lemma under third order smoothness

The results of Proposition D.5 can be refined if f satisfies condition $(\mathcal{T}_3)$.

Proposition D.6. *Let $f(\boldsymbol{\upsilon})$ be a strongly concave function with $f(\boldsymbol{\upsilon}^*) = \max_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon})$ and $\mathbb{F} = -\nabla^2 f(\boldsymbol{\upsilon}^*)$. Let $g(\boldsymbol{\upsilon})$ fulfill (D.6) with some vector $\boldsymbol{A}$. Suppose that $f(\boldsymbol{\upsilon})$ follows $(\mathcal{T}_3)$ at $\boldsymbol{\upsilon}^*$ with $\mathbb{D}^2$, $\mathtt{r}$, and τ_3 such that*

$$\mathbb{D}^2 \le \varkappa^2\, \mathbb{F}, \quad \mathtt{r} \ge \frac{4\varkappa}{3}\, \|\mathbb{F}^{-1/2}\boldsymbol{A}\|, \quad \varkappa^3 \tau_3\, \|\mathbb{F}^{-1/2}\boldsymbol{A}\| < \frac{1}{4}. \tag{D.16}$$

Then $\boldsymbol{\upsilon}^\circ = \operatorname{argmax}_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon})$ satisfies

$$\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)\| \le \frac{4}{3}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|, \qquad \|D(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)\| \le \frac{4\varkappa}{3}\, \|\mathbb{F}^{-1/2}\boldsymbol{A}\|.$$

Moreover,

$$\Bigl| 2g(\boldsymbol{\upsilon}^\circ) - 2g(\boldsymbol{\upsilon}^*) - \|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 \Bigr| \le \frac{\tau_3}{2}\, \|\mathbb{D}\, \mathbb{F}^{-1}\boldsymbol{A}\|^3. \tag{D.17}$$

Proof. W.l.o.g. assume $\varkappa = 1$. The first statement follows from Proposition D.4 with $\nu = 3/4$ because $(\mathcal{T}_3)$ ensures (D.1) with $\omega'(\boldsymbol{\upsilon}) = \tau_3\, \mathtt{r}/2$ and (D.16) implies (D.9).

As $\nabla f(\boldsymbol{\upsilon}^*) = 0$, $(\mathcal{T}_3)$ implies for any $\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})$

$$\Bigl| f(\boldsymbol{\upsilon}^*) - f(\boldsymbol{\upsilon}) - \frac{1}{2}\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 \Bigr| \le \frac{\tau_3}{6}\|\mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^3. \tag{D.18}$$

Further,

$$\begin{aligned}
& g(\boldsymbol{\upsilon}) - g(\boldsymbol{\upsilon}^*) - \frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 \\
&\quad = \bigl\langle \boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*, \boldsymbol{A} \bigr\rangle + f(\boldsymbol{\upsilon}) - f(\boldsymbol{\upsilon}^*) - \frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 \\
&\quad = -\frac{1}{2}\bigl\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \mathbb{F}^{-1/2}\boldsymbol{A}\bigr\|^2 + f(\boldsymbol{\upsilon}) - f(\boldsymbol{\upsilon}^*) + \frac{1}{2}\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2.
\end{aligned}$$

As $\boldsymbol{\upsilon}^\circ \in \mathcal{A}(\mathtt{r})$ and it maximizes $g(\boldsymbol{\upsilon})$, we derive by (D.18) and Lemma D.7 with $\mathbb{U} = \mathbb{F}^{1/2}\mathbb{D}^{-1}$ and $\boldsymbol{s} = \mathbb{D}\, \mathbb{F}^{-1}\boldsymbol{A}$

$$\begin{aligned}
& 2g(\boldsymbol{\upsilon}^\circ) - 2g(\boldsymbol{\upsilon}^*) - \|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 = \max_{\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})} \Bigl\{ 2g(\boldsymbol{\upsilon}) - 2g(\boldsymbol{\upsilon}^*) - \|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 \Bigr\} \\
&\quad \le \max_{\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})} \Bigl\{ -\bigl\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \mathbb{F}^{-1/2}\boldsymbol{A}\bigr\|^2 + \frac{\tau_3}{3}\|\mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^3 \Bigr\} \le \frac{\tau_3}{2}\, \|\mathbb{D}\, \mathbb{F}^{-1}\boldsymbol{A}\|^3.
\end{aligned}$$

Similarly

$$
\begin{aligned}
&2g(\boldsymbol{\upsilon}^{\circ}) - 2g(\boldsymbol{\upsilon}^{*}) - \|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 \\
&\quad \geq \max_{\boldsymbol{\upsilon}\in\mathcal{A}(\mathtt{r})} \Bigl\{ -\bigl\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*}) - \mathbb{F}^{-1/2}\boldsymbol{A}\bigr\|^2 - \frac{\tau_3}{3}\|\mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*})\|^3 \Bigr\} \geq -\frac{\tau_3}{2}\,\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3 .
\end{aligned}
$$

This implies (D.17). □

Lemma D.7. *Let* $\mathbb{U} \geq \mathbb{I}$ *. Fix some* $\mathtt{r}$ *and let* $\boldsymbol{s} \in \mathbb{R}^p$ *satisfy* $(3/4)\mathtt{r} \leq \|\boldsymbol{s}\| \leq \mathtt{r}$ *. If* $\tau\,\mathtt{r} \leq 1/3$ *, then*

$$
\max_{\|\boldsymbol{u}\|\leq\mathtt{r}} \Bigl(\frac{\tau}{3}\|\boldsymbol{u}\|^3 - (\boldsymbol{u} - \boldsymbol{s})^{\top}\mathbb{U}(\boldsymbol{u} - \boldsymbol{s}) \Bigr) \leq \frac{\tau}{2}\,\|\boldsymbol{s}\|^3 , \tag{D.19}
$$

$$
\min_{\|\boldsymbol{u}\|\leq\mathtt{r}} \Bigl(\frac{\tau}{3}\|\boldsymbol{u}\|^3 + (\boldsymbol{u} - \boldsymbol{s})^{\top}\mathbb{U}(\boldsymbol{u} - \boldsymbol{s}) \Bigr) \leq \frac{\tau}{2}\,\|\boldsymbol{s}\|^3 . \tag{D.20}
$$

Proof. Replacing $\|\boldsymbol{u}\|^3$ with $\mathtt{r}\|\boldsymbol{u}\|^2$ reduces the problem to quadratic programming. It holds with $\rho \stackrel{\text{def}}{=} \tau\mathtt{r}/3$ and $\boldsymbol{s}_{\rho} \stackrel{\text{def}}{=} (\mathbb{U} - \rho\mathbb{I})^{-1}\mathbb{U}\boldsymbol{s}$

$$
\begin{aligned}
\frac{\tau}{3}\|\boldsymbol{u}\|^3 - (\boldsymbol{u} - \boldsymbol{s})^{\top}\mathbb{U}(\boldsymbol{u} - \boldsymbol{s}) &\leq \frac{\tau\mathtt{r}}{3}\|\boldsymbol{u}\|^2 - (\boldsymbol{u} - \boldsymbol{s})^{\top}\mathbb{U}(\boldsymbol{u} - \boldsymbol{s}) \\
&= -\boldsymbol{u}^{\top}\bigl(\mathbb{U} - \rho\mathbb{I}\bigr)\boldsymbol{u} + 2\boldsymbol{u}^{\top}\mathbb{U}\boldsymbol{s} - \boldsymbol{s}^{\top}\mathbb{U}\boldsymbol{s} \\
&= -(\boldsymbol{u} - \boldsymbol{s}_{\rho})^{\top}(\mathbb{U} - \rho\mathbb{I})(\boldsymbol{u} - \boldsymbol{s}_{\rho}) + \boldsymbol{s}_{\rho}^{\top}(\mathbb{U} - \rho\mathbb{I})\boldsymbol{s}_{\rho} - \boldsymbol{s}^{\top}\mathbb{U}\boldsymbol{s} \\
&\leq \boldsymbol{s}^{\top}\bigl\{\mathbb{U}(\mathbb{U} - \rho\mathbb{I})^{-1}\mathbb{U} - \mathbb{U}\bigr\}\boldsymbol{s} = \rho\boldsymbol{s}^{\top}\mathbb{U}(\mathbb{U} - \rho\mathbb{I})^{-1}\boldsymbol{s}.
\end{aligned}
$$

This implies in view of $\mathbb{U} \geq \mathbb{I}$, $\mathtt{r} \leq (4/3)\|\boldsymbol{s}\|$, and $\rho \leq 1/9$

$$
\begin{aligned}
&\max_{\|\boldsymbol{u}\|\leq\mathtt{r}} \Bigl(\frac{\tau}{3}\|\boldsymbol{u}\|^3 - (\boldsymbol{u} - \boldsymbol{s})^{\top}\mathbb{U}(\boldsymbol{u} - \boldsymbol{s}) \Bigr) \\
&\qquad \leq \frac{\rho}{1-\rho}\|\boldsymbol{s}\|^2 \leq \frac{\tau\mathtt{r}}{3(1-\rho)}\|\boldsymbol{s}\|^2 \leq \frac{4\tau}{9(1-\rho)}\|\boldsymbol{s}\|^3 \leq \frac{\tau}{2}\|\boldsymbol{s}\|^3 ,
\end{aligned}
$$

and (D.19) follows. For (D.20) note that

$$
\begin{aligned}
&\min_{\|\boldsymbol{u}\|\leq\mathtt{r}} \Bigl(\frac{\tau}{3}\|\boldsymbol{u}\|^3 + (\boldsymbol{u} - \boldsymbol{s})^{\top}\mathbb{U}(\boldsymbol{u} - \boldsymbol{s}) \Bigr) \leq \min_{\boldsymbol{u}} \Bigl(\frac{\tau\mathtt{r}}{3}\|\boldsymbol{u}\|^2 + (\boldsymbol{u} - \boldsymbol{s})^{\top}\mathbb{U}(\boldsymbol{u} - \boldsymbol{s}) \Bigr) \\
&\quad \leq \boldsymbol{s}^{\top}\bigl\{\mathbb{U} - \mathbb{U}(\mathbb{U} + \rho\mathbb{I})^{-1}\mathbb{U}\bigr\}\boldsymbol{s} = \rho\boldsymbol{s}^{\top}\mathbb{U}(\mathbb{U} + \rho\mathbb{I})^{-1}\boldsymbol{s} \leq \frac{\tau\mathtt{r}}{3}\|\boldsymbol{s}\|^2 \leq \frac{4\tau}{9}\|\boldsymbol{s}\|^3 ,
\end{aligned}
$$

and (D.20) follows as well. □

D.2.2 Advanced approximation under locally uniform smoothness

The bounds of Proposition D.6 can be made more accurate if f follows $(\mathcal{T}_3^*)$ and $(\mathcal{T}_4^*)$ and one can apply the Taylor expansion around any point close to $\boldsymbol{\upsilon}^*$. In particular, the improved results do not involve the value $\|\mathbb{F}^{-1/2}\boldsymbol{A}\|$ which can be large or even infinite in some situation; see Section D.2.3.

Proposition D.8. *Let $f(\boldsymbol{\upsilon})$ be a strongly concave function with $f(\boldsymbol{\upsilon}^*) = \max_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon})$ and $\mathbb{F} = -\nabla^2 f(\boldsymbol{\upsilon}^*)$. Assume $(\mathcal{T}_3^*)$ at $\boldsymbol{\upsilon}^*$ with $\mathbb{D}^2$, $\mathtt{r}$, and τ_3 such that*

$$\mathbb{D}^2 \le \varkappa^2\, \mathbb{F}, \quad \mathtt{r} \ge \frac{3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|, \quad \varkappa^2\tau_3\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| < \frac{4}{9}.$$

Then $\|\mathbb{D}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)\| \le (3/2)\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|$ and moreover,

$$\|\mathbb{D}^{-1}\mathbb{F}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^* - \mathbb{F}^{-1}\boldsymbol{A})\| \le \frac{3\tau_3}{4}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2. \tag{D.21}$$

Proof. W.l.o.g. assume $\varkappa = 1$. If the function f is quadratic and concave with the maximum at $\boldsymbol{\upsilon}^*$ then the linearly perturbed function g is also quadratic and concave with the maximum at $\breve{\boldsymbol{\upsilon}} = \boldsymbol{\upsilon}^* + \mathbb{F}^{-1}\boldsymbol{A}$. In general, the point $\breve{\boldsymbol{\upsilon}}$ is not the maximizer of g, however, it is very close to $\boldsymbol{\upsilon}^\circ$. We use that $\nabla f(\boldsymbol{\upsilon}^*) = 0$ and $-\nabla^2 f(\boldsymbol{\upsilon}^*) = \mathbb{F}$. The main step of the proof is given by the next lemma.

Lemma D.9. *Assume $(\mathcal{T}_3^*)$ at $\boldsymbol{\upsilon}$. With $\mathbb{D} = \mathbb{D}(\boldsymbol{\upsilon})$, let $\mathcal{U}_{\mathtt{r}} = \{\boldsymbol{u} \colon \|\mathbb{D}\boldsymbol{u}\| \le \mathtt{r}\}$. Then*

$$\bigl\|\mathbb{D}^{-1}\bigl\{\nabla f(\boldsymbol{\upsilon}+\boldsymbol{u}) - \nabla f(\boldsymbol{\upsilon}) - \langle \nabla^2 f(\boldsymbol{\upsilon}), \boldsymbol{u}\rangle\bigr\}\bigr\| \le \frac{\tau_3}{2}\,\|\mathbb{D}\boldsymbol{u}\|^2, \quad \boldsymbol{u} \in \mathcal{U}_{\mathtt{r}}. \tag{D.22}$$

Also for all $\boldsymbol{u}, \boldsymbol{u}_1 \in \mathcal{U}_{\mathtt{r}}$

$$\bigl\|\mathbb{D}^{-1}\bigl\{\nabla^2 f(\boldsymbol{\upsilon}+\boldsymbol{u}_1) - \nabla^2 f(\boldsymbol{\upsilon}+\boldsymbol{u})\bigr\}\mathbb{D}^{-1}\bigr\| \le \tau_3\,\|\mathbb{D}(\boldsymbol{u}_1 - \boldsymbol{u})\| \tag{D.23}$$

and

$$\bigl\|\mathbb{D}^{-1}\bigl\{\nabla f(\boldsymbol{\upsilon}+\boldsymbol{u}_1) - \nabla f(\boldsymbol{\upsilon}+\boldsymbol{u}) - \nabla^2 f(\boldsymbol{\upsilon})(\boldsymbol{u}_1 - \boldsymbol{u})\bigr\}\bigr\| \le \frac{3\tau_3}{2}\,\|\mathbb{D}(\boldsymbol{u}_1 - \boldsymbol{u})\|^2. \tag{D.24}$$

Moreover, under $(\mathcal{T}_4^*)$, for any $\boldsymbol{u} \in \mathcal{U}_{\mathtt{r}}$,

$$\bigl\|\mathbb{D}^{-1}\bigl\{\nabla f(\boldsymbol{\upsilon}+\boldsymbol{u}) - \nabla f(\boldsymbol{\upsilon}) - \langle \nabla^2 f(\boldsymbol{\upsilon}), \boldsymbol{u}\rangle - \frac{1}{2}\langle \nabla^3 f(\boldsymbol{\upsilon}), \boldsymbol{u}^{\otimes 2}\rangle\bigr\}\bigr\| \le \frac{\tau_4}{6}\,\|\mathbb{D}\boldsymbol{u}\|^3. \tag{D.25}$$

Proof. Denote

$$\boldsymbol{A} \stackrel{\text{def}}{=} \nabla f(\boldsymbol{\upsilon}+\boldsymbol{u}) - \nabla f(\boldsymbol{\upsilon}) - \langle \nabla^2 f(\boldsymbol{\upsilon}), \boldsymbol{u}\rangle.$$

For any vector $\boldsymbol{w} \in I\!\!R^p$, $(\mathcal{T}_3^*)$ and (D.3) imply

$$\bigl|\langle \boldsymbol{A}, \boldsymbol{w} \rangle\bigr| \le \frac{\tau_3}{2} \, \|\mathbb{D}\boldsymbol{u}\|^2 \, \|\mathbb{D}\boldsymbol{w}\|.$$

Therefore,

$$\|\mathbb{D}^{-1}\boldsymbol{A}\| = \sup_{\|\boldsymbol{w}\|=1} \bigl|\langle \mathbb{D}^{-1}\boldsymbol{A}, \boldsymbol{w} \rangle\bigr| = \sup_{\|\boldsymbol{w}\|=1} \bigl|\langle \boldsymbol{A}, \mathbb{D}^{-1}\boldsymbol{w} \rangle\bigr| \le \frac{\tau_3}{2} \, \|\mathbb{D}\boldsymbol{u}\|^2$$

which yields the first statement. For (D.25), apply

$$\boldsymbol{A} \stackrel{\text{def}}{=} \nabla f(\boldsymbol{v} + \boldsymbol{u}) - \nabla f(\boldsymbol{v}) - \langle \nabla^2 f(\boldsymbol{v}), \boldsymbol{u} \rangle - \frac{1}{2} \langle \nabla^3 f(\boldsymbol{v}), \boldsymbol{u}^{\otimes 2} \rangle$$

and use $(\mathcal{T}_4^*)$ and (D.4) instead of $(\mathcal{T}_3^*)$ and (D.3). Further, with $I\!\!B_1 \stackrel{\text{def}}{=} \nabla^2 f(\boldsymbol{v} + \boldsymbol{u}_1) - \nabla^2 f(\boldsymbol{v} + \boldsymbol{u})$ and $\Delta = \boldsymbol{u}_1 - \boldsymbol{u}$, by $(\mathcal{T}_3^*)$, for any $\boldsymbol{w} \in I\!\!R^p$ and some $t \in [0,1]$,

$$\begin{aligned}
&\bigl|\langle \mathbb{D}^{-1} \bigl\{ \nabla^2 f(\boldsymbol{v} + \boldsymbol{u}_1) - \nabla^2 f(\boldsymbol{v} + \boldsymbol{u}) \bigr\} \, \mathbb{D}^{-1}, \boldsymbol{w}^{\otimes 2} \rangle\bigr| = \bigl|\langle I\!\!B_1, (\mathbb{D}^{-1}\boldsymbol{w})^{\otimes 2} \rangle\bigr| \\
&\quad = \bigl|\bigl\langle \nabla^3 f(\boldsymbol{v} + \boldsymbol{u} + t\Delta), \Delta \otimes (\mathbb{D}^{-1}\boldsymbol{w})^{\otimes 2} \bigr\rangle\bigr| \le \tau_3 \|\mathbb{D}\Delta\| \, \|\boldsymbol{w}\|^2 .
\end{aligned}$$

This proves (D.23). Similarly, for some $t \in [0,1]$

$$\begin{aligned}
&\bigl|\bigl\langle \mathbb{D}^{-1} \bigl\{ \nabla f(\boldsymbol{v} + \boldsymbol{u}_1) - \nabla f(\boldsymbol{v} + \boldsymbol{u}) \bigr\} - \nabla^2 f(\boldsymbol{v} + \boldsymbol{u}) \Delta \bigr\}, \boldsymbol{w} \bigr\rangle\bigr| \\
&\quad = \frac{1}{2} \bigl|\bigl\langle \nabla^3 f(\boldsymbol{v} + \boldsymbol{u} + t\Delta), \Delta \otimes \Delta \otimes \mathbb{D}^{-1}\boldsymbol{w} \bigr\rangle\bigr| \le \frac{\tau_3}{2} \|\mathbb{D}\Delta\|^2 \, \|\boldsymbol{w}\|
\end{aligned}$$

and with $I\!\!B = \nabla^2 f(\boldsymbol{v} + \boldsymbol{u}) - \nabla^2 f(\boldsymbol{v})$, by (D.23),

$$\bigl\|\mathbb{D}^{-1} I\!\!B \Delta\bigr\| \le \|\mathbb{D}^{-1} I\!\!B \, \mathbb{D}^{-1}\| \; \|\mathbb{D}\Delta\| \le \tau_3 \|\mathbb{D}\Delta\|^2 .$$

This completes the proof of (D.24). □

Now (D.22) of Lemma D.9 yields

$$\bigl\|\mathbb{D}^{-1} \nabla g(\breve{\boldsymbol{v}})\bigr\| = \bigl\|\mathbb{D}^{-1} \{ \nabla f(\boldsymbol{v}^* + \mathbb{F}^{-1}\boldsymbol{A}) - \nabla f(\boldsymbol{v}^*) + \boldsymbol{A} \}\bigr\| \le \frac{\tau_3}{2} \|\mathbb{D} \, \mathbb{F}^{-1}\boldsymbol{A}\|^2 . \tag{D.26}$$

As $\|\mathbb{D} \, \mathbb{F}^{-1}\boldsymbol{A}\| \le 2\mathtt{r}/3$, condition $(\mathcal{T}_3^*)$ can be applied in the $\mathtt{r}/3$-vicinity of $\breve{\boldsymbol{v}}$. Fix any $\boldsymbol{v}$ with $\|\mathbb{D}(\boldsymbol{v} - \breve{\boldsymbol{v}})\| \le \mathtt{r}/3$ and define $\Delta = \boldsymbol{v} - \breve{\boldsymbol{v}}$. By (D.24) of Lemma D.9

$$\bigl\|\mathbb{D}^{-1} \{ \nabla g(\boldsymbol{v}) - \nabla g(\breve{\boldsymbol{v}}) + \mathbb{F}\Delta \}\bigr\| = \bigl\|\mathbb{D}^{-1} \{ \nabla f(\boldsymbol{v}) - \nabla f(\breve{\boldsymbol{v}}) + \mathbb{F}\Delta \}\bigr\| \le \frac{3\tau_3}{2} \|\mathbb{D}\Delta\|^2 .$$

In particular, this and (D.26) yield

$$\bigl\|\mathbb{D}^{-1} \{ \nabla g(\breve{\boldsymbol{v}} + \Delta) + \mathbb{F}\Delta \}\bigr\| \le 2\tau_3 \|\mathbb{D}\Delta\|^2 .$$

For any $\boldsymbol{u}$ with $\|\boldsymbol{u}\| = 1$, this implies

$$\bigl|\bigl\langle \nabla g(\breve{\boldsymbol{v}} + \Delta) + \mathbb{F}\Delta, \mathbb{D}^{-1}\boldsymbol{u} \bigr\rangle\bigr| \leq 2\tau_3 \|\mathbb{D}\Delta\|^2 . \tag{D.27}$$

Suppose now that $\|\mathbb{D}\Delta\| = \mathtt{r}/3$ and consider the function $h(t) = g(\breve{\boldsymbol{v}} + t\Delta)$. Then $h'(t) = \langle \nabla g(\breve{\boldsymbol{v}} + t\Delta), \Delta \rangle$ and (D.27) implies with $\boldsymbol{u} = \mathbb{D}\Delta / \|\mathbb{D}\Delta\|$

$$\bigl|\langle \nabla g(\breve{\boldsymbol{v}} + \Delta), \Delta \rangle + \|\mathbb{F}^{1/2}\Delta\|^2\bigr| \leq 2\tau_3 \|\mathbb{D}\Delta\|^3 .$$

As $\mathbb{F} \geq \mathbb{D}^2$, this yields

$$h'(1) \leq 2\tau_3 \|\mathbb{D}\Delta\|^3 - \|\mathbb{D}\Delta\|^2. \tag{D.28}$$

Similarly, (D.26) yields by $\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| = 2\mathtt{r}/3$

$$|h'(0)| = \bigl|\langle \nabla g(\breve{\boldsymbol{v}}), \Delta \rangle\bigr| \leq \frac{\tau_3}{2} \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 \, \|\mathbb{D}\Delta\| = \frac{2\tau_3}{9} \, \mathtt{r}^2 \, \|\mathbb{D}\Delta\| . \tag{D.29}$$

Concavity of $g(\cdot)$ ensures that $t^* = \operatorname{argmax}_t h(t)$ satisfies $|t^*| \leq 1$ provided that

$$h'(1) < -|h'(0)|, \qquad h'(-1) < |h'(0)|.$$

Due to (D.28), (D.29), and $\|\mathbb{D}\Delta\| = \mathtt{r}/3$, the latter condition reads

$$\frac{2\tau_3}{9}\mathtt{r}^2 \, \|\mathbb{D}\Delta\| + 2\tau_3 \|\mathbb{D}\Delta\|^3 - \|\mathbb{D}\Delta\|^2 = \|\mathbb{D}\Delta\| \, \mathtt{r} \Bigl(\frac{2\tau_3 \, \mathtt{r}}{9} + \frac{2\tau_3 \, \mathtt{r}}{9} - \frac{1}{3} \Bigr) < 0.$$

which is fulfilled because of $\tau_3 \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \leq 4/9$ and $\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| = 2\mathtt{r}/3$. We summarize that $\boldsymbol{v}^{\circ} = \operatorname{argmax}_{\boldsymbol{v}} g(\boldsymbol{v})$ satisfies $\|\mathbb{D}\,(\boldsymbol{v}^{\circ} - \breve{\boldsymbol{v}})\| \leq \mathtt{r}/3$ while $\|\mathbb{D}(\breve{\boldsymbol{v}} - \boldsymbol{v}^*)\| = \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| = 2\mathtt{r}/3$. Therefore, $\|\mathbb{D}(\boldsymbol{v}^{\circ} - \boldsymbol{v}^*)\| \leq \mathtt{r}$. This allows us to use $(\mathcal{T}_3^*)$ at this point for establishing (D.21). By definition $\nabla g(\boldsymbol{v}^{\circ}) = 0$ and hence,

$$\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{v}^* + \mathbb{F}^{-1}\boldsymbol{A}) - \nabla g(\boldsymbol{v}^{\circ})\}\| \leq \frac{\tau_3}{2} \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 . \tag{D.30}$$

By (D.24) of Lemma D.9, it holds with $\Delta = \boldsymbol{v}^* + \mathbb{F}^{-1}\boldsymbol{A} - \boldsymbol{v}^{\circ}$

$$\bigl\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{v}^* + \mathbb{F}^{-1}\boldsymbol{A}) - \nabla g(\boldsymbol{v}^{\circ}) - \nabla^2 g(\boldsymbol{v}^*)\Delta\}\bigr\| \leq \frac{3\tau_3}{2} \|\mathbb{D}\Delta\|^2 .$$

Combining with (D.30) yields

$$\|\mathbb{D}^{-1}\mathbb{F}\Delta\| \leq \frac{3\tau_3}{2} \|\mathbb{D}\Delta\|^2 + \frac{\tau_3}{2} \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 \leq \frac{3\tau_3}{2} \|\mathbb{D}^{-1}\mathbb{F}\Delta\|^2 + \frac{\tau_3}{2} \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 .$$

As $2x \le \alpha x^2 + \beta$ with $\alpha = 3\tau_3$, $\beta = \tau_3 \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2$, and $x = \|\mathbb{D}^{-1}\mathbb{F}\varDelta\| \in (0, 1/\alpha)$ implies $x \le \beta/(2-\alpha\beta)$, this yields

$$\|\mathbb{D}^{-1}\mathbb{F}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*} - \mathbb{F}^{-1}\boldsymbol{A})\| \le \frac{\tau_3}{2 - 3\tau_3^2 \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2} \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2$$

and (D.21) follows by $\tau_3 \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \le 4/9$. □

Remark D.2. As in Remark D.1, the roles of f and g can be exchanged. In particular, (D.21) applies with $\mathbb{F} = \mathbb{F}(\boldsymbol{\upsilon}^{\circ})$ provided that $(\mathcal{T}_3^*)$ is also fulfilled at $\boldsymbol{\upsilon}^{\circ}$.

If f is fourth-order smooth and $(\mathcal{T}_4^*)$ holds then expansion (D.21) can further be refined.

Proposition D.10. *Let $f(\boldsymbol{\upsilon})$ be a strongly concave function with $f(\boldsymbol{\upsilon}^*) = \max_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon})$ and $\mathbb{F} = -\nabla^2 f(\boldsymbol{\upsilon}^*)$, and let $f(\boldsymbol{\upsilon})$ follow $(\mathcal{T}_3^*)$ and $(\mathcal{T}_4^*)$ with some $\mathbb{D}^2$, τ_3, τ_4, and $\mathtt{r}$ satisfying*

$$\mathbb{D}^2 \le \varkappa^2\, \mathbb{F}, \quad \mathtt{r} = \frac{3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|, \quad \varkappa^2 \tau_3 \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| < \frac{4}{9}, \quad \varkappa^2 \tau_4 \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 < \frac{1}{3}. \tag{D.31}$$

Let $g(\boldsymbol{\upsilon})$ fulfill (D.6) with some vector $\boldsymbol{A}$ and $g(\boldsymbol{\upsilon}^{\circ}) = \max_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon})$. Then $\|\mathbb{D}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^*)\| \le (3/2)\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|$. Further, define

$$\boldsymbol{a} = \mathbb{F}^{-1}\{\boldsymbol{A} + \nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\}, \tag{D.32}$$

where $\mathcal{T}(\boldsymbol{u}) = \frac{1}{6}\langle \nabla^3 f(\boldsymbol{\upsilon}^*), \boldsymbol{u}^{\otimes 3}\rangle$ for $\boldsymbol{u} \in I\!\!R^p$. Then

$$\|\mathbb{D}^{-1}\mathbb{F}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^* - \boldsymbol{a})\| \le (\tau_4/2 + \varkappa^2\tau_3^2)\,\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3. \tag{D.33}$$

Also

$$\begin{aligned} &\Bigl| g(\boldsymbol{\upsilon}^{\circ}) - g(\boldsymbol{\upsilon}^*) - \frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 - \mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A}) \Bigr| \\ &\quad \le \frac{\tau_4 + 4\varkappa^2\tau_3^2}{8}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^4 + \frac{\varkappa^2\,(\tau_4 + 2\varkappa^2\tau_3^2)^2}{4}\,\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^6 \end{aligned} \tag{D.34}$$

and

$$\bigl|\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\bigr| \le \frac{\tau_3}{6}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3. \tag{D.35}$$

Proof. W.l.o.g. assume $\varkappa = 1$ and $\boldsymbol{\upsilon}^* = 0$. Proposition D.8 yields (D.21). By $(\mathcal{T}_3^*)$

$$\begin{aligned} &\|\mathbb{D}^{-1}\,\mathbb{F}(\boldsymbol{a} - \mathbb{F}^{-1}\boldsymbol{A})\| = \|\mathbb{D}^{-1}\,\nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\| \\ &\quad = \sup_{\|\boldsymbol{u}\|=1} 3\bigl|\langle \mathcal{T}, \mathbb{F}^{-1}\boldsymbol{A} \otimes \mathbb{F}^{-1}\boldsymbol{A} \otimes \mathbb{D}^{-1}\boldsymbol{u}\rangle\bigr| \le \frac{\tau_3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2. \end{aligned} \tag{D.36}$$

As $\mathbb{D}^{-1}\,\mathbb{F} \geq \mathbb{F}^{1/2} \geq \mathbb{D}$, this implies by $\tau_3 \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \leq 4/9$

$$\begin{aligned}\|\mathbb{D}\boldsymbol{a}\| &\leq \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| + \|\mathbb{D}\,\mathbb{F}^{-1}\,\nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\| \\ &\leq \Bigl(1 + \frac{\tau_3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|\Bigr)\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \leq \frac{11}{9}\,\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \end{aligned} \tag{D.37}$$

and

$$\|\mathbb{F}^{1/2}\boldsymbol{a} - \mathbb{F}^{-1/2}\boldsymbol{A}\| \leq \frac{\tau_3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 . \tag{D.38}$$

Next, again by $(\mathcal{T}_3^*)$, for any $\boldsymbol{w}$

$$\|\mathbb{D}^{-1}\,\nabla^2\mathcal{T}(\boldsymbol{w})\,\mathbb{D}^{-1}\| = \sup_{\|\boldsymbol{u}\|=1} 6\bigl|\langle\mathcal{T}, \boldsymbol{w}\otimes(\mathbb{D}^{-1}\boldsymbol{u})^{\otimes 2}\rangle\bigr| \leq \tau_3\|\mathbb{D}\boldsymbol{w}\| .$$

The tensor $\nabla^2\mathcal{T}(\boldsymbol{u})$ is linear in $\boldsymbol{u}$, hence

$$\begin{aligned}&\sup_{t\in[0,1]} \|\mathbb{D}^{-1}\,\nabla^2\mathcal{T}(t\boldsymbol{a} + (1-t)\mathbb{F}^{-1}\boldsymbol{A})\,\mathbb{D}^{-1}\| \\ &\quad = \max\{\|\mathbb{D}^{-1}\,\nabla^2\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\,\mathbb{D}^{-1}\|, \|\mathbb{D}^{-1}\nabla^2\mathcal{T}(\boldsymbol{a})\mathbb{D}^{-1}\|\} \leq \tau_3\,\max\{\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|, \|\mathbb{D}\boldsymbol{a}\|\} .\end{aligned}$$

Based on (D.37), assume $\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \leq \|\mathbb{D}\boldsymbol{a}\| \leq (11/9)\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|$. Then (D.36) yield

$$\begin{aligned}&\|\mathbb{D}^{-1}\nabla\mathcal{T}(\boldsymbol{a}) - \mathbb{D}^{-1}\nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\| \\ &\quad\leq \sup_{t\in[0,1]} \|\mathbb{D}^{-1}\,\nabla^2\mathcal{T}(t\boldsymbol{a} + (1-t)\mathbb{F}^{-1}\boldsymbol{A})\,\mathbb{D}^{-1}\|\;\|\mathbb{D}\,\mathbb{F}^{-1}(\boldsymbol{a} - \mathbb{F}^{-1}\boldsymbol{A})\| \\ &\quad\leq \frac{\tau_3^2}{2}\,\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2\,\|\mathbb{D}\boldsymbol{a}\| \leq \frac{2\tau_3^2}{3}\,\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3 .\end{aligned}$$

Further, $-\nabla^2 f(0) = \mathbb{F}$, $\nabla\mathcal{T}(\boldsymbol{a}) = \frac{1}{2}\langle\nabla^3 f(0), \boldsymbol{a}\otimes\boldsymbol{a}\rangle$. By (D.25) of Lemma D.9 and (D.37)

$$\bigl\|\mathbb{D}^{-1}\{\nabla f(\boldsymbol{a}) + \mathbb{F}\boldsymbol{a} - \nabla\mathcal{T}(\boldsymbol{a})\}\bigr\| \leq \frac{\tau_4}{6}\|\mathbb{D}\boldsymbol{a}\|^3 \leq \frac{(11/9)^3\tau_4}{6}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3 \leq \frac{\tau_4}{3}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3 .$$

Next we bound $\bigl\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{a}) - \nabla g(\boldsymbol{v}^\circ)\}\bigr\|$. As $\nabla g(\boldsymbol{v}^\circ) = 0$, (D.6) and (D.32) imply

$$\begin{aligned}&\bigl\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{a}) - \nabla g(\boldsymbol{v}^\circ)\}\bigr\| = \bigl\|\mathbb{D}^{-1}\nabla g(\boldsymbol{a})\bigr\| = \bigl\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{a}) + \mathbb{F}\boldsymbol{a} - \nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A}) - \boldsymbol{A}\}\bigr\| \\ &\quad\leq \bigl\|\mathbb{D}^{-1}\{\nabla f(\boldsymbol{a}) + \mathbb{F}\boldsymbol{a} - \nabla\mathcal{T}(\boldsymbol{a})\}\bigr\| + \|\mathbb{D}^{-1}\{\nabla\mathcal{T}(\boldsymbol{a}) - \nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\}\| \leq \diamondsuit_1 ,\end{aligned} \tag{D.39}$$

where

$$\diamondsuit_1 \stackrel{\text{def}}{=} \frac{\tau_4 + 2\tau_3^2}{3}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3$$

and by (D.31)

$$3\tau_3 \diamondsuit_1 = \tau_3 \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \| \, \tau_4 \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \|^2 + 2\tau_3^3 \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \|^3 < \frac{1}{3} . \tag{D.40}$$

Further, $\nabla^2 g(0) = \nabla^2 f(0) = -\mathbb{F}$, and (D.24) of Lemma D.9 implies

$$\begin{aligned} & \big\| \mathbb{D}^{-1} \{ \nabla g(\boldsymbol{a}) - \nabla g(\boldsymbol{v}^\circ) + \mathbb{F}(\boldsymbol{a} - \boldsymbol{v}^\circ) \} \big\| \\ & \quad = \big\| \mathbb{D}^{-1} \{ \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{v}^\circ) + \mathbb{F}(\boldsymbol{a} - \boldsymbol{v}^\circ) \} \big\| \leq \frac{3\tau_3}{2} \| \mathbb{D}(\boldsymbol{a} - \boldsymbol{v}^\circ) \|^2. \end{aligned}$$

Combining with (D.39) yields in view of $\mathbb{D}^2 \leq \mathbb{F}$

$$\| \mathbb{D}^{-1} \mathbb{F}(\boldsymbol{a} - \boldsymbol{v}^\circ) \| \leq \frac{3\tau_3}{2} \| \mathbb{D}(\boldsymbol{a} - \boldsymbol{v}^\circ) \|^2 + \diamondsuit_1 \leq \frac{3\tau_3}{2} \| \mathbb{D}^{-1} \mathbb{F}(\boldsymbol{a} - \boldsymbol{v}^\circ) \|^2 + \diamondsuit_1 .$$

As $2x \leq \alpha x^2 + \beta$ with $\alpha = 3\tau_3$, $\beta = 2\diamondsuit_1$, and $x \in (0, 1/\alpha)$ implies $x \leq \beta/(2 - \alpha\beta)$, we conclude by (D.40)

$$\| \mathbb{D}^{-1} \mathbb{F}(\boldsymbol{a} - \boldsymbol{v}^\circ) \| \leq \frac{\diamondsuit_1}{1 - 3\tau_3 \diamondsuit_1} \leq \frac{\tau_4 + 2\tau_3^2}{2} \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \|^3 , \tag{D.41}$$

and (D.33) follows.

Next we bound $g(\boldsymbol{v}^\circ) - g(0) = g(\boldsymbol{v}^\circ) - g(\boldsymbol{a}) + g(\boldsymbol{a}) - g(0)$. By (D.38) and $\mathbb{D}^2 \leq \mathbb{F}$

$$\frac{1}{2} \| \mathbb{F}^{-1/2} \boldsymbol{A} \|^2 - \langle \boldsymbol{A}, \boldsymbol{a} \rangle + \frac{1}{2} \| \mathbb{F}^{1/2} \boldsymbol{a} \|^2 = \frac{1}{2} \| \mathbb{F}^{1/2} \boldsymbol{a} - \mathbb{F}^{-1/2} \boldsymbol{A} \|^2 \leq \frac{\tau_3^2}{8} \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \|^4 .$$

This together with $\nabla f(0) = 0$, $-\nabla^2 f(0) = \mathbb{F} \geq \mathbb{D}^2$, $(\mathcal{T}_4^*)$, and (D.37) implies

$$\begin{aligned} & \Big| g(\boldsymbol{a}) - g(0) - \frac{1}{2} \| \mathbb{F}^{-1/2} \boldsymbol{A} \|^2 - \mathcal{T}(\boldsymbol{a}) \Big| \\ & \quad = \Big| f(\boldsymbol{a}) - f(0) + \langle \boldsymbol{A}, \boldsymbol{a} \rangle - \frac{1}{2} \| \mathbb{F}^{-1/2} \boldsymbol{A} \|^2 - \mathcal{T}(\boldsymbol{a}) \Big| \\ & \quad \leq \Big| f(\boldsymbol{a}) - f(0) + \frac{1}{2} \| \mathbb{F}^{1/2} \boldsymbol{a} \|^2 - \mathcal{T}(\boldsymbol{a}) \Big| + \frac{\tau_3^2}{8} \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \|^4 \\ & \quad \leq \frac{\tau_4}{24} \| \mathbb{D}\, \boldsymbol{a} \|^4 + \frac{\tau_3^2}{8} \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \|^4 \leq \Big(\frac{\tau_4}{10} + \frac{\tau_3^2}{8} \Big) \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \|^4 . \end{aligned}$$

Further, by $\nabla g(\boldsymbol{v}^\circ) = 0$ and $\nabla^2 g(\cdot) \equiv \nabla^2 f(\cdot)$, it holds for some $\boldsymbol{v} \in [\boldsymbol{a}, \boldsymbol{v}^\circ]$

$$2 \big| g(\boldsymbol{a}) - g(\boldsymbol{v}^\circ) \big| = \big| \langle \nabla^2 f(\boldsymbol{v}), (\boldsymbol{a} - \boldsymbol{v}^\circ)^{\otimes 2} \rangle \big| .$$

The use of $-\nabla^2 f(0) = \mathbb{F} \geq \mathbb{D}^2$ and (D.23) of Lemma D.9 yields by $\| \mathbb{D} \boldsymbol{v} \| \leq \mathtt{r} = \frac{3}{2} \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \|$, $\tau_3 \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \| < \frac{4}{9}$, and (D.41)

$$\begin{aligned} 2 \big| g(\boldsymbol{a}) - g(\boldsymbol{v}^\circ) \big| & \leq \| \mathbb{F}^{1/2} (\boldsymbol{a} - \boldsymbol{v}^\circ) \|^2 + \big| \big\langle \nabla^2 f(\boldsymbol{v}) - \nabla^2 f(0), (\boldsymbol{a} - \boldsymbol{v}^\circ)^{\otimes 2} \big\rangle \big| \\ & \leq (1 + \tau_3 \mathtt{r}) \| \mathbb{F}^{1/2} (\boldsymbol{a} - \boldsymbol{v}^\circ) \|^2 \leq \frac{(5/3)(\tau_4 + 2\tau_3^2)^2}{4} \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \|^6 . \end{aligned}$$

Moreover, it holds with $\varDelta \stackrel{\mathrm{def}}{=} \mathbb{F}^{-1}\nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})$ for some $t \in [0,1]$

$$
\begin{aligned}
\bigl|\mathcal{T}(\boldsymbol{a}) - \mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\bigr| &= \bigl|\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A} + \varDelta) - \mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\bigr| = \bigl|\bigl\langle \nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A} + t\varDelta), \varDelta\bigr\rangle\bigr| \\
&\le \frac{\tau_3}{2}\|\mathbb{D}(\mathbb{F}^{-1}\boldsymbol{A} + t\varDelta)\|^2 \, \|\mathbb{D}\varDelta\| = \frac{\tau_3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A} + t\,\mathbb{D}\varDelta)\|^2 \, \|\mathbb{D}\varDelta\| \,.
\end{aligned}
$$

As in (D.36) $\|\mathbb{D}\varDelta\| \le \|\mathbb{D}^{-1}\nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\| \le (\tau_3/2)\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2$, and by $\tau_3\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \le 1/2$

$$
\bigl|\mathcal{T}(\boldsymbol{a}) - \mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\bigr| \le \frac{(5/4)^2\tau_3^2}{4}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^4 \,.
$$

Summing up the obtained bounds yields (D.34). (D.35) follows from $(\mathcal{T}_3^*)$. □

D.2.3 Quadratic penalization

Here we discuss the case when $g(\boldsymbol{v}) - f(\boldsymbol{v})$ is quadratic. The general case can be reduced to the situation with $g(\boldsymbol{v}) = f(\boldsymbol{v}) - \|G\boldsymbol{v}\|^2/2$. To make the dependence of G more explicit, denote $f_G(\boldsymbol{v}) \stackrel{\mathrm{def}}{=} f(\boldsymbol{v}) - \|G\boldsymbol{v}\|^2/2$,

$$
\boldsymbol{v}^* = \operatorname*{argmax}_{\boldsymbol{v}} f(\boldsymbol{v}), \quad \boldsymbol{v}_G^* = \operatorname*{argmax}_{\boldsymbol{v}} f_G(\boldsymbol{v}) = \operatorname*{argmax}_{\boldsymbol{v}}\bigl\{f(\boldsymbol{v}) - \|G\boldsymbol{v}\|^2/2\bigr\}.
$$

We study the bias $\boldsymbol{v}_G^* - \boldsymbol{v}^*$ induced by this penalization. To get some intuition, consider first the case of a quadratic function $f(\boldsymbol{v})$.

Lemma D.11. *Let $f(\boldsymbol{v})$ be quadratic with $\mathbb{F} \equiv -\nabla^2 f(\boldsymbol{v})$ and $\mathbb{F}_G = \mathbb{F} + G^2$. Then*

$$
\begin{aligned}
\boldsymbol{v}_G^* - \boldsymbol{v}^* &= -\mathbb{F}_G^{-1}G^2\boldsymbol{v}^*, \\
f_G(\boldsymbol{v}_G^*) - f_G(\boldsymbol{v}^*) &= \frac{1}{2}\|\mathbb{F}_G^{-1/2}G^2\boldsymbol{v}^*\|^2 \,.
\end{aligned}
$$

Proof. Quadraticity of $f(\boldsymbol{v})$ implies quadraticity of $f_G(\boldsymbol{v})$ with $\nabla^2 f_G(\boldsymbol{v}) \equiv -\mathbb{F}_G$ and

$$
\nabla f_G(\boldsymbol{v}_G^*) - \nabla f_G(\boldsymbol{v}^*) = -\mathbb{F}_G\,(\boldsymbol{v}_G^* - \boldsymbol{v}^*).
$$

Further, $\nabla f(\boldsymbol{v}^*) = 0$ yielding $\nabla f_G(\boldsymbol{v}^*) = -G^2\boldsymbol{v}^*$. Together with $\nabla f_G(\boldsymbol{v}_G^*) = 0$, this implies $\boldsymbol{v}_G^* - \boldsymbol{v}^* = -\mathbb{F}_G^{-1}G^2\boldsymbol{v}^*$. The Taylor expansion of f_G at $\boldsymbol{v}_G^*$ yields

$$
f_G(\boldsymbol{v}^*) - f_G(\boldsymbol{v}_G^*) = -\frac{1}{2}\|\mathbb{F}_G^{1/2}(\boldsymbol{v}^* - \boldsymbol{v}_G^*)\|^2 = -\frac{1}{2}\|\mathbb{F}_G^{-1/2}G^2\boldsymbol{v}^*\|^2
$$

and the assertion follows. □

Now we turn to the general case with f satisfying $(\mathcal{T}_3^*)$.

Proposition D.12. *Let* $f_G(\boldsymbol{v}) = f(\boldsymbol{v}) - \|G\boldsymbol{v}\|^2/2$ *be concave and follow* $(\mathcal{T}_3^*)$ *with some* $\mathbb{D}^2$, τ_3, *and* $\mathtt{r}$ *satisfying for* $\varkappa > 0$

$$\mathbb{D}^2 \le \varkappa^2 \mathbb{F}_G , \qquad \mathtt{r} \ge 3\mathsf{b}_G/2 , \qquad \varkappa^2 \tau_3 \, \mathsf{b}_G < 4/9,$$

where

$$\mathsf{b}_G = \|\mathbb{D} \, \mathbb{F}_G^{-1} G^2 \boldsymbol{v}^*\| . \tag{D.42}$$

Then

$$\|\mathbb{D}(\boldsymbol{v}_G^* - \boldsymbol{v}^*)\| \le 3\mathsf{b}_G/2. \tag{D.43}$$

Moreover,

$$\begin{aligned} \bigl\|\mathbb{D}^{-1}\mathbb{F}_G(\boldsymbol{v}_G^* - \boldsymbol{v}^* + \mathbb{F}_G^{-1} G^2 \boldsymbol{v}^*)\bigr\| &\le \frac{3\tau_3}{4} \, \mathsf{b}_G^2 , \\ \Bigl| 2 f_G(\boldsymbol{v}_G^*) - 2 f_G(\boldsymbol{v}^*) - \frac{1}{2} \|\mathbb{F}_G^{-1/2} G^2 \boldsymbol{v}^*\|^2 \Bigr| &\le \frac{\tau_3}{2} \, \mathsf{b}_G^3 . \end{aligned}$$

Proof. Define $g_G(\boldsymbol{v})$ by

$$g_G(\boldsymbol{v}) - g_G(\boldsymbol{v}_G^*) = f_G(\boldsymbol{v}) - f_G(\boldsymbol{v}_G^*) + \langle G^2 \boldsymbol{v}^*, \boldsymbol{v} - \boldsymbol{v}_G^* \rangle. \tag{D.44}$$

The function f_G is concave, the same holds for g_G from (D.44). Hence, $\nabla g_G(\boldsymbol{v}^*) = 0$ implies $\boldsymbol{v}^* = \operatorname{argmax} g_G(\boldsymbol{v})$. By definition, $\nabla f_G(\boldsymbol{v}^*) = -G^2 \boldsymbol{v}^*$ yielding $\nabla g_G(\boldsymbol{v}^*) = -G^2 \boldsymbol{v}^* + G^2 \boldsymbol{v}^* = 0$. Now the results follow from Propositions D.8 and D.6 applied with $f(\boldsymbol{v}) = g_G(\boldsymbol{v}) = f_G(\boldsymbol{v}) - \langle \boldsymbol{A}, \boldsymbol{v} \rangle$, $g(\boldsymbol{v}) = f_G(\boldsymbol{v})$, and $\boldsymbol{A} = -G^2 \boldsymbol{v}^*$. □

The bound on the bias can be further improved under fourth-order smoothness of f using the results of Proposition D.10.

Proposition D.13. *Let* f *be concave and* $\boldsymbol{v}^* = \operatorname{argmax}_{\boldsymbol{v}} f(\boldsymbol{v})$. *Let also* $f(\boldsymbol{v})$ *follow* $(\mathcal{T}_3^*)$ *and* $(\mathcal{T}_4^*)$ *with* $\mathbb{F}_G = -\nabla^2 f(\boldsymbol{v}^*) + G^2$ *and some* $\mathbb{D}^2$, τ_3, τ_4, *and* $\mathtt{r}$ *satisfying*

$$\mathbb{D}^2 \le \varkappa^2 \mathbb{F}_G , \quad \mathtt{r} = \frac{3}{2} \mathsf{b}_G , \quad \varkappa^2 \tau_3 \, \mathsf{b}_G < \frac{4}{9} , \quad \varkappa^2 \tau_4 \, \mathsf{b}_G^2 < \frac{1}{3}$$

for b_G *from* (D.42). *Then* (D.43) *holds. Furthermore, define*

$$\boldsymbol{\mu}_G = \mathbb{F}_G^{-1} \{ G^2 \boldsymbol{v}^* + \nabla \mathcal{T}(\mathbb{F}_G^{-1} G^2 \boldsymbol{v}^*) \}$$

with $\mathcal{T}(\boldsymbol{u}) = \frac{1}{6} \langle \nabla^3 f(\boldsymbol{v}^*), \boldsymbol{u}^{\otimes 3} \rangle$ *and* $\nabla \mathcal{T} = \frac{1}{2} \langle \nabla^3 f(\boldsymbol{v}^*), \boldsymbol{u}^{\otimes 2} \rangle$. *Then*

$$\|\mathbb{D}(\boldsymbol{\mu}_G - \mathbb{F}_G^{-1} G^2 \boldsymbol{v}^*)\| \le \frac{\tau_3}{2} \mathsf{b}_G^2 \le \frac{\tau_3 \, \mathtt{r}_G}{3} \, \mathsf{b}_G ,$$

and

$$\|\mathbb{D}^{-1}\mathbb{F}_G(\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^* + \boldsymbol{\mu}_G)\| \le \frac{\tau_4 + 2\varkappa^2\tau_3^2}{2}\,\mathsf{b}_G^3\,,$$

$$\Bigl| f_G(\boldsymbol{\upsilon}_G^*) - f_G(\boldsymbol{\upsilon}^*) - \frac{1}{2}\|\mathbb{F}_G^{-1/2}G^2\boldsymbol{\upsilon}^*\|^2 - \mathcal{T}(\mathbb{F}_G^{-1}G^2\boldsymbol{\upsilon}^*)\Bigr| \le \frac{\tau_4 + 4\varkappa^2\tau_3^2}{8}\mathsf{b}_G^4 + \frac{\varkappa^2\,(\tau_4 + 2\varkappa^2\tau_3^2)^2}{4}\,\mathsf{b}_G^6\,.$$

D.2.4 A smooth penalty

The case of a general smooth penalty $\operatorname{pen}_G(\boldsymbol{\upsilon})$ can be studied similarly to the quadratic case. Denote $f_G(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} f(\boldsymbol{\upsilon}) - \operatorname{pen}_G(\boldsymbol{\upsilon})$,

$$\boldsymbol{\upsilon}^* = \operatorname*{argmax}_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon}), \quad \boldsymbol{\upsilon}_G^* = \operatorname*{argmax}_{\boldsymbol{\upsilon}} f_G(\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\upsilon}}\bigl\{f(\boldsymbol{\upsilon}) - \operatorname{pen}_G(\boldsymbol{\upsilon})\bigr\}.$$

We study the bias $\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*$ induced by this penalization. The statement of Proposition D.12 and its proof can be extended to this situation by redefining $\boldsymbol{M} \stackrel{\text{def}}{=} \nabla\operatorname{pen}_G(\boldsymbol{\upsilon}^*)$.

Proposition D.14. *Let $f_G(\boldsymbol{\upsilon}) = f(\boldsymbol{\upsilon}) - \operatorname{pen}_G(\boldsymbol{\upsilon})$ be concave and follow* **($\mathcal{T}_3^*$)** *with some $\mathbb{D}^2$, τ_3, and $\mathtt{r}$ satisfying for $\varkappa > 0$*

$$\mathbb{D}^2 \le \varkappa^2\,\mathbb{F}_G\,, \qquad \mathtt{r} \ge 3\mathsf{b}_G/2\,, \qquad \varkappa^2\tau_3\,\mathsf{b}_G < 4/9,$$

where

$$\mathsf{b}_G \stackrel{\text{def}}{=} \|\mathbb{D}\,\mathbb{F}_G^{-1}\boldsymbol{M}\|\,, \qquad \boldsymbol{M} \stackrel{\text{def}}{=} \nabla\operatorname{pen}_G(\boldsymbol{\upsilon}^*)\,.$$

Then

$$\|\mathbb{D}(\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*)\| \le 3\mathsf{b}_G/2.$$

Moreover,

$$\bigl\|\mathbb{D}^{-1}\mathbb{F}_G(\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^* + \mathbb{F}_G^{-1}\boldsymbol{M})\bigr\| \le \frac{3\tau_3}{4}\,\mathsf{b}_G^2\,,$$

$$\Bigl|2f_G(\boldsymbol{\upsilon}_G^*) - 2f_G(\boldsymbol{\upsilon}^*) - \frac{1}{2}\|\mathbb{F}_G^{-1/2}\boldsymbol{M}\|^2\Bigr| \le \frac{\tau_3}{2}\,\mathsf{b}_G^3\,.$$

If, in addition, $f_G(\boldsymbol{\upsilon})$ satisfies **($\mathcal{T}_4^*$)** *and $\varkappa^2\tau_4\,\mathsf{b}_G^2 < \frac{1}{3}$, then with $\mathcal{T}_G(\boldsymbol{u}) = \frac{1}{6}\langle\nabla^3 f_G(\boldsymbol{\upsilon}^*), \boldsymbol{u}^{\otimes 3}\rangle$, $\nabla\mathcal{T}_G = \frac{1}{2}\langle\nabla^3 f_G(\boldsymbol{\upsilon}^*), \boldsymbol{u}^{\otimes 2}\rangle$, and*

$$\boldsymbol{\mu}_G = \mathbb{F}_G^{-1}\{\boldsymbol{M} + \nabla\mathcal{T}_G(\mathbb{F}_G^{-1}\boldsymbol{M})\}\,,$$

it holds

$$
\|\mathbb{D}^{-1}\mathbb{F}_G(\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_G^* - \boldsymbol{\mu}_G)\| \le \frac{\tau_4 + 2\varkappa^2\tau_3^2}{2}\,\mathsf{b}_G^3\,,
$$

$$
\Bigl| f_G(\boldsymbol{\upsilon}_G^*) - f_G(\boldsymbol{\upsilon}^*) - \frac{1}{2}\|\mathbb{F}_G^{-1/2}\boldsymbol{M}\|^2 - \mathcal{T}_G(\mathbb{F}_G^{-1}\boldsymbol{M}) \Bigr| \le \frac{\tau_4 + 4\varkappa^2\tau_3^2}{8}\,\mathsf{b}_G^4 + \frac{\varkappa^2\,(\tau_4 + 2\varkappa^2\tau_3^2)^2}{4}\,\mathsf{b}_G^6\,.
$$

D.3 Conditional and marginal optimization

This section describes the problem of conditional/partial and marginal optimization. Consider a function $f(\boldsymbol{\upsilon})$ of a parameter $\boldsymbol{\upsilon} \in \mathbb{R}^{\bar{p}}$ which can be represented as $\boldsymbol{\upsilon} = (\boldsymbol{x}, \boldsymbol{s})$, where $\boldsymbol{x} \in \mathbb{R}^p$ is the target subvector while $\boldsymbol{s} \in \mathbb{R}^q$ is a nuisance variable. Our goal is to study the solution to the optimization problem $\boldsymbol{\upsilon}^* = (\boldsymbol{x}^*, \boldsymbol{s}^*) = \operatorname{argmax}_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon})$ and, in particular, its target component $\boldsymbol{x}^*$. Later we consider a localized setup with a set Υ of pairs $\boldsymbol{\upsilon} = (\boldsymbol{x}, \boldsymbol{s})$ to be fixed around $\boldsymbol{\upsilon}^*$.

D.3.1 Partial optimization

For any fixed value of the nuisance variable $\boldsymbol{s} \in \mathcal{S}$, consider $f_{\boldsymbol{s}}(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{s})$ as a function of $\boldsymbol{x}$ only. Below we assume that $f_{\boldsymbol{s}}(\boldsymbol{x})$ is concave in $\boldsymbol{x}$ for any $\boldsymbol{s} \in \mathcal{S}$. Define

$$
\boldsymbol{x}_{\boldsymbol{s}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{x}\colon (\boldsymbol{x},\boldsymbol{s}) \in \Upsilon} f_{\boldsymbol{s}}(\boldsymbol{x})\,.
$$

Our goal is to describe variability of the partial solution $\boldsymbol{x}_{\boldsymbol{s}}$ in $\boldsymbol{s}$ in terms of $\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^*$ and $f(\boldsymbol{\upsilon}^*) - f_{\boldsymbol{s}}(\boldsymbol{x}_{\boldsymbol{s}})$. Introduce

$$
\begin{aligned}
\boldsymbol{A}_{\boldsymbol{s}} &\stackrel{\text{def}}{=} \nabla f_{\boldsymbol{s}}(\boldsymbol{x}^*) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{s})\,, \\
\mathbb{F}_{\boldsymbol{s}} &\stackrel{\text{def}}{=} -\nabla^2 f_{\boldsymbol{s}}(\boldsymbol{x}^*) = -\nabla^2_{\boldsymbol{x}\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{s})\,.
\end{aligned}
\tag{D.45}
$$

Local smoothness of each function $f_{\boldsymbol{s}}(\cdot)$ around $\boldsymbol{x}_{\boldsymbol{s}}$ can be well described under the self-concordance property. Let for any $\boldsymbol{s} \in \mathcal{S}$, some radius $\mathtt{r}_{\boldsymbol{s}}$ be fixed. We also assume that a local metric on $\mathbb{R}^p$ for the target variable $\boldsymbol{x}$ is defined by a matrix $\mathbb{D}_{\boldsymbol{s}} \in \mathfrak{M}_p$ that may depend on $\boldsymbol{s} \in \mathcal{S}$. Later we assume $(\mathcal{T}_3^*)$ to be fulfilled for all $f_{\boldsymbol{s}}$, $\boldsymbol{s} \in \mathcal{S}$.

$(\boldsymbol{\mathcal{T}_{3|s}^*})$ *For* $\boldsymbol{s} \in \mathcal{S}$, *it holds*

$$
\sup_{\boldsymbol{u} \in \mathbb{R}^p\colon \|\mathbb{D}_{\boldsymbol{s}}\boldsymbol{u}\| \le \mathtt{r}_{\boldsymbol{s}}} \ \sup_{\boldsymbol{z} \in \mathbb{R}^p} \frac{\bigl|\langle \nabla^3 f_{\boldsymbol{s}}(\boldsymbol{x}_{\boldsymbol{s}} + \boldsymbol{u}), \boldsymbol{z}^{\otimes 3} \rangle\bigr|}{\|\mathbb{D}_{\boldsymbol{s}}\boldsymbol{z}\|^3} \le \tau_3\,.
$$

Our first result describes the *semiparametric bias* $\boldsymbol{x}^* - \boldsymbol{x_s}$ caused by using the value $\boldsymbol{s}$ of the nuisance variable in place of $\boldsymbol{s}^*$.

Proposition D.15. *Let* $f_{\boldsymbol{s}}(\boldsymbol{x})$ *be a strongly concave function with* $f_{\boldsymbol{s}}(\boldsymbol{x_s}) = \max_{\boldsymbol{x}} f_{\boldsymbol{s}}(\boldsymbol{x})$ *and* $\mathbb{F}_{\boldsymbol{s}} = -\nabla^2 f_{\boldsymbol{s}}(\boldsymbol{x_s})$. *Let* $\boldsymbol{A_s}$ *and* $\mathbb{F}_{\boldsymbol{s}}$ *be given by* (D.45). *Assume* $(\boldsymbol{\mathcal{T}^*_{3|s}})$ *at* $\boldsymbol{x_s}$ *with* $\mathbb{D}^2_{\boldsymbol{s}}$, $\mathtt{r}_{\boldsymbol{s}}$, *and* τ_3 *such that*

$$\mathbb{D}^2_{\boldsymbol{s}} \le \varkappa^2 \, \mathbb{F}_{\boldsymbol{s}}, \quad \mathtt{r}_{\boldsymbol{s}} \ge \frac{3}{2} \| \mathbb{D}_{\boldsymbol{s}} \, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s} \| , \quad \varkappa^2 \tau_3 \| \mathbb{D}_{\boldsymbol{s}} \, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s} \| < \frac{4}{9} .$$

Then $\| \mathbb{D}_{\boldsymbol{s}} (\boldsymbol{x_s} - \boldsymbol{x}^*) \| \le (3/2) \| \mathbb{D}_{\boldsymbol{s}} \, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s} \|$ *and moreover,*

$$\| \mathbb{D}_{\boldsymbol{s}}^{-1} \mathbb{F}_{\boldsymbol{s}} (\boldsymbol{x_s} - \boldsymbol{x}^* - \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s}) \| \le \frac{3\tau_3}{4} \| \mathbb{D}_{\boldsymbol{s}} \, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s} \|^2 \le \frac{\tau_3 \, \mathtt{r}_{\boldsymbol{s}}}{2} \| \mathbb{D}_{\boldsymbol{s}} \, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s} \| . \tag{D.46}$$

Moreover,

$$\bigl| 2 f_{\boldsymbol{s}}(\boldsymbol{x_s}) - 2 f_{\boldsymbol{s}}(\boldsymbol{x}^*) - \| \mathbb{F}_{\boldsymbol{s}}^{-1/2} \boldsymbol{A_s} \|^2 \bigr| \le \frac{5\tau_3}{2} \, \| \mathbb{D}_{\boldsymbol{s}} \, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s} \|^3 . \tag{D.47}$$

Proof. Define $g_{\boldsymbol{s}}(\boldsymbol{x}) = f_{\boldsymbol{s}}(\boldsymbol{x}) - \langle \boldsymbol{x}, \boldsymbol{A_s} \rangle$. Then $g_{\boldsymbol{s}}$ is concave and $\nabla g_{\boldsymbol{s}}(\boldsymbol{x}^*) = 0$ yielding $\boldsymbol{x}^* = \operatorname{argmax}_{\boldsymbol{x}} g_{\boldsymbol{s}}(\boldsymbol{x})$. Now Proposition D.8 yields (D.46). Also by (D.17)

$$\Bigl| 2 g_{\boldsymbol{s}}(\boldsymbol{x}^*) - 2 g_{\boldsymbol{s}}(\boldsymbol{x_s}) - \| \mathbb{F}_{\boldsymbol{s}}^{-1/2} \boldsymbol{A_s} \|^2 \Bigr| \le \frac{\tau_3}{2} \, \| \mathbb{D}_{\boldsymbol{s}} \, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s} \|^3$$

yielding

$$\bigl| 2 f_{\boldsymbol{s}}(\boldsymbol{x}^*) - 2 f_{\boldsymbol{s}}(\boldsymbol{x_s}) - 2 \langle \boldsymbol{x}^* - \boldsymbol{x_s}, \boldsymbol{A_s} \rangle - \| \mathbb{F}_{\boldsymbol{s}}^{-1/2} \boldsymbol{A_s} \|^2 \bigr| \le \frac{\tau_3}{2} \, \| \mathbb{D}_{\boldsymbol{s}} \, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s} \|^3 . \tag{D.48}$$

By (D.46)

$$\begin{aligned}
\bigl| \langle \boldsymbol{x}^* - \boldsymbol{x_s}, \boldsymbol{A_s} \rangle + \| \mathbb{F}_{\boldsymbol{s}}^{-1/2} \boldsymbol{A_s} \|^2 \bigr| &= \bigl| \langle \mathbb{F}_{\boldsymbol{s}}^{1/2} (\boldsymbol{x}^* - \boldsymbol{x_s}), \mathbb{F}_{\boldsymbol{s}}^{-1/2} \boldsymbol{A_s} \rangle + \langle \mathbb{F}_{\boldsymbol{s}}^{-1/2} \boldsymbol{A_s}, \mathbb{F}_{\boldsymbol{s}}^{-1/2} \boldsymbol{A_s} \rangle \bigr| \\
&= \bigl| \langle \mathbb{F}_{\boldsymbol{s}}^{1/2} (\boldsymbol{x}^* - \boldsymbol{x_s} + \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s}), \mathbb{F}_{\boldsymbol{s}}^{-1/2} \boldsymbol{A_s} \rangle \bigr| \\
&\le \| \mathbb{D}_{\boldsymbol{s}}^{-1} \, \mathbb{F}_{\boldsymbol{s}} (\boldsymbol{x}^* - \boldsymbol{x_s} + \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s}) \| \, \| \mathbb{D}_{\boldsymbol{s}} \, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s} \| \le \frac{3\tau_3}{4} \, \| \mathbb{D}_{\boldsymbol{s}} \, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A_s} \|^3
\end{aligned}$$

This and (D.48) imply (D.47). □

D.3.2 Conditional optimization under (semi)orthogonality

Here we study variability of the value $\boldsymbol{x_s} = \operatorname{argmax}_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{s})$ w.r.t. the nuisance parameter $\boldsymbol{s}$. It appears that local quadratic approximation of the function f in a vicinity of the extreme point $\boldsymbol{\upsilon}^*$ yields a nearly linear dependence of $\boldsymbol{x_s}$ on $\boldsymbol{s}$. We illustrate this fact on the case of a quadratic function $f(\cdot)$. Consider the negative Hessian $\mathscr{F} = -\nabla^2 f(\boldsymbol{v}^*)$ in the block form:

$$\mathscr{F} \stackrel{\text{def}}{=} -\nabla^2 f(\boldsymbol{v}^*) = \begin{pmatrix} \mathscr{F}_{\boldsymbol{x}\boldsymbol{x}} & \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} \\ \mathscr{F}_{\boldsymbol{s}\boldsymbol{x}} & \mathscr{F}_{\boldsymbol{s}\boldsymbol{s}} \end{pmatrix}$$

with $\mathscr{F}_{\boldsymbol{s}\boldsymbol{x}} = \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}}^\top$. If $f(\boldsymbol{v})$ is quadratic then $\mathscr{F}$ and its blocks do not depend on $\boldsymbol{v}$.

Lemma D.16. *Let $f(\boldsymbol{v})$ be quadratic, strongly concave, and $\nabla f(\boldsymbol{v}^*) = 0$. Then*

$$\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^* = -\mathscr{F}_{\boldsymbol{x}\boldsymbol{x}}^{-1} \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} (\boldsymbol{s} - \boldsymbol{s}^*). \tag{D.49}$$

Proof. The condition $\nabla f(\boldsymbol{v}^*) = 0$ implies $f(\boldsymbol{v}) = f(\boldsymbol{v}^*) - (\boldsymbol{v} - \boldsymbol{v}^*)^\top \mathscr{F} (\boldsymbol{v} - \boldsymbol{v}^*)/2$ with $\mathscr{F} = -\nabla^2 f(\boldsymbol{v}^*)$. For $\boldsymbol{s}$ fixed, the point $\boldsymbol{x}_{\boldsymbol{s}}$ maximizes $-(\boldsymbol{x} - \boldsymbol{x}^*)^\top \mathscr{F}_{\boldsymbol{x}\boldsymbol{x}} (\boldsymbol{x} - \boldsymbol{x}^*)/2 - (\boldsymbol{x} - \boldsymbol{x}^*)^\top \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} (\boldsymbol{s} - \boldsymbol{s}^*)$ and thus, $\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^* = -\mathscr{F}_{\boldsymbol{x}\boldsymbol{x}}^{-1} \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} (\boldsymbol{s} - \boldsymbol{s}^*)$. □

This observation (D.49) is in fact discouraging because the bias $\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^*$ has the same magnitude as the nuisance parameter $\boldsymbol{s} - \boldsymbol{s}^*$. However, the condition $\mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} = 0$ yields $\boldsymbol{x}_{\boldsymbol{s}} \equiv \boldsymbol{x}^*$ and the bias vanishes. If $f(\boldsymbol{v})$ is not quadratic, the *orthogonality* condition $\nabla_{\boldsymbol{s}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{s}) \equiv 0$ for all $(\boldsymbol{x}, \boldsymbol{s}) \in \mathcal{W}$ still ensures a vanishing bias.

Lemma D.17. *Let $f(\boldsymbol{x}, \boldsymbol{s})$ be continuously differentiable and $\nabla_{\boldsymbol{s}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{s}) \equiv 0$. Then the point $\boldsymbol{x}_{\boldsymbol{s}} = \operatorname{argmax}_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{s})$ does not depend on $\boldsymbol{s}$.*

Proof. The condition $\nabla_{\boldsymbol{s}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{s}) \equiv 0$ implies the decomposition $f(\boldsymbol{x}, \boldsymbol{s}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{s})$ for some functions f_1 and f_2. This in turn yields $\boldsymbol{x}_{\boldsymbol{s}} \equiv \boldsymbol{x}^*$. □

In some cases, one can check *semi-orthogonality* condition

$$\nabla_{\boldsymbol{s}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{s}) = 0, \qquad \forall \boldsymbol{s} \in \mathcal{S}. \tag{D.50}$$

Lemma D.18. *Assume (D.50). Then*

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{s}) \equiv 0, \qquad \nabla^2_{\boldsymbol{x}\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{s}) \equiv \nabla^2_{\boldsymbol{x}\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{s}^*), \qquad \boldsymbol{s} \in \mathcal{S}. \tag{D.51}$$

Moreover, if $f(\boldsymbol{x}, \boldsymbol{s})$ is concave in $\boldsymbol{x}$ given $\boldsymbol{s}$ then

$$\boldsymbol{x}_{\boldsymbol{s}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{s}) \equiv \boldsymbol{x}^*, \qquad \forall \boldsymbol{s} \in \mathcal{S}. \tag{D.52}$$

Proof. Consider the vector $\boldsymbol{A}_{\boldsymbol{s}} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{s})$. Obviously $\boldsymbol{A}_{\boldsymbol{s}^*} = 0$. Moreover, (D.50) implies that $\boldsymbol{A}_{\boldsymbol{s}}$ does not depend on $\boldsymbol{s}$ and thus, vanishes everywhere. As f is concave in $\boldsymbol{x}$, this implies $f(\boldsymbol{x}^*, \boldsymbol{s}) = \max_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{s})$ and $\boldsymbol{x}_{\boldsymbol{s}} = \operatorname{argmax}_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{s}) \equiv \boldsymbol{x}^*$. Similarly, by (D.50), it holds $\nabla_{\boldsymbol{s}} \nabla_{\boldsymbol{x}\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{s}) \equiv 0$ and (D.51) follows. Concavity of $f(\boldsymbol{x}, \boldsymbol{s})$ in $\boldsymbol{x}$ for $\boldsymbol{s}$ fixed and (D.51) imply (D.52). $\square$

Orthogonality or semi-orthogonality (D.50) conditions are rather restrictive and fulfilled only in special situations. A weaker condition of *one-point orthogonality* means $\nabla_{\boldsymbol{s}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{s}^*) = 0$. This condition is not restrictive and can always be enforced by a linear transform of the nuisance variable $\boldsymbol{s}$.

D.3.3 Semiparametric bias

This section explains how one can bound variability of $\mathbb{F}_{\boldsymbol{s}} = -\nabla^2_{\boldsymbol{x}\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{s})$ and of the norm of $\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^*$ under conditions on the cross-derivatives of $f(\boldsymbol{x}, \boldsymbol{s})$ for $\boldsymbol{s} = \boldsymbol{s}^*$ or $\boldsymbol{x} = \boldsymbol{x}^*$. Suppose that the nuisance variable $\boldsymbol{s}$ is already localized to a small vicinity $\mathcal{S}$ of $\boldsymbol{s}^*$. A typical example is given by the level set $\mathcal{S}$ of the form

$$\mathcal{S} = \bigl\{\boldsymbol{s} \colon \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*)\|_{\circ} \le \mathtt{r}_{\circ}\bigr\}, \tag{D.53}$$

where $\mathbb{H}$ is a metric tensor in $\mathbb{R}^q$, $\|\cdot\|_{\circ}$ is a norm in $\mathbb{R}^q$, and $\mathtt{r}_{\circ} > 0$. Often $\|\cdot\|_{\circ}$ is the usual ℓ_2-norm. However, in some situations, it is beneficial to use different topology for the target parameter $\boldsymbol{x}$ and the nuisance parameter $\boldsymbol{s}$. One example is given by estimation in $\sup$-norm with $\|\cdot\|_{\circ} = \|\cdot\|_{\infty}$.

Assume the following condition.

$(\boldsymbol{\mathcal{T}^*_{3,\mathcal{S}}})$ *It holds with some* τ_{12}, τ_{21}

$$\sup_{\boldsymbol{s} \in \mathcal{S}} \sup_{\boldsymbol{z} \in \mathbb{R}^p, \boldsymbol{w} \in \mathbb{R}^q} \frac{|\langle \nabla^3_{\boldsymbol{x}\boldsymbol{s}\boldsymbol{s}} f(\boldsymbol{x}^*, \boldsymbol{s}), \boldsymbol{z} \otimes \boldsymbol{w}^{\otimes 2} \rangle|}{\|\mathbb{D}\boldsymbol{z}\| \, \|\mathbb{H}\boldsymbol{w}\|_{\circ}^2} \le \tau_{12}, \tag{D.54}$$

$$\sup_{\boldsymbol{s} \in \mathcal{S}} \sup_{\boldsymbol{z} \in \mathbb{R}^p, \boldsymbol{w} \in \mathbb{R}^q} \frac{|\langle \nabla^3_{\boldsymbol{x}\boldsymbol{x}\boldsymbol{s}} f(\boldsymbol{x}^*, \boldsymbol{s}), \boldsymbol{z}^{\otimes 2} \otimes \boldsymbol{w} \rangle|}{\|\mathbb{D}\boldsymbol{z}\|^2 \, \|\mathbb{H}\boldsymbol{w}\|_{\circ}} \le \tau_{21}. \tag{D.55}$$

Remark D.3. Later we establish some bounds on the semiparametric bias assuming $(\tau_{12} + \tau_{21})\, \mathtt{r}_{\circ} \ll 1$.

Remark D.4. Condition $(\boldsymbol{\mathcal{T}^*_{3,\mathcal{S}}})$ only involves mixed derivatives $\nabla^3_{\boldsymbol{x}\boldsymbol{s}\boldsymbol{s}} f(\boldsymbol{x}^*, \boldsymbol{s})$ and $\nabla^3_{\boldsymbol{x}\boldsymbol{x}\boldsymbol{s}} f(\boldsymbol{x}^*, \boldsymbol{s})$ of $f(\boldsymbol{x}^*, \boldsymbol{s})$ for the fixed value $\boldsymbol{x} = \boldsymbol{x}^*$, while condition $(\boldsymbol{\mathcal{T}^*_{3|\boldsymbol{s}}})$ only concerns smoothness of $f(\boldsymbol{x}, \boldsymbol{s})$ w.r.t. $\boldsymbol{x}$ for $\boldsymbol{s}$ fixed. Therefore, the combination of $(\boldsymbol{\mathcal{T}^*_{3|\boldsymbol{s}}})$ and $(\boldsymbol{\mathcal{T}^*_{3,\mathcal{S}}})$ is much weaker than the full dimensional condition $(\boldsymbol{\mathcal{T}^*_{3}})$.

The next result provides an expansion of the semiparametric bias $\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^*$. For the norm $\|\cdot\|_\circ$, let $\|B\|_*$ be the corresponding dual norm of an operator $B\colon \mathbb{R}^q \to \mathbb{R}^p$:

$$\|B\|_* = \sup_{\boldsymbol{z}\colon \|\boldsymbol{z}\|_\circ \le 1} \|B\boldsymbol{z}\|. \tag{D.56}$$

If $p = 1$ and $\|\cdot\|_\circ$ is the sup-norm $\|\cdot\|_\infty$ then $\|B\|_* = \|B\|_1$. Define

$$\rho_* \stackrel{\text{def}}{=} \|\mathbb{D}^{-1}\mathscr{F}_{\boldsymbol{x}\boldsymbol{s}}\,\mathbb{H}^{-1}\|_*\,, \tag{D.57}$$

$$\rho_2 \stackrel{\text{def}}{=} \|\mathbb{D}^{-1}\mathscr{F}_{\boldsymbol{x}\boldsymbol{s}}\,\mathbb{H}^{-1}\|_* + \tau_{12}\,\mathtt{r}_\circ/2\,. \tag{D.58}$$

Proposition D.15 evaluates the semiparametric bias $\boldsymbol{x}^* - \boldsymbol{x}_{\boldsymbol{s}}$ in terms of the matrix $\mathbb{F}_{\boldsymbol{s}} = -\nabla^2_{\boldsymbol{x}\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{s})$ and the vector $\boldsymbol{A}_{\boldsymbol{s}} = \nabla f_{\boldsymbol{s}}(\boldsymbol{x}^*)$. The next result explains how these quantities can be controlled under $(\mathcal{T}^*_{3,\mathcal{S}})$. To simplify the formulation, we assume that $\mathbb{D}^2 \le \mathbb{F}$; see Remark D.5 for an extension.

Proposition D.19. *Assume $(\mathcal{T}^*_{3,\mathcal{S}})$ with $\mathbb{D}^2 \le \mathbb{F}$. Fix $\boldsymbol{s} \in \mathcal{S}$, set $\omega \stackrel{\text{def}}{=} \tau_{21}\|\mathbb{H}(\boldsymbol{s}-\boldsymbol{s}^*)\|_\circ$, and assume $(\mathcal{T}^*_{3|\boldsymbol{s}})$ with $\mathbb{D}_{\boldsymbol{s}} \equiv \mathbb{D}$, $\mathtt{r}_{\boldsymbol{s}} \equiv \mathtt{r}$, and τ_3, such that*

$$\omega < 1\,, \quad \mathtt{r} \ge \frac{3\rho_2}{2(1-\omega)}\,\|\mathbb{H}(\boldsymbol{s}-\boldsymbol{s}^*)\|_\circ\,, \quad \frac{\rho_2\,\tau_3}{1-\omega}\,\|\mathbb{H}(\boldsymbol{s}-\boldsymbol{s}^*)\|_\circ \le \frac{4}{9}\,,$$

for ρ_2 from (D.58). Then the partial solution $\boldsymbol{x}_{\boldsymbol{s}}$ obeys

$$\|\mathbb{D}(\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^*)\| \le \frac{3\rho_2}{2(1-\omega)}\,\|\mathbb{H}(\boldsymbol{s}-\boldsymbol{s}^*)\|_\circ\,. \tag{D.59}$$

Moreover, with ρ_* from (D.57), it holds

$$\|Q\{\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^* + \mathbb{F}^{-1}\mathscr{F}_{\boldsymbol{x}\boldsymbol{s}}(\boldsymbol{s}-\boldsymbol{s}^*)\}\| \le \|Q\,\mathbb{F}^{-1}\mathbb{D}\|\,\tau_\circ\,\|\mathbb{H}(\boldsymbol{s}-\boldsymbol{s}^*)\|_\circ^2\,, \tag{D.60}$$

$$\tau_\circ \stackrel{\text{def}}{=} \frac{1}{1-\omega}\left(\rho_*\,\tau_{21} + \frac{\tau_{12}}{2} + \frac{3\rho_2^2\,\tau_3}{4(1-\omega)^2}\right). \tag{D.61}$$

Proof. First we bound the variability of $\mathbb{F}_{\boldsymbol{s}}$ over $\mathcal{S}$. Let $\mathbb{F} = \mathbb{F}_{\boldsymbol{s}^*}$.

Lemma D.20. *Assume $(\mathcal{T}^*_{3,\mathcal{S}})$. Then for any $\boldsymbol{s} \in \mathcal{S}$*

$$\|\mathbb{D}^{-1}\,(\mathbb{F}_{\boldsymbol{s}} - \mathbb{F})\,\mathbb{D}^{-1}\| \le \tau_{21}\,\|\mathbb{H}(\boldsymbol{s}-\boldsymbol{s}^*)\|_\circ\,. \tag{D.62}$$

Proof. Let $\boldsymbol{s} \in \mathcal{S}$. By (D.55), for any $\boldsymbol{z} \in \mathbb{R}^p$, it holds

$$\bigl|\bigl\langle \mathbb{D}^{-1}\,(\mathbb{F}_{\boldsymbol{s}} - \mathbb{F})\,\mathbb{D}^{-1}, \boldsymbol{z}^{\otimes 2}\bigr\rangle\bigr| = \bigl|\bigl\langle \mathbb{F}_{\boldsymbol{s}} - \mathbb{F}, (\mathbb{D}^{-1}\boldsymbol{z})^{\otimes 2}\bigr\rangle\bigr|$$

$$\le \sup_{t\in[0,1]} \bigl|\langle \nabla^3_{\boldsymbol{x}\boldsymbol{x}\boldsymbol{s}} f(\boldsymbol{x}^*, \boldsymbol{s}^* + t(\boldsymbol{s}-\boldsymbol{s}^*)), (\mathbb{D}^{-1}\boldsymbol{z})^{\otimes 2} \otimes (\boldsymbol{s}-\boldsymbol{s}^*)\rangle\bigr| \le \tau_{21}\,\|\boldsymbol{z}\|^2\,\|\mathbb{H}(\boldsymbol{s}-\boldsymbol{s}^*)\|_\circ\,.$$

This yields (D.62). $\square$

The next result describes some corollaries of (D.62).

Lemma D.21. *Assume* $\mathbb{D}^2 \le \varkappa^2 \mathbb{F}$ *and let some other matrix* $\mathbb{F}_1 \in \mathfrak{M}_p$ *satisfy*

$$\|\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\,\mathbb{D}^{-1}\| \le \varkappa^{-2}\omega \tag{D.63}$$

with $\omega < 1$ *. Then*

$$\|\mathbb{F}^{-1/2}(\mathbb{F}_1 - \mathbb{F})\,\mathbb{F}^{-1/2}\| \;\le\; \omega\,, \tag{D.64}$$

$$\|\mathbb{F}^{1/2}(\mathbb{F}_1^{-1} - \mathbb{F}^{-1})\,\mathbb{F}^{1/2}\| \le \frac{\omega}{1-\omega}\,, \tag{D.65}$$

and

$$\frac{1}{1+\omega}\,\|\mathbb{D}\,\mathbb{F}^{-1}\,\mathbb{D}\| \le \|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{D}\| \le \frac{1}{1-\omega}\,\|\mathbb{D}\,\mathbb{F}^{-1}\,\mathbb{D}\|\,. \tag{D.66}$$

Furthermore, for any vector $\boldsymbol{u}$

$$(1-\omega)\|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\| \le \|\mathbb{D}^{-1}\mathbb{F}_1\boldsymbol{u}\| \le (1+\omega)\|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\|\,, \tag{D.67}$$

$$\frac{1-2\omega}{1-\omega}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{u}\| \le \|\mathbb{D}\,\mathbb{F}_1^{-1}\boldsymbol{u}\| \le \frac{1}{1-\omega}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{u}\|\,. \tag{D.68}$$

Proof. Statement (D.64) follows from (D.63) because of $\mathbb{F}^{-1} \le \varkappa^2\,\mathbb{D}^{-2}$. Define now $\boldsymbol{U} \stackrel{\text{def}}{=} \mathbb{F}^{-1/2}(\mathbb{F}_1 - \mathbb{F})\,\mathbb{F}^{-1/2}$. Then $\|\boldsymbol{U}\| \le \omega$ and

$$\|\mathbb{F}^{1/2}(\mathbb{F}_1^{-1} - \mathbb{F}^{-1})\,\mathbb{F}^{1/2}\| = \|(I\!\!I + \boldsymbol{U})^{-1} - I\!\!I\| \le \frac{1}{1-\omega}\|\boldsymbol{U}\|$$

yielding (D.65). Further,

$$\begin{aligned}
\|\mathbb{D}\,(\mathbb{F}_1^{-1} - \mathbb{F}^{-1})\,\mathbb{D}\| &= \|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{F}_1(\mathbb{F}_1^{-1} - \mathbb{F}^{-1})\,\mathbb{F}\,\mathbb{F}^{-1}\mathbb{D}\| \\
&= \|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{D}\,\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{D}^{-1}\,\mathbb{D}\,\mathbb{F}^{-1}\mathbb{D}\| \\
&\le \|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{D}\|\,\|\mathbb{D}\,\mathbb{F}^{-1}\mathbb{D}\|\,\|\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{D}^{-1}\| \le \omega\|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{D}\|\,.
\end{aligned}$$

This implies (D.66). Also, by $\mathbb{D}^2 \le \varkappa^2\mathbb{F}$

$$\begin{aligned}
\|\mathbb{D}^{-1}\mathbb{F}_1\boldsymbol{u}\| &\le \|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\| + \|\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{D}^{-1}\mathbb{D}\boldsymbol{u}\| \le \|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\| + \omega\,\|\mathbb{D}\boldsymbol{u}\| \\
&\le \|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\| + \omega\|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\| \le (1+\omega)\|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\|\,,
\end{aligned}$$

and (D.67) follows. Similarly

$$\begin{aligned}
\|\mathbb{D}\,(\mathbb{F}_1^{-1} - \mathbb{F}^{-1})\boldsymbol{u}\| &= \|\mathbb{D}\,\mathbb{F}_1^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{F}^{-1}\boldsymbol{u}\| = \|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{D}\;\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{D}^{-1}\,\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{u}\| \\
&\le \|\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\,\mathbb{D}^{-1}\|\;\|\mathbb{D}\,\mathbb{F}_1^{-1}\,\mathbb{D}\|\;\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{u}\| \le \frac{\omega}{1-\omega}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{u}\|
\end{aligned}$$

and (D.68) follows as well. $\square$

Definition (D.56) implies the following bound.

Lemma D.22. *For any* $\boldsymbol{s} \in \mathcal{S}$ *and any linear mapping* Q *, it holds*

$$\|Q \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} (\boldsymbol{s} - \boldsymbol{s}^*)\| \leq \|Q \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} \mathbb{H}^{-1}\|_* \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*)\|_\circ .$$

The next lemma shows that $\boldsymbol{A}_{\boldsymbol{s}} = \nabla f_{\boldsymbol{s}}(\boldsymbol{x}^*)$ is nearly linear in $\boldsymbol{s}$.

Lemma D.23. *Assume* (D.54). *Then*

$$\|\mathbb{D}^{-1}\{\boldsymbol{A}_{\boldsymbol{s}} - \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}}(\boldsymbol{s} - \boldsymbol{s}^*)\}\| \leq \frac{\tau_{12}}{2} \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*)\|_\circ^2 . \tag{D.69}$$

and with ρ_2 *from* (D.58), *it holds*

$$\begin{aligned} \|\mathbb{D}^{-1}\boldsymbol{A}_{\boldsymbol{s}}\| &\leq \Bigl(\|\mathbb{D}^{-1}\mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} \mathbb{H}^{-1}\|_* + \frac{\tau_{12}}{2} \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*)\|_\circ \Bigr) \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*)\|_\circ \\ &\leq \rho_2 \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*)\|_\circ . \end{aligned} \tag{D.70}$$

Proof. Fix $\boldsymbol{s} \in \mathcal{S}$ and define $\boldsymbol{a}_{\boldsymbol{s}}(t) \stackrel{\text{def}}{=} \boldsymbol{A}_{\boldsymbol{s}^* + t(\boldsymbol{s} - \boldsymbol{s}^*)}$. Then $\boldsymbol{a}_{\boldsymbol{s}}(0) = \boldsymbol{A}_{\boldsymbol{s}^*} = 0$, $\boldsymbol{a}_{\boldsymbol{s}}(1) = \boldsymbol{A}_{\boldsymbol{s}}$, and

$$\boldsymbol{a}_{\boldsymbol{s}}(1) - \boldsymbol{a}_{\boldsymbol{s}}(0) = \int_0^1 \boldsymbol{a}_{\boldsymbol{s}}'(t) \, dt ,$$

where $\boldsymbol{a}_{\boldsymbol{s}}'(t) = \frac{d}{dt}\boldsymbol{a}_{\boldsymbol{s}}(t)$ for $t \in [0,1]$. Similarly, by $\boldsymbol{a}_{\boldsymbol{s}}'(0) = \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}}(\boldsymbol{s} - \boldsymbol{s}^*)$, we derive

$$\boldsymbol{a}_{\boldsymbol{s}}(1) - \boldsymbol{a}_{\boldsymbol{s}}(0) - \boldsymbol{a}_{\boldsymbol{s}}'(0) = \int_0^1 (\boldsymbol{a}_{\boldsymbol{s}}'(t) - \boldsymbol{a}_{\boldsymbol{s}}'(0)) \, dt = \int_0^1 (1-t) \, \boldsymbol{a}_{\boldsymbol{s}}''(t) \, dt ,$$

where $\boldsymbol{a}_{\boldsymbol{s}}''(t) = \frac{d^2}{dt^2}\boldsymbol{a}_{\boldsymbol{s}}(t)$. By condition $(\boldsymbol{\mathcal{T}_{3,\mathcal{S}}^*})$

$$\bigl|\langle \boldsymbol{a}_{\boldsymbol{s}}''(t), \boldsymbol{z} \rangle\bigr| = \bigl|\bigl\langle \nabla^3_{\boldsymbol{x}\boldsymbol{s}\boldsymbol{s}} f(\boldsymbol{x}^*, \boldsymbol{s}^* + t(\boldsymbol{s} - \boldsymbol{s}^*)), \boldsymbol{z} \otimes (\boldsymbol{s} - \boldsymbol{s}^*)^{\otimes 2} \bigr\rangle\bigr| \leq \tau_{12} \, \|\mathbb{D}\boldsymbol{z}\| \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*)\|_\circ^2$$

and hence,

$$\begin{aligned} \|\mathbb{D}^{-1}\boldsymbol{a}_{\boldsymbol{s}}''(t)\| &= \sup_{\boldsymbol{z} \colon \|\boldsymbol{z}\| \leq 1} \bigl|\langle \mathbb{D}^{-1}\boldsymbol{a}_{\boldsymbol{s}}''(t), \boldsymbol{z} \rangle\bigr| = \sup_{\boldsymbol{z} \colon \|\boldsymbol{z}\| \leq 1} \bigl|\langle \boldsymbol{a}_{\boldsymbol{s}}''(t), \mathbb{D}^{-1}\boldsymbol{z} \rangle\bigr| \\ &\leq \tau_{12} \sup_{\boldsymbol{z} \colon \|\boldsymbol{z}\| \leq 1} \|\boldsymbol{z}\| \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*)\|_\circ^2 = \tau_{12} \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*)\|_\circ^2 . \end{aligned}$$

This yields

$$\|\mathbb{D}^{-1}\{\boldsymbol{A}_{\boldsymbol{s}} - \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}}(\boldsymbol{s} - \boldsymbol{s}^*)\}\| \leq \tau_{12} \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*)\|_\circ^2 \int_0^1 (1-t) \, dt \leq \frac{\tau_{12} \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*)\|_\circ^2}{2}$$

as claimed in (D.69). Lemma D.22 implies (D.70). $\square$

Now Proposition D.15 helps to show (D.59) and to bound $\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^{*} - \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A}_{\boldsymbol{s}}$.

Lemma D.24. *If* $\omega = \tau_{21} \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^{*})\|_{\circ} < 1$ *then*

$$\|\mathbb{D}(\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^{*})\| \leq \frac{3}{2} \|\mathbb{D}\, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A}_{\boldsymbol{s}}\| \leq \frac{3\rho_2}{2(1-\omega)} \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^{*})\|_{\circ} \, . \tag{D.71}$$

Moreover,

$$\|\mathbb{D}^{-1} \mathbb{F}_{\boldsymbol{s}} (\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^{*} + \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A}_{\boldsymbol{s}})\| \leq \frac{3\rho_2^2 \, \tau_3}{4(1-\omega)^2} \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^{*})\|_{\circ}^2 \, . \tag{D.72}$$

Proof. By $\mathbb{D}^2 \leq \mathbb{F}$, (D.62), (D.68) of Lemma D.21, and (D.70)

$$\|\mathbb{D}\, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A}_{\boldsymbol{s}}\| \leq \|\mathbb{D}\, \mathbb{F}_{\boldsymbol{s}}^{-1} \mathbb{D}\| \; \|\mathbb{D}^{-1} \boldsymbol{A}_{\boldsymbol{s}}\| \leq \frac{1}{1-\omega} \, \|\mathbb{D}^{-1} \boldsymbol{A}_{\boldsymbol{s}}\| \leq \frac{\rho_2}{1-\omega} \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^{*})\|_{\circ} \, .$$

This and the conditions of the proposition enable us to apply Proposition D.15 which implies (D.71) and

$$\|\mathbb{D}^{-1} \mathbb{F}_{\boldsymbol{s}} (\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^{*} - \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A}_{\boldsymbol{s}})\| \leq \frac{3\tau_3}{4} \, \|\mathbb{D}\, \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A}_{\boldsymbol{s}}\|^2 \leq \frac{3\rho_2^2 \, \tau_3}{4(1-\omega)^2} \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^{*})\|_{\circ}^2 \, .$$

This completes the proof. $\square$

Now we can finalize the proof of the proposition. Represent

$$\begin{aligned}
\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^{*} + \mathbb{F}^{-1} \varDelta_{\boldsymbol{s}} &= \boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^{*} + \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A}_{\boldsymbol{s}} - \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A}_{\boldsymbol{s}} + \mathbb{F}^{-1} \varDelta_{\boldsymbol{s}} \\
&= \boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^{*} + \mathbb{F}_{\boldsymbol{s}}^{-1} \boldsymbol{A}_{\boldsymbol{s}} - \mathbb{F}_{\boldsymbol{s}}^{-1} (\boldsymbol{A}_{\boldsymbol{s}} - \varDelta_{\boldsymbol{s}}) - (\mathbb{F}_{\boldsymbol{s}}^{-1} - \mathbb{F}^{-1}) \varDelta_{\boldsymbol{s}} \, .
\end{aligned}$$

The use of (D.62) and (D.57) yields

$$\begin{aligned}
&\|\mathbb{D}^{-1} \mathbb{F}_{\boldsymbol{s}} (\mathbb{F}^{-1} - \mathbb{F}_{\boldsymbol{s}}^{-1}) \varDelta_{\boldsymbol{s}}\| = \|\mathbb{D}^{-1} (\mathbb{F} - \mathbb{F}_{\boldsymbol{s}}) \mathbb{F}^{-1} \varDelta_{\boldsymbol{s}}\| \\
&\leq \|\mathbb{D}^{-1} (\mathbb{F} - \mathbb{F}_{\boldsymbol{s}}) \mathbb{D}^{-1}\| \; \|\mathbb{D}\, \mathbb{F}^{-1} \mathbb{D}\| \; \|\mathbb{D}^{-1} \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} \mathbb{H}^{-1}\|_{*} \; \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^{*})\|_{\circ} \\
&\leq \tau_{21} \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^{*})\|_{\circ} \; \rho_{*} \, \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^{*})\|_{\circ} .
\end{aligned}$$

This together with (D.70) and (D.72) implies

$$\|\mathbb{D}^{-1} \mathbb{F}_{\boldsymbol{s}} (\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^{*} + \mathbb{F}^{-1} \varDelta_{\boldsymbol{s}})\| \leq \biggl\{ \frac{3\rho_2^2 \, \tau_3}{4(1-\omega)^2} + \rho_{*} \, \tau_{21} + \frac{\tau_{12}}{2} \biggr\} \|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^{*})\|_{\circ}^2 \, .$$

The use of (D.67) allows to bound

$$\|\mathbb{D}^{-1} \mathbb{F} (\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^{*} + \mathbb{F}^{-1} \varDelta_{\boldsymbol{s}})\| \leq \frac{1}{1-\omega} \, \|\mathbb{D}^{-1} \mathbb{F}_{\boldsymbol{s}} (\boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^{*} + \mathbb{F}^{-1} \varDelta_{\boldsymbol{s}})\| \, ,$$

and (D.60) follows. $\square$

Remark D.5. An extension to the case $\mathbb{D}^2 \leq \varkappa^2 \mathbb{F}$ can be done by replacing everywhere τ_3, τ_{12}, τ_{21}, $\tau_\circ$, ρ_*, ρ_2 with $\varkappa^3 \tau_3$, $\varkappa \tau_{12}$, $\varkappa^2 \tau_{21}$, $\varkappa \tau_\circ$, $\varkappa^{-1} \rho_*$, $\varkappa^{-1} \rho_2$ respectively.

D.3.4 A linear perturbation

Let $g(\boldsymbol{x}, \boldsymbol{s})$ be a linear perturbation of $f(\boldsymbol{x}, \boldsymbol{s})$:

$$g(\boldsymbol{x}, \boldsymbol{s}) - g(\boldsymbol{x}^*, \boldsymbol{s}^*) = f(\boldsymbol{x}, \boldsymbol{s}) - g(\boldsymbol{x}^*, \boldsymbol{s}^*) + \langle \mathcal{A}, \boldsymbol{x} - \boldsymbol{x}^* \rangle + \langle \mathcal{C}, \boldsymbol{s} - \boldsymbol{s}^* \rangle;$$

cf. (D.6). Let also $\mathcal{S}$ be given by (D.53) and $\boldsymbol{s} \in \mathcal{S}$. We are interested in quantifying the distance between $\boldsymbol{x}^*$ and $\boldsymbol{x}_{\boldsymbol{s}}^\circ$, where

$$\boldsymbol{x}_{\boldsymbol{s}}^\circ \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{s}).$$

The linear perturbation $\langle \mathcal{C}, \boldsymbol{s} - \boldsymbol{s}^* \rangle$ does not depend on $\boldsymbol{x}$ and it can be ignored.

Proposition D.25. *For $\boldsymbol{s} \in \mathcal{S}$, assume the conditions of Proposition D.19 and let*

$$\omega \stackrel{\text{def}}{=} \tau_{21} \| \mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*) \|_\circ \leq 1/4 .$$

Then

$$\| \mathbb{D}(\boldsymbol{x}_{\boldsymbol{s}}^\circ - \boldsymbol{x}^*) \| \leq \frac{3}{2(1-\omega)} \bigl(\rho_2 \, \| \mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*) \|_\circ + \| \mathbb{D}^{-1} \mathcal{A} \| \bigr) \tag{D.73}$$

and for any linear mapping Q, it holds with $\tau_\circ$ from (D.61)

$$\begin{aligned}
& \| Q \{ \boldsymbol{x}_{\boldsymbol{s}}^\circ - \boldsymbol{x}^* + \mathbb{F}^{-1} \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} (\boldsymbol{s} - \boldsymbol{s}^*) - \mathbb{F}^{-1} \mathcal{A} \} \| \\
& \leq \| Q \, \mathbb{F}^{-1} \mathbb{D} \| \Bigl(\tau_\circ \, \| \mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*) \|_\circ^2 + 2 \tau_3 \, \| \mathbb{D} \, \mathbb{F}^{-1} \mathcal{A} \|^2 + \frac{4 \tau_{21}}{3} \, \| \mathbb{D} \, \mathbb{F}^{-1} \mathcal{A} \| \, \| \mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*) \|_\circ \Bigr) \\
& \leq \| Q \, \mathbb{F}^{-1} \mathbb{D} \| \Bigl\{ (\tau_\circ + \tau_{21}) \, \| \mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^*) \|_\circ^2 + (2 \tau_3 + \tau_{21}/2) \, \| \mathbb{D} \, \mathbb{F}^{-1} \mathcal{A} \|^2 \Bigr\} .
\end{aligned} \tag{D.74}$$

Proof. Proposition D.15 applied to $f_{\boldsymbol{s}}(\boldsymbol{x})$ and $g_{\boldsymbol{s}}(\boldsymbol{x}) = f_{\boldsymbol{s}}(\boldsymbol{x}) - \langle \mathcal{A}, \boldsymbol{x} \rangle$ implies by (D.68)

$$\| \mathbb{D}(\boldsymbol{x}_{\boldsymbol{s}}^\circ - \boldsymbol{x}_{\boldsymbol{s}}) \| \leq \frac{3}{2} \| \mathbb{D} \, \mathbb{F}_{\boldsymbol{s}}^{-1} \mathcal{A} \| \leq \frac{3}{2(1-\omega)} \| \mathbb{D}^{-1} \mathcal{A} \| .$$

This and (D.59) imply (D.73). Next we check (D.74) using the decomposition

$$\begin{aligned}
& \boldsymbol{x}_{\boldsymbol{s}}^\circ - \boldsymbol{x}^* + \mathbb{F}^{-1} \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} (\boldsymbol{s} - \boldsymbol{s}^*) - \mathbb{F}^{-1} \mathcal{A} \\
& \quad = (\boldsymbol{x}_{\boldsymbol{s}}^\circ - \boldsymbol{x}_{\boldsymbol{s}} - \mathbb{F}_{\boldsymbol{s}}^{-1} \mathcal{A}) + (\mathbb{F}_{\boldsymbol{s}}^{-1} \mathcal{A} - \mathbb{F}^{-1} \mathcal{A}) + \{ \boldsymbol{x}_{\boldsymbol{s}} - \boldsymbol{x}^* + \mathbb{F}^{-1} \mathscr{F}_{\boldsymbol{x}\boldsymbol{s}} (\boldsymbol{s} - \boldsymbol{s}^*) \} .
\end{aligned}$$

Proposition D.19 evaluates the last term. Lemma D.20 helps to bound

$$\|\mathbb{D}^{-1}\mathbb{F}(\mathbb{F}_{\boldsymbol{s}}^{-1} - \mathbb{F}^{-1})\mathbb{F}_{\boldsymbol{s}}\,\mathbb{D}^{-1}\| = \|\mathbb{D}^{-1}(\mathbb{F}_{\boldsymbol{s}} - \mathbb{F})\mathbb{D}^{-1}\| \le \tau_{21}\|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^{*})\|_{\circ}\,.$$

This yields

$$\begin{aligned}\|Q(\mathbb{F}_{\boldsymbol{s}}^{-1} - \mathbb{F}^{-1})\mathcal{A}\| &\le \|Q\,\mathbb{F}^{-1}\mathbb{D}\|\,\|\mathbb{D}^{-1}\mathbb{F}_{\boldsymbol{s}}(\mathbb{F}_{\boldsymbol{s}}^{-1} - \mathbb{F}^{-1})\mathbb{F}\,\mathbb{D}^{-1}\|\,\|\mathbb{D}\,\mathbb{F}_{\boldsymbol{s}}^{-1}\mathcal{A}\| \\ &\le \|Q\,\mathbb{F}^{-1}\mathbb{D}\|\,\tau_{21}\,\|\mathbb{H}(\boldsymbol{s} - \boldsymbol{s}^{*})\|_{\circ}\,\frac{1}{1-\omega}\|\mathbb{D}\,\mathbb{F}^{-1}\mathcal{A}\|\,.\end{aligned}$$

Moreover, for $\boldsymbol{\varepsilon}_{\boldsymbol{s}} \stackrel{\mathrm{def}}{=} \boldsymbol{x}_{\boldsymbol{s}}^{\circ} - \boldsymbol{x}_{\boldsymbol{s}} - \mathbb{F}_{\boldsymbol{s}}^{-1}\mathcal{A}$, it holds

$$\|\mathbb{D}^{-1}\mathbb{F}_{\boldsymbol{s}}\,\boldsymbol{\varepsilon}_{\boldsymbol{s}}\| \le \frac{3\tau_3}{4}\|\mathbb{D}\,\mathbb{F}_{\boldsymbol{s}}^{-1}\mathcal{A}\|^2 \le \frac{3\tau_3}{4(1-\omega)^2}\|\mathbb{D}\,\mathbb{F}^{-1}\mathcal{A}\|^2\,,$$

and by (D.68) and $\omega \le 1/4$

$$\begin{aligned}\|Q\,\boldsymbol{\varepsilon}_{\boldsymbol{s}}\| &\le \|Q\,\mathbb{F}^{-1}\mathbb{D}\|\;\|\mathbb{D}^{-1}\mathbb{F}\,\boldsymbol{\varepsilon}_{\boldsymbol{s}}\| \le \|Q\,\mathbb{F}^{-1}\mathbb{D}\|\,\frac{1-\omega}{1-2\omega}\,\|\mathbb{D}^{-1}\mathbb{F}_{\boldsymbol{s}}\,\boldsymbol{\varepsilon}_{\boldsymbol{s}}\| \\ &\le \|Q\,\mathbb{F}^{-1}\mathbb{D}\|\,\frac{3\tau_3}{4(1-\omega)(1-2\omega)}\|\mathbb{D}\,\mathbb{F}^{-1}\mathcal{A}\|^2 \le \|Q\,\mathbb{F}^{-1}\mathbb{D}\|\,2\tau_3\,\|\mathbb{D}\,\mathbb{F}^{-1}\mathcal{A}\|^2\,.\end{aligned}$$

The obtained bounds imply (D.74) in view of $(4/3)ab \le a^2 + b^2/2$ for any a, b. □

D.4 Linearly perturbed optimization. Sup-norm expansions

Let $f(\boldsymbol{\upsilon})$ be a concave function and $\boldsymbol{\upsilon}^{*} = \operatorname{argmax}_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon})$. Define $\mathscr{F} = -\nabla^2 f(\boldsymbol{\upsilon}^{*}) = (\mathscr{F}_{jm})$. Let also $g(\boldsymbol{\upsilon})$ be obtained by a linear perturbation of $f(\boldsymbol{\upsilon})$ as in (D.6) and let $\boldsymbol{\upsilon}^{\circ} = \operatorname{argmax}_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon})$. We intend to bound the corresponding change $\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*}$ in a sup-norm. The strategy of the study is to fix one component υ_j of $\boldsymbol{\upsilon}$ and treat the remaining entries as the nuisance parameter. Fix a metric tensor $\mathbb{D}$ in diagonal form:

$$\mathbb{D} \stackrel{\mathrm{def}}{=} \operatorname{diag}(\mathbb{D}_1, \ldots, \mathbb{D}_p).$$

Later we assume $\mathbb{D}_j^2 = \mathscr{F}_{jj}$. For each $j \le p$, we will use the representation $\boldsymbol{\upsilon} = (\upsilon_j, \boldsymbol{s}_j)$, where $\boldsymbol{s}_j \in \mathbb{R}^{p-1}$ collects all the remaining entries of $\boldsymbol{\upsilon}$. Similarly, define $\mathbb{D} = (\mathbb{D}_j, \mathbb{H}_j)$ with $\mathbb{H}_j = \operatorname{diag}(\mathbb{D}_m)$ for $m \ne j$. Define also for some radius $\mathtt{r}_{\infty}$

$$\varUpsilon^{\circ} = \{\boldsymbol{\upsilon} \colon \|\mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*})\|_{\infty} \le \mathtt{r}_{\infty}\} = \Bigl\{\boldsymbol{\upsilon} \colon \max_{j \le p}\bigl|\mathbb{D}_j(\upsilon_j - \upsilon_j^{*})\bigr| \le \mathtt{r}_{\infty}\Bigr\}.$$

Assume the following condition.

$(\mathcal{T}_{\infty}^{*})$ *For each $j \le p$, the function $f(\upsilon_j, \boldsymbol{s}_j)$ fulfills*

$$\sup_{\boldsymbol{s}_j \colon \|\mathbb{H}_j(\boldsymbol{s}_j - \boldsymbol{s}_j^*)\|_\infty \le \mathtt{r}_\infty} \; \sup_{\upsilon \colon \mathbb{D}_j|\upsilon - \upsilon_j^*| \le 2\mathtt{r}_\infty} \frac{\bigl|\nabla^{(3)}_{\upsilon_j\upsilon_j\upsilon_j} f(\upsilon_j, \boldsymbol{s}_j)\bigr|}{\mathbb{D}_j^3} \le \tau_3 ,$$

and

$$\sup_{\boldsymbol{\upsilon} = (\upsilon_j^*, \boldsymbol{s}_j) \in \Upsilon^\circ} \; \sup_{\boldsymbol{z}_j \in I\!\!R^{p-1}} \frac{|\langle \nabla^{(3)}_{\upsilon_j\upsilon_j\boldsymbol{s}_j} f(\upsilon_j^*, \boldsymbol{s}_j), \boldsymbol{z}_j \rangle|}{\mathbb{D}_j^2 \, \|\mathbb{H}_j \boldsymbol{z}_j\|_\infty} \le \tau_{21} ,$$

$$\sup_{\boldsymbol{\upsilon} = (\upsilon_j^*, \boldsymbol{s}_j) \in \Upsilon^\circ} \; \sup_{\boldsymbol{z}_j \in I\!\!R_j^p} \frac{|\langle \nabla^3_{\upsilon_j\boldsymbol{s}_j\boldsymbol{s}_j} f(\upsilon_j^*, \boldsymbol{s}_j), \boldsymbol{z}_j^{\otimes 2} \rangle|}{\mathbb{D}_j \, \|\mathbb{H}_j \boldsymbol{z}_j\|_\infty^2} \le \tau_{12} .$$

Define

$$\rho_1 \stackrel{\text{def}}{=} \max_{j=1,\ldots,p} \|\mathbb{D}_j^{-1} \mathscr{F}_{\upsilon_j, \boldsymbol{s}_j} \, \mathbb{H}_j^{-1}\|_* . \tag{D.75}$$

Lemma D.26. *It holds*

$$\rho_1^2 \le \max_{j=1,\ldots,p} \frac{1}{\mathbb{D}_j^2} \sum_{m \ne j} \frac{\mathscr{F}_{jm}^2}{\mathbb{D}_m^2} .$$

Proof. By definition,

$$\begin{aligned}
\rho_1^2 &= \max_{j=1,\ldots,p} \sup_{\|\boldsymbol{z}\|_\infty \le 1} \|\mathbb{D}_j^{-1} \mathscr{F}_{\upsilon_j, \boldsymbol{s}_j} \, \mathbb{H}_j^{-1} \boldsymbol{z}\|^2 \\
&\le \max_{j=1,\ldots,p} \sup_{\|\boldsymbol{z}\|_\infty \le 1} \sum_{m \ne j} \frac{1}{\mathbb{D}_j^2 \, \mathbb{D}_m^2} \mathscr{F}_{jm}^2 z_m^2 \le \max_{j=1,\ldots,p} \frac{1}{\mathbb{D}_j^2} \sum_{m \ne j} \frac{\mathscr{F}_{jm}^2}{\mathbb{D}_m^2}
\end{aligned}$$

as claimed. □

It is mandatory for the proposed approach that $\rho_1 < 1$ and $\omega = \tau_{21} \, \mathtt{r}_\infty < 1$. The value $\tau_\circ$ from (D.61) can be specified as

$$\tau_\circ \stackrel{\text{def}}{=} \frac{1}{1 - \omega} \left(\rho_1 \, \tau_{21} + \frac{\tau_{12}}{2} + \frac{3(\rho_1 + \omega/2)^2 \, \tau_3}{4(1 - \omega)^2} \right) . \tag{D.76}$$

The next result provides an upper bound on $\|\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*\|_\infty$.

Proposition D.27. *Let $f(\boldsymbol{\upsilon})$ be concave function and $g(\boldsymbol{\upsilon})$ be a linear perturbation of $f(\boldsymbol{\upsilon})$ with a vector $\mathcal{A}$. Assume $\rho_1 < 1$; see (D.75). Fix*

$$\begin{aligned}
\mathtt{r}_\infty &= \frac{\sqrt{2}}{1 - \rho_1} \|\mathbb{D}^{-1} \mathcal{A}\|_\infty , \\
\tau_\infty &\stackrel{\text{def}}{=} 2\tau_3 + \frac{\tau_{21}}{2} + \frac{2(\tau_\circ + \tau_{21})}{(1 - \rho_1)^2} ,
\end{aligned} \tag{D.77}$$

where $\tau_{\circ}$ *is from* (D.76). *Let* $(\mathcal{T}_{\infty}^{*})$ *hold with this* $\mathtt{r}_{\infty}$ *and* $\tau_{12}, \tau_{21}, \tau_{3}$ *satisfying*

$$\tau_{12}\, \mathtt{r}_{\infty} \leq 1/4\,, \qquad \tau_{\infty} \|\mathbb{D}^{-1}\mathcal{A}\|_{\infty} \leq \sqrt{2} - 1\,. \tag{D.78}$$

Then $\boldsymbol{\upsilon}^{\circ} = \operatorname{argmax}_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon})$ *satisfies*

$$\|\mathbb{D}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})\|_{\infty} \stackrel{\text{def}}{=} \max_{j \leq p} |\mathbb{D}_{j}(\upsilon_{j}^{\circ} - \upsilon_{j}^{*})| \leq \mathtt{r}_{\infty}\,. \tag{D.79}$$

Furthermore,

$$\begin{aligned} \bigl\|\mathbb{D}^{-1}\bigl\{\mathscr{F}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*}) - \mathcal{A}\bigr\}\bigr\|_{\infty} &\leq \tau_{\infty}\, \|\mathbb{D}^{-1}\mathcal{A}\|_{\infty}^{2}\,, \\ \bigl\|\mathbb{D}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*} - \mathscr{F}^{-1}\mathcal{A})\bigr\|_{\infty} &\leq \frac{\tau_{\infty}}{1 - \rho_{1}}\, \|\mathbb{D}^{-1}\mathcal{A}\|_{\infty}^{2}\,, \end{aligned} \tag{D.80}$$

and with $\varDelta \stackrel{\text{def}}{=} \mathbb{I}_{p} - \mathbb{D}^{-1}\mathscr{F}\,\mathbb{D}^{-1}$

$$\begin{aligned} \bigl\|\mathbb{D}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*}) - \mathbb{D}^{-1}\mathcal{A})\bigr\|_{\infty} &\leq \frac{\tau_{\infty}}{1 - \rho_{1}}\, \|\mathbb{D}^{-1}\mathcal{A}\|_{\infty}^{2} + \frac{\rho_{1}}{1 - \rho_{1}}\, \|\mathbb{D}^{-1}\mathcal{A}\|_{\infty}\,, \\ \bigl\|\mathbb{D}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*}) - (\mathbb{I}_{p} + \varDelta)\mathbb{D}^{-1}\mathcal{A})\bigr\|_{\infty} &\leq \frac{\tau_{\infty}}{1 - \rho_{1}}\, \|\mathbb{D}^{-1}\mathcal{A}\|_{\infty}^{2} + \frac{\rho_{1}^{2}}{1 - \rho_{1}}\, \|\mathbb{D}^{-1}\mathcal{A}\|_{\infty}\,. \end{aligned}$$

Remark D.6. To gain an intuition about the result of the proposition, consider a typical situation with $\tau_{12} \leq \tau_{3}$, $\tau_{21} \leq \tau_{3}$, $\omega \leq 1/4$, $1 - \rho_{1} \leq 1/\sqrt{2}$. Then $\tau_{\circ}$ from (D.76) fulfills $\tau_{\circ} \leq 1.37$, and it holds $\tau_{\infty} \leq 12\tau_{3}$; see (D.77).

Proof. Let us fix any $j \leq p$, e.g. $j = 1$. Represent any $\boldsymbol{\upsilon} \in \varUpsilon^{\circ}$ as $\boldsymbol{\upsilon} = (\upsilon_{1}, \boldsymbol{s}_{1})$, where $\boldsymbol{s}_{1} = (\upsilon_{2}, \ldots, \upsilon_{p})^{\top}$. By (D.75)

$$\bigl|\mathbb{D}_{1}\mathscr{F}_{11}^{-1}\mathscr{F}_{\upsilon_{1}\boldsymbol{s}_{1}}(\boldsymbol{s}_{1} - \boldsymbol{s}_{1}^{*})\bigr| \leq \bigl|\mathbb{D}_{1}^{-1}\mathscr{F}_{\upsilon_{1}\boldsymbol{s}_{1}}(\boldsymbol{s}_{1} - \boldsymbol{s}_{1}^{*})\bigr| \leq \rho_{1}\, \|\mathbb{H}_{1}(\boldsymbol{s}_{1} - \boldsymbol{s}_{1}^{*})\|_{\infty}\,. \tag{D.81}$$

For any $\boldsymbol{s}_{1}$, define

$$\upsilon_{1}^{\circ}(\boldsymbol{s}_{1}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\upsilon_{1}} g(\upsilon_{1}, \boldsymbol{s}_{1}).$$

Now we apply Proposition D.25 with $Q = \mathbb{D}_{1}$, $\mathbb{F} = \mathscr{F}_{11}$, $\mathbb{H} = \mathbb{H}_{1} = \operatorname{diag}(\mathbb{D}_{2}, \ldots, \mathbb{D}_{p})$, and $\|\mathbb{H}_{1}(\boldsymbol{s}_{1} - \boldsymbol{s}_{1}^{*})\|_{\infty} \leq \mathtt{r}_{\infty}$. As $\mathbb{D}_{1}\mathscr{F}_{11}^{-1} \leq \mathbb{D}_{1}^{-1}$, bound (D.74) yields

$$\begin{aligned} &\bigl|\mathbb{D}_{1}\{\upsilon_{1}^{\circ}(\boldsymbol{s}_{1}) - \upsilon_{1}^{*} - \mathscr{F}_{11}^{-1}\mathcal{A}_{1} + \mathscr{F}_{11}^{-1}\mathscr{F}_{\upsilon_{1}\boldsymbol{s}_{1}}(\boldsymbol{s}_{1} - \boldsymbol{s}_{1}^{*})\}\bigr| \\ &\quad \leq (\tau_{\circ} + \tau_{21})\, \|\mathbb{H}_{1}(\boldsymbol{s}_{1} - \boldsymbol{s}_{1}^{*})\|_{\circ}^{2} + \bigl(2\tau_{3} + \frac{\tau_{21}}{2}\bigr)\, \mathbb{D}_{1}^{-2}\mathcal{A}_{1}^{2} \\ &\quad \leq (\tau_{\circ} + \tau_{21})\mathtt{r}_{\infty}^{2} + \bigl(2\tau_{3} + \frac{\tau_{21}}{2}\bigr)\, \|\mathbb{D}^{-1}\mathcal{A}\|_{\infty}^{2} = \tau_{\infty}\, \|\mathbb{D}^{-1}\mathcal{A}\|_{\infty}^{2}\,. \end{aligned}$$

This, (D.81), and (D.78) imply by $\mathtt{r}_\infty = \mathtt{C}\|\mathbb{D}^{-1}\mathcal{A}\|_\infty$ and $\mathtt{C} = \sqrt{2}/(1-\rho_1)$

$$\begin{aligned} \bigl|\mathbb{D}_1\{\upsilon_1^\circ(\boldsymbol{s}) - \upsilon_1^*\}\bigr| &\le \|\mathbb{D}^{-1}\mathcal{A}\|_\infty + \rho_1\,\mathtt{r}_\infty + \tau_\infty\,\|\mathbb{D}^{-1}\mathcal{A}\|_\infty^2 \\ &\le \bigl(1 + \mathtt{C}\,\rho_1 + \sqrt{2} - 1\bigr)\|\mathbb{D}^{-1}\mathcal{A}\|_\infty = \mathtt{C}\,\|\mathbb{D}^{-1}\mathcal{A}\|_\infty = \mathtt{r}_\infty\,. \end{aligned}$$

The same bounds apply to each $j \le p$. Therefore, when started from any point $\boldsymbol{\upsilon} \in \varUpsilon^\circ$, the sequential optimization procedure which maximizes the objective function w.r.t. one coordinate while keeping all the remaining coordinates has its solution within $\varUpsilon^\circ$. As the function f is strongly concave, and its value improves at every step, the procedure converges to a unique solution $\boldsymbol{\upsilon}^\circ \in \varUpsilon^\circ$. This implies (D.79). Further, with $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}^\circ$,

$$\mathscr{F}_{11}\{\upsilon_1(\boldsymbol{s}_1) - \upsilon_1^*\} + \mathscr{F}_{\upsilon_1\boldsymbol{s}_1}(\boldsymbol{s}_1 - \boldsymbol{s}_1^*) = \mathscr{F}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)\,.$$

This yields (D.80). Moreover, with $\boldsymbol{u} \stackrel{\text{def}}{=} \mathbb{D}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)$ and $B \stackrel{\text{def}}{=} \mathbb{D}^{-1}\mathscr{F}\,\mathbb{D}^{-1}$, it holds

$$\mathbb{D}^{-1}\bigl\{\mathscr{F}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*) - \mathcal{A}\bigr\} = B(\boldsymbol{u} - \mathbb{D}\,\mathscr{F}^{-1}\mathcal{A})\,,$$

and by (D.80) and Lemma D.28

$$\begin{aligned} &\bigl\|\mathbb{D}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^* - \mathscr{F}^{-1}\mathcal{A})\bigr\|_\infty = \bigl\|\boldsymbol{u} - \mathbb{D}\,\mathscr{F}^{-1}\mathcal{A}\bigr\|_\infty \\ &\quad = \bigl\|B^{-1}\mathbb{D}^{-1}\{\mathscr{F}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*) - \mathcal{A}\}\bigr\|_\infty \le \frac{\tau_\infty}{1-\rho_1}\,\|\mathbb{D}^{-1}\mathcal{A}\|_\infty^2\,. \end{aligned} \tag{D.82}$$

Finally, by (D.84) of Lemma D.28 and by $\mathbb{D}\,\mathscr{F}^{-1}\mathcal{A} = B^{-1}\mathbb{D}^{-1}\mathcal{A}$

$$\begin{aligned} \|\mathbb{D}\,\mathscr{F}^{-1}\mathcal{A} - \mathbb{D}^{-1}\mathcal{A}\|_\infty &= \bigl\|(B^{-1} - I\!\!I_p)\mathbb{D}^{-1}\mathcal{A}\bigr\|_\infty \le \frac{\rho_1}{1-\rho_1}\bigl\|\mathbb{D}^{-1}\mathcal{A}\bigr\|_\infty\,, \\ \|\mathbb{D}\,\mathscr{F}^{-1}\mathcal{A} - (I\!\!I_p + \Delta)\mathbb{D}^{-1}\mathcal{A}\|_\infty &= \bigl\|(B^{-1} - I\!\!I_p - \Delta)\mathbb{D}^{-1}\mathcal{A}\bigr\|_\infty \le \frac{\rho_1^2}{1-\rho_1}\bigl\|\mathbb{D}^{-1}\mathcal{A}\bigr\|_\infty\,. \end{aligned}$$

This and (D.82) imply the final inequalities of the proposition. □

Lemma D.28. *Let* $B = (B_{ij}) \in \mathfrak{M}_p$ *with* $B_{ii} = 1$ *and*

$$\sup_{\boldsymbol{u}\colon \|\boldsymbol{u}\|_\infty \le 1} \|(B - I\!\!I_p)\boldsymbol{u}\|_\infty \le \rho_1 < 1. \tag{D.83}$$

Then $\|B\boldsymbol{u}\|_\infty \le (1-\rho_1)^{-1}\|\boldsymbol{u}\|_\infty$ *for any* $\boldsymbol{u} \in I\!\!R^p$. *Similarly,*

$$\|(B^{-1} - I\!\!I_p)\boldsymbol{u}\|_\infty \le \frac{\rho_1}{1-\rho_1}\,\|\boldsymbol{u}\|_\infty\,, \quad \|(B^{-1} - 2I\!\!I_p + B)\boldsymbol{u}\|_\infty \le \frac{\rho_1^2}{1-\rho_1}\,\|\boldsymbol{u}\|_\infty\,. \tag{D.84}$$

Proof. Represent $B = \mathbb{I}_p - \Delta$. Then (D.83) implies $\|\Delta \boldsymbol{u}\|_\infty \le \rho_1 \|\boldsymbol{u}\|_\infty$. The use of $B^{-1} = \mathbb{I}_p + \Delta + \Delta^2 + \ldots$ yields for any $\boldsymbol{u} \in \mathbb{R}^p$

$$\|B^{-1}\boldsymbol{u}\|_\infty \le \sum_{m=0}^{\infty} \|\Delta^m \boldsymbol{u}\|_\infty .$$

Further, with $\boldsymbol{u}_m \stackrel{\text{def}}{=} \Delta^m \boldsymbol{u}$, it holds

$$\|\Delta^{m+1}\boldsymbol{u}\|_\infty = \|\Delta \boldsymbol{u}_m\|_\infty \le \rho_1 \|\boldsymbol{u}_m\|_\infty .$$

By induction, this yields $\|\Delta^m \boldsymbol{u}\|_\infty \le \rho_1^m \|\boldsymbol{u}\|_\infty$ and thus,

$$\|B^{-1}\boldsymbol{u}\|_\infty \le \sum_{m=0}^{\infty} \rho_1^m \|\boldsymbol{u}\|_\infty = \frac{1}{1-\rho_1} \|\boldsymbol{u}\|_\infty$$

as claimed. The proof of (D.84) is similar in view of $\mathbb{I}_p + \Delta = 2\mathbb{I}_p - B$. □

E Deviation bounds for Bernoulli vector sums

Let Y_i be independent $\operatorname{Be}(\theta_i^*)$, $i = 1, \ldots, n$. We denote $\boldsymbol{Y} = (Y_i) \in \mathbb{R}^n$. Weighted sums of the Y_i naturally appear in various statistical tasks including classification, binary response models, logistic regression etc. Recent applications include e.g. stochastic block modeling; see e.g. Gao et al. (2017), Abbe (2018) and references therein, or ranking from pairwise comparison Chen et al. (2022) among many others. We show how the general bounds of Section **??** can be used for vector sums of Bernoulli r.v.s. For a linear mapping $\boldsymbol{\Psi}\colon \mathbb{R}^n \to \mathbb{R}^p$, define $\boldsymbol{\xi} = \boldsymbol{\Psi}(\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y})$. Below we state some deviation bounds on the squared norm $\|\boldsymbol{\xi}\|^2$ starting from the univariate case.

E.1 Weighted sums of Bernoulli r.v.'s: univariate case

Given a collections of weights (w_i), define

$$\begin{aligned} S &= \sum_{i=1}^{n} Y_i w_i , \\ V^2 &= \operatorname{Var}(S) = \sum_{i=1}^{n} \theta_i^*(1-\theta_i^*) w_i^2 , \\ w^* &= \max_i |w_i|. \end{aligned}$$

First, we state a deviation bound for a centered sum $S - \mathbb{E}S$.

Proposition E.1. *Let* Y_i *be independent* $\mathrm{Be}(\theta_i^*)$ *and* $w_i \in \mathbb{R}$, $i = 1, \ldots, n$. *Then* $S = \sum_{i=1}^n Y_i w_i$ *satisfies*

$$\log \mathbb{E} \exp\Bigl\{ \frac{\lambda(S - \mathbb{E}S)}{V} \Bigr\} \leq \lambda^2, \qquad \lambda \leq \frac{\log(2) V}{w^*} . \tag{E.1}$$

Furthermore, suppose that given $\mathtt{x} \geq 0$,

$$V \geq \frac{3}{2}\, w^* \sqrt{\mathtt{x}} . \tag{E.2}$$

Then

$$\mathbb{P}\bigl(V^{-1} |S - \mathbb{E}S| \geq 2\sqrt{\mathtt{x}} \bigr) \leq 2 \mathrm{e}^{-\mathtt{x}}. \tag{E.3}$$

Without (E.2), *the bound* (E.3) *applies with* V *replaced by* $V_{\mathtt{x}} = V \vee (3\, w^* \sqrt{\mathtt{x}}\,/2)$.

Proof. Without loss of generality assume $w^* = 1$, otherwise just rescale all the weights by the factor $1/w^*$. We use that

$$f(u) \stackrel{\mathrm{def}}{=} \log \mathbb{E} \exp\Bigl\{ u(S - \mathbb{E}S) \Bigr\} = \sum_{i=1}^N \Bigl[\log\bigl(\theta_i^* \mathrm{e}^{u w_i} + 1 - \theta_i^* \bigr) - u w_i \theta_i^* \Bigr] .$$

This is an analytic function of u for $|u| \leq \log 2$ satisfying $f(0) = 0$, $f'(0) = 0$, and, with $\upsilon_i^* = \log \theta_i^* - \log(1 - \theta_i^*)$,

$$f''(u) = \sum_{i=1}^N \frac{w_i^2\, \theta_i^* (1 - \theta_i^*)\, \mathrm{e}^{u w_i}}{(\theta_i^* \mathrm{e}^{u w_i} + 1 - \theta_i^*)^2} = \sum_{i=1}^N \frac{w_i^2\, \mathrm{e}^{\upsilon_i^* + u w_i}}{(\mathrm{e}^{\upsilon_i^* + u w_i} + 1)^2} = \sum_{i=1}^N \theta_i(u) \bigl\{ 1 - \theta_i(u) \bigr\} w_i^2$$

for $\theta_i(u) = \mathrm{e}^{\upsilon_i^* + u w_i} / (\mathrm{e}^{\upsilon_i^* + u w_i} + 1)$. Clearly $\theta_i(u)$ and thus, $\theta_i(u)\bigl\{1 - \theta_i(u)\bigr\}$ monotonously increases with u and it holds for $\theta_i^* = \theta_i(0)$

$$\theta_i(u) \bigl\{ 1 - \theta_i(u) \bigr\} \leq \mathrm{e}^{|u|}\, \theta_i^*\, (1 - \theta_i^*) \leq 2\, \theta_i^*\, (1 - \theta_i^*), \qquad |u| \leq \log 2.$$

This yields

$$f(u) \leq V^2\, u^2 \qquad |u| \leq \log 2.$$

As $\mathtt{x} \leq 4V^2/9$, the value $\lambda = \sqrt{\mathtt{x}}$ fulfills $\lambda/V = \sqrt{\mathtt{x}}/V \leq \log 2 \leq 2^{-1/2}$. Now by the exponential Chebyshev inequality

$$\begin{aligned} \mathbb{P}\Bigl(V^{-1}(S - \mathbb{E}S) \geq 2\sqrt{\mathtt{x}} \Bigr) &\leq \exp\bigl\{ -2\lambda\sqrt{\mathtt{x}} + f(\lambda/V) \bigr\} \\ &\leq \exp\bigl(-2\lambda\sqrt{\mathtt{x}} + \lambda^2 \bigr) = \mathrm{e}^{-\mathtt{x}}. \end{aligned}$$

Similarly one can bound $\mathbb{E}S - S$. $\square$

E.2 Deviation bounds for Bernoulli vector sums

Now we present an upper bound on the norm of a vector $\boldsymbol{\xi} = \boldsymbol{\Psi}(\boldsymbol{Y} - I\!\!E\boldsymbol{Y})$, where $\boldsymbol{\Psi}$ is a linear mapping $\boldsymbol{\Psi}\colon I\!\!R^n \to I\!\!R^p$. It holds

$$\operatorname{Var}(\boldsymbol{\xi}) = \operatorname{Var}(\boldsymbol{\Psi}\boldsymbol{Y}) = \boldsymbol{\Psi}\operatorname{Var}(\boldsymbol{Y})\boldsymbol{\Psi}^{\top}.$$

We aim at bounding the squared norm $\|Q\boldsymbol{\xi}\|^2$ for another linear mapping $Q\colon I\!\!R^p \to I\!\!R^q$.

Theorem E.2. *Let* $Y_i \sim \operatorname{Be}(\theta_i^*)$, $i = 1,\ldots,n$. *Consider* $\boldsymbol{\xi} = \boldsymbol{\Psi}(\boldsymbol{Y} - I\!\!E\boldsymbol{Y})$, *and let* $\mathbb{V}^2 \geq 2\operatorname{Var}(\boldsymbol{\xi})$. *Define*

$$w^* = \max_{i \leq n} \|\mathbb{V}^{-1}\boldsymbol{\Psi}_i\|, \qquad \mathtt{g} = \log(2)/w^*.$$

Then with $B = Q\,\mathbb{V}^2 Q^{\top}$ *and* $z_c(B,\mathtt{x})$ *from* (**??**), *it holds*

$$I\!\!P\bigl(\|Q\boldsymbol{\xi}\| \geq z_c(B,\mathtt{x})\bigr) \leq 3\mathrm{e}^{-\mathtt{x}}.$$

Proof. We apply the general result of Corollary **??** under conditions (**??**). For any vector $\boldsymbol{u}$, consider the scalar product $\langle \mathbb{V}^{-1}\boldsymbol{\xi}, \boldsymbol{u}\rangle = \langle \mathbb{V}^{-1}\boldsymbol{\Psi}(\boldsymbol{Y} - I\!\!E\boldsymbol{Y}), \boldsymbol{u}\rangle$. It is obviously a weighted centered sum of the Bernoulli r.v.'s $Y_i - \theta_i^*$ with

$$\operatorname{Var}\langle \mathbb{V}^{-1}\boldsymbol{\xi}, \boldsymbol{u}\rangle \leq \|\boldsymbol{u}\|^2/2.$$

One can write with $\varepsilon_i = Y_i - \theta_i^*$ and $\boldsymbol{\varepsilon} = (\varepsilon_i)$

$$\langle \mathbb{V}^{-1}\boldsymbol{\xi}, \boldsymbol{u}\rangle = \bigl\langle \boldsymbol{\varepsilon}, \boldsymbol{\Psi}^{\top}\mathbb{V}^{-1}\boldsymbol{u}\bigr\rangle.$$

By the Cauchy-Schwarz inequality, it holds

$$\|\boldsymbol{\Psi}^{\top}\mathbb{V}^{-1}\boldsymbol{u}\|_{\infty} = \max_i \bigl|(\mathbb{V}^{-1}\boldsymbol{\Psi}_i)^{\top}\boldsymbol{u}\bigr| \leq w^*\|\boldsymbol{u}\|.$$

Bound (E.1) of Proposition E.1 on the exponential moments of $\langle \mathbb{V}^{-1}\boldsymbol{\xi}, \boldsymbol{u}\rangle$ implies

$$\log I\!\!E\exp\bigl\{\langle \mathbb{V}^{-1}\boldsymbol{\xi}, \boldsymbol{u}\rangle\bigr\} \leq \|\boldsymbol{u}\|^2/2, \qquad \|\boldsymbol{u}\| \leq \log(2)/w^*.$$

Therefore, (**??**) is fulfilled with $\mathtt{g} = \log(2)/w^*$. The deviation bound (**??**) of Corollary **??** yields the assertion. $\square$

References

Abbe, E. (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86.

Andresen, A. and Spokoiny, V. (2016). Convergence of an alternating maximization procedure. *Journal of Machine Learning Research*, 17(63):1–53.

Banach, S. (1938). Über homogene Polynome in (L^2). *Studia Mathematica*, 7(1):36–44.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Chen, P., Gao, C., and Zhang, A. Y. (2020). Partial recovery for top-k ranking: Optimality of mle and sub-optimality of spectral method.

Chen, P., Gao, C., and Zhang, A. Y. (2022). Optimal full ranking from pairwise comparisons. *The Annals of Statistics*, 50(3):1775 – 1805.

Chen, Y., Fan, J., Ma, C., and Wang, K. (2019). Spectral method and regularized mle are both optimal for top-k ranking. *Ann. Statist.*, 47(4):2204–2235.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *Journal of Machine Learning Research*, 18(60):1–45.

Gao, C., Shen, Y., and Zhang, A. Y. (2023). Uncertainty quantification in the BradleyTerryLuce model. *Information and Inference: A Journal of the IMA*, 12(2):1073–1140.

Katsevich, A. (2023). Tight bounds on the laplace approximation accuracy in high dimensions. https://arxiv.org/abs/2305.17604.

Luce, R. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley.

Nesterov, Y. E. (1988). Polynomial methods in the linear and quadratic programming. *Sov. J. Comput. Syst. Sci.*, 26(5):98–101.

Ostrovskii, D. M. and Bach, F. (2021). Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1):326 – 391.

Spokoiny, V. (2017). Penalized maximum likelihood estimation and effective dimension. *AIHP*, 53(1):389–429. arXiv:1205.0498.

Spokoiny, V. (2023a). Dimension free non-asymptotic bounds on the accuracy of high dimensional Laplace approximation. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3):1044–1068. in print, https://arxiv.org/abs/2204.11038.

Spokoiny, V. (2023b). Nonlinear regression: finite sample guarantees. https://arxiv.org/abs/2305.08193.

Spokoiny, V. (2024a). Deviation bounds for the norm of a random vector under exponential moment conditions with applications. https://arxiv.org/abs/2309.02302v1.

Spokoiny, V. (2024b). Estimation for SLS models: finite sample guarantees. https://arxiv.org/abs/2404.14227.

Spokoiny, V. (2024c). Sharp deviation bounds and concentration phenomenon for the squared norm of a sub-gaussian vector. https://arxiv.org/abs/2305.07885.

Spokoiny, V. and Panov, M. (2021). Accuracy of Gaussian approximation for high-dimensional posterior distributions. *Bernoulli*. in print. arXiv:1910.06028.