Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2017, Vol. 53, No. 1, 389–429
DOI: 10.1214/15-AIHP720

ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES
www.imstat.org/aihp

Penalized maximum likelihood estimation and effective dimension

Vladimir Spokoiny[1]

Weierstrass-Institute and Humboldt University Berlin, Higher School of Economics, IITP RAS, MIPT, Mohrenstr. 39, 10117 Berlin, Germany.
E-mail: spokoiny@wias-berlin.de

Received 29 July 2014; revised 24 August 2015; accepted 5 October 2015

Abstract. This paper extends some prominent statistical results including *Fisher Theorem and Wilks phenomenon* to the penalized maximum likelihood estimation with a quadratic penalization. It appears that sharp expansions for the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ and for the penalized maximum likelihood can be obtained without involving any asymptotic arguments, the results only rely on smoothness and regularity properties of the of the considered log-likelihood function. The error of estimation is specified in terms of the effective dimension $\mathtt{p}_G$ of the parameter set which can be much smaller than the true parameter dimension and even allows an infinite dimensional functional parameter. In the i.i.d. case, the Fisher expansion for the penalized MLE can be established under the constraint "$\mathtt{p}_G^2/n$ is small" while the remainder in the Wilks result is of order $\sqrt{\mathtt{p}_G^3/n}$.

Résumé. Cet article généralise certains résultats statistiques importants dont le *Théorème de Fisher* et le *phénomène de Wilks* à l'estimation du maximum de vraisemblance pénalisée de façon quadratique. Il apparaît que des développements précis pour l'EMV pénalisée $\widetilde{\boldsymbol{\theta}}_G$ et le maximum de vraisemblance pénalisé peuvent être obtenus sans arguments asymptotiques, les résultats reposent alors seulement sur la régularité et les propriétés de la fonction de log-vraisemblance. L'erreur d'estimation est spécifiée en fonction de la dimension effective $\mathtt{p}_G$ de l'ensemble des paramètres qui peut être beaucoup plus petite que la véritable dimension et permet ainsi de considérer un cas infini dimensionnel. Dans le cas i.i.d., le développement de Fisher pour l'EMV pénalisée peut être établi sous la contrainte « $\mathtt{p}_G^2/n$ est petit » tandis que le reste dans le résultat de Wilks est d'ordre $\sqrt{\mathtt{p}_G^3/n}$.

MSC: Primary 62F10; secondary 62J12; 62F25; 62H12

Keywords: Penalty; Wilks and Fisher expansions

1. Introduction

The Fisher and Wilks theorems belong to the short list of most fascinating results in the statistical theory. In particular, the Wilks result in its simple form claims that the likelihood ratio test statistic is close in distribution to the χ^2_p distribution as the sample size increases, where p means the parameter dimension. So, the limiting distribution of this test statistic only depends on the dimension of the parameter space whatever the parametric model is. This explains why this result is sometimes called the *Wilks phenomenon*. This paper aims at reconsidering the mentioned results from different viewpoints. One important issue is that the presented results are stated for *finite samples*. There are only few general finite-sample results in statistical inference; see Boucheron and Massart [7] and references therein in context of i.i.d. modeling. The novel approach from Spokoiny [25] offered a general framework for a *finite sample*

[1]The research was partly supported by the Russian Science Foundation grant (project 14-50-00150). Financial support by the German Research Foundation (DFG) through the Collaborative Research Center 649 "Economic Risk" is gratefully acknowledged.

theory, and the present paper makes a further step in this direction: the classical large sample results are extended to the finite sample case with *explicit and sharp* error bounds.

Another important point is a possible *model misspecification*. The classical parametric theory requires the parametric assumption to be exactly fulfilled. Any violation of the parametric specification may destroy the Fisher and Wilks results; cf. Huber [13]. This study admits from the very beginning that the parametric specification is probably wrong. This automatically extends the applicability of the proposed approach.

The further issue is the use of *penalization* for reducing the *model complexity*. If the parameter dimension is too large, the classical statistical results become almost intractable because the corresponding error is proportional to the dimension of parameter space. Sieve parametric approach is often used to replace the an infinite dimensional problem with a finite dimensional one; see e.g. Shen and Wong [24], Shen [23], Van de Geer [29], Birgé and Massart [4], Barron et al. [2], and references therein. Some asymptotic results for generalized regression models are available in Fan et al. [8].

Another standard way of reducing the complexity of the model is by introducing some penalty in the likelihood function. In this paper we focus on quadratic-type penalization. Roughness penalty approach provides a popular example; cf. Green and Silverman [12]. Koenker et al. [16] explained how roughness penalty works in context of quantile regression. Tikhonov regularization and ridge regression are the other examples which are often used in linear inverse problems. It is well known that the use of a penalization in context of an inverse problem provides regularization and uncertainty reduction at the same time. Our results show that the use of penalization indeed leads to some improvement in the obtained error bounds. Namely, one can replace the original parameter dimension p by the so called *effective dimension* $\mathtt{p}_G$ which can be much smaller than p. Even the case of a functional parameter $\boldsymbol{\theta}$ with $p = \infty$ can be included. In this paper the penalty term is supposed to be given in advance. In general, a model selection procedure based on a proper choice of penalization is a high topic, one of the central in nonparametric statistics. We refer to Shen [23], Birgé and Massart [4], van de Geer [28] for the general models and to Birgé and Massart [5,6] for Gaussian model selection where one can find an extensive overview of the vast literature on this problem.

The final issue is the *critical parameter dimension* which is measured by the effective dimension $\mathtt{p}_G$. The problem of statistical inference for models with growing parameter dimension is quite involved. There are some specific issues even if a simple linear or exponential model is considered, the results from Portnoy [19,20] requires "p^2/n small" for asymptotic normality of the MLE. Depending on the considered problem and the model at hand, the conditions on the critical parameter dimension p may differ. For instance, Portnoy [22] obtained the Fisher and Wilks results for a generalized linear model under $p^{3/2}/n \to 0$, Mammen [18] established similar results for high-dimensional linear models. A general Wilks result can be stated under the condition that p^3/n is small; see e.g. Belloni and Chernozhukov [3]. Below we show that the conditions on the critical dimension in penalized ML estimation can be given in terms of the effective dimension $\mathtt{p}_G$ rather than the parameter dimension p. In particular, in the i.i.d. case, the Fisher expansion can be stated under "$\mathtt{p}_G^2/n$ small" and "$\mathtt{p}_G^3/n$ small" is sufficient for the Wilks result.

First we specify our set-up. Let $\mathbf{Y}$ denote the observed data and $\mathbb{P}$ mean their distribution. A general parametric assumption (PA) means that $\mathbb{P}$ belongs to p-dimensional family $(\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p)$ dominated by a measure $\boldsymbol{\mu}_0$. This family yields the log-likelihood function $L(\boldsymbol{\theta}) = L(\mathbf{Y}, \boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \log \frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\boldsymbol{\mu}_0}(\mathbf{Y})$. The PA can be misspecified, so, in general, $L(\boldsymbol{\theta})$ is a *quasi log-likelihood*. The classical likelihood principle suggests to estimate $\boldsymbol{\theta}$ by maximizing the function $L(\boldsymbol{\theta})$:

$$\widetilde{\boldsymbol{\theta}} \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}). \tag{1.1}$$

If $\mathbb{P} \notin (\mathbb{P}_{\boldsymbol{\theta}})$, then the quasi MLE estimate $\widetilde{\boldsymbol{\theta}}$ from (1.1) is still meaningful and it can be viewed as estimate of the value $\boldsymbol{\theta}^*$ defined by maximizing the expected value of $L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E} L(\boldsymbol{\theta})$$

which is the true value in the parametric situation and can be viewed as the parameter of the best parametric fit in the general case. The classical *Fisher Theorem* claims the expansion for the MLE $\widetilde{\boldsymbol{\theta}}$:

$$D\big(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big) - \boldsymbol{\xi} \stackrel{\mathbb{P}}{\longrightarrow} 0,$$

where $D^2 = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}^*)$ and $\boldsymbol{\xi} \stackrel{\mathrm{def}}{=} D^{-1} \nabla L(\boldsymbol{\theta}^*)$. Under the correct model specification, D^2 is the total Fisher information matrix and the vector $\boldsymbol{\xi}$ is centered and standardized. So, it is asymptotically standard normal under general CLT conditions.

It is well known that many important properties of the quasi MLE $\widetilde{\boldsymbol{\theta}}$ like concentration or coverage probability can be described in terms of the *excess* or *quasi maximum likelihood* $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \stackrel{\mathrm{def}}{=} L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) = \max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$, which is the difference between the maximum of the process $L(\boldsymbol{\theta})$ and its value at the "true" point $\boldsymbol{\theta}^*$. The *Wilks phenomenon* claims that the distribution of the twice excess $2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ can be approximated by $\|\boldsymbol{\xi}\|^2$ which is asymptotically χ^2_p, where p is the dimension of the parameter space:

$$2L\bigl(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*\bigr) - \|\boldsymbol{\xi}\|^2 \xrightarrow{\mathbb{P}} 0, \qquad \|\boldsymbol{\xi}\|^2 \xrightarrow{w} \chi^2_p.$$

This fact is very attractive and yields asymptotic confidence and concentration sets as well as the limiting critical values for the likelihood ratio tests. However, practical applications of all mentioned results are limited: they require true parametric distribution, large samples and a fixed parameter dimension.

Modern applications stimulate a further extension of the classical theory beyond the classical parametric assumptions. Spokoiny [25] offers a general approach which appears to be very useful for such an extension. The whole approach is based on the following local bracketing result:

$$\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}\bigl(\boldsymbol{\theta}, \boldsymbol{\theta}^*\bigr) - \diamondsuit_{\underline{\boldsymbol{\epsilon}}} \le L(\boldsymbol{\theta}) - L\bigl(\boldsymbol{\theta}^*\bigr) \le \mathbb{L}_{\boldsymbol{\epsilon}}\bigl(\boldsymbol{\theta}, \boldsymbol{\theta}^*\bigr) + \diamondsuit_{\boldsymbol{\epsilon}}, \quad \boldsymbol{\theta} \in \Theta_0. \tag{1.2}$$

Here $\mathbb{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ are quadratic in $\boldsymbol{\theta} - \boldsymbol{\theta}^*$ expressions and Θ_0 is a local vicinity of the central point $\boldsymbol{\theta}^*$. This result can be viewed as an extension of the famous Le Cam *local asymptotic normality* (LAN) condition. The LAN condition considers just one quadratic process for approximating the log-likelihood $L(\boldsymbol{\theta})$. The use of bracketing with two different quadratic expressions allows one to keep control of the error terms $\diamondsuit_{\boldsymbol{\epsilon}}, \diamondsuit_{\underline{\boldsymbol{\epsilon}}}$ even for relatively large neighborhoods Θ_0 of $\boldsymbol{\theta}^*$ while the LAN approach is essentially restricted to a root-n vicinity of $\boldsymbol{\theta}^*$. It also allows to incorporate a large parameter dimension and a model misspecification. However, the approach from Spokoiny [25] has natural limitations: the parameter dimension p cannot be too large. For instance, in the i.i.d. case, the error terms $\diamondsuit_{\boldsymbol{\epsilon}}$ and $\diamondsuit_{\underline{\boldsymbol{\epsilon}}}$ are of order $\sqrt{p^3/n}$ which destroys the Wilks result if $p > n^{1/3}$.

A standard way of overcoming this difficulty is to impose a kind of smoothness assumption on the unknown parameter value $\boldsymbol{\theta}^*$. Here we discuss one general way to deal with such smoothness assumptions using a quadratic penalization. Section 2 offers a new approach to studying the properties of the penalized MLE which is based on a linear approximation of the gradient of the log-likelihood process. Compared to the bracketing approach (1.2), it allows to establish a Fisher type expansion for the penalized MLE under weaker conditions on the critical dimension of the problem. Another important novelty of the approach is the systematic use of the *effective dimension* $\mathtt{p}_G$ in place of the original dimension p of the parameter space. Usually $\mathtt{p}_G$ is much smaller than p. It is even possible to treat the case of a functional parameter if the effective dimension of the parameter set remains finite. Our main results include the Fisher and Wilks expansions for the penalized MLE. In the important special case of an i.i.d. model, the error term in the Wilks expansion is of order $\mathtt{p}_G^3/n$, while the Fisher expansion requires $\mathtt{p}_G^2/n$ small.

Also we discuss an implication of these results to the bias-variance decomposition of the squared risk of the penalized MLE. In all our results, the error terms only depend on the effective dimension $\mathtt{p}_G$.

The paper is organized as follows. Section 2 states the analog of Fisher and Wilks results for the penalized MLE procedure. Section 2.5 explains how the Fisher expansion can be used for obtaining the bias-variance decomposition of the quadratic risk of the penalized MLE. Section 3 specifies the general conditions and results to i.i.d. and generalized linear models. Appendix A collects some deviation bounds for general quadratic forms. Appendix B contains technical results on the maximum of a stochastic process under exponential moment conditions.

2. Fisher and Wilks theorems under quadratic penalization

Let $\mathrm{pen}(\boldsymbol{\theta})$ be a penalty function on Θ. A big value of $\mathrm{pen}(\boldsymbol{\theta})$ corresponds to a large degree of roughness or a small amount of smoothness of $\boldsymbol{\theta}$. The underlying assumption on the model is that the true value $\boldsymbol{\theta}^*$ is smooth in the sense

that $\mathrm{pen}(\boldsymbol{\theta}^*)$ is relatively small. A penalized (quasi) MLE approach leads to maximizing the penalized log-likelihood:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \bigl\{ L(\boldsymbol{\theta}) - \mathrm{pen}(\boldsymbol{\theta}) \bigr\}.$$

Below we discuss an important special case of a quadratic penalty $\mathrm{pen}(\boldsymbol{\theta}) = \|G\boldsymbol{\theta}\|^2/2$ for a given symmetric matrix G; see e.g. Green and Silverman [12] or Koenker et al. [16] for particular examples. Denote

$$L_G(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2,$$
$$\widetilde{\boldsymbol{\theta}}_G \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} L_G(\boldsymbol{\theta}).$$

The use of a penalty changes the target of estimation which is now defined as

$$\boldsymbol{\theta}^*_G \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E} L_G(\boldsymbol{\theta}). \tag{2.1}$$

So, introducing a penalty leads to some estimation bias: the new target $\boldsymbol{\theta}^*_G$ may be different from $\boldsymbol{\theta}^*$. At the same time, similarly to linear modeling, the use of penalization reduces the variability of the estimate $\widetilde{\boldsymbol{\theta}}_G$ and improves its concentration properties. An interesting question is the total impact and a possible gain of using the penalized procedure. A preliminary answer is that the penalty term $\|G\boldsymbol{\theta}^*\|^2$ at the true point should not be too large relative to the squared error of estimation for the penalized model. This rule is known under the name "bias-variance trade-off."

Another important message of this study is that the use of penalization allows to reduce the parameter dimension to the *effective dimension* which characterizes the entropy of the penalized parameter space. The resulting confidence and concentration sets depend on the effective dimension rather than on the real parameter dimension and they can be much more narrow than in the non-penalized case.

The principle steps of the study are as follows. The *concentration* step claims that the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ is concentrated in a local vicinity $\Theta_{0,G}(\mathtt{r}_G)$ of the point $\boldsymbol{\theta}^*_G$. It is based on the upper function method which bounds the penalized log-likelihood $L_G(\boldsymbol{\theta})$ from above by a deterministic function. Theorem 2.1 states that $\widetilde{\boldsymbol{\theta}}_G$ belongs to the local set $\Theta_{0,G}(\mathtt{r}_G)$ with a dominating probability, and this local set can be much smaller than the similar set for the non-penalized results. As the next step, Spokoiny [25] applied the *bracketing* approach to bound from above and from below the log-likelihood process $L(\boldsymbol{\theta})$ by two quadratic in $\boldsymbol{\theta} - \boldsymbol{\theta}^*$ expressions. Here the bracketing step is changed essentially by using a local linear approximation of the vector gradient process $\nabla L(\boldsymbol{\theta})$. This helps to get a sharper bound on the error of approximation and improve the quality of the Fisher expansion. Similarly to Spokoiny [25], the obtained results are stated for finite samples and do not involve any asymptotic arguments. An advantage of the proposed approach is that it combines an accurate local approximation with rather rough large deviation arguments and allows one to obtain usual asymptotic statements including asymptotic normality of the penalized MLE; see Section 3.1 for the i.i.d. case.

2.1. *Effective dimension*

Let V^2 be the matrix shown in condition (E_0G) in Section 2.2. Typically $V^2 = \mathrm{Var}\{\nabla L(\boldsymbol{\theta}^*_G)\}$ and this matrix measures the local variability of the process $L_G(\cdot)$. Let also D^2_G be a penalized information matrix defined as

$$D^2_G = -\nabla^2 \mathbb{E} L_G\bigl(\boldsymbol{\theta}^*_G\bigr) = D^2 + G^2$$

with $D^2 = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}^*_G)$. One can redefine $D^2 = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}^*)$ under condition $(\mathcal{L}_0G)$ below and the so called small modeling bias condition; see Section 2.5. The *effective dimension* $\mathtt{p}_G$ is defined as the trace of the matrix $B_G \stackrel{\mathrm{def}}{=} D_G^{-1} V^2 D_G^{-1}$:

$$\mathtt{p}_G \stackrel{\mathrm{def}}{=} \mathrm{tr}(B_G). \tag{2.2}$$

Below we show that the use of penalization enables us to replace the original dimension p in our risk bounds with the effective dimension $\mathtt{p}_G$ which can be much smaller than p depending on relations between the matrices D^2, V^2, and G^2.

In our results the value $\mathtt{p}_G$ will be used via another quantity $z(B_G, \mathtt{x})$ which also depends on a fixed constant $\mathtt{x}$ and for moderate values of $\mathtt{x}$ can be defined as

$$z(B_G, \mathtt{x}) = \sqrt{\mathtt{p}_G} + \sqrt{2\mathtt{x}\lambda_G}, \tag{2.3}$$

where $\lambda_G \stackrel{\text{def}}{=} \lambda_{\max}(B_G)$ is the largest eigenvalue of B_G; see (A.4) for a precise definition.

Now we present a couple of typical examples of using the quadratic penalty: blockwise penalization and estimation under a Sobolev smoothness constraint. For simplicity of presentation we assume that $V^2 = D^2 = n\mathbb{I}_p$, while G^2 is diagonal with non-decreasing eigenvalues g_j^2. Then $D_G^2 = D^2 + G^2 = \operatorname{diag}\{n + g_1^2, \ldots, n + g_p^2\}$. It holds that $B_G = \operatorname{diag}\{(1 + n^{-1}g_1^2)^{-1}, \ldots, (1 + n^{-1}g_p^2)^{-1}\}$, and we apply (2.2) for computing the effective dimension $\mathtt{p}_G$.

Block penalization

Consider the case when G is of a simple two-block structure: $G = \operatorname{diag}\{0, G_1\}$. Many blocks can be considered in the similar way. The first block of dimension p_0 corresponds to the unconstrained part of the parameter vector while the second block of dimension p_1 corresponds to the low energy component. An interesting question is the minimal penalization G_1 making the impact of the low energy part inessential. Assume for simplicity that $G_1 = g\mathbb{I}_{p_1}$. Then

$$\mathtt{p}_G = \operatorname{tr} B_G = p_0 + p_1/\bigl(1 + n^{-1}g^2\bigr).$$

One can see that the impact of the second block G_1 in the effective dimension is inessential if $g^2/n \gg p_1/p_0$.

Sobolev smoothness constraint

Consider the case with $D^2 = V^2 = n\mathbb{I}_p$ and $G^2 = \operatorname{diag}\{g_1^2, \ldots, g_p^2\}$ with $g_j = Lj^\beta$ for $\beta > 1/2$. The value β is usually considered as the Sobolev smoothness parameter. It holds

$$\mathtt{p}_G = \sum_{j=1}^{p} \frac{1}{1 + L^2 j^{2\beta}/n}.$$

Define also the index $\mathtt{p}_e$ as the largest j satisfying $L^2 j^{2\beta} \le n$. It is straightforward to see that $\beta > 1/2$ yields $\mathtt{p}_G \le \mathtt{C}(\beta, L)\mathtt{p}_e$ for a constant $\mathtt{C}(\beta, L)$ depending on β, L only.

Linear inverse problem

The next example corresponds to the case of a linear inverse problem. Assume for simplicity of notation the sequence space representation, the noise is inhomogeneous with increasing eigenvalues $V^2 = \operatorname{diag}\{\mathtt{v}_1^2, \ldots, \mathtt{v}_p^2\}$ and the information matrix D^2 is proportional to identity, that is, $D^2 = n\mathbb{I}_p$. Then the effective dimension is given by the sum

$$\mathtt{p}_G = \sum_{j=1}^{p} \frac{\mathtt{v}_j^2}{n + g_j^2}.$$

To keep the effective dimension small, one has to compensate the increase of the eigenvalues $\mathtt{v}_j^2$ by the penalization g_j^2.

2.2. *Conditions*

This section presents the list of conditions which are similar to ones from the non-penalized case in Spokoiny [25]. However, the use of penalization leads to some change in each condition. Most important fact is that the use of penalization helps to state the large deviation (LD) result for much smaller local neighborhoods than in the non-penalized case. Spokoiny [25] presented the LD result for local sets of the form $\Theta_0(\mathtt{r}) = \{\boldsymbol{\theta} : \|V(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}\}$ with

a proper $\mathtt{r} \asymp p^{1/2}$. Now we redefine this set by using D_G^2 in place of V^2 and $\boldsymbol{\theta}_G^*$ in place of $\boldsymbol{\theta}^*$:

$$\Theta_{0,G}(\mathtt{r}) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\theta}: \bigl\| D_G\bigl(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*\bigr)\bigr\| \le \mathtt{r}\bigr\}.$$

Moreover, the radius $\mathtt{r}$ can be selected of order $\mathtt{p}_G^{1/2}$, which can be very useful for large or infinite p. Our conditions mainly assume some regularity and smoothness of the penalized log-likelihood process $L_G(\boldsymbol{\theta})$. The first condition states some smoothness properties of the expected log-likelihood $\mathbb{E}L_G(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ in a vicinity $\Theta_{0,G}(\mathtt{r})$ of $\boldsymbol{\theta}_G^*$. More precisely, it effectively means that the expected log-likelihood $\mathbb{E}L(\boldsymbol{\theta})$ is twice continuously differentiable on the local set $\Theta_{0,G}(\mathtt{r})$.

Below each condition is given in penalized and non-penalized form for the sake of comparison. Already now it is worth saying that the use of penalization helps to relax most of conditions. Define

$$\mathbb{F}_G(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}L_G(\boldsymbol{\theta}) = -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}) + G^2.$$

Then $D_G^2 = \mathbb{F}_G(\boldsymbol{\theta}_G^*)$. The conditions involve a radius $\mathtt{r}_G$ which separates the local zone and the zone of large deviations. This value will be made precise in Theorem 2.1.

First we consider the stochastic component of the log-likelihood process $L_G(\boldsymbol{\theta})$ which is the same as in the non-penalized case:

$$\zeta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L_G(\boldsymbol{\theta}) - \mathbb{E}L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta}).$$

We assume that it is twice differentiable and denote by $\nabla\zeta(\boldsymbol{\theta})$ its gradient and by $\nabla^2\zeta(\boldsymbol{\theta})$ its Hessian matrix. The next two conditions are to ensure that the random vector $\nabla\zeta(\boldsymbol{\theta}_G^*)$ and the random processes $\nabla^2\zeta(\boldsymbol{\theta})$ are stochastically bounded with exponential moments. The conditions involve a $p \times p$-matrix V which normalizes the vector $\nabla\zeta(\boldsymbol{\theta}_G^*)$, and a similar matrix V_2 normalizing $\nabla^2\zeta(\boldsymbol{\theta})$.

(E_0G) There exist a positively semi-definite symmetric matrix V^2, and constants $\mathtt{g} > 0$, $\nu_0 \ge 1$ such that $\operatorname{Var}\{\nabla\zeta(\boldsymbol{\theta}_G^*)\} \le V^2$ and

$$\sup_{\boldsymbol{\gamma}\in\mathbb{R}^p} \log \mathbb{E}\exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta(\boldsymbol{\theta}_G^*)}{\|V\boldsymbol{\gamma}\|}\right\} \le \frac{\nu_0^2\lambda^2}{2}, \quad |\lambda| \le \mathtt{g}.$$

(E_2G) There exist a positively semi-definite symmetric matrix V_2^2 and a value $\omega > 0$ such that it holds for any $\boldsymbol{\theta} \in \Theta_{0,G}(\mathtt{r}_0)$:

$$\sup_{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\in\mathbb{R}^p} \log \mathbb{E}\exp\left\{\frac{\lambda}{\omega} \frac{\boldsymbol{\gamma}_1^\top \nabla^2\zeta(\boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|V_2\boldsymbol{\gamma}_1\| \cdot \|V_2\boldsymbol{\gamma}_2\|}\right\} \le \frac{\nu_0^2\lambda^2}{2}, \quad |\lambda| \le \mathtt{g}.$$

Below we only need that the constant $\mathtt{g}(\mathtt{r})$ is larger than $\mathtt{C}\mathtt{p}_G$ for a fixed constant $\mathtt{C}$. This allows to reduce the condition to the case with a fixed $\mathtt{g}$ which does not depend on the distance $\mathtt{r}$.

Their non-penalized versions are almost identical: one has to replace $\boldsymbol{\theta}_G^*$ with $\boldsymbol{\theta}^*$ and $\Theta_{0,G}(\mathtt{r})$ with $\Theta_0(\mathtt{r})$.

(E_0) $$\sup_{\boldsymbol{\gamma}\in\mathbb{R}^p} \log \mathbb{E}\exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta(\boldsymbol{\theta}^*)}{\|V\boldsymbol{\gamma}\|}\right\} \le \frac{\nu_0^2\lambda^2}{2}, \quad |\lambda| \le \mathtt{g}.$$

(E_2) $$\sup_{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\in\mathbb{R}^p} \log \mathbb{E}\exp\left\{\frac{\lambda}{\omega} \frac{\boldsymbol{\gamma}_1^\top \nabla^2\zeta(\boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|V_2\boldsymbol{\gamma}_1\| \cdot \|V_2\boldsymbol{\gamma}_2\|}\right\} \le \frac{\nu_0^2\lambda^2}{2}, \quad |\lambda| \le \mathtt{g}(\mathtt{r}).$$

The conditions (E_0) and (E_0G) are very similar while (E_2G) is restricted to the vicinity $\Theta_{0,G}(\mathtt{r})$ which can be much smaller than $\Theta_0(\mathtt{r})$.

The *identifiability condition* relates the matrices V^2 and V_2^2 and to D_G^2.

$(\mathcal{I}G)$ There is a constant $\mathfrak{a}_G > 0$ such that

$$\mathfrak{a}_G^2 D_G^2 \ge V^2, \qquad \mathfrak{a}_G^2 D_G^2 \ge V_2^2.$$

In the non-penalized case of Spokoiny [25], this condition reads as

$(\mathcal{I})$ $\mathfrak{a}^2 D^2 \geq V^2$ with $D^2 = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}^*)$.

Therefore, the use of regularization helps to improve the identifiability in the regularized problem relative to the non-penalized one as $D^2 \leq D_G^2$.

Finally, we consider the expected log-likelihood $\mathbb{E} L_G(\boldsymbol{\theta})$. The local condition requires that it is nearly quadratic in the vicinity $\Theta_{0,G}(\mathtt{r}_G)$ of $\boldsymbol{\theta}_G^*$ while the global condition assumes a linear growth in the complement of this vicinity. Here and below $\|A\|_{\mathrm{op}}$ means the operator norm of a matrix A.

$(\mathcal{L}_0 G)$ For each $\mathtt{r} \leq \mathtt{r}_G$, there is a constant $\delta_G(\mathtt{r}) \leq 1/2$ such that

$$\bigl\| D_G^{-1} \mathbb{F}_G(\boldsymbol{\theta}) D_G^{-1} - \mathbb{I}_p \bigr\|_{\mathrm{op}} \leq \delta_G(\mathtt{r}), \quad \boldsymbol{\theta} \in \Theta_{0,G}(\mathtt{r}). \tag{2.4}$$

Under condition $(\mathcal{L}_0 G)$, it follows from the second order Taylor expansion at $\boldsymbol{\theta}_G^*$:

$$\bigl| -2\mathbb{E} L_G\bigl(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*\bigr) - \bigl\| D_G\bigl(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*\bigr) \bigr\|^2 \bigr| \leq \delta_G(\mathtt{r}) \bigl\| D_G\bigl(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*\bigr) \bigr\|^2, \quad \boldsymbol{\theta} \in \Theta_{0,G}(\mathtt{r}).$$

A non-penalized version of (2.4) claims a similar approximation of $\mathbb{F}(\boldsymbol{\theta}) = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta})$ by $D^2 \stackrel{\mathrm{def}}{=} \mathbb{F}(\boldsymbol{\theta}^*)$ in the vicinity $\Theta_0(\mathtt{r}_0)$ centered at $\boldsymbol{\theta}^*$ instead of $\boldsymbol{\theta}_G^*$:

$$(\mathcal{L}_0) \qquad \bigl\| D^{-1} \mathbb{F}(\boldsymbol{\theta}) D^{-1} - \mathbb{I}_p \bigr\|_{\mathrm{op}} \leq \delta(\mathtt{r}_0), \qquad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0) = \bigl\{ \boldsymbol{\theta} \colon \bigl\| D\bigl(\boldsymbol{\theta} - \boldsymbol{\theta}^*\bigr) \bigr\| \leq \mathtt{r}_0 \bigr\}.$$

As the quadratic penalty $\|G\boldsymbol{\theta}\|^2$ does not change the smoothness properties of the expected contrast $\mathbb{E} L_G(\boldsymbol{\theta})$, the conditions $(\mathcal{L}_0 G)$ and $(\mathcal{L}_0)$ are essentially equivalent provided that the points $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_G^*$ are not too far from each others.

The local condition $(\mathcal{L}_0 G)$ describes the behavior of $\mathbb{E} L_G(\boldsymbol{\theta})$ within $\Theta_{0,G}(\mathtt{r}_G)$. In particular, $\mathbb{E} L_G(\boldsymbol{\theta}_G^*) - \mathbb{E} L_G(\boldsymbol{\theta}) \approx \mathtt{r}_G^2/2$ on the boundary of this local set. The global condition means that $\mathbb{E} L_G(\boldsymbol{\theta}_G^*) - \mathbb{E} L_G(\boldsymbol{\theta})$ can be lower bounded by a linear function on the complement of this set.

$(\mathcal{L} G)$ For each $\boldsymbol{\theta}$ with $\mathtt{r} = \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \geq \mathtt{r}_G$

$$\mathbb{E} L_G\bigl(\boldsymbol{\theta}_G^*\bigr) - \mathbb{E} L_G(\boldsymbol{\theta}) \geq \bigl(1 - \delta(\mathtt{r}_G)\bigr) \biggl(\mathtt{r}_G \mathtt{r} - \frac{\mathtt{r}_G^2}{2} \biggr) + \mathtt{c}_1 \mathtt{r}^2, \tag{2.5}$$

for a small constant $\mathtt{c}_1$; see Theorem 2.1 below for a precise bound.

A non-penalized version of this condition is obtained by letting $G^2 = 0$.

$$(\mathcal{L}) \qquad \mathbb{E} L\bigl(\boldsymbol{\theta}^*\bigr) - \mathbb{E} L(\boldsymbol{\theta}) \geq \bigl(1 - \delta(\mathtt{r}_0)\bigr)\bigl(\mathtt{r}_0 \mathtt{r} - \mathtt{r}_0^2/2\bigr) + \mathtt{c}_1 \mathtt{r}^2 \quad \text{for } \mathtt{r} = \bigl\| D\bigl(\boldsymbol{\theta} - \boldsymbol{\theta}^*\bigr) \bigr\|.$$

Remark 2.1. *Conditions $(\mathcal{L}_0 G)$ and $(\mathcal{L} G)$ can be effectively checked if the function $f(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} -\mathbb{E} L_G(\boldsymbol{\theta})$ is smooth and convex in $\boldsymbol{\theta}$. Continuity of the second derivative $\nabla^2 f(\boldsymbol{\theta})$ in $\Theta_{0,G}(\mathtt{r}_G)$ implies $(\mathcal{L}_0 G)$. Convexity of f implies for any $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}_G^* + \rho(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)$ with $\rho = \mathtt{r}_G / \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \leq 1$*

$$f(\boldsymbol{\theta}) \geq f\bigl(\boldsymbol{\theta}^\circ\bigr) + \bigl(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\bigr)^\top \nabla f\bigl(\boldsymbol{\theta}^\circ\bigr) \geq f\bigl(\boldsymbol{\theta}^\circ\bigr) + \bigl\{ D_G\bigl(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\bigr) \bigr\}^\top D_G^{-1} \nabla f\bigl(\boldsymbol{\theta}^\circ\bigr).$$

Condition $(\mathcal{L}_0 G)$ implies in view of $\|D_G(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}_G^*)\| = \mathtt{r}_G$ that $f(\boldsymbol{\theta}^\circ) \geq (1 - \delta)\mathtt{r}_G^2/2$ and $\nabla^2 f(\boldsymbol{\theta}^\circ) \geq (1 - \delta) D_G^2$ for $\delta = \delta(\mathtt{r}_G)$. As $\|D_G(\boldsymbol{\theta}^\circ - \boldsymbol{\theta})\| = \mathtt{r} - \mathtt{r}_G$, we conclude that

$$f(\boldsymbol{\theta}) \geq (1 - \delta) \biggl\{ \frac{\mathtt{r}_G^2}{2} + \mathtt{r}_G(\mathtt{r} - \mathtt{r}_G) \biggr\}$$

and (2.5) follows for $\mathtt{c}_1 = 0$. The case $\mathtt{c}_1 > 0$ can be similarly checked under a strong convexity of $-\mathbb{E} L_G(\boldsymbol{\theta})$. In the case of linear of generalized linear models, one can use $\mathtt{c}_1 = 0$, in regular situations, it suffices that $(\mathcal{L} G)$ holds with $\mathtt{c}_1$ of order $n^{-1/2}$.

We briefly comment on examples for which the conditions can be easily verified. Conditions (E_0G) and (E_2G) require some exponential moments of log-likelihood ratio and its derivatives. Usually one assumes some finite moments of the normalized increments of the likelihood function; cf. Ibragimov and Khas'minskij [14], Chapter 2. Our conditions (E_0G) and (E_2G) a bit more restrictive but it allows one to obtain some finite sample bounds. Note that majority of finite samples results are stated under Gaussian or sub-Gaussian stochastic errors. The sub-Gaussian case corresponds to $\mathtt{g} = \infty$ in (E_0G) and (E_2G). Our results apply for sub-exponential errors with $\mathtt{g} < \infty$ as well. Condition $(\mathcal{L}_0G)$ only requires some regularity of the considered parametric family and is not restrictive. As already mentioned, condition $(\mathcal{L}G)$ can be easily checked if $\mathbb{E}L(\boldsymbol{\theta})$ is smooth and concave in $\boldsymbol{\theta}$. It is also easy to verify if $\nabla^2\mathbb{E}L_G(\boldsymbol{\theta})$ is bounded from below by a positive matrix.

The i.i.d. case will be considered in details below in Section 3.1. Section 3.6 demonstrates how the conditions can be checked for generalized linear models in terms of design regularity, smoothness of the link function, and exponential moment conditions on the errors. The regression and generalized regression models are included as well; cf. Ghosal [9,10] or Kim [15]. Spokoiny [25], Section 5.2, argued that (E_0G) and (E_2G) are fulfilled when regression errors fulfill some exponential moments condition. If this condition is too restrictive and a more stable (robust) estimation procedure is desirable, one can apply the LAD-type contrast leading to median regression. Spokoiny [25], Section 5.3, showed for the case of linear median regression that all the required conditions are fulfilled automatically if the sample size n exceeds $\mathtt{C}p$ for a fixed constant $\mathtt{C}$. Spokoiny et al. [26] applied this approach for local polynomial quantile regression. Zaitsev et al. [31] applied the approach to the problem of regression with Gaussian process where the unknown parameters enter in the likelihood function in a rather complicated way. We conclude that the imposed conditions are quite general and can be verified for many classical examples met in the statistical literature.

2.3. *Concentration and a large deviation bound*

This section demonstrates that the use of the penalty term helps to strengthen the concentration properties of the penalized quasi maximum likelihood estimator (qMLE) $\widetilde{\boldsymbol{\theta}}_G$. Namely, we show that $\widetilde{\boldsymbol{\theta}}_G$ belongs with a dominating probability to a set $\Theta_{0,G}(\mathtt{r}_G)$ which can be much smaller than a similar set from the non-penalized case; see Remark 2.2. All our results involve a value $\mathtt{x}$. We say that a generic random set $\Omega(\mathtt{x})$ is of a *dominating probability* if $\mathbb{P}(\Omega(\mathtt{x})) \geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$ for a fixed constant $\mathtt{C}$ like 1 or 2. We also use two growing functions $z(B_G, \mathtt{x})$ and $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ of the argument $\mathtt{x}$. The functions $z(B_G, \mathtt{x})$ already mentioned in (2.3) and it describes the quantiles of the norm of the normalized score vector $\boldsymbol{\xi}_G$; see (2.8) below. The formal definition of $z(B_G, \mathtt{x})$ is given in (A.4). The function $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ is related to the penalized entropy of the parameter space and it is given by (B.3). In typical situations one can use the upper bounds $z^2(B_G, \mathtt{x}) \leq \mathtt{C}(\mathtt{p}_G + \mathtt{x})$ and $\mathfrak{z}^2_{\mathbb{H}}(\mathtt{x}) \leq \mathtt{C}(\mathtt{p}_G + \mathtt{x})$ for both functions.

Theorem 2.1. *Let* (E_0G), (E_2G), $(\mathcal{I}G)$, $(\mathcal{L}_0G)$, *and* $(\mathcal{L}G)$ *hold with*

$$\bigl\{1 - \delta(\mathtt{r}_G)\bigr\}\mathtt{r}_G \geq 2z(B_G, \mathtt{x}), \tag{2.6}$$

where $z(B_G, \mathtt{x})$ *is from* (A.4), *and let the constant* $\mathtt{C}_1$ *in* $(\mathcal{L}G)$ *satisfy*

$$\mathtt{C}_1 \geq \sup_{\mathtt{r} > \mathtt{r}_G} \varrho_G(\mathtt{r}, \mathtt{x}), \quad \textit{with } \varrho_G(\mathtt{r}, \mathtt{x}) \stackrel{\mathrm{def}}{=} \sqrt{8}\nu_0\mathfrak{a}_G\mathfrak{z}_{\mathbb{H}}\bigl(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_G)\bigr)\omega$$

with the function $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ *given by* (B.3). *Then*

$$\mathbb{P}\bigl(\widetilde{\boldsymbol{\theta}}_G \notin \Theta_{0,G}(\mathtt{r}_G)\bigr) \leq 3\mathrm{e}^{-\mathtt{x}}. \tag{2.7}$$

Remark 2.2. *This result explains a proper* $\mathtt{r}_G$ *ensuring* (2.7). *In the non-penalized case of Spokoiny* [25], *a similar condition reads as* $\mathtt{r}_0 \geq \mathtt{C}(\sqrt{p} + \sqrt{2\mathtt{x}})$, *so the use of penalization helps to improve the concentration properties of the penalized MLE. We conclude that the use of penalization leads to weaker conditions and to a stronger concentration property. The only problem is that the corresponding estimate* $\widetilde{\boldsymbol{\theta}}_G$ *concentrates around* $\boldsymbol{\theta}^*_G$ *instead of* $\boldsymbol{\theta}^*$. *This can yield a bias effect*; *see Section 2.5 below.*

Proof of Theorem 2.1. By definition $\sup_{\boldsymbol{\theta}\in\Theta_{0,G}(\mathtt{r}_G)} L_G(\boldsymbol{\theta},\boldsymbol{\theta}^*_G) \geq 0$. So, it suffices to check that $L_G(\boldsymbol{\theta},\boldsymbol{\theta}^*_G) < 0$ for all $\boldsymbol{\theta}\in\Theta\setminus\Theta_{0,G}(\mathtt{r}_G)$. The proof is based on the following bound: for each $\mathtt{r}$

$$\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_{0,G}(\mathtt{r})} \big|\zeta\big(\boldsymbol{\theta},\boldsymbol{\theta}^*_G\big) - \big(\boldsymbol{\theta}-\boldsymbol{\theta}^*_G\big)^\top \nabla\zeta\big(\boldsymbol{\theta}^*_G\big)\big| \geq \sqrt{8}\nu_0\mathfrak{a}_G\mathfrak{z}_{\mathbb{H}}(\mathtt{x})\omega\mathtt{r}^2\Big) \leq \mathrm{e}^{-\mathtt{x}}.$$

This bound follows from Theorem B.15; see (2.18) for more details. It implies by Theorem B.3 with $\rho = 1/2$ on a set of dominating probability at least $1-\mathrm{e}^{-\mathtt{x}}$ that for all $\mathtt{r}\geq\mathtt{r}_G$ and all $\boldsymbol{\theta}$ with $\|D_G(\boldsymbol{\theta}-\boldsymbol{\theta}^*_G)\|\leq\mathtt{r}$

$$\big|\zeta\big(\boldsymbol{\theta},\boldsymbol{\theta}^*_G\big) - \big(\boldsymbol{\theta}-\boldsymbol{\theta}^*_G\big)^\top \nabla\zeta\big(\boldsymbol{\theta}^*_G\big)\big| \leq \varrho_G(\mathtt{r},\mathtt{x})\mathtt{r}^2,$$

where $\varrho_G(\mathtt{r},\mathtt{x}) = \nu_0\mathfrak{a}_G\mathfrak{z}_{\mathbb{H}}(\mathtt{x}+\log(2\mathtt{r}/\mathtt{r}_G))\omega$. The use of $\nabla\mathbb{E}L_G(\boldsymbol{\theta}^*_G)=0$ yields

$$\sup_{\boldsymbol{\theta}\in\Theta_{0,G}(\mathtt{r})} \big|L_G\big(\boldsymbol{\theta},\boldsymbol{\theta}^*_G\big) - \mathbb{E}L_G\big(\boldsymbol{\theta},\boldsymbol{\theta}^*_G\big) - \big(\boldsymbol{\theta}-\boldsymbol{\theta}^*_G\big)^\top \nabla L_G\big(\boldsymbol{\theta}^*_G\big)\big| \leq \varrho_G(\mathtt{r},\mathtt{x})\mathtt{r}^2.$$

Also the vector $\boldsymbol{\xi}_G = D_G^{-1}\nabla L_G(\boldsymbol{\theta}^*_G) = D_G^{-1}\nabla\zeta(\boldsymbol{\theta}^*_G)$ can be bounded with a dominating probability: by Theorem A.1 $\mathbb{P}(\|\boldsymbol{\xi}_G\|\geq z(B_G,\mathtt{x}))\leq 2\mathrm{e}^{-\mathtt{x}}$. We ignore here the negligible term $\mathtt{C}\mathrm{e}^{-\mathtt{x}_c}$. The condition $\|\boldsymbol{\xi}_G\|\leq z(B_G,\mathtt{x})$ implies for each $\mathtt{r}\geq\mathtt{r}_G$

$$\begin{aligned}&\sup_{\boldsymbol{\theta}\in\Theta_{0,G}(\mathtt{r})} \big|\big(\boldsymbol{\theta}-\boldsymbol{\theta}^*_G\big)^\top \nabla L_G\big(\boldsymbol{\theta}^*_G\big)\big| \\ &\quad\leq \sup_{\boldsymbol{\theta}\in\Theta_{0,G}(\mathtt{r})} \big\|D_G\big(\boldsymbol{\theta}-\boldsymbol{\theta}^*_G\big)\big\| \times \big\|D_G^{-1}\nabla\zeta\big(\boldsymbol{\theta}^*_G\big)\big\| = \mathtt{r}\|\boldsymbol{\xi}_G\| \leq z(B_G,\mathtt{x})\mathtt{r}.\end{aligned}$$

Condition $(\mathcal{L}G)$ implies for each $\boldsymbol{\theta}$ with $\|D_G(\boldsymbol{\theta}-\boldsymbol{\theta}^*_G)\|=\mathtt{r}>\mathtt{r}_0$ that

$$-\mathbb{E}L_G\big(\boldsymbol{\theta},\boldsymbol{\theta}^*_G\big) > (1-\delta)\Big(\mathtt{r}_0\mathtt{r}-\frac{1}{2}\mathtt{r}_0^2\Big) + \mathtt{C}_1\mathtt{r}^2 \geq z(B_G,\mathtt{x})\mathtt{r} + \varrho_G(\mathtt{r},\mathtt{x})\mathtt{r}^2$$

provided that $\mathtt{r}_0\geq 2(1-\delta)^{-1}z(B_G,\mathtt{x})$ and $\mathtt{C}_1\geq\varrho_G(\mathtt{r},\mathtt{x})$. This ensures that $L_G(\boldsymbol{\theta},\boldsymbol{\theta}^*_G)<0$ for all $\boldsymbol{\theta}\notin\Theta_{0,G}(\mathtt{r}_G)$ with a dominating probability. □

2.4. *Wilks and Fisher expansions*

This section collects the main results of the paper. Let $\boldsymbol{\theta}^*_G$ be the point of concentration from (2.1) and let $\zeta(\boldsymbol{\theta}) = L_G(\boldsymbol{\theta}) - \mathbb{E}L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{E}L(\boldsymbol{\theta})$. Define a random p-vector

$$\boldsymbol{\xi}_G \stackrel{\mathrm{def}}{=} D_G^{-1}\nabla\zeta\big(\boldsymbol{\theta}^*_G\big) = D_G^{-1}\big\{\nabla L\big(\boldsymbol{\theta}^*_G\big) - G^2\boldsymbol{\theta}^*_G\big\}. \tag{2.8}$$

Theorem 2.2. *Suppose that $\mathtt{r}_G$ is selected to ensure (2.6). Suppose also that the conditions (E_0G), (E_2G), $(\mathcal{I}G)$ hold. On a random set $\Omega(\mathtt{x})$ of a dominating probability at least $1-4\mathrm{e}^{-\mathtt{x}}$, it holds*

$$\big\|D_G\big(\widetilde{\boldsymbol{\theta}}_G-\boldsymbol{\theta}^*_G\big) - \boldsymbol{\xi}_G\big\| \leq \diamondsuit_G(\mathtt{x}), \tag{2.9}$$

where $\diamondsuit_G(\mathtt{x})$ is given by

$$\diamondsuit_G(\mathtt{x}) \stackrel{\mathrm{def}}{=} \big\{\delta_G(\mathtt{r}_G) + \sqrt{8}\nu_0\mathfrak{a}_G\mathfrak{z}_{\mathbb{H}}(\mathtt{x})\omega\big\}\mathtt{r}_G \tag{2.10}$$

for $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ given by (B.3).

The proof of this and the next result is based on a linear expansion of the gradient $\nabla L_G(\boldsymbol{\theta})$ and will be given in Section 2.6.

Now we present a result on the excess $L_G(\widetilde{\boldsymbol{\theta}}_G,\boldsymbol{\theta}^*_G) = L_G(\widetilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}^*_G)$. The classical Wilks result claims that the twice excess is nearly χ^2_p. Our result describes the quality of its approximation by a quadratic form $\|\boldsymbol{\xi}_G\|^2$.

Theorem 2.3. *Suppose that* $(\mathcal{L}_0 G)$, $(E_0 G)$, *and* $(E_2 G)$ *hold. Suppose also that* $\mathtt{r}_G$ *is selected to ensure* (2.6). *On a random set* $\Omega(\mathtt{x})$ *of a dominating probability at least* $1 - 5\mathrm{e}^{-\mathtt{x}}$, *it holds with* $\diamondsuit_G(\mathtt{x})$ *from* (2.10)

$$\bigl|2L_G\bigl(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G\bigr) - \|\boldsymbol{\xi}_G\|^2\bigr| \le 2\mathtt{r}_G \diamondsuit_G(\mathtt{x}) + \diamondsuit^2_G(\mathtt{x}), \tag{2.11}$$

$$\Bigl|\sqrt{2L_G\bigl(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G\bigr)} - \|\boldsymbol{\xi}_G\|\Bigr| \le 3\diamondsuit_G(\mathtt{x}). \tag{2.12}$$

One can see that the Fisher expansion (2.9) and the square root Wilks expansion (2.12) require $\diamondsuit_G(\mathtt{x})$ small, while the standard Wilks expansion (2.11) is accurate if $\mathtt{r}_G \diamondsuit_G(\mathtt{x})$ is small. This makes some difference if the parameter dimension is large. Below we address this question for the important special case of an i.i.d. likelihood.

The classical Fisher and Wilks results include some statements about the limiting behavior of the vector $\boldsymbol{\xi}_G$ and of the quadratic form $\|\boldsymbol{\xi}_G\|^2$. In the i.i.d. case, one can easily show that the vector $\boldsymbol{\xi}_G$ is asymptotically standard normal as $n \to \infty$; see Section 3.5 below. However, it is well known that the convergence of $\|\boldsymbol{\xi}_G\|^2$ to the χ^2-distribution is quite slow even in the case of a fixed dimension p. For finite sample inference, we recommend to combine the approximations (2.9) to (2.12) with any resampling technique which mimics the specific behavior of the quadratic form $\|\boldsymbol{\xi}_G\|^2$; see Spokoiny and Zhilova [27].

2.5. *Quadratic risk bound and modeling bias*

This section demonstrates the applicability of the obtained general results to bounding the quadratic risk of estimation. For the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ of the parameter $\boldsymbol{\theta}$, consider the quadratic loss of estimation $\|W(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2$ for a given non-negative symmetric matrix W. A special case includes the usual quadratic loss $\|\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\|^2$. Here the point $\boldsymbol{\theta}^* \in \Theta$ is a proxy for the true parameter value which describes the best parametric fit of the true measure $\mathbb{P}$ by the family $(\mathbb{P}_{\boldsymbol{\theta}})$:

$$\boldsymbol{\theta}^* \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E} L(\boldsymbol{\theta}).$$

The use of penalization $\|G\boldsymbol{\theta}\|^2/2$ introduces some estimation bias: the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ estimates $\boldsymbol{\theta}^*_G$ from (2.1) rather than $\boldsymbol{\theta}^*$. The value $\|W(\boldsymbol{\theta}^* - \boldsymbol{\theta}^*_G)\|^2$ is called the modeling bias and it describes the modeling error caused by using the penalization. The variance term $\|W(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G)\|^2$ describes the error *within the penalized model*, and it can be studied with the help of the Fisher expansion of Theorem 2.2: $\|D_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G) - \boldsymbol{\xi}_G\| \le \diamondsuit_G(\mathtt{x})$ on a set $\Omega(\mathtt{x})$ of dominating probability for $\boldsymbol{\xi}_G = D_G^{-1}\nabla\zeta(\boldsymbol{\theta}^*_G)$. This yields the following result on $\Omega(\mathtt{x})$:

$$\bigl\|D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^* - b_G\bigr) - \boldsymbol{\xi}_G\bigr\| \le \diamondsuit_G(\mathtt{x})$$

with the *bias* $b_G = \boldsymbol{\theta}^*_G - \boldsymbol{\theta}^*$. For any positive symmetric $p \times p$ matrix W satisfying $W^2 \le D_G^2$, it implies the probability bound for the squared loss

$$\bigl\|W\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\bigr)\bigr\| = \bigl\|Wb_G + WD_G^{-1}\boldsymbol{\xi}_G\bigr\| \pm \diamondsuit_G(\mathtt{x}).$$

One can see that analysis of the quadratic risk of the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ can be reduced to the analysis of $\|Wb_G + WD_G^{-1}\boldsymbol{\xi}_G\|^2$. Now we consider an implication of this bound to the squared risk $\mathbb{E}\|W(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2$. The use of the identity $\mathbb{E}\nabla\zeta(\boldsymbol{\theta}^*_G) = 0$ and $\operatorname{Var}(\nabla\zeta(\boldsymbol{\theta}^*_G)) \le V^2$ yields

$$\begin{aligned}
\mathbb{E}\bigl\|Wb_G + WD_G^{-1}\boldsymbol{\xi}_G\bigr\|^2 &= \|Wb_G\|^2 + \mathbb{E}\bigl\|WD_G^{-2}\nabla\zeta\bigl(\boldsymbol{\theta}^*_G\bigr)\bigr\|^2 \\
&= \|Wb_G\|^2 + \operatorname{tr}\bigl(WD_G^{-2}\operatorname{Var}\bigl\{\nabla\zeta\bigl(\boldsymbol{\theta}^*_G\bigr)\bigr\}D_G^{-2}W\bigr) \\
&\le \|Wb_G\|^2 + \operatorname{tr}\bigl(WD_G^{-2}V^2D_G^{-2}W\bigr).
\end{aligned}$$

Denote $\mathcal{X}_G \stackrel{\mathrm{def}}{=} \operatorname{tr}(WD_G^{-2}V^2D_G^{-2}W)$ and

$$\mathcal{R}_G \stackrel{\mathrm{def}}{=} \|Wb_G\|^2 + \mathcal{X}_G = \|Wb_G\|^2 + \operatorname{tr}\bigl(WD_G^{-2}V^2D_G^{-2}W\bigr). \tag{2.13}$$

Theorem 2.4. *Let* (E_0G), (E_2G), $(\mathcal{L}_0G)$, $(\mathcal{I}G)$, *and* $(\mathcal{L}G)$ *hold. If* $W^2 \le D_G^2$, *then it holds with* $\mathcal{R}_G$ *from* (2.13)

$$\mathbb{E}\bigl\|W\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\bigr)\bigr\|^2 \le \bigl\{\mathcal{R}_G^{1/2} + \diamondsuit_G^*\bigr\}^2, \tag{2.14}$$

where

$$\diamondsuit_G^* = 4\bigl\{\delta_G(\mathtt{r}_G)\mathtt{r}_G + 2\nu_0\mathfrak{a}_G\mathtt{r}_G(\mathbb{H}_1 + \mathbb{H}_2/\mathtt{g} + 4)\omega\bigr\}.$$

Remark 2.3. *If the error term* $\diamondsuit_G^*$ *in* (2.14) *is relatively small, this result implies* $\mathbb{E}\|W(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2 \approx \mathcal{R}_G = \|D_G b_G\|^2 + \mathcal{X}_G$. *This is the usual decomposition of the quadratic risk in term of the squared bias* $\|W(\boldsymbol{\theta}_G^* - \boldsymbol{\theta}^*)\|^2$ *and the variance term* $\mathcal{X}_G$. *The condition "*$\|Wb_G\|^2/\mathcal{X}_G$ *is small" yields* $\mathcal{R}_G \approx \mathcal{X}_G$. *This condition can be naturally called the* small modeling bias (*SMB*) *condition*, *often it is referred to as* undersmoothing. *The bias-variance trade-off corresponds to the situation with* $\|Wb_G\|^2 \asymp \mathcal{X}_G$. Oversmoothing *means that the bias terms* $\|Wb_G\|^2$ *dominates*.

Remark 2.4. *As already mentioned, the result* (2.14) *is informative if the remainder* $\diamondsuit_G^*$ *is relatively small and can be ignored. For the special case* $W^2 = D_G^2$, *it holds* $\mathcal{X}_G = \mathtt{p}_G \asymp \mathtt{r}_G^2$. *In the i.i.d. situation* (*see Section* 3.5 *below*)

$$\mathtt{r}_G^{-1}\diamondsuit_G^* \le \mathtt{C}\sqrt{\mathtt{p}_G/n}$$

which yields a sharp risk bound $\mathbb{E}\|W(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\|^2 = \mathcal{R}_G(1 + o(1))$ *under "*$\mathtt{p}_G/n$ *small."*

Remark 2.5. *The bias induced by penalization can be measured in terms of the value* $\|G\boldsymbol{\theta}^*\|^2$. *To be more precise, consider the case with* $W^2 = D^2$, *where* $D^2 = -\nabla^2\mathbb{E}L(\boldsymbol{\theta}^*)$ *is the non-penalized Fisher information matrix. The definition of* $\boldsymbol{\theta}^*$ *and* $\boldsymbol{\theta}_G^*$ *implies*

$$\mathbb{E}L\bigl(\boldsymbol{\theta}^*\bigr) - \bigl\|G\boldsymbol{\theta}^*\bigr\|^2/2 \le \mathbb{E}L\bigl(\boldsymbol{\theta}_G^*\bigr) - \bigl\|G\boldsymbol{\theta}_G^*\bigr\|^2/2 \le \mathbb{E}L\bigl(\boldsymbol{\theta}_G^*\bigr).$$

Condition $(\mathcal{L}_0G)$ *implies* $\mathbb{E}L(\boldsymbol{\theta}^*) - \mathbb{E}L(\boldsymbol{\theta}_G^*) \approx \|D(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*)\|^2/2$ *and*

$$\bigl\|D\bigl(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*\bigr)\bigr\|^2 \le \bigl\|G\boldsymbol{\theta}^*\bigr\|^2 - \bigl\|G\boldsymbol{\theta}_G^*\bigr\|^2 \le \bigl\|G\boldsymbol{\theta}^*\bigr\|^2.$$

So, if the true point is "smooth" in there sense that $\|G\boldsymbol{\theta}^*\|^2$ *is small, then the squared bias* $\|D(\boldsymbol{\theta}^* - \boldsymbol{\theta}_G^*)\|^2$ *caused by penalization is small as well.*

Proof of Theorem 2.4. The Fisher expansion from Theorem 2.2 can be written as

$$\mathbb{P}\bigl(\bigl\|D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\bigr) - D_G b_G - \boldsymbol{\xi}_G\bigr\| \ge \diamondsuit_G(\mathtt{x})\bigr) \le 4\mathrm{e}^{-\mathtt{x}}.$$

The definition (2.10) of $\diamondsuit_G(\mathtt{x})$ and (B.4) of Theorem B.1 imply

$$\mathbb{E}^{1/2}\bigl\|D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\bigr) - D_G b_G - \boldsymbol{\xi}_G\bigr\|^2 \le 4\bigl\{\delta_G(\mathtt{r}_G)\mathtt{r}_G + 2\nu_0\mathfrak{a}_G\mathtt{r}_G(\mathbb{H}_1 + \mathbb{H}_2/\mathtt{g} + 4)\omega\bigr\}.$$

By the result follows by the triangle inequality

$$\mathbb{E}^{1/2}\bigl\|D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\bigr)\bigr\|^2 \le \mathbb{E}^{1/2}\bigl\|D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*\bigr) - D_G b_G - \boldsymbol{\xi}_G\bigr\|^2 + \mathbb{E}^{1/2}\bigl\|D_G b_G + \boldsymbol{\xi}_G\bigr\|^2.$$

This yields the assertion of the theorem. □

2.6. *Proofs of the Fisher and Wilks expansions*

This section presents the proofs of the main results and some additional statements which can be of independent interest. The principle step of the proof is a bound on the local linear approximation of the gradient $\nabla L_G(\boldsymbol{\theta})$. Below we

study separately its stochastic and deterministic components coming from the decomposition $L(\boldsymbol{\theta}) = \mathbb{E}L(\boldsymbol{\theta}) + \zeta(\boldsymbol{\theta})$. With $D_G^2 = -\nabla^2 \mathbb{E}L_G(\boldsymbol{\theta}_G^*)$, this leads to the decomposition

$$\begin{aligned}\chi(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) &\stackrel{\mathrm{def}}{=} D_G^{-1}\{\nabla L_G(\boldsymbol{\theta}) - \nabla L_G(\boldsymbol{\theta}_G^*)\} + D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*) \\ &= D_G^{-1}\{\nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}_G^*)\} \\ &\quad + D_G^{-1}\{\nabla \mathbb{E}L_G(\boldsymbol{\theta}) - \nabla \mathbb{E}L_G(\boldsymbol{\theta}_G^*)\} + D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*).\end{aligned}$$

First we check the deterministic part. For any $\boldsymbol{\theta}$ with $\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \le \mathtt{r}$ and any unit vector $\mathbf{u} \in \mathbb{R}^p$, it holds

$$\begin{aligned}\mathbf{u}^\top \mathbb{E}\chi(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) &= \mathbf{u}^\top D_G^{-1}\{\nabla \mathbb{E}L_G(\boldsymbol{\theta}) - \nabla \mathbb{E}L_G(\boldsymbol{\theta}_G^*) + D_G^2(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\} \\ &= \mathbf{u}^\top \{\mathbb{I}_p - D_G^{-1}\mathbb{F}_G(\boldsymbol{\theta}^\circ) D_G^{-1}\} D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*),\end{aligned}$$

where $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^\circ(\mathbf{u})$ is a point on the line connecting $\boldsymbol{\theta}_G^*$ and $\boldsymbol{\theta}$. This implies by $(\mathcal{L}_0 G)$

$$\|\mathbb{E}\chi(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*)\| \le \|\mathbb{I}_p - D_G^{-1}\mathbb{F}_G(\boldsymbol{\theta}^\circ) D_G^{-1}\|_{\mathrm{op}} \mathtt{r} \le \delta_G(\mathtt{r})\mathtt{r}. \tag{2.15}$$

Now we study the stochastic part. Consider the vector process

$$\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*) \stackrel{\mathrm{def}}{=} D_G^{-1}\{\nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}_G^*)\}. \tag{2.16}$$

Further, define $\boldsymbol{\upsilon} = V_2(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)$ and introduce a vector process $\mathcal{Y}(\boldsymbol{\upsilon})$ with

$$\mathcal{Y}(\boldsymbol{\upsilon}) \stackrel{\mathrm{def}}{=} V_2^{-1}[\nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}_G^*)].$$

It obviously holds $\nabla \mathcal{Y}(\boldsymbol{\upsilon}) = V_2^{-1} \nabla^2 \zeta(\boldsymbol{\theta}) V_2^{-1}$. Moreover, for any $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}^p$ with $\|\boldsymbol{\gamma}_1\| = \|\boldsymbol{\gamma}_2\| = 1$, condition $(E_2 G)$ implies for $|\lambda| \le \mathtt{g}(\mathtt{r})$

$$\log \mathbb{E}\exp\left\{\frac{\lambda}{\omega} \boldsymbol{\gamma}_1^\top \nabla \mathcal{Y}(\boldsymbol{\upsilon}) \boldsymbol{\gamma}_2\right\} = \log \mathbb{E}\exp\left\{\frac{\lambda}{\omega} \boldsymbol{\gamma}_1^\top V_2^{-1} \nabla^2 \zeta(\boldsymbol{\theta}) V_2^{-1} \boldsymbol{\gamma}_2\right\} \le \frac{\nu_0^2 \lambda^2}{2}.$$

Define $\Upsilon_\circ(\mathtt{r}) \stackrel{\mathrm{def}}{=} \{\boldsymbol{\upsilon}: \|\boldsymbol{\upsilon}\| \le \mathtt{r}, \|S\boldsymbol{\upsilon}\| \le \mathtt{r}\}$ for $S^{-2} = \mathfrak{a}_G^{-2} D_G^{-1} V_2^2 D_G^{-1}$. Then

$$\sup_{\boldsymbol{\theta} \in \Theta_{0,G}(\mathtt{r})} \|\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*)\| \le \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \|A\mathcal{Y}(\boldsymbol{\upsilon})\| \tag{2.17}$$

for $A = \mathfrak{a}_G^{-1} D_G^{-1} V_2$. Theorem B.15 yields

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \|A\mathcal{Y}(\boldsymbol{\upsilon})\| \le \sqrt{8}\nu_0 \mathfrak{z}_{\mathbb{H}}(\mathtt{x}) \mathfrak{a}_G \omega \mathtt{r} \tag{2.18}$$

on a set of a dominating probability at least $1 - \mathrm{e}^{-\mathtt{x}}$, where the function $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ is given by (B.3). Putting together the bounds (2.15) and (2.17) imply the following result.

Theorem 2.5. *Suppose that the matrix $\mathbb{F}_G(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} -\nabla^2 \mathbb{E}L_G(\boldsymbol{\theta})$ fulfills the condition $(\mathcal{L}_0 G)$ and let $(E_0 G)$ and $(E_2 G)$ be fulfilled on $\Theta_{0,G}(\mathtt{r})$ for any fixed $\mathtt{r} \le \mathtt{r}^*$. Then*

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta} \in \Theta_{0,G}(\mathtt{r})} \|D_G^{-1}\{\nabla L_G(\boldsymbol{\theta}) - \nabla L_G(\boldsymbol{\theta}_G^*)\} + D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \ge \diamondsuit_G(\mathtt{r}, \mathtt{x})\right\} \le \mathrm{e}^{-\mathtt{x}},$$

where

$$\diamondsuit_G(\mathtt{r}, \mathtt{x}) \stackrel{\mathrm{def}}{=} \{\delta_G(\mathtt{r}) + \sqrt{8}\nu_0 \mathfrak{z}_{\mathbb{H}}(\mathtt{x}) \mathfrak{a}_G \omega\}\mathtt{r}. \tag{2.19}$$

The result of Theorem 2.5 can be extended to the increments of the process $\mathcal{U}(\boldsymbol{\theta})$: on a random set of probability at least $1-\mathrm{e}^{-\mathtt{x}}$, it holds for any $\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\in\Theta_{0,G}(\mathtt{r})$ and $\chi(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ})=D_G^{-1}\{\nabla L_G(\boldsymbol{\theta})-\nabla L_G(\boldsymbol{\theta}^{\circ})\}+D_G(\boldsymbol{\theta}-\boldsymbol{\theta}^{\circ})$

$$\begin{aligned}&\mathbb{E}\bigl[\chi\bigl(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\bigr)\bigr]\le\delta_G(\mathtt{r})\bigl\|D_G\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^{\circ}\bigr)\bigr\|\le 2\mathtt{r}\delta_G(\mathtt{r}),\\&\bigl\|\chi\bigl(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\bigr)\bigr\|\le 2\diamond_G(\mathtt{r},\mathtt{x}).\end{aligned}\tag{2.20}$$

Now we present the proof of Theorem 2.2 about the Fisher expansion for the qMLE $\widetilde{\boldsymbol{\theta}}_G$ defined by maximization of $L_G(\boldsymbol{\theta})$. Let $\mathtt{r}_G$ be selected to ensure that $\mathbb{P}\{\widetilde{\boldsymbol{\theta}}_G\notin\Theta_{0,G}(\mathtt{r}_G)\}\le\mathrm{e}^{-\mathtt{x}}$. Furthermore, the definition of $\widetilde{\boldsymbol{\theta}}_G$ yields $\nabla L_G(\widetilde{\boldsymbol{\theta}}_G)=0$ and

$$\chi\bigl(\widetilde{\boldsymbol{\theta}}_G,\boldsymbol{\theta}_G^*\bigr)=-D_G^{-1}\nabla L_G\bigl(\boldsymbol{\theta}_G^*\bigr)+D_G\bigl(\widetilde{\boldsymbol{\theta}}_G-\boldsymbol{\theta}_G^*\bigr).$$

By Theorem 2.5, it holds on a set of a dominating probability

$$\bigl\|D_G\bigl(\widetilde{\boldsymbol{\theta}}_G-\boldsymbol{\theta}_G^*\bigr)-\boldsymbol{\xi}_G\bigr\|\le\diamond_G(\mathtt{x})\tag{2.21}$$

as required.

As the next step, we apply the obtained results to evaluate the quality of the Wilks expansion $2L_G(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}_G^*)\approx\|\boldsymbol{\xi}_G\|^2$. For this we derive a uniform deviation bound on the error of a quadratic approximation

$$\alpha\bigl(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\bigr)\stackrel{\mathrm{def}}{=}L_G(\boldsymbol{\theta})-L_G\bigl(\boldsymbol{\theta}^{\circ}\bigr)-\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^{\circ}\bigr)^{\top}\nabla L_G\bigl(\boldsymbol{\theta}^{\circ}\bigr)+\frac{1}{2}\bigl\|D_G\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^{\circ}\bigr)\bigr\|^2$$

in all $\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\in\Theta_0$, where Θ_0 is some vicinity of a fixed point $\boldsymbol{\theta}_G^*$. With $\boldsymbol{\theta}^{\circ}$ fixed, the gradient $\nabla\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ})\stackrel{\mathrm{def}}{=}\frac{d}{d\boldsymbol{\theta}}\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ})$ fulfills

$$\nabla\alpha\bigl(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\bigr)=\nabla L_G(\boldsymbol{\theta})-\nabla L_G\bigl(\boldsymbol{\theta}^{\circ}\bigr)+D_G^2\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^{\circ}\bigr)=D_G\chi\bigl(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\bigr);$$

cf. (2.16). This implies

$$\alpha\bigl(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\bigr)=\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^{\circ}\bigr)^{\top}\nabla\alpha\bigl(\boldsymbol{\theta}',\boldsymbol{\theta}^{\circ}\bigr),$$

where $\boldsymbol{\theta}'$ is a point on the line connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{\circ}$. Further,

$$\bigl|\alpha\bigl(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\bigr)\bigr|=\bigl|\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^{\circ}\bigr)^{\top}D_GD_G^{-1}\nabla\alpha\bigl(\boldsymbol{\theta}',\boldsymbol{\theta}^{\circ}\bigr)\bigr|\le\bigl\|D_G\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^{\circ}\bigr)\bigr\|\sup_{\boldsymbol{\theta}'\in\Theta_{0,G}(\mathtt{r})}\bigl|\chi\bigl(\boldsymbol{\theta}',\boldsymbol{\theta}^{\circ}\bigr)\bigr|,$$

and one can apply (2.20). This yields the following result.

Theorem 2.6. *Suppose* $(\mathcal{L}_0G)$, (E_0G), *and* (E_2G). *For each* $\mathtt{r}$, *it holds on a random set* $\Omega(\mathtt{x})$ *of a dominating probability at least* $1-\mathrm{e}^{-\mathtt{x}}$, *it holds with any* $\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\in\Theta_{0,G}(\mathtt{r})$ *and* $\diamond_G(\mathtt{r},\mathtt{x})$ *is from* (2.19)

$$\begin{aligned}&\frac{|\alpha(\boldsymbol{\theta},\boldsymbol{\theta}_G^*)|}{\|D_G(\boldsymbol{\theta}-\boldsymbol{\theta}_G^*)\|}\le\diamond_G(\mathtt{r},\mathtt{x}),\quad\bigl|\alpha\bigl(\boldsymbol{\theta},\boldsymbol{\theta}_G^*\bigr)\bigr|\le\mathtt{r}\diamond_G(\mathtt{r},\mathtt{x}),\\&\frac{|\alpha(\boldsymbol{\theta}_G^*,\boldsymbol{\theta})|}{\|D_G(\boldsymbol{\theta}-\boldsymbol{\theta}_G^*)\|}\le 2\diamond_G(\mathtt{r},\mathtt{x}),\quad\bigl|\alpha\bigl(\boldsymbol{\theta}_G^*,\boldsymbol{\theta}\bigr)\bigr|\le 2\mathtt{r}\diamond_G(\mathtt{r},\mathtt{x}),\\&\frac{|\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ})|}{\|D_G(\boldsymbol{\theta}-\boldsymbol{\theta}^{\circ})\|}\le 2\diamond_G(\mathtt{r},\mathtt{x}),\quad\bigl|\alpha\bigl(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\bigr)\bigr|\le 4\mathtt{r}\diamond_G(\mathtt{r},\mathtt{x}).\end{aligned}$$

The result of Theorem 2.6 for the special case with $\boldsymbol{\theta}=\boldsymbol{\theta}_G^*$ and $\boldsymbol{\theta}^{\circ}=\widetilde{\boldsymbol{\theta}}_G$ yields in view of $\nabla L_G(\widetilde{\boldsymbol{\theta}}_G)=0$ for $\mathtt{r}=\mathtt{r}_G$ and $\diamond_G(\mathtt{x})=\diamond_G(\mathtt{r}_G,\mathtt{x})$ under the condition $\widetilde{\boldsymbol{\theta}}_G\in\Theta_{0,G}(\mathtt{r}_G)$

$$\bigl|L_G\bigl(\widetilde{\boldsymbol{\theta}}_G,\boldsymbol{\theta}_G^*\bigr)-\bigl\|D_G\bigl(\widetilde{\boldsymbol{\theta}}_G-\boldsymbol{\theta}_G^*\bigr)\bigr\|^2/2\bigr|=\bigl|\alpha\bigl(\boldsymbol{\theta}_G^*,\widetilde{\boldsymbol{\theta}}_G\bigr)\bigr|\le 2\mathtt{r}_G\diamond_G(\mathtt{x}).$$

Furthermore, with $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}_G$ and $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^*_G$

$$\bigl| L_G\bigl(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G\bigr) - \boldsymbol{\xi}_G^\top D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G\bigr) + \bigl\| D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G\bigr)\bigr\|^2/2 \bigr| = \bigl|\alpha\bigl(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G\bigr)\bigr| \le \mathtt{r}_G \diamondsuit_G(\mathtt{x})$$

which implies

$$\bigl| L\bigl(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G\bigr) - \|\boldsymbol{\xi}_G\|^2 + \bigl\| D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G\bigr) - \boldsymbol{\xi}_G\bigr\|^2 \bigr| \le 2\mathtt{r}_G \diamondsuit_G(\mathtt{x}).$$

Now it follows by (2.21) that

$$\bigl| L\bigl(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G\bigr) - \bigl\|\boldsymbol{\xi}_G\bigr\|^2/2 \bigr| \le \mathtt{r}_G \diamondsuit_G(\mathtt{x}) + \diamondsuit^2_G(\mathtt{x})/2.$$

The error term can be improved if the squared root of the excess is considered. Indeed, if $\widetilde{\boldsymbol{\theta}}_G \in \Theta_{0,G}(\mathtt{r}_G)$

$$\begin{aligned}\bigl| \bigl\{2L_G\bigl(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G\bigr)\bigr\}^{1/2} - \bigl\| D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G\bigr)\bigr\| \bigr| &\le \frac{|2L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G) - \|D_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G)\|^2|}{\|D_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G)\|} \\ &\le \frac{2|\alpha(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G)|}{\|D_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G)\|} \le \sup_{\boldsymbol{\theta} \in \Theta_{0,G}(\mathtt{r}_G)} \frac{2|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*_G)|}{\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}^*_G)\|} \le 2\diamondsuit_G(\mathtt{x}).\end{aligned}$$

The Fisher expansion (2.21) allows to replace here the norm of the standardized error $D_G(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G)$ with the norm of the normalized score $\boldsymbol{\xi}_G$. This completes the proof of Theorem 2.3.

3. Examples

This section illustrates the general results for two particularly important cases of i.i.d. and generalized linear models. The primary focus of the study is to compare the penalized and non-penalized cases and to quantify the impact of penalization.

3.1. *I.i.d. case*

The model with independent identically distributed (i.i.d.) observations is one of the most popular setups in statistical literature and in statistical applications. The essential and the most developed part of the statistical theory is designed for the i.i.d. modeling. Especially, the classical asymptotic parametric theory is almost complete including asymptotic root-n normality and efficiency of the MLE and Bayes estimators under rather mild assumptions; see e.g. Chapter 2 and 3 in Ibragimov and Khas'minskij [14]. So, the i.i.d. model can naturally serve as a benchmark for any extension of the statistical theory: being applied to the i.i.d. setup, the new approach should lead to essentially the same conclusions as in the classical theory. Similar reasons apply to the regression model and its extensions. Below we try demonstrate that the proposed non-asymptotic viewpoint is able to reproduce the existing brilliant and well established results of the classical parametric theory. Surprisingly, the majority of classical efficiency results can be easily derived from the obtained general non-asymptotic bounds.

3.2. *Quasi MLE in an i.i.d. model*

The basic i.i.d. parametric model means that the observations $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are independent identically distributed from a distribution P from a given parametric family $(P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta)$ on the observation space $\mathcal{Y}_1$. Each $\boldsymbol{\theta} \in \Theta$ clearly yields the product data distribution $\mathbb{P}_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta}}^{\otimes n}$ on the product space $\mathcal{Y} = \mathcal{Y}_1^n$. This section illustrates how the obtained general results can be applied to this type of modeling under possible model misspecification. Different types of misspecification can be considered. Each of the assumptions, namely, data independence, identical distribution, parametric form of the marginal distribution can be violated. To be specific, we assume the observations Y_i independent and identically distributed. However, we admit that the distribution of each Y_i does not necessarily belong to the

parametric family $(P_{\boldsymbol{\theta}})$. The case of non-identically distributed observations can be done similarly at cost of more complicated notation.

In what follows the parametric family $(P_{\boldsymbol{\theta}})$ is supposed to be dominated by a measure μ_0, and each density $p(y,\boldsymbol{\theta}) = dP_{\boldsymbol{\theta}}/d\mu_0(y)$ is two times continuously differentiable in $\boldsymbol{\theta}$ for all y. Denote $\ell(y,\boldsymbol{\theta}) = \log p(y,\boldsymbol{\theta})$. The parametric assumption $Y_i \sim P_{\boldsymbol{\theta}^*} \in (P_{\boldsymbol{\theta}})$ leads to the log-likelihood

$$L(\boldsymbol{\theta}) = \sum \ell(Y_i,\boldsymbol{\theta}), \tag{3.1}$$

where the summation is taken over $i = 1,\ldots,n$. The quasi MLE $\widetilde{\boldsymbol{\theta}}$ maximizes this sum over $\boldsymbol{\theta} \in \Theta$:

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \sum \ell(Y_i,\boldsymbol{\theta}).$$

The target of estimation $\boldsymbol{\theta}^*$ maximizes the expectation of $L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \sum \mathbb{E}\ell(Y_i,\boldsymbol{\theta}).$$

Let $\zeta_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell(Y_i,\boldsymbol{\theta}) - \mathbb{E}\ell(Y_i,\boldsymbol{\theta})$. Then $\zeta(\boldsymbol{\theta}) = \sum \zeta_i(\boldsymbol{\theta})$. The equation $\mathbb{E}\nabla L(\boldsymbol{\theta}^*) = 0$ implies

$$\nabla\zeta\bigl(\boldsymbol{\theta}^*\bigr) = \sum \nabla\zeta_i\bigl(\boldsymbol{\theta}^*\bigr) = \sum \nabla\ell_i\bigl(\boldsymbol{\theta}^*\bigr). \tag{3.2}$$

3.3. *Conditions in the i.i.d. case*

I.i.d. structure of the Y_i's allows for rewriting the conditions (E_0), (E_2), $(\mathcal{I})$, $(\mathcal{L}_0)$, and $(\mathcal{L})$ in terms of the marginal distribution. In the following conditions the index i runs from 1 to n.

(ed_0) There exists a positive symmetric matrix v_0, such that for all $|\lambda| \le \mathtt{g}_1$

$$\sup_{\boldsymbol{\gamma}\in\mathcal{S}^p} \log \mathbb{E}\exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta_i(\boldsymbol{\theta}^*)}{\|\mathsf{v}_0\boldsymbol{\gamma}\|}\right\} \le \nu_0^2\lambda^2/2.$$

A natural candidate on v_0^2 is given by the variance of the gradient $\nabla\ell(Y_1,\boldsymbol{\theta}^*)$, that is, $\mathsf{v}_0^2 = \operatorname{Var}\nabla\ell(Y_1,\boldsymbol{\theta}) = \operatorname{Var}\nabla\zeta_1(\boldsymbol{\theta})$. Note that (ed_0) is automatically fulfilled if the model is correctly specified and $P = P_{\boldsymbol{\theta}^*}$ because $E_{\boldsymbol{\theta}^*}\exp\{\ell(Y_1,\boldsymbol{\theta}) - \ell(Y_1,\boldsymbol{\theta}^*)\} \equiv 1$.

Next consider the local sets

$$\Theta_0(\mathtt{r}) = \bigl\{\boldsymbol{\theta} : \bigl\|\mathsf{v}_0\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)\bigr\| \le \mathtt{r}/n^{1/2}\bigr\}.$$

The local smoothness conditions (E_2) and $(\mathcal{L}_0)$ require to specify the functions $\delta(\mathtt{r})$ and $\varrho(\mathtt{r})$. If the log-likelihood function $\ell(y,\boldsymbol{\theta})$ is sufficiently smooth in $\boldsymbol{\theta}$, these functions can be selected proportional to $\mathtt{r}$.

(ed_2) There exist a value $\omega^* > 0$ and for each $\mathtt{r} > 0$, a constant $\mathtt{g}(\mathtt{r}) > 0$ such that

$$\sup_{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2\in\mathbb{R}^p} \log \mathbb{E}\exp\left\{\frac{\lambda}{\omega^*}\frac{\boldsymbol{\gamma}_1^\top\nabla^2\zeta_i(\boldsymbol{\theta})\boldsymbol{\gamma}_2}{\|\mathsf{v}_0\boldsymbol{\gamma}_1\|\cdot\|\mathsf{v}_0\boldsymbol{\gamma}_2\|}\right\} \le \frac{\nu_0^2\lambda^2}{2}, \quad |\lambda| \le \mathtt{g}(\mathtt{r}).$$

Further we restate the local regularity condition $(\mathcal{L}_0)$ in terms of the expected value $\bar{\ell}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{E}\ell(Y_i,\boldsymbol{\theta})$ of each $\ell(Y_i,\boldsymbol{\theta})$. We suppose that $\bar{\ell}(\boldsymbol{\theta})$ is two times differentiable and define the matrix function $\mathsf{H}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2\bar{\ell}(\boldsymbol{\theta})$.

(ℓ_0) The function $\bar{\ell}(\boldsymbol{\theta})$ is two times differentiable and the matrix function $\mathsf{H}(\boldsymbol{\theta}) = -\nabla^2\mathbb{E}\ell(Y_1,\boldsymbol{\theta})$ fulfills with $\mathsf{H}_0 \stackrel{\text{def}}{=} \mathsf{H}(\boldsymbol{\theta}^*)$ for some constant δ^*:

$$\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \bigl\|\mathsf{H}_0^{-1/2}\mathsf{H}(\boldsymbol{\theta})\mathsf{H}_0^{-1/2} - \mathbb{I}_p\bigr\|_{\mathrm{op}} \le \frac{\delta^*\mathtt{r}}{\sqrt{n}}.$$

In the regular parametric case with $P \in (P_{\boldsymbol{\theta}})$, the matrices $\mathtt{v}_0^2$ and $\mathtt{H}_0$ coincide with the Fisher information matrix $\mathtt{H}_0 = \mathtt{H}(\boldsymbol{\theta}^*)$ of the family $(P_{\boldsymbol{\theta}})$ at the point $\boldsymbol{\theta}^*$.

The consistency result for $\widetilde{\boldsymbol{\theta}}$ requires certain growth of the value $\bar{\ell}(\boldsymbol{\theta}^*) - \bar{\ell}(\boldsymbol{\theta})$ as $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|$ grows. The marginal version of the global condition $(\mathcal{L})$ reads as follows:

$(\bar{\ell})$ There exists $\mathtt{C}_1 \geq 0$ such that with $\delta = \delta^*/\sqrt{n}$ and $\mathtt{r} = \sqrt{n}\|\mathtt{H}_0^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$

$$n\bigl\{\bar{\ell}\bigl(\boldsymbol{\theta}^*\bigr) - \bar{\ell}(\boldsymbol{\theta})\bigr\} \geq (1-\delta)\bigl(\mathtt{r}\mathtt{r}_0 - \mathtt{r}_0^2/2\bigr) - \mathtt{C}_1\mathtt{r}^2.$$

Remark 3.1. *If the parametric i.i.d. model is correct, then*

$$\bar{\ell}\bigl(\boldsymbol{\theta}^*\bigr) - \bar{\ell}(\boldsymbol{\theta}) = \mathcal{K}\bigl(\boldsymbol{\theta}^*, \boldsymbol{\theta}\bigr) = E_{\boldsymbol{\theta}^*} \log \frac{d P_{\boldsymbol{\theta}^*}}{d P_{\boldsymbol{\theta}}}(Y_1)$$

is the Kullback–Leibler divergence for the family $(P_{\boldsymbol{\theta}})$. *Condition* $(\bar{\ell})$ *is fulfilled automatically if* $\bar{\ell}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) > 0$ *for* $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$ *and* Θ *is a compact set. Then*

$$\inf_{\boldsymbol{\theta} \in \Theta} \frac{\bar{\ell}(\boldsymbol{\theta}^*) - \bar{\ell}(\boldsymbol{\theta})}{\|\mathtt{H}_0^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} \geq \mathtt{b} > 0.$$

This and (ℓ_0) *imply* $(\bar{\ell})$.

The *identifiability condition* relates the matrices $\mathtt{v}_0^2$ and $\mathtt{H}_0$.

(ι) There is a constant $\mathfrak{a} > 0$ such that $\mathfrak{a}^2\mathtt{H}_0 \geq \mathtt{v}_0^2$.

Lemma 3.1. *Let* $Y_1, \ldots, Y_n$ *be i.i.d. Then* (ed_0), (ed_2), (ℓ_0), $(\bar{\ell})$, *and* (ι) *imply* (E_0), (E_2), $(\mathcal{L}_0)$, $(\mathcal{L})$, $(\mathcal{I})$, *with* $V^2 = n\mathtt{v}_0^2$, $D^2 = n\mathtt{H}_0$, $\omega = \omega^*/\sqrt{n}$, $\delta(\mathtt{r}) = \delta^*\mathtt{r}/\sqrt{n}$, *and the same constants* ν_0, $\mathfrak{a}$, $\mathtt{g} \stackrel{\text{def}}{=} \mathtt{g}_1\sqrt{n}$.

Proof. The identities $V^2 = n\mathtt{v}_0^2$, $D^2 = n\mathtt{H}_0$ follow from the i.i.d. structure of the observations Y_i. We briefly comment on condition (E_0). The use once again of the i.i.d. structure yields by (3.2) in view of $V^2 = n\mathtt{v}_0^2$

$$\log \mathbb{E}\exp\left\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla\zeta(\boldsymbol{\theta}^*)}{\|V\boldsymbol{\gamma}\|}\right\} = n\mathbb{E}\exp\left\{\frac{\lambda}{n^{1/2}} \frac{\boldsymbol{\gamma}^\top \nabla\zeta_1(\boldsymbol{\theta}^*)}{\|\mathtt{v}\boldsymbol{\gamma}\|}\right\} \leq \nu_0^2\lambda^2/2$$

as long as $\lambda \leq n^{1/2}\mathtt{g}_1 \leq \mathtt{g}$. Similarly one can check (E_2). The conditions $(\mathcal{L}_0)$, $(\mathcal{L})$, and $(\mathcal{I})$ follow from (ℓ_0) and $(\bar{\ell})$, and (ι) due to $D^2 = n\mathtt{H}_0$ and $\mathbb{E}L(\boldsymbol{\theta}) = n\bar{\ell}(\boldsymbol{\theta})$. □

Below we specify the obtained general results to the i.i.d. setup.

3.4. *Results in the non-penalized i.i.d. case*

Here we specify the general results of previous chapters to the i.i.d. case. In particular, we explicitly state the large deviation bound and show that it yields a root-n consistency of the qMLE $\widetilde{\boldsymbol{\theta}}$. Then we comment on the Fisher and Wilks theorems.

First we describe the large deviation probability for the event $\{\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)\}$ for a fixed $\mathtt{r}_0$. The next result specifies the general large deviation statement of Theorem 2.1 to the finite dimensional non-penalized i.i.d. case and states the inference results.

Theorem 3.2. *Let* (ed_0), (ed_2), (ℓ_0), (ι), *and* $(\bar{\ell})$ *hold with*

$$\bigl\{1 - \delta(\mathtt{r}_0)\bigr\}\mathtt{r}_0 \geq 2z(B, \mathtt{x}), \qquad \mathtt{C}_1 \geq \varrho(\mathtt{r}, \mathtt{x}), \quad \mathtt{r} > \mathtt{r}_0, \tag{3.3}$$

where $B = \mathsf{H}_0^{-1/2}\mathsf{v}_0^2\mathsf{H}_0^{-1/2} = D^{-1}V^2D^{-1}$, $z(B,\mathtt{x})$ *is given by* (A.4), *and*

$$\varrho(\mathtt{r},\mathtt{x}) \stackrel{\mathrm{def}}{=} \nu_0 \mathfrak{z}_{\mathbb{H}}\bigl(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)\bigr)\omega^*/\sqrt{n}$$

with $\mathfrak{z}_{\mathbb{H}}(\mathtt{x}) \le \mathtt{C}\sqrt{p+\mathtt{x}}$. *Then it holds on a set* $\Omega(\mathtt{x})$ *with* $\mathbb{P}(\Omega(\mathtt{x})) \ge 1 - 5\mathrm{e}^{-\mathtt{x}}$

$$\sqrt{n}\bigl\|\mathsf{H}_0^{1/2}\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr)\bigr\| \le \mathtt{r}_0. \tag{3.4}$$

Furthermore, *on this set* $\Omega(\mathtt{x})$, *it holds*

$$\bigl\|\sqrt{n\mathsf{H}_0}\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \boldsymbol{\xi}\bigr\| \le \mathtt{C}\sqrt{(p+\mathtt{x})^2/n},$$
$$\bigl|\sqrt{2L\bigl(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*\bigr)} - \|\boldsymbol{\xi}\|\bigr| \le \mathtt{C}\sqrt{(p+\mathtt{x})^2/n},$$
$$\bigl|2L\bigl(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*\bigr) - \|\boldsymbol{\xi}\|^2\bigr| \le \mathtt{C}\sqrt{(p+\mathtt{x})^3/n}.$$

The constant $\mathtt{C}$ *here depends in an explicit way on the constants* $\mathfrak{a}_G$, $\mathfrak{g}_1$, *and* ν_0 *from our conditions*, *and*

$$\boldsymbol{\xi} \stackrel{\mathrm{def}}{=} (n\mathsf{H}_0)^{-1/2}\sum_{i=1}^{n}\nabla\ell\bigl(Y_i,\boldsymbol{\theta}^*\bigr). \tag{3.5}$$

Proof. Condition (ι) implies $B = \mathsf{H}_0^{-1/2}\mathsf{v}_0^2\mathsf{H}_0^{-1/2} \le \mathfrak{a}^2\mathbb{I}_p$ and thus, $\mathrm{tr}(B) \le \mathfrak{a}^2 p$. Therefore, the value $z(B,\mathtt{x})$ fulfills $z^2(B,\mathtt{x}) \le \mathtt{C}(p+\mathtt{x})$. The same bound holds for $\mathfrak{z}^2_{\mathbb{H}}(\mathtt{x})$. Condition (3.3) with $\mathtt{b}(\mathtt{r}_0) \approx 1$ yields $\mathtt{r}_0^2 \approx 4z^2(B,\mathtt{x}) \approx \mathtt{C}(p+\mathtt{x})$. This yields in view of $\delta(\mathtt{r}_0) \le \delta^*\mathtt{r}_0/\sqrt{n}$ and $\omega = \omega^* n^{-1/2}$

$$\diamondsuit(\mathtt{r}_0,\mathtt{x}) \le \bigl\{\delta(\mathtt{r}_0) + \nu_0\mathfrak{z}_{\mathbb{H}}(\mathtt{x})\omega\bigr\}\mathtt{r}_0 \le \mathtt{C}(p+\mathtt{x})/\sqrt{n}.$$

Similarly

$$\Delta(\mathtt{r}_0,\mathtt{x}) \le \bigl\{\delta(\mathtt{r}_0) + \nu_0\mathfrak{z}_{\mathbb{H}}(\mathtt{x})\omega\bigr\}\mathtt{r}_0^2 \le \mathtt{C}\sqrt{(p+\mathtt{x})^3/n}.$$

The results follow now from general theorems of Section 2. □

For the classical asymptotic setup when n tends to infinity, the random vector $\boldsymbol{\xi}$ from (3.5) fulfills $\mathrm{Var}(\boldsymbol{\xi}) \le \mathsf{H}_0^{-1/2}\mathsf{v}_0^2\mathsf{H}_0^{-1/2} = B$ and by the central limit theorem $\boldsymbol{\xi}$ is asymptotically normal $\mathcal{N}(0,B)$. This yields by Theorem 3.2 that $\sqrt{n\mathsf{H}_0}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is asymptotically normal $\mathcal{N}(0,B)$ as well. The correct model specification implies $B \equiv \mathbb{I}_p$ and hence $\widetilde{\boldsymbol{\theta}}$ is asymptotically efficient; see Ibragimov and Khas'minskij [14]. Also $2L(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*) \approx \|\boldsymbol{\xi}\|^2$ which is nearly χ^2 r.v. with p degrees of freedom. This result is known as asymptotic Wilks theorem.

In the non-asymptotic framework of this paper, the error terms still depend on n and they can only be small if n is large. However, we show in explicit way how these error terms depend on the parameter dimension. It appears that the root-n consistency result (3.4) requires "p/n small." The Fisher and square root Wilks results apply if "p^2/n is small." Finally, the Wilks expansion is valid under "p^3/n small." Existing statistical literature addresses the issue of a growing parameter dimension in different set-ups. The classical results by Portnoy [19–21] provide some constraints on parameter dimension for consistency and asymptotic normality of the M-estimator for regression models. Our results are consistent with the conclusion of that papers. We refer to Andresen and Spokoiny [1] for a version of such result in context of semiparametric profile estimation. That paper also provides an example of an i.i.d. model in which the Fisher expansion of Theorem 3.2 fails for $p^2 \ge n$. The next section demonstrates how these constraints on the parameter dimension can be relaxed by using a penalization.

3.5. *Roughness penalization for an i.i.d. sample*

This section discusses the impact of penalization in the case of an i.i.d. model with n observations. For penalty term $\operatorname{pen}(\boldsymbol{\theta}) = \|G\boldsymbol{\theta}\|^2/2$, the penalized log-likelihood is given by $L_G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \|G\boldsymbol{\theta}\|^2/2$, where $L(\boldsymbol{\theta})$ is from (3.1). With $\boldsymbol{\theta}^*_G = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}L_G(\boldsymbol{\theta})$, define

$$D_G^2 = n\mathsf{H}\bigl(\boldsymbol{\theta}^*_G\bigr) + G^2, \qquad V^2 = n\mathsf{v}_0^2, \qquad \boldsymbol{\xi}_G = D_G^{-1}\sum_{i=1}^n \nabla\ell\bigl(Y_i, \boldsymbol{\theta}^*_G\bigr),$$

where $\mathsf{H}(\boldsymbol{\theta}) = -\nabla^2\mathbb{E}\ell(Y_1, \boldsymbol{\theta})$, $\mathsf{v}_0^2 = \operatorname{Var}\{\ell(Y_1, \boldsymbol{\theta}^*_G)\}$. The value $\mathtt{p}_G$ is defined as previously by (2.2).

Note that all the introduced quantities including the parameter set Θ, the parameter dimension p, and the effective dimension $\mathtt{p}_G$, may depend on n. Here we also allow a functional parameter $\boldsymbol{\theta}$ with $p = \infty$. The main goal is to show that the presented general approach yields sharp results in this special case.

Suppose that the conditions of Section 3.3 are fulfilled. One can easily check the conditions from Section 2.2 with $\delta_G(\mathtt{r}) = \mathtt{C}\mathtt{r}/\sqrt{n}$ and $\omega = \mathtt{C}/\sqrt{n}$; cf. Lemma 3.1. The large deviation bound of Theorem 2.1 applies for $\mathtt{r}_G \approx 2z(B_G, \mathtt{x}) \asymp \sqrt{\mathtt{p}_G + \mathtt{x}}$. The general statements of Theorems 2.2 and 2.3 apply with $\diamondsuit_G(\mathtt{x}) \le \mathtt{C}(\mathtt{p}_G + \mathtt{x})/\sqrt{n}$ yielding the following expansions.

Theorem 3.3. *Suppose also that the conditions* (ed_0), (ed_2), (ℓ_0), ($\bar{\ell}$), *and* (ι) *are fulfilled, and* $\mathtt{r}_G$ *and* $\mathtt{C}_1$ *fulfill*

$$\bigl\{1 - \delta_G(\mathtt{r}_G)\bigr\}\mathtt{r}_G \ge 2z(B_G, \mathtt{x}), \qquad \mathtt{C}_1 \ge \varrho_G(\mathtt{r}, \mathtt{x}), \quad \mathtt{r} > \mathtt{r}_G,$$

with $B_G = D_G^{-1}V^2D_G^{-1/2}$, *then on a set of dominating probability* $1 - 5\mathrm{e}^{-\mathtt{x}}$, *it holds*

$$\begin{aligned}
&\bigl\|D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G\bigr) - \boldsymbol{\xi}_G\bigr\| \le \mathtt{C}\sqrt{(\mathtt{p}_G + \mathtt{x})^2/n},\\
&\Bigl|\sqrt{2L_G\bigl(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G\bigr)} - \|\boldsymbol{\xi}_G\|\Bigr| \le \mathtt{C}\sqrt{(\mathtt{p}_G + \mathtt{x})^2/n},\\
&\bigl|2L_G\bigl(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G\bigr) - \|\boldsymbol{\xi}_G\|^2\bigr| \le \mathtt{C}\sqrt{(\mathtt{p}_G + \mathtt{x})^3/n}.
\end{aligned}$$

The constant $\mathtt{C}$ *here depends in an explicit way on the constants* $\mathfrak{a}_G$, $\mathtt{g}_1$, *and* ν_0 *from our conditions.*

A short look at the results for non-penalized and penalized estimates indicates that the quality of the penalized MLE $\widetilde{\boldsymbol{\theta}}_G$ improves relative to the non-penalized case because the matrix D_G^2 can be much larger than D^2, the variance of the stochastic term $\boldsymbol{\xi}_G$ is of order $\mathtt{p}_G$ instead of p for the variance of $\boldsymbol{\xi}$, and, simultaneously, the error terms in the Fisher and Wilks expansions become smaller due to reduction of the effective dimension $\mathtt{p}_G$ in place of the full dimension p.

Andresen and Spokoiny [1] provides a simple example of estimating the squared norm $\|\boldsymbol{\theta}\|^2$ which shows that the Fisher expansion fails if $p/\sqrt{n}$ is not small. The result can be easily adjusted to the penalized case.

3.6. *Generalized linear models* (*GLM*)

Generalized linear models (GLM) are frequently used for modeling the data with special structure: categorical data, binary data, Poissonian and exponential data, volatility models, etc. All these examples can be treated in a unified way by a GLM approach. This section specifies the results and conditions to this case.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top \sim \mathbb{P}$ be a sample of independent r.v.'s. The parametric GLM is given by $Y_i \sim P_{\Psi_i^\top\boldsymbol{\theta}} \in (P_\upsilon)$, where Ψ_i are given factors in $\mathbb{R}^p$, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the unknown parameter in $\mathbb{R}^p$, and (P_υ) is an exponential family with canonical parametrization yielding the log-density $\ell(y, \upsilon) = y\upsilon - g(\upsilon)$ for a convex function $g(\upsilon)$. Below we suppose that the function $g(\upsilon)$ is sufficiently smooth, in particular, three times differentiable.

The (quasi) log-likelihood $L(\boldsymbol{\theta})$ can be represented in the form

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \bigl\{ Y_i \Psi_i^\top \boldsymbol{\theta} - g\bigl(\Psi_i^\top \boldsymbol{\theta}\bigr) \bigr\} = S^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) \tag{3.6}$$

with a random p-vector S and a function $A(\boldsymbol{\theta})$ given by

$$S \stackrel{\mathrm{def}}{=} \sum_{i=1}^{n} Y_i \Psi_i, \qquad A(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \sum_{i} g\bigl(\Psi_i^\top \boldsymbol{\theta}\bigr).$$

The MLE $\widetilde{\boldsymbol{\theta}}$ and the target $\boldsymbol{\theta}^*$ for this GLM read as

$$\begin{aligned} \widetilde{\boldsymbol{\theta}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \bigl\{ S^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) \bigr\}, \\ \boldsymbol{\theta}^* &= \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \bigl\{ \mathbb{E} S^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) \bigr\}, \end{aligned} \tag{3.7}$$

where

$$\mathbb{E} S = \sum_{i=1}^{n} \mathbb{E} Y_i \Psi_i.$$

The definition of $\boldsymbol{\theta}^*$ implies the identity $\nabla \mathbb{E} L(\boldsymbol{\theta}^*) = 0$ which yields

$$\mathbb{E} S = \nabla A\bigl(\boldsymbol{\theta}^*\bigr).$$

An important feature of a GLM is that the stochastic component $\zeta(\boldsymbol{\theta})$ of $L(\boldsymbol{\theta})$ is *linear in* $\boldsymbol{\theta}$: with $\varepsilon_i = Y_i - \mathbb{E} Y_i$

$$\begin{aligned} \zeta(\boldsymbol{\theta}) &= L(\boldsymbol{\theta}) - \mathbb{E} L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \varepsilon_i \Psi_i^\top \boldsymbol{\theta}, \\ \nabla \zeta(\boldsymbol{\theta}) &= S - \mathbb{E} S = \sum_{i=1}^{n} \varepsilon_i \Psi_i. \end{aligned} \tag{3.8}$$

In the contrary to the linear case, the Fisher information matrix $D^2 = \mathbb{F}(\boldsymbol{\theta}^*)$ for

$$\mathbb{F}(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Psi_i \Psi_i^\top g''\bigl(\Psi_i^\top \boldsymbol{\theta}\bigr) \tag{3.9}$$

depends on the true data distribution via the target $\boldsymbol{\theta}^*$. As $g(\cdot)$ is convex, it holds $g''(u) \geq 0$ for any u and thus $\mathbb{F}(\boldsymbol{\theta}) \geq 0$.

Linearity in $\boldsymbol{\theta}$ of the stochastic component $\zeta(\boldsymbol{\theta})$ and concavity of the deterministic part $\mathbb{E} L(\boldsymbol{\theta})$ allow for a simple and straightforward proof of the result about concentration of the MLE $\widetilde{\boldsymbol{\theta}}$. Recall the definition of the local vicinity $\Theta_0(\mathtt{r})$ of $\boldsymbol{\theta}^*$:

$$\Theta_0(\mathtt{r}) \stackrel{\mathrm{def}}{=} \bigl\{ \boldsymbol{\theta} : \bigl\| D\bigl(\boldsymbol{\theta} - \boldsymbol{\theta}^*\bigr) \bigr\| \leq \mathtt{r} \bigr\}.$$

Also define

$$B \stackrel{\mathrm{def}}{=} D^{-1} \operatorname{Var}(S) D^{-1}.$$

Theorem 3.4. *If for some* $\mathtt{r}_0 > 0$, $\mathbb{F}(\boldsymbol{\theta})$ *from* (3.9) *fulfill for* $D^2 = \mathbb{F}(\boldsymbol{\theta}^*)$

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)} \bigl\| D^{-1} \mathbb{F}(\boldsymbol{\theta}) D^{-1} - \mathbb{I}_p \bigr\|_{\mathrm{op}} \le \delta(\mathtt{r}_0) \tag{3.10}$$

with $\delta(\mathtt{r}_0) < 1$, *and if* S *from* (3.8) *follows for* $\mathtt{x} > 0$ *the probability bound*

$$\mathbb{P}\bigl(\bigl\| D^{-1} (S - \mathbb{E}S) \bigr\| > z(B, \mathtt{x}) \bigr) \le 2\mathrm{e}^{-\mathtt{x}}, \tag{3.11}$$

then the solution $\widetilde{\boldsymbol{\theta}}$ *of* (3.7) *satisfies*

$$\mathbb{P}\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0) \bigr) \le 2\mathrm{e}^{-\mathtt{x}}$$

provided that

$$\mathtt{r}_0 \bigl\{ 1 - \delta(\mathtt{r}_0) \bigr\} \ge 2 z(B, \mathtt{x}).$$

Proof. The function $L(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$ because

$$-\nabla^2 L(\boldsymbol{\theta}) = \mathbb{F}(\boldsymbol{\theta}) \ge 0. \tag{3.12}$$

If $\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)$, denote by $\check{\boldsymbol{\theta}}$ the point at which the line connecting $\boldsymbol{\theta}^*$ and $\widetilde{\boldsymbol{\theta}}$ crosses the boundary of $\Theta_0(\mathtt{r}_0)$. It is easy to see that

$$\check{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = \frac{\| D(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \|}{\| D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \|} \bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \bigr) = \frac{\mathtt{r}_0}{\| D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \|} \bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \bigr).$$

Concavity of $L(\boldsymbol{\theta})$ implies for the point of maximum $\widetilde{\boldsymbol{\theta}}$ that

$$L(\widetilde{\boldsymbol{\theta}}) - L\bigl(\boldsymbol{\theta}^*\bigr) \ge L(\check{\boldsymbol{\theta}}) - L\bigl(\boldsymbol{\theta}^*\bigr).$$

Therefore, it suffices to check that for each $\boldsymbol{\theta}$ with $\| D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| = \mathtt{r}_0$ that

$$L\bigl(\boldsymbol{\theta}^*\bigr) - L(\boldsymbol{\theta}) > 0$$

on a set $\Omega(\mathtt{x})$ of probability $1 - 2\mathrm{e}^{-\mathtt{x}}$. Then the event $\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)$ is impossible on $\Omega(\mathtt{x})$. For any such $\boldsymbol{\theta}$, we apply the second order Taylor expansion of $L(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$. By definition of $\boldsymbol{\theta}^*$, it holds $\nabla \mathbb{E} L(\boldsymbol{\theta}^*) = 0$ and thus $\nabla L(\boldsymbol{\theta}^*) = \nabla \zeta(\boldsymbol{\theta}^*) = (S - \mathbb{E}S)$. The use of (3.12), (3.10) yields now for $\boldsymbol{\xi} = D^{-1}(S - \mathbb{E}S)$ and for $\boldsymbol{\theta}$ with $\| D(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| = \mathtt{r}_0$

$$\begin{aligned} L\bigl(\boldsymbol{\theta}^*\bigr) - L(\boldsymbol{\theta}) &= \bigl(\boldsymbol{\theta} - \boldsymbol{\theta}^*\bigr)^\top \nabla L\bigl(\boldsymbol{\theta}^*\bigr) + \frac{1}{2} \bigl\| \sqrt{\mathbb{F}\bigl(\boldsymbol{\theta}^\circ\bigr)} \bigl(\boldsymbol{\theta} - \boldsymbol{\theta}^*\bigr) \bigr\|^2 \\ &\ge (S - \mathbb{E}S)^\top \bigl(\boldsymbol{\theta} - \boldsymbol{\theta}^*\bigr) + \frac{1 - \delta(\mathtt{r}_0)}{2} \bigl\| D\bigl(\boldsymbol{\theta} - \boldsymbol{\theta}^*\bigr) \bigr\|^2 \\ &= \boldsymbol{\xi}^\top D\bigl(\boldsymbol{\theta} - \boldsymbol{\theta}^*\bigr) + \frac{1 - \delta(\mathtt{r}_0)}{2} \mathtt{r}_0^2 \ge -\| \boldsymbol{\xi} \| \mathtt{r}_0 + \frac{1 - \delta(\mathtt{r}_0)}{2} \mathtt{r}_0^2. \end{aligned}$$

Here $\boldsymbol{\theta}^\circ$ is a point from $\Omega(\mathtt{x})$ on the interval connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. If $\| \boldsymbol{\xi} \| \le \mathtt{r}_0 \{ 1 - \delta(\mathtt{r}_0) \}/2$, then this implies $L(\boldsymbol{\theta}^*) - L(\boldsymbol{\theta}) > 0$, and the result follows. □

As a corollary, we obtain Fisher and Wilks expansions for the quasi MLE $\widetilde{\boldsymbol{\theta}}$ in a generalized linear model.

Theorem 3.5. *Suppose the conditions of Theorem* 3.4 *for some* $\mathtt{r}_0$. *Then it holds on a set* $\Omega(\mathtt{x})$ *with* $\mathbb{P}(\Omega(\mathtt{x})) \ge 1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\bigl\| D\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) - \boldsymbol{\xi} \bigr\| \le \mathtt{r}_0 \delta(\mathtt{r}_0),$$

$$\bigl|2L\bigl(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*\bigr)-\|\boldsymbol{\xi}\|^2\bigr|\le 2\mathtt{r}_0^2\delta(\mathtt{r}_0)+\mathtt{r}_0^2\delta^2(\mathtt{r}_0),$$

$$\bigl|\sqrt{2L\bigl(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}^*\bigr)}-\|\boldsymbol{\xi}\|\bigr|\le 3\mathtt{r}_0\delta(\mathtt{r}_0).$$

Proof. The large deviation bound of Theorem 3.4 allows to restrict the whole parameter space to the local vicinity $\Theta_0(\mathtt{r}_0)$. In this vicinity, the log-likelihood ratio $L(\boldsymbol{\theta},\boldsymbol{\theta}^*)=L(\boldsymbol{\theta})-L(\boldsymbol{\theta}^*)$ can be well approximated by the quadratic expansion $\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$:

$$L\bigl(\boldsymbol{\theta},\boldsymbol{\theta}^*\bigr)=(S-\mathbb{E}S)^\top\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)+\mathbb{E}S^\top\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)-A(\boldsymbol{\theta})+A\bigl(\boldsymbol{\theta}^*\bigr),$$

$$\mathbb{L}\bigl(\boldsymbol{\theta},\boldsymbol{\theta}^*\bigr)\stackrel{\mathrm{def}}{=}(S-\mathbb{E}S)^\top\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)-\frac{1}{2}\bigl\|D\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)\bigr\|^2.$$

Lemma 3.6. *Suppose* (3.10) *for some* $\mathtt{r}_0$. *The difference* $L(\boldsymbol{\theta})-\mathbb{L}(\boldsymbol{\theta})$ *is deterministic and it holds for each* $\boldsymbol{\theta}\in\Theta_0(\mathtt{r}_0)$

$$\begin{aligned}&\bigl|L(\boldsymbol{\theta})-L\bigl(\boldsymbol{\theta}^*\bigr)-\mathbb{L}\bigl(\boldsymbol{\theta},\boldsymbol{\theta}^*\bigr)\bigr|\le\frac{\delta(\mathtt{r}_0)}{2}\bigl\|D\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)\bigr\|^2\le\frac{\delta(\mathtt{r}_0)}{2}\mathtt{r}_0^2,\\&\bigl\|D^{-1}\bigl\{\nabla L(\boldsymbol{\theta})-\nabla\mathbb{L}(\boldsymbol{\theta})\bigr\}\bigr\|\le\mathtt{r}_0\delta(\mathtt{r}_0).\end{aligned}\tag{3.13}$$

Proof. The linear stochastic terms $(S-\mathbb{E}S)^\top\boldsymbol{\theta}$ are the same for $L(\boldsymbol{\theta})$ and $\mathbb{L}(\boldsymbol{\theta})$. For the deterministic terms $\mathbb{E}S^\top\boldsymbol{\theta}-A(\boldsymbol{\theta})$ we use the Taylor formula of the second order at $\boldsymbol{\theta}^*$, the extreme point equation $\nabla A(\boldsymbol{\theta}^*)=\mathbb{E}S$, and the definition $D^2=\mathbb{F}(\boldsymbol{\theta}^*)$:

$$\begin{aligned}\bigl|\mathbb{E}L(\boldsymbol{\theta})-\mathbb{E}\mathbb{L}(\boldsymbol{\theta})\bigr|&=\bigl|A(\boldsymbol{\theta})-A\bigl(\boldsymbol{\theta}^*\bigr)-\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)^\top\nabla A\bigl(\boldsymbol{\theta}^*\bigr)-\bigl\|D\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)\bigr\|^2/2\bigr|\\&=\frac{1}{2}\bigl|\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)^\top\bigl\{\mathbb{F}\bigl(\boldsymbol{\theta}^*\bigr)-\mathbb{F}\bigl(\boldsymbol{\theta}^\circ\bigr)\bigr\}\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)\bigr|,\end{aligned}$$

where $\boldsymbol{\theta}^\circ$ is a point on the interval between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. Now the condition (3.10) implies

$$\bigl|\mathbb{E}L(\boldsymbol{\theta})-\mathbb{E}\mathbb{L}(\boldsymbol{\theta})\bigr|\le\frac{\delta(\mathtt{r}_0)}{2}\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)^\top D^2\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)=\frac{\delta(\mathtt{r}_0)}{2}\bigl\|D\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*\bigr)\bigr\|^2\le\frac{\delta(\mathtt{r}_0)}{2}\mathtt{r}_0^2$$

and the first assertion follows. The second one can be proved similarly. □

With the approximation (3.13), all the statements of the theorem follow from the general results of Theorem 2.6. □

To complete the study of a generalized linear model, we translate the general conditions of Theorem 3.4 into conditions on the design Ψ and on individual errors ε_i.

- *Design regularity* is measured by the value

$$\delta_\Psi\stackrel{\mathrm{def}}{=}\max_i\bigl\|D^{-1}\Psi_i\bigr\|.$$

  In the case of a regular or random design, the Fisher design matrix $D^2=\mathbb{F}(\boldsymbol{\theta}^*)$ is proportional to the sample size and thus, the value δ_Ψ is of order $n^{-1/2}$. Our results only apply if this value is small, in particular, the condition $\delta_\Psi<1/2$ has to be fulfilled.
- *Exponential moments of the errors* $\varepsilon_i=Y_i-\mathbb{E}Y_i$. Suppose that for some values $\mathtt{s}_i$ and fixed constants $\mathtt{C}_0,\lambda_0>0$

$$\mathbb{E}\exp\bigl(\lambda_0\mathtt{s}_i^{-1}\varepsilon_i\bigr)\le\mathtt{C}_0,\quad i=1,\ldots,n.\tag{3.14}$$

  This condition means that the errors ε_i have exponential moments. In most of cases one can use $\mathtt{s}_i^2=\mathrm{Var}(Y_i)$. Condition (3.14) implies that there are another constants $\mathtt{g}_1\le\lambda_0$ and ν_0 such that the following condition is fulfilled:

$$\mathbb{E}\exp\bigl(\lambda_0\mathtt{s}_i^{-1}\varepsilon_i\bigr)\le\frac{1}{2}\nu_0^2\lambda^2,\quad i=1,\ldots,n,\,|\lambda|\le\mathtt{g}_1.\tag{3.15}$$

This follows from the fact that each function $\log \mathbb{E} \exp(\lambda_0 \mathtt{s}_i^{-1} \varepsilon_i)$ analytic in λ in a vicinity of the point zero and can be well approximated by $\lambda^2/2$; see Golubev and Spokoiny [11] for more details.

- *Noise homogeneity* is measured by the variability of the values $\mathtt{s}_i$:

$$a_{\mathtt{s}} \stackrel{\text{def}}{=} \max_{i,j=1,\ldots,n} \mathtt{s}_i / \mathtt{s}_j. \tag{3.16}$$

- *Smoothness of the link function* $g(\upsilon)$ can be measured by its third derivative. It will be assumed that given $\mathtt{r}$, there is a constant $a_g(\mathtt{r})$

$$\frac{|g'''(\Psi_i^\top \boldsymbol{\theta})|}{g''(\Psi_i^\top \boldsymbol{\theta}^*)} \le a_g(\mathtt{r}), \quad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}), i = 1, \ldots, n. \tag{3.17}$$

- *Signal-to-noise ratio* is measured by relationship between the matrices D^2 and V^2, where the matrix V^2 defined as

$$V^2 \stackrel{\text{def}}{=} \sum_{i=1}^n \mathtt{s}_i^2 \Psi_i \Psi_i^\top. \tag{3.18}$$

If the observation Y_i follow the GLM assumption P_{υ_i} for $\upsilon_i = \Psi_i^\top \boldsymbol{\theta}^*$, that is, the model is correctly specified, then $\operatorname{Var}(Y_i) = g''(\upsilon_i)$ and the matrices V^2 and D^2 coincide. In the general case under a possible model misspecification, the matrices V^2 and D^2 may be different. In this case we need an identifiability condition

$$V^2 \le \mathfrak{a}^2 D^2 \tag{3.19}$$

for some $\mathfrak{a} > 0$. This condition can be spelled out as

$$\sum_{i=1}^n \mathtt{s}_i^2 \Psi_i \Psi_i^\top \le \mathfrak{a}^2 \sum_{i=1}^n g''\bigl(\Psi_i^\top \boldsymbol{\theta}^*\bigr) \Psi_i \Psi_i^\top.$$

First we discuss a deviation bound for the norm of the vector $\boldsymbol{\xi}$ given by

$$\boldsymbol{\xi} = D^{-1}(S - \mathbb{E}S) = D^{-1} \sum_{i=1}^n \varepsilon_i \Psi_i.$$

The squared norm $\|\boldsymbol{\xi}\|^2$ is a quadratic form of the ε_i's and one can directly apply general results for quadratic forms from Appendix A.

Theorem 3.7. *Suppose* (3.15), (3.16), (3.17), *and* (3.19). *For* $z(p, \mathtt{x})$ *from* (A.3) *with* $z(p, \mathtt{x}) \le \sqrt{p} + \sqrt{2\mathtt{x}}$, *fix*

$$\mathtt{r}_0 = 4 \nu_0 z(p, \mathtt{x}), \tag{3.20}$$

and suppose that δ_Ψ *is small enough to ensure*

$$a_g(\mathtt{r}_0) \delta_\Psi \mathtt{r}_0 < 1/2. \tag{3.21}$$

Then the conditions of Theorem 3.4 *are fulfilled with* $\delta(\mathtt{r}_0) \le a_g(\mathtt{r}_0) \delta_\Psi \mathtt{r}_0$ *and the results of this theorem continue to apply.*

Proof. Let $\mathtt{r}_0$ be fixed by (3.20). First we bound the value $\delta(\mathtt{r}_0)$.

Lemma 3.8. *The condition* (3.10) *is fulfilled with* $\delta(\mathtt{r}_0) = a_g(\mathtt{r}_0) \delta_\Psi \mathtt{r}_0$.

Proof. For each $\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)$ and $i \le n$, it holds by (3.23)

$$\bigl|\Psi_i^\top \boldsymbol{\theta} - \Psi_i^\top \boldsymbol{\theta}^*\bigr| = \bigl|\bigl(D^{-1}\Psi_i\bigr)^\top D\bigl(\boldsymbol{\theta} - \boldsymbol{\theta}^*\bigr)\bigr| \le \bigl\|D^{-1}\Psi_i\bigr\|\mathtt{r}_0 \le \delta_\Psi \mathtt{r}_0. \tag{3.22}$$

This implies for the difference $\mathbb{F}(\boldsymbol{\theta}) - \mathbb{F}(\boldsymbol{\theta}^*)$

$$\mathbb{F}(\boldsymbol{\theta}) - \mathbb{F}\bigl(\boldsymbol{\theta}^*\bigr) = \sum_{i=1}^n \bigl\{g''\bigl(\Psi_i^\top \boldsymbol{\theta}\bigr) - g''\bigl(\Psi_i^\top \boldsymbol{\theta}^*\bigr)\bigr\}\Psi_i\Psi_i^\top.$$

Next, for each $i \le n$, there exists a point $\boldsymbol{\theta}^\circ$ on the interval between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$ (possibly depending on i) such that

$$g''\bigl(\Psi_i^\top \boldsymbol{\theta}\bigr) - g''\bigl(\Psi_i^\top \boldsymbol{\theta}^*\bigr) = \frac{g'''(\Psi_i^\top \boldsymbol{\theta}^\circ)}{g''(\Psi_i^\top \boldsymbol{\theta}^*)}\bigl(\Psi_i^\top \boldsymbol{\theta} - \Psi_i^\top \boldsymbol{\theta}\bigr)g''\bigl(\Psi_i^\top \boldsymbol{\theta}^*\bigr).$$

Now (3.17) and (3.22) imply

$$\max_{i \le n}\left|\frac{g'''(\Psi_i^\top \boldsymbol{\theta}^\circ)}{g''(\Psi_i^\top \boldsymbol{\theta}^*)}\bigl(\Psi_i^\top \boldsymbol{\theta} - \Psi_i^\top \boldsymbol{\theta}\bigr)\right| \le a_g(\mathtt{r}_0)\delta_\Psi \mathtt{r}_0$$

and

$$\pm\bigl\{\mathbb{F}(\boldsymbol{\theta}) - \mathbb{F}\bigl(\boldsymbol{\theta}^*\bigr)\bigr\} \le a_g(\mathtt{r}_0)\delta_\Psi \mathtt{r}_0 \sum_{i=1}^n g''\bigl(\Psi_i^\top \boldsymbol{\theta}^*\bigr)\Psi_i\Psi_i^\top = a_g(\mathtt{r}_0)\delta_\Psi \mathtt{r}_0 D^2$$

yielding (3.10) in an obvious way. □

This lemma and (3.21) imply $\delta(\mathtt{r}_0) < 1/2$. Now we show that (3.15) implies (3.11).

Lemma 3.9. *Let the errors $\varepsilon_i = Y_i - \mathbb{E}Y_i$ be independent and follow (3.15). Then*

$$\log \mathbb{E}\exp\bigl\{\mathbf{u}^\top V^{-1}(S - \mathbb{E}S)\bigr\} \le \frac{\nu_0^2}{2}\|\mathbf{u}\|^2, \qquad \|\mathbf{u}\| \le \mathtt{g} \stackrel{\text{def}}{=} \frac{\mathtt{g}_1}{\delta_\Psi a_{\mathtt{s}}}, \tag{3.23}$$

where V^2 is from (3.18) and $a_{\mathtt{s}}$ from (3.16).

Proof. The formula (3.8) and independence of the ε_i's imply for any vector $\mathbf{u} \in \mathbb{R}^p$ with $\|\mathbf{u}\| \le \mathtt{g}$

$$\log \mathbb{E}\exp\bigl\{\mathbf{u}^\top V^{-1}(S - \mathbb{E}S)\bigr\} = \sum_{i=1}^n \log \mathbb{E}\exp\bigl(\lambda_i \mathtt{s}_i^{-1}\varepsilon_i\bigr),$$

where the definition (3.23) of $\mathtt{g}$ and condition (3.16) imply for $\lambda_i = \mathtt{s}_i \mathbf{u}^\top V^{-1}\Psi_i$

$$|\lambda_i| = \bigl|\mathbf{u}^\top V^{-1}\Psi_i\bigr|\mathtt{s}_i \le \mathtt{g}\bigl\|V^{-1}\Psi_i\bigr\|\mathtt{s}_i \le \mathtt{g}_1.$$

Therefore, by (3.15) and the definition of V^2

$$\log \mathbb{E}\exp\bigl\{\mathbf{u}^\top V^{-1}(S - \mathbb{E}S)\bigr\} \le \frac{\nu_0^2}{2}\sum_{i=1}^n \lambda_i^2 = \frac{\nu_0^2}{2}\sum_{i=1}^n \mathbf{u}^\top V^{-1}\bigl(\Psi_i\Psi_i^\top \mathtt{s}_i^2\bigr)V^{-1}\mathbf{u} = \frac{\nu_0^2}{2}\|\mathbf{u}\|^2,$$

and the assertion follows. □

The result of Lemma 3.9 provides exponential moments of $\boldsymbol{\xi}$ and one can apply Theorem A.1 from Appendix A yielding the bound (3.11) under the condition

$$\frac{1-\delta(\mathtt{r}_0)}{2}\mathtt{r}_0 \geq \nu_0 z(p,\mathtt{x})$$

which is obviously fulfilled for our choice of $\mathtt{r}_0 = 4\nu_0 z(p,\mathtt{x})$ in view of $\delta(\mathtt{r}_0) < 1/2$. This will also provide (3.11). All the conditions of Theorem 3.4 have been checked. □

3.7. *Estimation for a penalized GLM*

This section briefly discusses what will be changed if the GLM (3.6) is penalized by a roughness penalty term $\|G\boldsymbol{\theta}\|^2/2$. The corresponding penalized log-likelihood $L_G(\boldsymbol{\theta})$ reads as

$$L_G(\boldsymbol{\theta}) = S^\top\boldsymbol{\theta} - A(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2.$$

The penalized MLE and its target are defined by maximizing $L_G(\boldsymbol{\theta})$ and its expectation:

$$\begin{aligned}\widetilde{\boldsymbol{\theta}}_G &\stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta}\bigl\{S^\top\boldsymbol{\theta} - A(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2\bigr\},\\ \boldsymbol{\theta}^*_G &\stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta}\bigl\{\mathbb{E}S^\top\boldsymbol{\theta} - A(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2\bigr\}.\end{aligned} \tag{3.24}$$

Further, define the matrix D_G by $D_G^2 = \mathbb{F}_G(\boldsymbol{\theta}^*_G)$ for

$$\mathbb{F}_G(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{F}(\boldsymbol{\theta}) + G^2 = \sum_{i=1}^n \Psi_i\Psi_i^\top g''\bigl(\Psi_i^\top\boldsymbol{\theta}\bigr) + G^2. \tag{3.25}$$

One can see that the use of penalization leads to a growth of the "information matrix" D_G^2 relative to the non-penalized case. The stochastic term $(S-\mathbb{E}S)^\top\boldsymbol{\theta}$ of $L_G(\boldsymbol{\theta})$ remains the same as in the non-penalized case, thus, the matrix V^2 from (3.18) can be used here as well and the identifiability condition (3.19) continues to hold.

The local vicinity $\Theta_{0,G}(\mathtt{r})$ of $\boldsymbol{\theta}^*_G$ is now defined as

$$\Theta_{0,G}(\mathtt{r}) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\theta}\colon \bigl\|D_G\bigl(\boldsymbol{\theta}-\boldsymbol{\theta}^*_G\bigr)\bigr\| \leq \mathtt{r}\bigr\}.$$

The concentration result for $\widetilde{\boldsymbol{\theta}}_G$ can be easily extended to the penalized case.

Theorem 3.10. *Let, for some $\mathtt{r}_G > 0$, the matrix function $\mathbb{F}_G(\boldsymbol{\theta})$ from (3.25) fulfill with $D_G^2 = \mathbb{F}_G(\boldsymbol{\theta}^*_G)$*

$$\sup_{\boldsymbol{\theta}\in\Theta_{0,G}(\mathtt{r}_0)} \bigl\|D_G^{-1}\mathbb{F}_G(\boldsymbol{\theta})D_G^{-1} - \mathbb{I}_p\bigr\|_{\text{op}} \leq \delta(\mathtt{r}_G)$$

for $\delta(\mathtt{r}_G) < 1$. Let also S from (3.8) follow for $\mathtt{x} > 0$ the probability bound

$$\mathbb{P}\bigl(\bigl\|D_G^{-1}(S-\mathbb{E}S)\bigr\| > z(B_G,\mathtt{x})\bigr) \leq 2\mathrm{e}^{-\mathtt{x}}.$$

If

$$\mathtt{r}_G\bigl\{1-\delta(\mathtt{r}_G)\bigr\} \geq 2z(B_G,\mathtt{x}),$$

with $z(B_G,\mathtt{x}) \leq \sqrt{\mathtt{p}_G} + \sqrt{2\mathtt{x}}$ from (A.4), then the solution $\widetilde{\boldsymbol{\theta}}_G$ of (3.25) satisfies

$$\mathbb{P}\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_{0,G}(\mathtt{r}_G)\bigr) \leq 2\mathrm{e}^{-\mathtt{x}}.$$

Then all the statements of Theorem 3.5 hold for the pair $\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G$ *with* $\boldsymbol{\xi}_G \stackrel{\text{def}}{=} D_G^{-1}(S - \mathbb{E}S)$. *Then it holds on a set* $\Omega_G(\mathtt{x})$ *with* $\mathbb{P}(\Omega_G(\mathtt{x})) \geq 1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\begin{aligned}
&\bigl\| D_G\bigl(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G\bigr) - \boldsymbol{\xi}_G \bigr\| \leq \mathtt{r}_G \delta(\mathtt{r}_G), \\
&\bigl| 2L_G\bigl(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G\bigr) - \|\boldsymbol{\xi}_G\|^2 \bigr| \leq 2\mathtt{r}_G^2 \delta(\mathtt{r}_G) + \mathtt{r}_G^2 \delta^2(\mathtt{r}_G), \\
&\Bigl| \sqrt{2L_G\bigl(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}^*_G\bigr)} - \|\boldsymbol{\xi}_G\| \Bigr| \leq 3\mathtt{r}_G \delta(\mathtt{r}_G).
\end{aligned}$$

The proof of the non-penalized case applies here with obvious changes in notation. However, at one place the difference is essential. Namely, the radius $\mathtt{r}_G$ can be much smaller and it depends on the effective dimension $\mathtt{p}_G = \operatorname{tr}(B_G) = \operatorname{tr}(D_G^{-1} V^2 D_G^{-1})$ rather than on the total dimension p.

Appendix A: Deviation bounds for quadratic forms

This section collects some probability bounds for non-Gaussian quadratic forms. The presented results can be viewed as a slight improvement of the bounds from Spokoiny [25] using the deviation bound from Laurent and Massart [17]. The proofs are very similar to ones from Spokoiny [25] and are omitted by the space reasons.

Let a random vector $\boldsymbol{\xi} \in \mathbb{R}^p$ has some exponential moments. More exactly, suppose for some fixed $\mathtt{g} > 0$ that

$$\log \mathbb{E} \exp\bigl(\boldsymbol{\gamma}^\top \boldsymbol{\xi}\bigr) \leq \|\boldsymbol{\gamma}\|^2/2, \qquad \boldsymbol{\gamma} \in \mathbb{R}^p, \|\boldsymbol{\gamma}\| \leq \mathtt{g}. \tag{A.1}$$

For ease of presentation, assume below that $\mathtt{g}$ is sufficiently large, namely, $0.3\mathtt{g} \geq \sqrt{p}$. In typical examples of an i.i.d. sample, $\mathtt{g} \asymp \sqrt{n}$. Define

$$\begin{aligned}
&\mathtt{x}_c \stackrel{\text{def}}{=} \mathtt{g}^2/4, \\
&z_c^2 \stackrel{\text{def}}{=} p + \sqrt{p\mathtt{g}^2} + \mathtt{g}^2/2 = \mathtt{g}^2\bigl(1/2 + \sqrt{p/\mathtt{g}^2} + p/\mathtt{g}^2\bigr), \\
&\mathtt{g}_c \stackrel{\text{def}}{=} \frac{\mathtt{g}(1/2 + \sqrt{p/\mathtt{g}^2} + p/\mathtt{g}^2)^{1/2}}{1 + \sqrt{p/\mathtt{g}^2}}.
\end{aligned}$$

Note that with $\alpha = \sqrt{p/\mathtt{g}^2} \leq 0.3$, one has

$$\begin{aligned}
&z_c^2 = \mathtt{g}^2\bigl(1/2 + \alpha + \alpha^2\bigr), \\
&\mathtt{g}_c = \mathtt{g}\frac{(1/2 + \alpha + \alpha^2)^{1/2}}{1 + \alpha}
\end{aligned}$$

so that $z_c^2/\mathtt{g}^2 \in [1/2, 1]$ and $\mathtt{g}_c^2/\mathtt{g}^2 \in [1/2, 1]$.

Theorem A.1. *Let* (A.1) *hold and* $0.3\mathtt{g} \geq \sqrt{p}$. *Then for each* $\mathtt{x} > 0$

$$\mathbb{P}\bigl(\|\boldsymbol{\xi}\| \geq z(p, \mathtt{x})\bigr) \leq 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c} \mathbb{1}(\mathtt{x} < \mathtt{x}_c), \tag{A.2}$$

where $z(p, \mathtt{x})$ *is defined by*

$$z(p, \mathtt{x}) \stackrel{\text{def}}{=} \begin{cases} (p + 2\sqrt{p\mathtt{x}} + 2\mathtt{x})^{1/2}, & \mathtt{x} \leq \mathtt{x}_c, \\ z_c + 2\mathtt{g}_c^{-1}(\mathtt{x} - \mathtt{x}_c), & \mathtt{x} > \mathtt{x}_c. \end{cases} \tag{A.3}$$

Depending on the value $\mathtt{x}$, we have two types of tail behavior of the quadratic form $\|\boldsymbol{\xi}\|^2$. For $\mathtt{x} \leq \mathtt{x}_c = \mathtt{g}^2/4$, we have the same deviation bounds as in the Gaussian case with the extra-factor two in the deviation probability. Remind that one can use a simplified expression $(p + 2\sqrt{p\mathtt{x}} + 2\mathtt{x})^{1/2} \leq \sqrt{p} + \sqrt{2\mathtt{x}}$. For $\mathtt{x} > \mathtt{x}_c$, we switch to the special

regime driven by the exponential moment condition (A.1). Usually $\mathtt{g}^2$ is a large number (of order n in the i.i.d. setup) and the second term in (A.2) can be simply ignored.

Next we present a bound for a quadratic form $\boldsymbol{\xi}^\top B\boldsymbol{\xi}$, where $\boldsymbol{\xi}$ satisfies (A.1) and B is a given symmetric positive $p \times p$ matrix. Define

$$\mathtt{p} \stackrel{\text{def}}{=} \operatorname{tr}(B), \qquad \mathtt{v}^2 \stackrel{\text{def}}{=} \operatorname{tr}\bigl(B^2\bigr), \qquad \lambda \stackrel{\text{def}}{=} \lambda_{\max}(B).$$

For ease of presentation, suppose that $0.3\mathtt{g} \geq \sqrt{\mathtt{p}}$ so that $\alpha = \sqrt{\mathtt{p}/\mathtt{g}^2} \leq 0.3$. The other case only changes the constants in the inequalities. Define also

$$\begin{aligned}
&\mathtt{x}_c \stackrel{\text{def}}{=} \mathtt{g}^2/4,\\
&z_c^2 \stackrel{\text{def}}{=} \mathtt{p} + \mathtt{v}\mathtt{g} + \lambda\mathtt{g}^2/2,\\
&\mathtt{g}_c \stackrel{\text{def}}{=} \frac{\sqrt{\mathtt{p}/\lambda + \mathtt{g}\mathtt{v}/\lambda + \mathtt{g}^2/2}}{1 + \mathtt{v}/(\lambda\mathtt{g})}.
\end{aligned}$$

Theorem A.2. *Let* (A.1) *hold and* $0.3\mathtt{g} \geq \sqrt{\mathtt{p}/\lambda}$. *Then for each* $\mathtt{x} > 0$

$$\mathbb{P}\bigl(\bigl\|B^{1/2}\boldsymbol{\xi}\bigr\| \geq z(B,\mathtt{x})\bigr) \leq 2\mathrm{e}^{-\mathtt{x}} + 8.4\mathrm{e}^{-\mathtt{x}_c}\mathbb{1}(\mathtt{x} < \mathtt{x}_c),$$

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

$$z(B,\mathtt{x}) \stackrel{\text{def}}{=} \begin{cases} \sqrt{\mathtt{p} + 2\mathtt{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}}, & \mathtt{x} \leq \mathtt{x}_c, \\ z_c + 2\lambda(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c, & \mathtt{x} > \mathtt{x}_c. \end{cases} \tag{A.4}$$

Similarly to the case $B = \mathbb{I}_p$, the upper quantile $z(B,\mathtt{x}) = \sqrt{\mathtt{p} + 2\mathtt{v}\mathtt{x}^{1/2} + 2\lambda\mathtt{x}}$ can be upper bounded by $\sqrt{\mathtt{p}} + \sqrt{2\lambda\mathtt{x}}$:

$$z(B,\mathtt{x}) \leq \begin{cases} \sqrt{\mathtt{p}} + \sqrt{2\lambda\mathtt{x}}, & \mathtt{x} \leq \mathtt{x}_c, \\ z_c + 2\lambda(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c, & \mathtt{x} > \mathtt{x}_c. \end{cases}$$

Appendix B: Deviation bounds for random processes

This chapter presents some general results of the theory of empirical processes. We assume some exponential moment conditions on the increments of the process which allow to apply the well developed chaining arguments in Orlicz spaces; see e.g. van der Vaart and Wellner [30], Chapter 2.2. We state the results in a slightly different form and present an independent and self-contained proof.

B.1. *Chaining and covering numbers*

An important step in the whole construction is an exponential bound on the maximum of a random process $\mathcal{U}(\boldsymbol{\upsilon})$ under the exponential moment conditions on its increments. Let $d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}')$ be a semi-distance on Υ. We suppose the following condition to hold:

$(\mathcal{E}d)$ There exist $\mathtt{g} > 0$, $\mathtt{r}_0 > 0$, $\nu_0 \geq 1$, such that for any $\lambda \leq \mathtt{g}$ and $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \Upsilon$ with $d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}') \leq \mathtt{r}_0$

$$\log \mathbb{E}\exp\biggl\{\lambda\frac{\mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}')}{d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}')}\biggr\} \leq \nu_0^2\lambda^2/2.$$

By $\mathcal{B}_{\mathtt{r}}(\boldsymbol{\upsilon})$ we denote the d-ball centered at $\boldsymbol{\upsilon}$ of radius $\mathtt{r}$:

$$\mathcal{B}_{\mathtt{r}}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\upsilon}' \in \Upsilon : d\bigl(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}'\bigr) \leq \mathtt{r}\bigr\}.$$

Let Υ° be a subset of a ball in Υ with center at υ^* and radius $\mathtt{r}_0$, and let a sequence $\mathtt{r}_k$ be fixed with $\mathtt{r}_k = \mathtt{r}_0 2^{-k}$. For each k, by $\mathcal{M}_k$ we denote a $\mathtt{r}_k$-net in Υ°, so that

$$\Upsilon^\circ \subseteq \bigcup_{\upsilon \in \mathcal{M}_k} \mathcal{B}_{\mathtt{r}_k}(\upsilon).$$

Let also $\Pi_k \upsilon$ be the closest to υ point from $\mathcal{M}_k$, so that $d(\upsilon, \Pi_k \upsilon) \le \mathtt{r}_k$. We assume that $\mathcal{M}_0$ consists of one point υ^*, that is, $\Pi_0 \upsilon = \upsilon^*$. Let $\mathbb{N}_k \stackrel{\text{def}}{=} |\mathcal{M}_k|$ denote the cardinality of $\mathcal{M}_k$. Finally set $c_k = 2^{-k}$ for $k \ge 1$, and define the values $\mathbb{Q}_1(\Upsilon^\circ)$ and $\mathbb{Q}_2(\Upsilon^\circ)$ by

$$\begin{aligned} \mathbb{Q}_1\bigl(\Upsilon^\circ\bigr) &\stackrel{\text{def}}{=} \sum_{k=1}^{\infty} c_k \sqrt{2 \log(2\mathbb{N}_k)} = \sum_{k=1}^{\infty} 2^{-k} \sqrt{2 \log(2\mathbb{N}_k)}, \\ \mathbb{Q}_2\bigl(\Upsilon^\circ\bigr) &\stackrel{\text{def}}{=} \sum_{k=1}^{\infty} 2 c_k \log(2\mathbb{N}_k) = \sum_{k=1}^{\infty} 2^{-k+1} \log(2\mathbb{N}_k). \end{aligned} \tag{B.1}$$

By the Cauchy–Schwarz inequality $\mathbb{Q}_1^2(\Upsilon^\circ) \le \mathbb{Q}_2(\Upsilon^\circ)$. The inverse relation is not generally true and one can build some examples with $\mathbb{Q}_1(\Upsilon^\circ)$ finite and $\mathbb{Q}_2(\Upsilon^\circ)$ infinite. If the process $\mathcal{U}(\upsilon)$ is sub-Gaussian and $(\mathcal{E}d)$ is fulfilled with $\mathtt{g} = \infty$, then one can only operate with $\mathbb{Q}_1(\Upsilon^\circ)$ which is equivalent to the Dudley integral.

Theorem B.1. *Let $\mathcal{U}$ be a separable process and Υ° be a ball in Υ with center υ° and radius $\mathtt{r}_0$ for the distance $d(\cdot,\cdot)$, i.e. $d(\upsilon, \upsilon^\circ) \le \mathtt{r}_0$ for all $\upsilon \in \Upsilon^\circ$. If $(\mathcal{E}d)$ holds with $\mathtt{g} = \infty$ then for any $\mathtt{x} \ge 1/2$, it holds with $\mathbb{Q}_1 = \mathbb{Q}_1(\Upsilon^\circ)$ and $\mathbb{Q}_2 = \mathbb{Q}_2(\Upsilon^\circ)$*

$$\mathbb{P}\biggl(\frac{1}{\nu_0 \mathtt{r}_0} \sup_{\upsilon \in \Upsilon^\circ} \mathcal{U}\bigl(\upsilon, \upsilon^*\bigr) \ge \mathfrak{z}_{\mathbb{H}}(\mathtt{x}) \biggr) \le \mathrm{e}^{-\mathtt{x}} \tag{B.2}$$

with

$$\mathfrak{z}_{\mathbb{H}}(\mathtt{x}) \stackrel{\text{def}}{=} 2\mathbb{Q}_1 + \sqrt{8\mathtt{x}}.$$

If $\mathtt{g} < \infty$ in $(\mathcal{E}d)$ then (B.2) *holds with $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ given by one of the following rules*:

$$\begin{aligned} &\mathfrak{z}_{\mathbb{H}}(\mathtt{x}) = 2\mathbb{Q}_1 + \sqrt{8\mathtt{x}} + 2\mathtt{g}^{-1}\bigl(\mathtt{g}^{-2}\mathtt{x} + 1\bigr)\mathbb{Q}_2, \\ &\mathfrak{z}_{\mathbb{H}}(\mathtt{x}) = \begin{cases} 2\sqrt{\mathbb{Q}_2 + 2\mathtt{x}}, & \text{if } \mathbb{Q}_2 + 2\mathtt{x} \le \mathtt{g}^2, \\ 2\mathtt{g}^{-1}\mathtt{x} + \mathtt{g}^{-1}\mathbb{Q}_2 + \mathtt{g}, & \text{if } \mathbb{Q}_2 + 2\mathtt{x} > \mathtt{g}^2. \end{cases} \end{aligned} \tag{B.3}$$

Moreover, the r.v. $\mathcal{U}^*(\mathtt{r}_0) \stackrel{\text{def}}{=} \sup_{\upsilon \in \Upsilon^\circ} \mathcal{U}(\upsilon, \upsilon^*)$ fulfills

$$\begin{aligned} &\mathbb{E}\mathcal{U}^*(\mathtt{r}_0) \le 2\nu_0 \mathtt{r}_0 (\mathbb{Q}_1 + \mathbb{Q}_2/\mathtt{g} + 3), \\ &\bigl\{\mathbb{E}\bigl|\mathcal{U}^*(\mathtt{r}_0)\bigr|^2\bigr\}^{1/2} \le 2\nu_0 \mathtt{r}_0 (\mathbb{Q}_1 + \mathbb{Q}_2/\mathtt{g} + 4). \end{aligned} \tag{B.4}$$

Proof. We start the proof by stating some general facts for a convex combinations of sub-exponential r.v.'s ζ_k such that

$$\log \mathbb{E} \exp(\lambda \zeta_k) \le \bigl(q_k^2 + \lambda^2\bigr)/2, \quad |\lambda| \le \mathtt{g}, k = 0, 1, 2, \ldots, \tag{B.5}$$

where $q_k \ge 1$ are fixed numbers, and $\mathtt{g}$ is some positive value or infinity. We aim at bounding a sum S of the form $S = \sum_k c_k \zeta_k$ for a sequence of positive weights c_k satisfying $\sum_k c_k = 1$. We implicitly assume that the numbers q_k grow with k in a way that $\sum_k \exp(-q_k) \le 1$. Define

$$\mathbb{H}_1 \stackrel{\text{def}}{=} \sum_k c_k q_k, \qquad \mathbb{H}_2 \stackrel{\text{def}}{=} \sum_k c_k q_k^2.$$

Lemma B.2. *Suppose that random variables ζ_k follow (B.5) with $\mathtt{g} = \infty$ and $\sum_k \exp(-q_k) \le 1$. Let also $\sum_k c_k = 1$. Then it holds for the sum $S = \sum_k c_k \zeta_k$ and $\mathtt{x} \ge 1/2$*

$$\begin{aligned} &\log \mathbb{E} \exp(S) \le \mathbb{H}_1, \\ &\mathbb{P}(S \ge \mathbb{H}_1 + \sqrt{2\mathtt{x}}) \le \mathrm{e}^{-\mathtt{x}}. \end{aligned} \tag{B.6}$$

If (B.5) holds for $\mathtt{g} < \infty$, then for each $\lambda > 0$ with $|\lambda| \le \mathtt{g}$ and $\mathtt{x} \ge 1/2$

$$\log \mathbb{E} \exp(\lambda S) \le \bigl(\mathbb{H}_2 + \lambda^2\bigr)/2, \tag{B.7}$$

$$\mathbb{P}\bigl\{S \ge \mathfrak{z}_{\mathbb{H}}(\mathtt{x})\bigr\} \le \mathrm{e}^{-\mathtt{x}}, \tag{B.8}$$

where $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ is given by (B.3). Moreover, if $\mathtt{g}^2 \ge \mathbb{H}_2 + 1$, then

$$\mathbb{E} S \le \mathbb{H}_1 + \mathbb{H}_2/\mathtt{g} + 3, \qquad \bigl\{\mathbb{E} S^2\bigr\}^{1/2} \le \mathbb{H}_1 + \mathbb{H}_2/\mathtt{g} + 4.$$

Proof. Consider first the sub-Gaussian case with $\mathtt{g} = \infty$. Define $\alpha_k = c_k/q_k$. Obviously $\sum_k \alpha_k \le \sum_k c_k = 1$. By the Hölder inequality and (B.5), it holds

$$\begin{aligned} \log \mathbb{E} \exp\Bigl(\sum_k c_k \zeta_k\Bigr) &= \log \mathbb{E} \exp\Bigl(\sum_k \alpha_k q_k \zeta_k\Bigr) \le \sum_k \alpha_k \log \mathbb{E} \exp(q_k \zeta_k) \\ &\le \frac{1}{2} \sum_k \alpha_k \bigl(q_k^2 + q_k^2\bigr) \le \sum_k c_k q_k. \end{aligned}$$

Further, by the same arguments, it holds

$$\log \mathbb{E} \exp(\lambda S) \le \sum_k c_k \log \mathbb{E} \exp(\lambda \zeta_k) \le \frac{1}{2} \sum_k c_k \bigl(q_k^2 + \lambda^2\bigr)$$

and the assertion (B.7) follows as well.

Let $\mathtt{x} \ge 1/2$ be fixed. With $z_k = q_k + \sqrt{2\mathtt{x}}$, it follows by (B.5) for $\lambda_k = z_k$ in view of $\sum_k \mathrm{e}^{-q_k} \le 1$

$$\begin{aligned} &\mathbb{P}\Bigl(\sum_k c_k(\zeta_k - z_k) \ge 0\Bigr) \\ &\quad \le \sum_k \mathbb{P}(\zeta_k - z_k \ge 0) \le \sum_k \mathbb{E} \exp\bigl\{\lambda_k(\zeta_k - z_k)\bigr\} \\ &\quad \le \sum_k \exp\Bigl(-\lambda_k z_k + \frac{\lambda_k^2}{2} + \frac{q_k^2}{2}\Bigr) = \sum_k \exp\Bigl(-\frac{z_k^2}{2} + \frac{q_k^2}{2}\Bigr) = \sum_k \mathrm{e}^{-\mathtt{x} - q_k\sqrt{2\mathtt{x}}} \le \mathrm{e}^{-\mathtt{x}}. \end{aligned} \tag{B.9}$$

This implies (B.6) by

$$\sum_k c_k z_k = \sum_k c_k (q_k + \sqrt{2\mathtt{x}}) = \mathbb{H}_1 + \sqrt{2\mathtt{x}}. \tag{B.10}$$

Now we briefly discuss how the condition (B.5) can be relaxed to the case of a finite $\mathtt{g}$. Suppose that (B.5) holds for all $\lambda \le \mathtt{g} < \infty$. Define $k(\mathtt{x})$ as the largest index k, for which $\lambda_k = q_k + \sqrt{2\mathtt{x}} \le \mathtt{g}$. For $k > k(\mathtt{x})$, define $\lambda_k = \mathtt{g}$ and

$$z_k = \frac{\mathtt{x} + q_k}{\mathtt{g}} + \frac{\mathtt{g}}{2} + \frac{q_k^2}{2\mathtt{g}}. \tag{B.11}$$

The above arguments yield for $k > k(\mathtt{x})$

$$\mathbb{P}(\zeta_k \geq z_k) \leq \exp\bigl\{-\mathtt{g} z_k + \bigl(q_k^2 + \mathtt{g}^2\bigr)/2\bigr\} = \exp(-\mathtt{x} - q_k).$$

This and (B.9) yield

$$\sum_k \mathbb{P}(\zeta_k \geq z_k) \leq \sum_{k \leq k(\mathtt{x})} \mathrm{e}^{-\mathtt{x} - q_k \sqrt{2\mathtt{x}}} + \sum_{k > k(\mathtt{x})} \mathrm{e}^{-\mathtt{x} - q_k} \leq \sum_k \mathrm{e}^{-\mathtt{x} - q_k} \leq \mathrm{e}^{-\mathtt{x}}.$$

Further, as $q_k > \mathtt{g}$ for $k > k(\mathtt{x})$, it follows from the definition (B.11)

$$\begin{aligned}\sum_{k > k(\mathtt{x})} c_k z_k &= \frac{1}{\mathtt{g}} \sum_{k > k(\mathtt{x})} c_k (\mathtt{x} + q_k) + \frac{\mathtt{g}}{2} \sum_{k > k(\mathtt{x})} c_k + \frac{1}{2\mathtt{g}} \sum_{k > k(\mathtt{x})} c_k q_k^2 \\ &\leq \frac{1}{\mathtt{g}} \sum_{k > k(\mathtt{x})} c_k q_k + \left(\frac{\mathtt{x}}{\mathtt{g}^3} + \frac{1}{\mathtt{g}} \right) \sum_{k > k(\mathtt{x})} c_k q_k^2.\end{aligned}$$

This and (B.10) imply due to $\mathtt{g} \geq 1$

$$\sum_k c_k z_k \leq \sum_k c_k q_k + \left(\frac{\mathtt{x}}{\mathtt{g}^3} + \frac{1}{\mathtt{g}} \right) \sum_k c_k q_k^2 + \sqrt{2\mathtt{x}} \leq \mathbb{H}_1 + \left(\frac{\mathtt{x}}{\mathtt{g}^3} + \frac{1}{\mathtt{g}} \right) \mathbb{H}_2 + \sqrt{2\mathtt{x}}.$$

In particular, if $\mathtt{x} \leq \mathtt{g}^2$, then

$$\sum_k c_k z_k \leq \mathbb{H}_1 + \frac{2}{\mathtt{g}} \mathbb{H}_2 + \sqrt{2\mathtt{x}}.$$

Now (B.8) with $\mathfrak{z}(\mathtt{x}) = \mathbb{H}_1 + \sqrt{2\mathtt{x}} + \mathtt{g}^{-1}(\mathtt{g}^{-2}\mathtt{x} + 1)\mathbb{H}_2$ follows similarly to (B.6). Further, if $\mathfrak{z}(\mathtt{x}) = \sqrt{\mathbb{H}_2 + 2\mathtt{x}} \leq \mathtt{g}$, then (B.7) with $\lambda = \mathfrak{z}(\mathtt{x})$ and the exponential Chebyshev inequality implies again

$$\mathbb{P}\bigl(S \geq \mathfrak{z}(\mathtt{x})\bigr) \leq \exp\left(-\lambda \mathfrak{z}(\mathtt{x}) + \frac{\mathbb{H}_2 + \lambda^2}{2} \right) = \exp\left(\frac{-\mathfrak{z}^2(\mathtt{x}) + \mathbb{H}_2}{2} \right) = \exp(-\mathtt{x}).$$

Similarly one can check the case with $\lambda = \mathtt{g}$ and $\mathfrak{z}(\mathtt{x}) = \mathtt{x}/\mathtt{g} + (\mathbb{H}_2/\mathtt{g} + \mathtt{g})/2 > \mathtt{g}$.

To bound the moments of S, we apply the following technical result: if $\mathbb{P}(S \geq \mathfrak{z}(\mathtt{x})) \leq \mathrm{e}^{-\mathtt{x}}$ for all $\mathtt{x} \geq \mathtt{x}_0$ and if $\mathfrak{z}(\cdot)$ is absolutely continuous, then

$$\mathbb{E} S \leq \mathfrak{z}(\mathtt{x}_0) + \int_{\mathtt{x}_0}^{\infty} \mathfrak{z}'(\mathtt{x}) \mathrm{e}^{-\mathtt{x}} \, d\mathtt{x}, \qquad \mathbb{E} S^2 \leq \mathfrak{z}^2(\mathtt{x}_0) + 2 \int_{\mathtt{x}_0}^{\infty} \mathfrak{z}(\mathtt{x}) \mathfrak{z}'(\mathtt{x}) \mathrm{e}^{-\mathtt{x}} \, d\mathtt{x}.$$

For $\mathfrak{z}(\mathtt{x}) = \mathbb{H}_1 + \sqrt{2\mathtt{x}} + \mathtt{g}^{-1}(\mathtt{g}^{-2}\mathtt{x} + 1)\mathbb{H}_2$, it holds $\mathfrak{z}'(\mathtt{x}) \leq 1 + \mathtt{g}^{-3}$. In view of $\mathtt{g}^2 \geq \mathbb{H}_2 + 1$

$$\mathbb{E} S \leq \mathbb{H}_1 + 1 + (\mathbb{H}_2 + 1/2)/\mathtt{g} + \int_{1/2}^{\infty} \bigl(1 + \mathtt{g}^{-3}\bigr) \mathrm{e}^{-\mathtt{x}} \, d\mathtt{x} \leq \mathbb{H}_1 + \mathbb{H}_2/\mathtt{g} + 3.$$

Similarly one can bound

$$\mathbb{E} S^2 \leq (\mathbb{H}_1 + \mathbb{H}_2/\mathtt{g} + 3/2)^2 + 2 \int_{1/2}^{\infty} \left(\frac{1}{\sqrt{2\mathtt{x}}} + \mathtt{g}^{-3} \right) \mathfrak{z}(\mathtt{x}) \mathrm{e}^{-\mathtt{x}} \, d\mathtt{x} \leq (\mathbb{H}_1 + \mathbb{H}_2/\mathtt{g} + 4)^2$$

as required. □

Now we show how the statement of the theorem can be reduced to the bounds of Lemma B.2. Denote for $i < k$ by Π_i^k the product $\Pi_i^k = \Pi_i \Pi_{i+1} \cdots \Pi_k$. As $\Pi_0 \upsilon \equiv \upsilon^*$, the telescopic sum devices yields

$$\bigl|\mathcal{U}(\Pi_k \upsilon) - \mathcal{U}(\upsilon^*)\bigr| \le \sum_{i=1}^{k} \bigl|\mathcal{U}\bigl(\Pi_{i-1}^k \upsilon\bigr) - \mathcal{U}\bigl(\Pi_i^k \upsilon\bigr)\bigr|.$$

Separability of $\mathcal{U}(\cdot)$ implies that $\lim_{k\to\infty} \mathcal{U}(\Pi_k \upsilon) = \mathcal{U}(\upsilon)$. Therefore, for any $\upsilon \in \Upsilon^\circ$

$$\bigl|\mathcal{U}(\upsilon) - \mathcal{U}(\upsilon^*)\bigr| = \lim_{k\to\infty} \bigl|\mathcal{U}(\Pi_k \upsilon) - \mathcal{U}(\upsilon^*)\bigr| \le \sum_{k=1}^{\infty} \xi_k^*,$$

where

$$\xi_k^* \stackrel{\mathrm{def}}{=} \max_{\upsilon \in \mathcal{M}_k} \bigl|\mathcal{U}(\upsilon) - \mathcal{U}(\Pi_{k-1}\upsilon)\bigr|.$$

For each $\upsilon \in \mathcal{M}_k$, it holds $d(\upsilon, \Pi_{k-1}\upsilon) \le \mathtt{r}_{k-1}$ and

$$\bigl|\mathcal{U}(\upsilon) - \mathcal{U}(\Pi_{k-1}\upsilon)\bigr| \le \mathtt{r}_{k-1} \frac{|\mathcal{U}(\upsilon) - \mathcal{U}(\Pi_{k-1}\upsilon)|}{d(\upsilon, \Pi_{k-1}\upsilon)}.$$

This implies by the Jensen inequality and $(\mathcal{E}d)$ in view of $e^{|x|} \le e^x + e^{-x}$ for each $k \ge 1$ and $|\lambda| \le \mathtt{g}$

$$\mathbb{E}\exp\biggl(\frac{\lambda}{\mathtt{r}_{k-1}} \xi_k^*\biggr) \le 2 \sum_{\upsilon \in \mathcal{M}_k} \mathbb{E}\exp\biggl(\lambda \frac{|\mathcal{U}(\upsilon) - \mathcal{U}(\Pi_{k-1}\upsilon)|}{d(\upsilon, \Pi_{k-1}\upsilon)}\biggr) \le 2\mathbb{N}_k \exp\bigl(\lambda^2/2\bigr). \tag{B.12}$$

For $k \ge 1$, define $q_k^2/2 = \log(2\mathbb{N}_k)$, $c_k = 2^{-k}$, and $\zeta_k = \xi_k^*/\mathtt{r}_{k-1} = c_k^{-1}\xi_k^*/(2\mathtt{r}_0)$. Then (B.12) implies by $\mathtt{r}_{k-1} = 2^{-k+1}\mathtt{r}_0$

$$\log \mathbb{E}\exp(\lambda\zeta_k) \le \log(2\mathbb{N}_k) + \lambda^2/2 = \bigl(q_k^2 + \lambda^2\bigr)/2.$$

Now we apply Lemma B.2 with $c_k = 2^{-k}$. By construction

$$\sum_{k=1}^{\infty} c_k \zeta_k = \frac{1}{2\mathtt{r}_0} \sum_{k=1}^{\infty} \xi_k^*$$

and the results follow with $\mathbb{H}_1 = \mathbb{Q}_1(\Upsilon^\circ)$, $\mathbb{H}_2 = \mathbb{Q}_2(\Upsilon^\circ)$. □

B.2. A large deviation bound

Due to the result of Theorem B.1, the bound for the maximum of $\mathcal{U}(\upsilon, \upsilon^*)$ over $\upsilon \in \mathcal{B}_{\mathtt{r}}(\upsilon^*)$ grows linearly in $\mathtt{r}$. So, its applications to situations with $\mathtt{r} \gg \mathbb{Q}_1(\Upsilon^\circ)$ are limited. The next result shows that introducing a negative drift helps to state a uniform in $\mathtt{r}$ local probability bound. Namely, the bound for the process $\mathcal{U}(\upsilon, \upsilon^*) - f(d(\upsilon, \upsilon^*))$ for some function $f(\mathtt{r})$ over a ball $\mathcal{B}_{\mathtt{r}}(\upsilon^*)$ around the point υ^* does not depend on $\mathtt{r}$. Here the generic chaining arguments are accomplished with the slicing technique. The idea is for a given $\mathtt{r}^* > 1$ to split the ball $\mathcal{B}_{\mathtt{r}^*}(\upsilon^*)$ into the slices $\mathcal{B}_{\mathtt{r}_k}(\upsilon^*) \setminus \mathcal{B}_{\mathtt{r}_{k-1}}(\upsilon^*)$ and to apply Theorem B.1 to each slice separately.

Theorem B.3. *Let $\mathtt{r}^*$ be such that $(\mathcal{E}d)$ holds on $\mathcal{B}_{\mathtt{r}^*}(\upsilon^*)$. Let also $\mathbb{Q}_1(\mathcal{B}_{\mathtt{r}}(\upsilon^*)) \le \mathbb{H}_1$ and $\mathbb{Q}_2(\mathcal{B}_{\mathtt{r}}(\upsilon^*)) \le \mathbb{H}_2$ for $\mathtt{r} \le \mathtt{r}^*$. Given $\mathtt{r}_0 < \mathtt{r}^*$, let a monotonous function $f(\mathtt{r}, \mathtt{r}_0)$ fulfill for some $\rho < 1$*

$$f(\mathtt{r}, \mathtt{r}_0) \ge \nu_0 \mathtt{r} \mathfrak{z}_{\mathbb{H}}\bigl(\mathtt{x} + \log(\mathtt{r}/\mathtt{r}_0)\bigr), \quad \mathtt{r}_0 \le \mathtt{r} \le \mathtt{r}^*, \tag{B.13}$$

where the function $\mathfrak{z}_{\mathbb{H}}(\cdot)$ is given by (B.3). *Then it holds*

$$\mathbb{P}\Big(\sup_{\mathtt{r}_0 \le \mathtt{r} \le \mathtt{r}^*} \sup_{\upsilon \in \mathcal{B}_{\mathtt{r}}(\upsilon^*)} \big\{ \mathcal{U}\big(\upsilon, \upsilon^*\big) - f\big(\rho^{-1}\mathtt{r}, \mathtt{r}_0\big) \big\} \ge 0 \Big) \le \frac{\rho}{1-\rho} \mathrm{e}^{-\mathtt{x}}.$$

Remark B.1. *Formally the bound applies even with $\mathtt{r}^* = \infty$ provided that* ($\mathcal{E}d$) *is fulfilled on the whole set Υ°.*

Remark B.2. *If $\mathtt{g} = \infty$, then $\mathfrak{z}_{\mathbb{H}}(\mathtt{x}) = 2\mathbb{H}_1 + \sqrt{8\mathtt{x}}$ and the condition* (B.13) *on the drift simplifies to $(2\nu_0\mathtt{r})^{-1} f(\mathtt{r}, \mathtt{r}_0) \ge \mathbb{H}_1 + \sqrt{2\mathtt{x} + 2\log(\mathtt{r}/\mathtt{r}_0)}$.*

Proof of Theorem B.3. By (B.13) and Theorem B.1 for any $\mathtt{r} > \mathtt{r}_0$

$$\begin{aligned} &\mathbb{P}\Big(\sup_{\upsilon \in \mathcal{B}_{\mathtt{r}}(\upsilon^*) \setminus \mathcal{B}_{\rho\mathtt{r}}(\upsilon^*)} \big\{ \mathcal{U}\big(\upsilon, \upsilon^*\big) - f(\mathtt{r}, \mathtt{r}_0) \big\} \ge 0 \Big) \\ &\quad \le \mathbb{P}\Big(\frac{1}{\nu_0 \mathtt{r}} \sup_{\upsilon \in \mathcal{B}_{\mathtt{r}}(\upsilon^*)} \mathcal{U}\big(\upsilon, \upsilon^*\big) \ge \mathfrak{z}\big(\mathtt{x} + \log(\mathtt{r}/\mathtt{r}_0)\big) \Big) \le \frac{\mathtt{r}_0}{\mathtt{r}} \mathrm{e}^{-\mathtt{x}}. \end{aligned} \tag{B.14}$$

Now defined $\mathtt{r}_k = \mathtt{r}_0 \rho^{-k}$ for $k = 0, 1, 2, \ldots$. Define also $k^* \stackrel{\text{def}}{=} \log(\mathtt{r}^*/\mathtt{r}_0) + 1$. It follows from (B.14) that

$$\begin{aligned} &\mathbb{P}\Big(\sup_{\upsilon \in \mathcal{B}_{\mathtt{r}^*}(\upsilon^*) \setminus \mathcal{B}_{\mathtt{r}_0}(\upsilon^*)} \big\{ \mathcal{U}\big(\upsilon, \upsilon^*\big) - f\big(\rho^{-1} d\big(\upsilon, \upsilon^*\big), \mathtt{r}_0\big) \big\} \ge 0 \Big) \\ &\quad \le \sum_{k=1}^{k*} \mathbb{P}\Big(\frac{1}{\mathtt{r}_k} \sup_{\upsilon \in \mathcal{B}_{\mathtt{r}_k}(\upsilon^*) \setminus \mathcal{B}_{\mathtt{r}_{k-1}}(\upsilon^*)} \big\{ \mathcal{U}\big(\upsilon, \upsilon^*\big) - f(\mathtt{r}_k, \mathtt{r}_0) \big\} \ge 0 \Big) \le \mathrm{e}^{-\mathtt{x}} \sum_{k=1}^{k^*} \rho^k \le \frac{\rho}{1-\rho} \mathrm{e}^{-\mathtt{x}} \end{aligned}$$

as required. □

B.3. Finite-dimensional smooth case

Here we discuss the special case when Υ is an open subset in $\mathbb{R}^p$, the stochastic process $\mathcal{U}(\upsilon)$ is Fréchet differentiable and its gradient $\nabla \mathcal{U}(\upsilon) \stackrel{\text{def}}{=} d\mathcal{U}(\upsilon)/d\upsilon$ has bounded exponential moments.

($\mathcal{E}D$) There exist $\mathtt{g} > 0$, $\nu_0 \ge 1$, and for each $\upsilon \in \Upsilon$, a symmetric non-negative matrix $\mathbb{V}(\upsilon)$ such that for any $\lambda \le \mathtt{g}$ and any unit vector $\gamma \in \mathbb{R}^p$, it holds

$$\log \mathbb{E} \exp\Big\{ \lambda \frac{\gamma^\top \nabla \mathcal{U}(\upsilon)}{\| \mathbb{V}(\upsilon)\gamma \|} \Big\} \le \frac{\nu_0^2 \lambda^2}{2}.$$

A natural candidate for $\mathbb{V}^2(\upsilon)$ is the covariance matrix $\mathrm{Var}(\nabla \mathcal{U}(\upsilon))$ provided that this matrix is well posed. Then the constant ν_0 can be taken close to one by reducing the value $\mathtt{g}$.

In what follows we fix a subset Υ° of Υ and establish a bound for the maximum of the process $\mathcal{U}(\upsilon, \upsilon^\circ) = \mathcal{U}(\upsilon) - \mathcal{U}(\upsilon^\circ)$ on Υ° for a fixed point υ°. We assume existence of a matrix $\mathbb{V} = \mathbb{V}(\Upsilon^\circ)$ such that $\mathbb{V}(\upsilon) \preceq \mathbb{V}$ for all $\upsilon \in \Upsilon^\circ$. We also assume that π is the Lebesgue measure on Υ. First we show that ($\mathcal{E}D$) implies ($\mathcal{E}d$).

Lemma B.4. *Assume that* ($\mathcal{E}D$) *holds with some $\mathtt{g}$ and $\mathbb{V}(\upsilon) \preceq \mathbb{V}$ for $\upsilon \in \Upsilon^\circ$. Consider any $\upsilon, \upsilon^\circ \in \Upsilon^\circ$. Then it holds for $|\lambda| \le \mathtt{g}$*

$$\log \mathbb{E} \exp\Big\{ \lambda \frac{\mathcal{U}(\upsilon, \upsilon^\circ)}{\| \mathbb{V}(\upsilon - \upsilon^\circ) \|} \Big\} \le \frac{\nu_0^2 \lambda^2}{2}.$$

Proof. Denote $\delta = \|\upsilon - \upsilon^\circ\|$, $\gamma = (\upsilon - \upsilon^\circ)/\delta$. Then

$$\mathcal{U}(\upsilon, \upsilon^\circ) = \delta \gamma^\top \int_0^1 \nabla \mathcal{U}\bigl(\upsilon^\circ + t\delta\gamma\bigr)\, dt$$

and $\|\mathbb{V}(\upsilon - \upsilon^\circ)\| = \delta \|\mathbb{V}\gamma\|$. Now the Hölder inequality and $(\mathcal{E}D)$ yield

$$\begin{aligned}\mathbb{E}\exp\left\{\lambda \frac{\mathcal{U}(\upsilon, \upsilon^\circ)}{\|\mathbb{V}(\upsilon - \upsilon^\circ)\|} - \frac{\nu_0^2\lambda^2}{2}\right\} &= \mathbb{E}\exp\left\{\int_0^1 \left[\lambda \frac{\gamma^\top \nabla \mathcal{U}(\upsilon^\circ + t\delta\gamma)}{\|\mathbb{V}\gamma\|} - \frac{\nu_0^2\lambda^2}{2}\right] dt\right\} \\ &\le \int_0^1 \mathbb{E}\exp\left\{\lambda \frac{\gamma^\top \nabla \mathcal{U}(\upsilon^\circ + t\delta\gamma)}{\|\mathbb{V}\gamma\|} - \frac{\nu_0^2\lambda^2}{2}\right\} dt \le 1\end{aligned}$$

as required. □

The result of Lemma B.4 enables us to define $d(\upsilon, \upsilon') = \|\mathbb{V}(\upsilon - \upsilon')\|$ so that the corresponding d-ball coincides with the following ellipsoidal set $\mathcal{B}(\mathtt{r}, \upsilon^\circ)$:

$$\mathcal{B}\bigl(\mathtt{r}, \upsilon^\circ\bigr) \stackrel{\text{def}}{=} \bigl\{\upsilon: \bigl\|\mathbb{V}\bigl(\upsilon - \upsilon^\circ\bigr)\bigr\| \le \mathtt{r}\bigr\}.$$

Now we bound the value $\mathbb{Q}(\Upsilon^\circ)$ for $\Upsilon^\circ = \mathcal{B}(\mathtt{r}, \upsilon^\circ)$. Note that by change of variable one can reduce the study to the case $\mathbb{V} = \mathbb{I}_p$ and consider the entropy of the unit ball in $\mathbb{R}^p$ w.r.t. the Euclidean distance. We use the following general result which allows to upperbound the covering number of a convex set in $\mathbb{R}^p$ for the Euclidean metric.

Lemma B.5. *Let Υ° be a convex set in $\mathbb{R}^p$, $\delta > 0$, and B be the unit ball in $\mathbb{R}^p$. Then the covering number $\mathbb{N}(\Upsilon^\circ, \delta)$ fulfills*

$$\mathbb{N}\bigl(\Upsilon^\circ, \delta\bigr) \le \frac{\mathrm{vol}(\Upsilon^\circ + (\delta/2)B)}{\mathrm{vol}(B)} (2/\delta)^p.$$

Proof. Let $(\upsilon^{(i)}, i = 1, \ldots, \mathbb{N})$ be a maximal subset of Υ° such that $\|\upsilon^{(i)} - \upsilon^{(j)}\| \ge \delta$ for all $i \ne j$. By maximality, $(\upsilon^{(i)})$ is a δ-net of Υ°. Let also B be the unit ball in $\mathbb{R}^p$. Note that the balls $\upsilon^{(i)} + (\delta/2)B$ are disjoint and included in $\Upsilon^\circ + (\delta/2)B$. Therefore,

$$\sum_{i \le \mathbb{N}} \mathrm{vol}\left(\upsilon^{(i)} + \frac{\delta}{2}B\right) \le \mathrm{vol}\left(\Upsilon^\circ + \frac{\delta}{2}B\right),$$

where $\mathrm{vol}(A)$ means the Lebesgue measure of the set A. This yields

$$\mathbb{N}(\delta/2)^p \,\mathrm{vol}(B) \le \mathrm{vol}\bigl(\Upsilon^\circ + (\delta/2)B\bigr)$$

and the claim of the lemma follows. □

Lemma B.6 (Entropy of a ball). *Let $\Upsilon^\circ = \mathcal{B}(\mathtt{r}_\circ, \upsilon^*)$ and $\mathtt{r}_k = 2^{-k}\mathtt{r}_\circ$. Then the covering numbers $\mathbb{N}_k$ fulfill with $\delta = \mathtt{r}_k/\mathtt{r}_\circ = 2^{-k}$*

$$\mathbb{N}_k \le (1 + 2/\delta)^p = \bigl(1 + 2^{k+1}\bigr)^p.$$

Moreover, with $\mathfrak{c}_2 = 4.67$,

$$\begin{aligned}&\mathbb{Q}_2\bigl(\Upsilon^\circ\bigr) \le 2\log 2 + \mathfrak{c}_2 p \le 6p, \\ &\mathbb{Q}_1\bigl(\Upsilon^\circ\bigr) \le \sqrt{2\log 2 + \mathfrak{c}_2 p} \le \sqrt{6p}.\end{aligned} \tag{B.15}$$

Proof. A change of variable reduces the statement to the case $\mathbb{V} = \mathbb{I}_p$ and $\mathtt{r}_\circ = 1$. For $\delta = 2^{-k}$, this implies by Lemma B.5 in view of $\Upsilon^\circ = B$

$$\mathrm{vol}\Big(\Upsilon^\circ + \frac{\delta}{2} B\Big) = (1 + \delta/2)^p \,\mathrm{vol}(B),$$

that $\mathbb{N}_k \le (1 + 2/\delta)^p$ as claimed. Now we derive

$$\begin{aligned}\mathbb{Q}_2\big(\Upsilon^\circ\big) &\le \sum_{k=1}^{\infty} 2^{-k+1} \log(2\mathbb{N}_k) \le \sum_{k=1}^{\infty} 2^{-k+1} \big\{\log 2 + 2p \log\big(1 + 2^{k+1}\big)\big\} \\ &\le 2\log 2 + p \sum_{k=0}^{\infty} 2^{-k+1} \log\big(1 + 2^k\big) \le 2\log 2 + \mathfrak{c}_2 p\end{aligned}$$

as required. □

Now we specify the local bounds of Theorem B.1 to the smooth case. We consider the local sets of the elliptic form $\Upsilon_\circ(\mathtt{r}) \stackrel{\mathrm{def}}{=} \{\boldsymbol{\upsilon} : \|\mathbb{V}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}\}$, where $\mathbb{V}$ dominates $\mathbb{V}(\boldsymbol{\upsilon})$ on this set: $\mathbb{V}(\boldsymbol{\upsilon}) \preceq \mathbb{V}$.

Theorem B.7. *Let ($\mathcal{E}D$) hold with some $\mathtt{g} > 0$, and matrices $\mathbb{V}(\boldsymbol{\upsilon})$ such that $\mathbb{V}(\boldsymbol{\upsilon}) \preceq \mathbb{V}$ for all $\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})$ and a fixed $\mathtt{r}$. For any $\mathtt{x} \ge 1/2$*

$$\mathbb{P}\bigg\{\frac{1}{\nu_0 \mathtt{r}} \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \big|\mathcal{U}\big(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*\big)\big| \ge \mathfrak{z}_{\mathbb{H}}(\mathtt{x})\bigg\} \le \mathrm{e}^{-\mathtt{x}},$$

where $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ is given by (B.3) with $\mathbb{Q}_1(\Upsilon^\circ)$ and $\mathbb{Q}_2(\Upsilon^\circ)$ from (B.15).

Proof. Lemma B.6 implies ($\mathcal{E}d$) with $d(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) = \|\mathbb{V}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|$. Now the result follows from Theorem B.1. □

B.4. Entropy of an ellipsoid

Let H be a positive self adjoint operator in $\mathbb{R}^\infty$. We are interested to describe the entropy of the elliptic set

$$\mathcal{E}_H(\mathtt{r}_\circ) \stackrel{\mathrm{def}}{=} \big\{\boldsymbol{\upsilon}: \big\|H\big(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\big)\big\| \le \mathtt{r}_\circ\big\} \tag{B.16}$$

for given $\boldsymbol{\upsilon}^\circ \in \mathbb{R}^\infty$ and $\mathtt{r}_\circ > 0$ with respect to the usual Euclidean distance in $\mathbb{R}^\infty$. Below we evaluate the entropy of this set assuming that $\|H^{-1}\|_{\mathrm{op}} = 1$ and H^{-2} is a trace operator, i.e., $h_1 = 1$ and

$$\mathtt{p}_H \stackrel{\mathrm{def}}{=} \mathrm{tr}\big(H^{-2}\big) = \sum_{j=1}^{\infty} h_j^{-2} < \infty, \tag{B.17}$$

where $h_1 \le h_2 \le \cdots$ are the ordered eigenvalues of H.

Theorem B.8. *Suppose that for some $\alpha > 1$*

$$\mathtt{p}_H(\alpha) \stackrel{\mathrm{def}}{=} \sum_{j=1}^{\infty} h_j^{-2} \log^\alpha\big(h_j^2\big) < \infty. \tag{B.18}$$

Then for $\mathcal{E} = \mathcal{E}_H(\mathtt{r}_\circ)$

$$\mathbb{Q}_1(\mathcal{E}) \le \mathtt{C}(\alpha - 1)^{-1/2} \sqrt{\mathtt{p}_H(\alpha)}, \tag{B.19}$$

where $\mathtt{C}$ *is an absolute constant. Furthermore,*

$$\mathbb{Q}_2(\mathcal{E}) \le \mathtt{C}\mathtt{p}_H^* = \mathtt{C}\sum_{j=1}^{\infty} h_j^{-1}.$$

Remark B.3. *The log-factor in the definition of* $\mathtt{p}_H(\alpha)$ *can be removed by using a more advanced generic chaining and majorising measure technique. However, in most of situations, the bound in terms of* $\mathtt{p}_H(\alpha)$ *is also sharp.*

The term $\mathtt{p}_H^*$ *only appears in the sub-exponential case when* $\mathtt{g} < \infty$*. In this case we need the condition* $\mathtt{p}_H^* < \infty$ *which requires* $\sum_j h_j^{-1} < \infty$*, that is, a more rapid growth of the values* h_j *is necessary than in* (B.18).

Proof of Theorem B.8. We begin by a general lemma which bounds the covering numbers for the elliptic set $\mathcal{E}$ for the Euclidean distance.

Lemma B.9 (Entropy of the ellipsoid). *Let* $\mathcal{E} = \mathcal{E}_H(\mathtt{r}_\circ)$ *be an elliptic set from* (B.16) *with* $\|H^{-1}\|_{\mathrm{op}} = 1$ *and* $\mathrm{tr}(H^{-2}) < \infty$*. Let also* $d(\upsilon, \upsilon') = \|\upsilon - \upsilon'\|$*. Then for* $\mathtt{r}_k = 2^{-k}\mathtt{r}_\circ$*, the value* $\mathbb{Q}_1(\mathcal{E})$ *from* (B.1) *satisfies*

$$\mathbb{Q}_1(\mathcal{E}) \le \sum_{k=1}^{\infty} 2^{-k}\sqrt{\log 2 + 2L_H(m_k)}, \tag{B.20}$$

where m_k *is the index* j *for which* $h_{m_k}^2 = 2^{2k+1}$ *and hence,*

$$h_j^2 \le 2^{2k+1}, \quad j \le m_k, \tag{B.21}$$

and

$$L_H(m) \stackrel{\mathrm{def}}{=} \sum_{j=1}^{m} \log(3h_m/h_j).$$

Remark B.4. *For the ease of presentation, we supposed in the lemma that for each* $k \ge 1$*, there exists some* m_k *with* $h_{m_k} = 2^{k+1/2}$*. The results easily extend to the case when this equality is approximate.*

Proof of Lemma B.9. Without loss of generality assume $\upsilon^\circ = 0$. A basis transform reduces the study to the case when H is diagonal: $H = \mathrm{diag}\{h_1, h_2, \ldots\}$. We only have to evaluate the covering numbers $\mathbb{N}_k$. Let us fix $k \ge 1$ and let m_k be given by (B.21). For any point $\upsilon = (\upsilon_1, \upsilon_2, \ldots)^\top$ in $\mathcal{E}$, it holds

$$\begin{aligned}\sum_{j=m_k+1}^{\infty} \upsilon_j^2 &= \sum_{j=m_k+1}^{\infty} h_j^{-2} h_j^2 \upsilon_j^2 \le h_{m_k+1}^{-2} \sum_{j=m_k+1}^{\infty} h_j^2 \upsilon_j^2 \\ &\le h_{m_k+1}^{-2} \sum_{j=1}^{\infty} h_j^2 \upsilon_j^2 \le 2^{-2k-1}\mathtt{r}_\circ^2 \le \mathtt{r}_k^2/2. \end{aligned} \tag{B.22}$$

Consider the elliptic set $\mathcal{E}_k$ in $\mathbb{R}^{m_k}$ obtained by projection Π_k of $\mathcal{E}$ on the first m_k coordinates:

$$\mathcal{E}_k \stackrel{\mathrm{def}}{=} \left\{ (\upsilon_1, \ldots, \upsilon_{m_k})^\top : \sum_{j=1}^{m_k} h_j^2 \upsilon_j^2 \le \mathtt{r}_\circ^2 \right\}.$$

Let $\mathcal{M}_k$ be a ϵ_k-net in $\mathcal{E}_k$ for $\epsilon_k^2 = \mathtt{r}_k^2/2$. A $\mathtt{r}_k$-net in $\mathcal{E}$ can be constructed from $\mathcal{M}_k$ in a simple way: just fix to zero the remaining coordinates $\upsilon_j = 0$ for $j > m_k$. If υ° is constructed in this way, then $\|H\upsilon^\circ\| = \|H\Pi_k\upsilon^\circ\| \le 1$, that is, $\upsilon^\circ \in \mathcal{E}$. Moreover, for any other point $\upsilon \in \mathcal{E}$, take υ° such that their projections satisfy $\|\Pi_k(\upsilon - \upsilon^\circ)\| \le \epsilon_k$. Then

by (B.22)

$$\bigl\|\upsilon-\upsilon^\circ\bigr\|^2=\bigl\|\Pi_k\bigl(\upsilon-\upsilon^\circ\bigr)\bigr\|^2+\bigl\|(\mathbb{I}-\Pi_k)\upsilon\bigr\|^2\le \mathtt{r}_k^2/2+\mathtt{r}_k^2/2=\mathtt{r}_k^2.$$

Therefore, the covering number $\mathbb{N}(\mathcal{E},\mathtt{r}_k)$ of the infinite dimensional elliptic set $\mathcal{E}$ does not exceed the covering number $\mathbb{N}(\mathcal{E}_k,\epsilon_k)$ for the m_k-dimensional ellipsoid $\mathcal{E}_k$. By Lemma B.5 with $\delta=\epsilon_k$,

$$\mathbb{N}(\mathcal{E}_k,\epsilon_k)\le\frac{\mathrm{vol}(\mathcal{E}_k+(\epsilon_k/2)B_k)}{\mathrm{vol}(B_k)}(2/\epsilon_k)^{m_k},$$

where B_k is the unit ball in $\mathbb{R}^{m_k}$. The bound $h_j^{-2}\ge 2^{-2k-1}$ for $j\le m_k$ implies that $\mathcal{E}_k+(\epsilon_k/2)B_k$ is contained in the elliptic set $(3/2)\mathcal{E}_k$.

The definition implies due to $h_{m_k}^2=2^{2k+1}$

$$\mathbb{N}(\mathcal{E},\mathtt{r}_k)\le\log\frac{\mathrm{vol}((3/2)\mathcal{E}_k)}{(\epsilon_k/2)^{m_k}\,\mathrm{vol}(B_k)}\le\sum_{j=1}^{m_k}\log\frac{3h_j^{-1}}{\epsilon_k}\le\sum_{j=1}^{m_k}\log\biggl(\frac{3h_{m_k}}{h_j}\biggr)=L_H(m_k). \tag{B.23}$$

Now the result (B.20) follows by the definition of $\mathbb{Q}_1(\mathcal{E})$. □

Denote $\mathbb{N}_k=\mathbb{N}(\mathcal{E},\mathtt{r}_k)$. By the Cauchy–Schwarz inequality for $\alpha>1$

$$\mathbb{Q}_1(\mathcal{E})=\sum_{k=1}^{\infty}2^{-k}\sqrt{2\log(2\mathbb{N}_k)}\le\biggl\{\sum_{k=1}^{\infty}k^{-\alpha}\sum_{k=1}^{\infty}k^{\alpha}2^{-2k}2\log(2\mathbb{N}_k)\biggr\}^{1/2}. \tag{B.24}$$

The use of $h_{m_\ell}^{-2}=2^{2\ell+1}$ and $h_j^2\ge 4h_{m_{\ell-1}}^2$ for $j\in(m_{\ell-1},m_\ell]$ yields by (B.23) with $n_\ell\stackrel{\mathrm{def}}{=}m_\ell-m_{\ell-1}$

$$2\log(2\mathbb{N}_k)=\sum_{\ell=1}^{k}\sum_{j=m_{\ell-1}+1}^{m_\ell}\log\frac{9h_{m_k}^2}{h_j^2}\le\sum_{\ell=1}^{k}\bigl\{k-\ell+\log(36)\bigr\}n_\ell.$$

Further, in view of $h_{m_k}=2^k$

$$\begin{aligned}\sum_{k=1}^{\infty}k^{\alpha}2^{-2k}2\mathbb{N}_k&\le\sum_{k=1}^{\infty}k^{\alpha}2^{-2k}\sum_{\ell=1}^{k}\bigl\{k-\ell+\log(36)\bigr\}n_\ell=\sum_{\ell=1}^{\infty}\sum_{k\ge\ell}k^{\alpha}2^{-2k}\bigl\{k-\ell+\log(36)\bigr\}n_\ell\\&=\sum_{\ell=1}^{\infty}n_\ell 2^{-2\ell}\sum_{k\ge\ell}k^{\alpha}2^{-2(k-\ell)}\bigl\{k-\ell+\log(36)\bigr\}=\mathtt{C}\sum_{\ell=1}^{\infty}n_\ell 2^{-2\ell}\ell^{\alpha}.\end{aligned}$$

It remains to note that $2^{2\ell-1}\le h_j^2\le 2^{2\ell+1}$ for $m_{\ell-1}<j\le m_\ell$ and

$$\sum_{\ell=1}^{\infty}n_\ell 2^{-2\ell}\ell^{\alpha}\le\sum_{\ell=1}^{\infty}\sum_{j=m_{\ell-1}+1}^{m_\ell}h_j^{-2}\log^{\alpha}\bigl(h_j^2\bigr)=\sum_{j=1}^{\infty}h_j^{-2}\log^{\alpha}\bigl(h_j^2\bigr)=\mathtt{p}_H(\alpha). \tag{B.25}$$

The assertion (B.19) now follows from (B.24) in view of $\sum_{k\ge1}k^{-\alpha}\le\mathtt{C}(\alpha-1)^{-1}$.

The result on $\mathbb{Q}_2(\mathcal{E})$ requires to bound the sum of $2^{-k}\log\mathbb{N}_k$. Similarly to the above, one easily derives

$$\begin{aligned}\sum_{k=1}^{\infty}2^{-k}\mathbb{N}_k&\le\sum_{k=1}^{\infty}2^{-k}\sum_{\ell=1}^{k}\bigl\{k-\ell+\log(36)\bigr\}n_\ell=\sum_{\ell=1}^{\infty}\sum_{k\ge\ell}2^{-k}\bigl\{k-\ell+\log(36)\bigr\}n_\ell\\&=\sum_{\ell=1}^{\infty}n_\ell 2^{-\ell}\sum_{k\ge\ell}2^{-(k-\ell)}\bigl\{k-\ell+\log(36)\bigr\}=\mathtt{C}\sum_{\ell=1}^{\infty}n_\ell 2^{-\ell}\le\mathtt{C}\sum_{j=1}^{\infty}h_j^{-1}=\mathtt{C}\mathtt{p}_H^*.\end{aligned}$$

Theorem is proved. □

Now we present a special case for which the entropy can be bounded via the effective dimension $\mathtt{p}_H$ of Υ° defined in (B.17).

Theorem B.10. *Let $h_j^2 = f(j)$ for a monotonously increasing smooth function $f(x) > 0$. If $xf'(x)/f(x) \le \beta$, then*

$$\begin{aligned} &\mathbb{Q}_2(\mathcal{E}) \le \mathtt{C}\beta\mathtt{p}_H, \\ &\mathbb{Q}_1(\mathcal{E}) \le \mathtt{C}\sqrt{\beta\mathtt{p}_H}, \end{aligned} \tag{B.26}$$

where the effective dimension $\mathtt{p}_H$ is defined in (B.17).

Proof. Obviously

$$\sum_{j=1}^{m} \log\Big(\frac{h_m^2}{h_j^2}\Big) \le \int_0^m \log\Big(\frac{f(m)}{f(t)}\Big)\, dt.$$

Now we note that the function

$$F(x) \stackrel{\text{def}}{=} \int_0^x \log\Big(\frac{f(x)}{f(t)}\Big)\, dt$$

fulfills $F(0) = 0$ and $F'(x) = xf'(x)/f(x)$ yielding

$$\sum_{j=1}^{m} \log\Big(\frac{h_m^2}{h_j^2}\Big) \le \int_0^m \log\Big(\frac{f(m)}{f(t)}\Big)\, dt = \int_0^m \frac{xf'(x)}{f(x)}\, dx.$$

Moreover, in particular, if $F'(x) \le \beta$, then $F(x) \le \beta x$ and thus, $L_H(m_k) \le \beta m_k$. Now it holds similarly to (B.25)

$$\sum_{k=1}^{\infty} 2^{-k} m_k = \sum_{k=1}^{\infty} 2^{-k} \sum_{\ell=1}^{k} n_\ell \le \sum_{\ell=1}^{\infty} n_\ell \sum_{k \ge \ell}^{\infty} 2^{-k} = \sum_{\ell=1}^{\infty} n_\ell 2^{-\ell} \le 2 \sum_{j=1}^{\infty} h_j^{-2} = 2\mathtt{p}_H,$$

and the statement (B.19) follows. □

Now we evaluate the entropy for the cases when h_j grow polynomially.

Theorem B.11. *Let $h_j^2 = 1 + \varkappa^2 j^{2\beta}$ for $\beta > 1/2$ and some small value $\varkappa$. Then*

$$\begin{aligned} &\mathbb{Q}_1(\mathcal{E}) \le \mathtt{C}(2\beta - 1)^{-1/2} \varkappa^{-1/(2\beta)}, \\ &\mathbb{Q}_2(\mathcal{E}) \le \mathtt{C}(2\beta - 1)^{-1} \varkappa^{-1/\beta}, \end{aligned}$$

where $\mathtt{C}$ is an absolute constant.

Proof. For $f(x) = 1 + \varkappa^2 x^{2\beta}$, it holds $xf'(x)/f(x) \le 2\beta$ and we can apply the result of Theorem B.10. With $\beta > 1/2$, the effective dimension $\mathtt{p}_H$ from (B.17) fulfills

$$\mathtt{p}_H \le \sum_{j=1}^{\infty} h_j^{-2} = \sum_{j=1}^{\infty} \frac{1}{1 + \varkappa^2 j^{2\beta}} \le \int_0^{\infty} \frac{1}{1 + \varkappa^2 x^{2\beta}}\, dx = \mathtt{C}\varkappa^{-1/\beta} \frac{1}{2\beta - 1}$$

and the result follows by (B.26). □

B.5. Roughness constraints for dimension reduction

The local bounds of Theorems B.1 and B.3 can be extended in several directions. Here we briefly discuss one extension related to the use of a smoothness condition on the parameter υ. Let $\operatorname{pen}(\upsilon)$ be a non-negative *penalty* function on Υ. A particular example of such penalty function is the *roughness penalty* $\operatorname{pen}(\upsilon) = \|G\upsilon\|^2$ for a given p-matrix G^2. Let $\mathtt{r}$ be fixed. Consider the intersection of the ball $\mathcal{B}_{\mathtt{r}}(\upsilon^\circ)$ with the set Υ given by the constraint $\operatorname{pen}(\upsilon) \le 1$:

$$\Upsilon_{\mathrm{pen}}(\mathtt{r}) = \bigl\{\upsilon \in \Upsilon : d\bigl(\upsilon, \upsilon^\circ\bigr) \le \mathtt{r};\, \operatorname{pen}(\upsilon) \le 1\bigr\},$$

for a fixed central point υ° and the radius $\mathtt{r}$. Here and below we assume that the central point υ° is "smooth" in the sense that $\operatorname{pen}(\upsilon^\circ) < 1$. One can easily check that the results of Theorems B.1 and B.3 and their corollaries extend to this situation without any change. The only difference is in the definition of the values $\mathbb{Q}_1(\Upsilon_\circ)$ and $\mathbb{Q}_2(\Upsilon_\circ)$ for $\Upsilon_\circ = \Upsilon_{\mathrm{pen}}(\mathtt{r})$. Examples below show that the use of the penalization can substantially reduce these values relative to the non-penalized case.

We consider the case of a smooth process $\mathcal{U}$ given on a local set $\Upsilon_G(\mathtt{r})$ of the form

$$\Upsilon_G(\mathtt{r}) = \bigl\{\upsilon \in \Upsilon : \bigl\|\mathbb{V}\bigl(\upsilon - \upsilon^\circ\bigr)\bigr\| \le \mathtt{r};\, \|G\upsilon\| \le 1\bigr\}, \tag{B.27}$$

with the distance $d(\upsilon, \upsilon^\circ) = \|\mathbb{V}(\upsilon - \upsilon^\circ)\|$ and a smoothness constraint $\|G\upsilon\|^2 \le 1$. Then the set $\Upsilon_G(\mathtt{r})$ is contained in an elliptic set

$$\Upsilon_\circ \stackrel{\mathrm{def}}{=} \bigl\{\boldsymbol{\theta} : \|G\upsilon\|^2 + \bigl\|\mathbb{V}\bigl(\upsilon - \upsilon^\circ\bigr)\bigr\|^2 \le 1 + \mathtt{r}^2\bigr\}. \tag{B.28}$$

Define

$$\mathbb{V}_G^2 = \mathbb{V}^2 + G^2, \qquad \upsilon_G = \mathbb{V}_G^{-2}\mathbb{V}^2\upsilon^\circ.$$

Then $\upsilon^\circ - \upsilon_G = (\mathbb{I}_p - \mathbb{V}_G^{-2}\mathbb{V}^2)\upsilon^\circ = \mathbb{V}_G^{-2}G^2\upsilon^\circ$ and one can get by simple algebra

$$\begin{aligned}\|G\upsilon\|^2 + \bigl\|\mathbb{V}\bigl(\upsilon - \upsilon^\circ\bigr)\bigr\|^2 &= \bigl\|\mathbb{V}_G(\upsilon - \upsilon_G)\bigr\|^2 + \|G\upsilon_G\|^2 + \bigl\|\mathbb{V}\bigl(\upsilon_G - \upsilon^\circ\bigr)\bigr\|^2 \\ &= \bigl\|\mathbb{V}_G(\upsilon - \upsilon_G)\bigr\|^2 + \upsilon^{\circ\top}G^2\mathbb{V}_G^{-2}\mathbb{V}^2\upsilon^\circ = \bigl\|\mathbb{V}_G\bigl(\upsilon - \upsilon^\circ\bigr)\bigr\|^2 + d_G\end{aligned}$$

with $d_G = \upsilon^{\circ\top}G^2\mathbb{V}_G^{-2}\mathbb{V}^2\upsilon^\circ \le \|G\upsilon^\circ\|^2 < 1$. A change of variables $\upsilon \to \mathbb{V}(\upsilon - \upsilon_G)$ allows us to reduce the study to the case of an ellipsoid considered in Section B.4. For H defined by $H^{-2} = \mathbb{V}\mathbb{V}_G^{-2}\mathbb{V}$, the set $\Upsilon_\circ$ from (B.28) is transferred into the elliptic set

$$\Upsilon_H(\mathtt{r}) = \bigl\{\upsilon : \|H\upsilon\|^2 \le 1 + \mathtt{r}^2 - d_G\bigr\},$$

whose entropy for the Euclidean distance is given via the trace $\mathtt{p}_H = \operatorname{tr}(H^{-2})$.

Now we are prepared to state the penalized bound for the process $\mathcal{U}(\cdot)$ over $\Upsilon_\circ$ which naturally generalizes the result of Theorem B.7 to the non-penalized case.

Theorem B.12. *Let $\Upsilon_\circ = \Upsilon_{\mathrm{pen}}(\mathtt{r})$ be given by (B.27) and $\|G\upsilon^\circ\| \le 1$. Let also ($\mathcal{E}D$) hold with some $\mathtt{g}$ and a matrix $\mathbb{V}(\upsilon) \preceq \mathbb{V}$ for all $\upsilon \in \Upsilon_\circ$. For H defined by $H^{-2} = \mathbb{V}\mathbb{V}_G^{-2}\mathbb{V}$, let the entropy values $\mathbb{Q}_1(\Upsilon^\circ)$ and $\mathbb{Q}_2(\Upsilon^\circ)$ for the elliptic set $\Upsilon_H(\mathtt{r})$ from (B.28) be given in Section B.4. Then for any $\mathtt{x} \ge 1/2$*

$$\mathbb{P}\biggl\{\frac{1}{\nu_0\mathtt{r}}\sup_{\upsilon\in\Upsilon_{\mathrm{pen}}(\mathtt{r})}\bigl|\mathcal{U}\bigl(\upsilon, \upsilon^\circ\bigr)\bigr| \ge \mathfrak{z}_{\mathbb{H}}(\mathtt{x})\biggr\} \le \mathrm{e}^{-\mathtt{x}},$$

where $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ is from (B.3) with these values $\mathbb{Q}_1(\Upsilon^\circ)$ and $\mathbb{Q}_2(\Upsilon^\circ)$.

B.6. Bound for a bivariate process

Consider a smooth bivariate process $\mathcal{U}(\upsilon) = \mathcal{U}(\upsilon_1, \upsilon_2)$ over a product set $\Upsilon = \Upsilon_1 \times \Upsilon_2$, where $\Upsilon_j \subseteq \mathbb{R}^{p_j}$ for $j = 1, 2$. We suppose that partial derivatives of $\mathcal{U}$ have uniform exponential moments.

$(\mathcal{E}D_p)$ There exist $\mathtt{g} > 0$, $\nu_0 \geq 1$, and for each $\upsilon = (\upsilon_1, \upsilon_2) \in \Upsilon = \Upsilon_1 \times \Upsilon_2$, symmetric non-negative $p_j \times p_j$ matrices $\mathbb{V}_j$, $j = 1, 2$, such that for any $\lambda \leq \mathtt{g}$ and any unit vector $\boldsymbol{\gamma} \in \mathbb{R}^p$, it holds

$$\log \mathbb{E} \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla_j \mathcal{U}(\upsilon)}{\| \mathbb{V}_j \gamma \|} \right\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad j = 1, 2.$$

Here $\nabla_j \mathcal{U}$ denotes the partial derivative $\partial \mathcal{U} / \partial \upsilon_j$ for $j = 1, 2$.

This allows to establish an exponential bound for the process $\mathcal{U}(\upsilon)$. Let us fix the central point $\upsilon^\circ = (\upsilon_1^\circ, \upsilon_2^\circ)$ and a radius $\mathtt{r}$. As usual,

$$\Upsilon_j(\mathtt{r}) = \bigl\{ \upsilon_j \in \Upsilon_j \colon \bigl\| \mathbb{V}_j \bigl(\upsilon_j - \upsilon_j^\circ \bigr) \bigr\| \leq \mathtt{r} \bigr\}$$

denotes the ball in Υ_j with this radius.

Theorem B.13. *Let a bivariate random process $\mathcal{U}(\upsilon)$ on $\Upsilon = \Upsilon_1 \times \Upsilon_2$ satisfy $(\mathcal{E}D_p)$. Then for any $\mathtt{r}_\circ$ and $\mathtt{x} \geq 1/2$, it holds on the product set $\Upsilon_\circ = \Upsilon_1(\mathtt{r}_\circ) \times \Upsilon_2(\mathtt{r}_\circ)$*

$$\mathbb{P}\left\{ \frac{1}{\sqrt{8} \nu_0 \mathtt{r}_\circ} \sup_{\upsilon \in \Upsilon_\circ} \bigl| \mathcal{U}\bigl(\upsilon, \upsilon^\circ \bigr) \bigr| \geq \mathfrak{z}_{\mathbb{H}}(\mathtt{x}) \right\} \leq \mathrm{e}^{-\mathtt{x}},$$

with $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ from (B.3) for $\mathbb{Q}_1(\Upsilon^\circ) = \mathbb{Q}_1(\Upsilon_1) + \mathbb{Q}_1(\Upsilon_2)$ and $\mathbb{Q}_2(\Upsilon^\circ) = \mathbb{Q}_2(\Upsilon_1) + \mathbb{Q}_2(\Upsilon_2)$.

Proof. By the Hölder inequality, (B.31), and (B.30), it holds for $\| \boldsymbol{\gamma}_1 \| = \| \boldsymbol{\gamma}_2 \| = 1$ and $\upsilon \in \Upsilon_\circ$

$$\begin{aligned}
&\log \mathbb{E} \exp\left\{ \frac{\lambda}{2} (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)^\top \nabla \mathcal{U}(\upsilon) \right\} \\
&\quad \leq \frac{1}{2} \log \mathbb{E} \exp\bigl\{ \lambda \boldsymbol{\gamma}_1^\top \nabla_1 \mathcal{U}(\upsilon) \bigr\} + \frac{1}{2} \log \mathbb{E} \exp\bigl\{ \lambda \boldsymbol{\gamma}_2^\top \nabla_2 \mathcal{U}(\upsilon) \bigr\} \\
&\quad \leq \frac{1}{2} \log \mathbb{E} \exp\bigl\{ \lambda \boldsymbol{\gamma}_1^\top \nabla_1 \mathcal{U}(\upsilon) \bigr\} + \frac{1}{2} \log \mathbb{E} \exp\bigl\{ \lambda \boldsymbol{\gamma}_2^\top \nabla_2 \mathcal{U}(\upsilon) \bigr\} \leq \frac{\nu_0^2 \lambda^2}{2}, \quad |\lambda| \leq \mathtt{g}.
\end{aligned}$$

This means that the bivariate process $\mathcal{U}(\upsilon)/2$ fulfills the full dimensional condition $(\mathcal{E}D)$ with $\mathbb{V} = \mathrm{block}(\mathbb{V}_1, \mathbb{V}_2)$. Let $\upsilon = (\upsilon_1, \upsilon_2)$ and $\upsilon^\circ = (\upsilon_1^\circ, \upsilon_2^\circ)$ be a couple of points in Υ such that $\| \mathbb{V}_j (\upsilon_j - \upsilon_j^\circ) \| \leq \varepsilon$ for $j = 1, 2$. Then obviously

$$\bigl\| \mathbb{V} \bigl(\upsilon - \upsilon^\circ \bigr) \bigr\|^2 \leq 2 \varepsilon^2. \tag{B.29}$$

Therefore, the direct product of two ε-nets $\mathcal{M}_j(\varepsilon)$ in Υ_j for $j = 1, 2$ yield a $\sqrt{2}\varepsilon$-net $\mathcal{M}(\varepsilon) = \mathcal{M}_1(\varepsilon) \times \mathcal{M}_2(\varepsilon)$ in the product space Υ.

Due to (B.29), the product set $\Upsilon_\circ \stackrel{\mathrm{def}}{=} \Upsilon_1(\mathtt{r}_\circ) \times \Upsilon_2(\mathtt{r}_\circ)$ has the radius $\mathtt{r}_\circ$. Now we can easily bound the entropy of the product set $\Upsilon_\circ$ via the entropy of Υ_1 and Υ_2. Indeed, with $\mathtt{r}_k = 2^{-k} \mathtt{r}_\circ$, the cardinality $\mathbb{N}_k$ of $\mathcal{M}_k = \mathcal{M}(\mathtt{r}_k)$ fulfills $\mathbb{N}_k = \mathbb{N}_k(\Upsilon_1) \mathbb{N}_k(\Upsilon_2)$ and

$$\begin{aligned}
\mathbb{Q}_2(\Upsilon_\circ) &\leq \sum_{k=1}^{\infty} 2^{-k+1} \log(2 \mathbb{N}_k) \\
&\leq \sum_{k=1}^{\infty} 2^{-k+1} \log\bigl(2 \mathbb{N}_k(\Upsilon_1) \bigr) + \sum_{k=1}^{\infty} 2^{-k+1} \log\bigl(2 \mathbb{N}_k(\Upsilon_2) \bigr) \leq \mathbb{Q}_2(\Upsilon_1) + \mathbb{Q}_2(\Upsilon_2).
\end{aligned}$$

Similarly

$$\mathbb{Q}_1(\Upsilon_\circ) \le \sum_{k=1}^{\infty} 2^{-k}\sqrt{2\log(2\mathbb{N}_k)}$$
$$\le \sum_{k=1}^{\infty} 2^{-k}\sqrt{2\log\bigl(2\mathbb{N}_k(\Upsilon_1)\bigr) + 2\log\bigl(2\mathbb{N}_k(\Upsilon_2)\bigr)} \le \mathbb{Q}_1(\Upsilon_1) + \mathbb{Q}_1(\Upsilon_2).$$

Now we just apply the assertion of Theorem B.7 to the process $\mathcal{U}(\upsilon)/2$ and account for the fact that by (B.29) the radius of $\Upsilon_\circ$ is $\sqrt{2}\mathtt{r}_\circ$. □

B.7. A bound for the norm of a vector random process

Let $\mathcal{Y}(\upsilon)$, $\upsilon \in \Upsilon$, be a smooth centered random vector process with values in $\mathbb{R}^q$, where $\Upsilon \subseteq \mathbb{R}^p$. Let also $\mathcal{Y}(\upsilon^*) = 0$ for a fixed point $\upsilon^* \in \Upsilon$. Without loss of generality assume $\upsilon^* = 0$. We aim to bound the maximum of the norm $\|\mathcal{Y}(\upsilon)\|$ over a vicinity $\Upsilon_\circ$ of υ^*. By $\nabla\mathcal{U}(\upsilon)$ we denote the $p \times q$ matrix with entries $\nabla_{\upsilon_i}\mathcal{U}_j$, $i \le p$, $j \le q$. Suppose that $\mathcal{Y}(\upsilon)$ satisfies for each $\boldsymbol{\gamma}_1 \in \mathbb{R}^p$ and $\boldsymbol{\gamma}_2 \in \mathbb{R}^q$ with $\|\boldsymbol{\gamma}_1\| = \|\boldsymbol{\gamma}_2\| = 1$

$$\sup_{\upsilon\in\Upsilon} \log \mathbb{E}\exp\bigl\{\lambda\boldsymbol{\gamma}_1^\top\nabla\mathcal{Y}(\upsilon)\boldsymbol{\gamma}_2\bigr\} \le \frac{\nu_0^2\lambda^2}{2}, \quad |\lambda| \le \mathtt{g}. \tag{B.30}$$

Condition (B.30) implies for any $\upsilon \in \Upsilon_\circ$ with $\|\upsilon\| \le \mathtt{r}$ and $\boldsymbol{\gamma} \in \mathbb{R}^q$ with $\|\boldsymbol{\gamma}\| = 1$ in view of $\mathcal{Y}(\upsilon^*) = 0$ by Lemma B.4

$$\log \mathbb{E}\exp\left\{\frac{\lambda}{\mathtt{r}}\mathcal{Y}(\upsilon)^\top\boldsymbol{\gamma}\right\} \le \frac{\nu_0^2\lambda^2\|\upsilon\|^2}{2\mathtt{r}^2}, \quad |\lambda| \le \mathtt{g}. \tag{B.31}$$

In what follows, we use the representation

$$\bigl\|\mathcal{Y}(\upsilon)\bigr\| = \sup_{\|\mathbf{u}\|\le\mathtt{r}} \frac{1}{\mathtt{r}}\mathbf{u}^\top\mathcal{Y}(\upsilon). \tag{B.32}$$

This implies for $\Upsilon_\circ(\mathtt{r}) = \{\upsilon \in \Upsilon : \|\upsilon - \upsilon^*\| \le \mathtt{r}\}$

$$\sup_{\upsilon\in\Upsilon_\circ(\mathtt{r})} \bigl\|\mathcal{Y}(\upsilon)\bigr\| = \sup_{\upsilon\in\Upsilon_\circ(\mathtt{r})} \sup_{\|\mathbf{u}\|\le\mathtt{r}} \frac{1}{\mathtt{r}}\mathbf{u}^\top\mathcal{Y}(\upsilon).$$

Consider a bivariate process $\mathbf{u}^\top\mathcal{Y}(\upsilon)$ of $\mathbf{u} \in \mathbb{R}^q$ and $\upsilon \in \Upsilon \subset \mathbb{R}^p$. By definition $\mathbb{E}\mathbf{u}^\top\mathcal{Y}(\upsilon) = 0$. Further, $\nabla_{\mathbf{u}}[\mathbf{u}^\top\mathcal{Y}(\upsilon)] = \mathcal{Y}(\upsilon)$ while $\nabla_\upsilon[\mathbf{u}^\top\mathcal{Y}(\upsilon)] = \mathbf{u}^\top\nabla\mathcal{Y}(\upsilon) = \|\mathbf{u}\|\boldsymbol{\gamma}^\top\nabla\mathcal{Y}(\upsilon)$ for $\boldsymbol{\gamma} = \mathbf{u}/\|\mathbf{u}\|$. Suppose that $\mathbf{u} \in \mathbb{R}^q$ and $\upsilon \in \Upsilon$ are such that $\|\mathbf{u}\| \le \mathtt{r}$ and $\|\upsilon\| \le \mathtt{r}$. By (B.30), it holds for $\boldsymbol{\gamma} \in \mathbb{R}^p$ with $\|\boldsymbol{\gamma}\| = 1$ and $\upsilon \in \Upsilon_\circ(\mathtt{r})$

$$\log \mathbb{E}\exp\left\{\frac{\lambda}{\mathtt{r}}\nabla_\upsilon\bigl[\mathbf{u}^\top\mathcal{Y}(\upsilon)\bigr]\boldsymbol{\gamma}\right\} \le \log \mathbb{E}\exp\left\{\frac{\lambda}{\mathtt{r}}\mathbf{u}^\top\nabla\mathcal{Y}(\upsilon)\boldsymbol{\gamma}\right\} \le \frac{\nu_0^2\lambda^2}{2},$$

and by (B.31) for a unit vector $\boldsymbol{\gamma} \in \mathbb{R}^q$

$$\log \mathbb{E}\exp\left\{\frac{\lambda}{\mathtt{r}}\nabla_{\mathbf{u}}\bigl[\mathbf{u}^\top\mathcal{Y}(\upsilon)\bigr]\boldsymbol{\gamma}\right\} \le \log \mathbb{E}\exp\left\{\frac{\lambda}{\mathtt{r}}\mathcal{Y}(\upsilon)\boldsymbol{\gamma}\right\} \le \frac{\nu_0^2\lambda^2}{2}.$$

Therefore, $(\mathcal{E}D_p)$ is fulfilled for $\mathbf{u}^\top\mathcal{Y}(\upsilon)$ and Theorem B.7 applies. We summarize our findings in the following theorem.

Theorem B.14. *Let a random p-vector process $\mathcal{Y}(\upsilon)$ for $\upsilon \in \Upsilon \subseteq \mathbb{R}^p$ fulfill $\mathcal{Y}(\upsilon^*) = 0$, $\mathbb{E}\mathcal{Y}(\upsilon) \equiv 0$, and the condition* (B.30) *be satisfied. Then for each $\mathtt{r}$ and any $\mathtt{x} \geq 1/2$, it holds for $\Upsilon_\circ = \Upsilon_\circ(\mathtt{r})$*

$$\mathbb{P}\Big\{ \sup_{\upsilon \in \Upsilon_\circ(\mathtt{r})} \big\| \mathcal{Y}(\upsilon) \big\| \geq \sqrt{8} \nu_0 \mathtt{r} \mathfrak{z}_{\mathbb{H}}(\mathtt{x}) \Big\} \leq \mathrm{e}^{-\mathtt{x}}, \tag{B.33}$$

where $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ is given by (B.3) *with $\mathbb{Q}_1 = \mathbb{Q}_1(\Upsilon_\circ) + \sqrt{6q}$ and $\mathbb{Q}_1 = \mathbb{Q}_1(\Upsilon_\circ) + 6q$.*

B.8. A bound for a family of quadratic forms

Now we consider an extension of the previous result with a quadratic form $\|A\mathcal{Y}(\upsilon)\|^2$ to be bounded under the conditions (B.30) and (B.31) on $\mathcal{Y}(\upsilon)$ for $\upsilon \in \Upsilon \subset \mathbb{R}^p$. Here $\mathcal{Y}(\cdot)$ is a vector process with values in $\mathbb{R}^q$ and A is a $q \times q$ matrix with $\|A^\top A\|_{\mathrm{op}} \leq 1$. The idea is to use the representation (B.32) in which we replace $\mathbf{u}$ with $A\mathbf{u}$. The bound (B.33) implies for any $\mathtt{r}$

$$\mathbb{P}\Big\{ \sup_{\upsilon \in \Upsilon_\circ(\mathtt{r}), \|A\mathbf{u}\| \leq \mathtt{r}} \mathbf{u}^\top A \mathcal{Y}(\upsilon) > \sqrt{8} \nu_0 \mathtt{r} \mathfrak{z}_{\mathbb{H}}(\mathtt{x}) \Big\} \leq \mathrm{e}^{-\mathtt{x}},$$

where $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ corresponds to $\mathbb{Q}_1 = \sqrt{\mathbb{Q}_2} = \sqrt{6p + \mathbb{Q}_2(\Upsilon_\circ)}$.

Now we discuss how this bound can be refined if A is a smoothing operator. For simplicity assume that A fulfills the condition of Theorem B.10. One can expect that the dimension q can be replaced by the effective dimension $\mathtt{p}_A$. The arguments similar to the above yield

$$\big\| A\mathcal{Y}(\upsilon) \big\| = \sup_{\mathbf{u} \in \mathbb{R}^q : \|\mathbf{u}\| \leq \mathtt{r}} \frac{1}{\mathtt{r}} \mathbf{u}^\top A \mathcal{Y}(\upsilon),$$

and we again consider a bivariate process $\mathbf{u}^\top A\mathcal{Y}(\upsilon)$ of $\mathbf{u} \in \mathbb{R}^q$ and $\upsilon \in \Upsilon \subset \mathbb{R}^p$. The conditions (B.30) and (B.31) imply for any two unit vectors $\boldsymbol{\gamma}_1 \in \mathbb{R}^q$ and $\boldsymbol{\gamma}_2 \in \mathbb{R}^p$ and any points $\mathbf{u} \in \mathbb{R}^q$ with $\|A\mathbf{u}\| \leq \mathtt{r}$ and $\upsilon \in \Upsilon_\circ(\mathtt{r})$, it holds

$$\log \mathbb{E} \exp\Big\{ \frac{\lambda}{\mathtt{r}} \nabla_\upsilon \big[\mathbf{u}^\top A \mathcal{Y}(\upsilon) \big] \boldsymbol{\gamma}_2 \Big\} = \log \mathbb{E} \exp\Big\{ \frac{\lambda}{\mathtt{r}} \mathbf{u}^\top A \nabla \mathcal{Y}(\upsilon) \boldsymbol{\gamma}_2 \Big\} \leq \frac{\nu_0^2 \lambda^2}{2},$$

and by (B.31) with $\mathbb{V}_1^2 = A^\top A$

$$\log \mathbb{E} \exp\Big\{ \frac{\lambda}{\|\mathbb{V}_1 \boldsymbol{\gamma}_1\|} \boldsymbol{\gamma}_1^\top \nabla_{\mathbf{u}} \big[\mathbf{u}^\top A \mathcal{Y}(\upsilon) \big] \Big\} \leq \log \mathbb{E} \exp\Big\{ \frac{\lambda}{\|\mathbb{V}_1 \boldsymbol{\gamma}_1\|} (A\boldsymbol{\gamma}_1)^\top \mathcal{Y}(\upsilon) \Big\} \leq \frac{\nu_0^2 \lambda^2}{2}.$$

Therefore, $(\mathcal{E}D_p)$ is fulfilled for $\mathbf{u}^\top A\mathcal{Y}(\upsilon)$. Now we apply the bound from Theorem B.13 and the entropy bound for the elliptic set $\|A\mathbf{u}\| \leq \mathtt{r}$ from Theorem B.10.

Theorem B.15. *Let a random vector process $\mathcal{Y}(\upsilon) \in \mathbb{R}^q$ for $\upsilon \in \Upsilon \subseteq \mathbb{R}^p$ fulfill $\mathcal{Y}(\upsilon^*) = 0$, $\mathbb{E}\mathcal{Y}(\upsilon) \equiv 0$, and the condition* (B.30) *be satisfied. Let A fulfill $1/2 \leq \|AA^\top\|_{\mathrm{op}} \leq 1$. Then for each $\mathtt{r}$, it holds*

$$\mathbb{P}\Big\{ \sup_{\upsilon \in \Upsilon_\circ(\mathtt{r})} \big\| A\mathcal{Y}(\upsilon) \big\| > \sqrt{8} \nu_0 \mathtt{r} \mathfrak{z}_{\mathbb{H}}(\mathtt{x}) \Big\} \leq \mathrm{e}^{-\mathtt{x}},$$

where $\mathfrak{z}_{\mathbb{H}}(\mathtt{x})$ is given by (B.3) *with $\mathbb{Q}_2 = \mathtt{C}\mathtt{p}_A + \mathbb{Q}_2(\Upsilon_\circ(\mathtt{r}))$ and $\mathbb{Q}_1 = \mathtt{C}\sqrt{\mathtt{p}_A} + \mathbb{Q}_1(\Upsilon_\circ(\mathtt{r}))$.*

References

[1] A. Andresen and V. Spokoiny. Critical dimension in profile semiparametric estimation. *Electron. J. Stat.* **8** (2) (2014) 3077–3125. MR3301302

[2] A. Barron, L. Birgé and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** (3) (1999) 301–413. MR1679028

[3] A. Belloni and V. Chernozhukov. On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.* **37** (4) (2009) 2011–2055. MR2533478

[4] L. Birgé and P. Massart. Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* **4** (3) (1998) 329–375. MR1653272

[5] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** (3) (2001) 203–268. MR1848946

[6] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** (1–2) (2007) 33–73. MR2288064

[7] S. Boucheron and P. Massart. A high-dimensional Wilks phenomenon. *Probab. Theory Related Fields* **150** (2011) 405–433. MR2824862

[8] J. Fan, C. Zhang and J. Zhang. Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29** (1) (2001) 153–193. MR1833962

[9] S. Ghosal. Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli* **5** (2) (1999) 315–331. MR1681701

[10] S. Ghosal. Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.* **74** (1) (2000) 49–68. MR1790613

[11] Y. Golubev and V. Spokoiny. Exponential bounds for minimum contrast estimators. *Electron. J. Stat.* **3** (2009) 712–746. MR2534199

[12] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London, 1994. MR1270012

[13] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1* 221–233. Univ. California Press, Berkeley, CA, 1967. MR0216620

[14] I. A. Ibragimov and R. Z. Khas'minskij. *Statistical Estimation. Asymptotic Theory*. Springer, New York, 1981. Transl. from the Russian by Samuel Kotz. MR0620321

[15] Y. Kim. The Bernstein–von Mises theorem for the proportional hazard model. *Ann. Statist.* **34** (4) (2006) 1678–1700. MR2283713

[16] R. Koenker, P. Ng and S. Portnoy. Quantile smoothing splines. *Biometrika* **81** (4) (1994) 673–680. MR1326417

[17] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** (5) (2000) 1302–1338.

[18] E. Mammen. Empirical process of residuals for high-dimensional linear models. *Ann. Statist.* **24** (1) (1996) 307–335. MR1389892

[19] S. Portnoy. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12** (4) (1984) 1298–1309. MR0760690

[20] S. Portnoy. Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** (4) (1985) 1403–1417. MR0811499

[21] S. Portnoy. Asymptotic behavior of the empiric distribution of M-estimated residuals from a regression model with many parameters. *Ann. Statist.* **14** (1986) 1152–1170. MR0856812

[22] S. Portnoy. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** (1) (1988) 356–366. MR0924876

[23] X. Shen. On methods of sieves and penalization. *Ann. Statist.* **25** (6) (1997) 2555–2591. MR1604416

[24] X. Shen and W. H. Wong. Convergence rate of sieve estimates. *Ann. Statist.* **22** (2) (1994) 580–615. MR1292531

[25] V. Spokoiny. Parametric estimation. Finite sample theory. *Ann. Statist.* **40** (6) (2012) 2877–2909. MR3097963

[26] V. Spokoiny, W. Wang and W. Härdle. Local quantile regression (with rejoinder). *J. Statist. Plann. Inference* **143** (7) (2013) 1109–1129. MR3049611

[27] V. Spokoiny and M. Zhilova. Bootstrap confidence sets under model misspecification. *Ann. Statist.* **43** (2015) 2653–2675. MR3405607

[28] S. van de Geer. M-estimation using penalties or sieves. *J. Statist. Plann. Inference* **108** (1–2) (2002) 55–69. MR1947391

[29] S. A. Van de Geer. *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000. MR1739079

[30] A. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes. With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York, 1996. MR1385671

[31] A. Zaitsev, E. Burnaev and V. Spokoiny. Properties of the posterior distribution of a regression model based on Gaussian random fields. *Autom. Remote Control* **74** (10) (2013) 1645–1655. MR3219856