Finite-sample performance of the maximum likelihood estimator in logistic regression

Hugo Chardon* Matthieu Lerasle* Jaouad Mourtada*

December 5, 2024

Abstract

Logistic regression is a classical model for describing the probabilistic dependence of binary responses to multivariate covariates. We consider the predictive performance of the maximum likelihood estimator (MLE) for logistic regression, assessed in terms of logistic risk. We consider two questions: first, that of the existence of the MLE (which occurs when the dataset is not linearly separated), and second that of its accuracy when it exists. These properties depend on both the dimension of covariates and on the signal strength. In the case of Gaussian covariates and a well-specified logistic model, we obtain sharp non-asymptotic guarantees for the existence and excess logistic risk of the MLE. We then generalize these results in two ways: first, to non-Gaussian covariates satisfying a certain two-dimensional margin condition, and second to the general case of statistical learning with a possibly misspecified logistic model. Finally, we consider the case of a Bernoulli design, where the behavior of the MLE is highly sensitive to the parameter direction.

Contents

*Department of Statistics, CREST/ENSAE Paris, Palaiseau, France

1 Introduction

Logistic regression [Ber44, MN89] is a classical model describing the dependence of binary outcomes on multivariate features. In this work, we investigate the predictive performance of the most standard method for fitting this model, namely the maximum likelihood estimator (MLE). Our emphasis is placed on the dependence of the estimation error on the various parameters of the problem, as well as on the conditions under which the MLE performs well.

1.1 Problem setting and main questions

To set the stage for the discussion, we start by recalling the definition of the logistic (logit) model. Given a dimension $d \geqslant 1$, the *logistic model* is the family of conditional distributions on the outcome $y \in \{-1, 1\}$ given the covariates $x \in \mathbb{R}^d$ defined by:

$$\mathcal{P}_{\text{logit}} = \left\{ p_\theta : \theta \in \mathbb{R}^d \right\}, \quad \text{where} \quad p_\theta(y|x) = \sigma(y\langle \theta, x\rangle) \quad \text{for } (x, y) \in \mathbb{R}^d \times \{-1, 1\}, \tag{1}$$

where we let $\sigma(s) = e^s/(e^s + 1)$ for $s \in \mathbb{R}$ be the sigmoid function, and where $\langle \cdot, \cdot \rangle$ denotes the usual scalar product on $\mathbb{R}^d$. We say that a random pair (X, Y) on $\mathbb{R}^d \times \{-1, 1\}$ follows the logistic model if the conditional distribution of Y given X belongs to $\mathcal{P}_{\text{logit}}$.

In short, the logistic model is appealing because it constitutes a natural "linear" model for binary outcomes: indeed, the conditional probability $\mathbb{P}(Y = 1|X = x)$ is obtained by applying the link function $\sigma : \mathbb{R} \to [0, 1]$ to a linear function of x. Note that the specific choice of the link function σ in the logistic model is not arbitrary: it corresponds to the "canonical link function" for the Bernoulli parameter, in the sense of exponential families [Bro86].

In the statistical setting, the true distribution P of the random pair (X, Y) is unknown, but one has access to a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of i.i.d. random variables with distribution P. Using this sample, one can compute the MLE, defined by

$$\widehat{\theta}_n = \operatorname*{arg\,max}_{\theta \in \mathbb{R}^d} \prod_{i=1}^n p_\theta(Y_i|X_i) = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \log\left(1 + e^{-Y_i\langle \theta, X_i\rangle}\right). \tag{2}$$

In this work, we will be concerned with the following two questions:

1. *Existence*: When does the MLE exist?
2. *Performance*: When the MLE exists, how accurate is it?

To make these two questions precise, some discussion is in order.

First, we must clarify the geometric meaning of existence (and uniqueness) of the MLE; we refer to [AA84] and to the introduction of [CS20] for an interesting discussion of this point, with thorough references. Uniqueness of the MLE is in fact a straightforward question: whenever the points $X_1, \dots, X_n$ span $\mathbb{R}^d$ (a property that holds with high probability for $n \gtrsim d$, under suitable assumptions on X), the second function in (2) that the MLE minimizes is strictly convex on $\mathbb{R}^d$, and thus admits at most one minimizer. The property of existence of the MLE has a richer geometric content. Assume again to simplify that $X_1, \dots, X_n$ span $\mathbb{R}^d$, so that for every $\theta \neq 0$, there exists $i \in \{1, \dots, n\}$ such that $\langle \theta, X_i\rangle \neq 0$. Then, *the MLE exists if and only if the dataset is not linearly separated*, by which we mean that there is no $\theta \neq 0$ such that $\{X_i : 1 \leqslant i \leqslant n, Y_i = 1\} \subset \mathcal{H}_\theta^+ = \{x \in \mathbb{R}^d : \langle \theta, x\rangle \geqslant 0\}$ and $\{X_i : 1 \leqslant i \leqslant n, Y_i = -1\} \subset \mathcal{H}_\theta^- = \{x \in \mathbb{R}^d : \langle \theta, x\rangle \leqslant 0\}$—or, in more succinct form, if there is no $\theta \neq 0$ such that $Y_i\langle \theta, X_i\rangle \geqslant 0$ for every $i = 1, \dots, n$. Indeed, if such a θ exists, then the second function in (2) evaluated at $t\theta$ remains upper bounded as $t \to +\infty$; since a strictly convex function admitting a global minimizer diverges at infinity, the objective function admits no global minimizer. Conversely, if no such θ exists, then simple compactness arguments show that the function in (2) diverges at infinity and is continuous, hence admits a global minimizer.

Second, in order to assess the performance of the MLE, one must specify a notion of accuracy. In this work, we will mainly focus on the predictive performance of the MLE, as measured by its risk for prediction under logistic loss. Specifically, we consider the problem of assigning probabilities to the possible values ± 1 of Y, given the knowledge of the associated covariate vector X. Each parameter $\theta \in \mathbb{R}^d$ gives rise to the conditional distribution p_θ defined in (1). We can then define the logistic loss ℓ (at a point $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$) and risk L of θ by, respectively:

$$\ell(\theta, (x, y)) = -\log p_\theta(y|x) = \log(1 + e^{-y\langle \theta, x \rangle}) \quad \text{and} \quad L(\theta) = \mathbb{E}[\ell(\theta, (X, Y))] \,. \tag{3}$$

Hence, the logistic loss corresponds to the negative log-likelihood (or logarithmic loss) for the logistic model. The logarithmic loss is a classical way to assess the quality of probabilistic forecasts: it enforces calibrated predictions by penalizing both overconfident and under-confident probabilities. In particular, assigning a probability of 0 to a label y that does appear leads to an infinite loss. In addition, this criterion is closely related to the MLE, which corresponds to the minimizer of the *empirical risk* $\widehat{L}_n : \mathbb{R}^d \to \mathbb{R}$ under logistic loss, defined by

$$\widehat{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\theta, (X_i, Y_i)) = \frac{1}{n} \sum_{i=1}^n \log(1 + e^{-Y_i \langle \theta, X_i \rangle}) \,. \tag{4}$$

Finally, the logistic loss also naturally arises in statistical learning theory [BBL05, Kol11, Bac24], as a convex surrogate of the classification error [Zha04, BJM06]. With these definitions at hand, one can measure the prediction accuracy of the MLE by its excess risk under logistic loss, namely $L(\widehat{\theta}_n) - L(\theta^*)$, where $\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} L(\theta)$ (assuming this set is nonempty).

Thirdly, the existence of the MLE depends on the dataset and is thus a random event, and likewise the excess risk $L(\widehat{\theta}_n) - L(\theta^*)$ is a random quantity. As such, both the existence and the accuracy of the MLE depend on the joint distribution P of (X, Y). To give a precise meaning to the questions above, we must therefore specify which distributions P we consider. Note that the joint distribution P is characterized by (a) the marginal distribution P_X of X, and (b) the conditional distribution $P_{Y|X}$ of Y given X.

We will actually consider three different settings of increasing generality, depending on the respective assumptions on P_X and $P_{Y|X}$, but for concreteness and in order to compare with previous results, we will in this introduction start with the simplest one:

(a) The design follows a Gaussian distribution: $X \sim \mathsf{N}(0, \Sigma)$ for some positive matrix Σ. By invariance of the problem under invertible linear transformations of X, we may assume that $\Sigma = I_d$ is the identity matrix, which we will do in what follows.

(b) The model is *well-specified*, in that the conditional distribution $P_{Y|X}$ belongs to the logistic model $\mathcal{P}_{\text{logit}}$. In other words, there exists $\theta^* \in \mathbb{R}^d$ such that $\mathbb{P}(Y = 1|X) = \sigma(\langle \theta^*, X \rangle)$.

Besides its natural character, the appeal of this setting is that the problem only depends on a small number of parameters. These are: the sample size n, the data dimension d, the probability $1 - \delta$ with which the guarantees hold, and importantly the *signal strength* (or signal-to-noise ratio, or inverse temperature) $B = \max\{e, \|\theta^*\|\}$, where $\| \cdot \|$ stands for the Euclidean norm.

It is worth commenting on the role of the dimension d and of the signal strength B. Intuitively, there are two distinct effects that may lead the dataset to be linearly separated. First, the larger the dimension d, the more degrees of freedom there are to linearly separate the dataset. But another effect comes from the signal strength: the stronger the signal B, the more the labels Y_i tend to be of the same sign as $\langle \theta^*, X_i \rangle$—and thus, the more likely it is for the dataset to be separated by θ^*, or by a "close" direction. As we will see, the "dimensionality" and the "signal strength" effects interact with each other. We also note that, intuitively, a stronger signal should make the *classification* problem (of predicting the value of the label Y, and minimizing the fraction of errors) easier. This amounts to saying that the larger B is, the smaller the estimation error for the *direction* $u^* = \theta^*/\|\theta^*\|$ of the parameter θ^* should be. On the other hand, under a stronger signal, the MLE is known (see, e.g., [CS20, SC19] and references therein) to tend to underestimate the uncertainty in the labels, that is, to return overconfident conditional probabilities for Y given X. This holds, for instance, if the dataset is nearly linearly separated, in which case the MLE predicts conditional probabilities close to 0 or 1. Hence, for the *conditional density estimation* problem we consider, a stronger signal may degrade the performance of the MLE. This should manifest itself by the fact that the *norm* of the MLE (as opposed to its direction) may be far from that of θ^*, so the overall estimation error of θ^* may be larger.

To summarize, we are interested in explicit and *non-asymptotic* guarantees for the existence and accuracy of the MLE, in terms of the relevant parameters B, d, n, δ—ideally, in the general situation where these parameters may take arbitrary values. Our aim is twofold: first, to obtain the optimal dependence on all parameters in the case of a Gaussian design and a well-specified model; second, to investigate to which extent these results extend to more general distributions.

1.2 Existing results

Before describing our contributions, we first provide an overview of known results on the questions we consider. As a basic statistical method, logistic regression has been studied extensively in the literature, hence we focus on those results that are most directly relevant to our setting. Again, for the sake of comparison, we will mainly focus on the case of a Gaussian design and a well-specified model, although extensions will also be discussed.

1.2.1 Classical asymptotics

The behavior of the MLE is well-understood in the context of classical parametric asymptotics [LCY00, vdV98]. In this setting, the distribution P is fixed (and thus, so are the dimension d and signal strength B) while the sample size n goes to infinity. In this case, as $n \to \infty$, the MLE $\widehat{\theta}_n$ exists with probability converging to 1, converges to θ^* at a $1/\sqrt{n}$ rate, and is asymptotically normal, with asymptotic covariance given by the inverse of the Fisher information matrix [vdV98, §5.2–5.6]. This implies that the excess risk converges to 0 at a rate $1/n$, and more precisely that

$$2n\big\{L(\widehat{\theta}_n) - L(\theta^*)\big\} \xrightarrow{\text{(d)}} \chi^2(d)\,, \tag{5}$$

where $\xrightarrow{\text{(d)}}$ denotes convergence in distribution and $\chi^2(d)$ denotes the χ^2 distribution with d degrees of freedom. Together with a tail bound for the χ^2 distribution, this implies the following: for fixed $d \geqslant 1$, $\theta^* \in \mathbb{R}^d$, and $\delta \in (0,1)$, we have

$$\liminf_{n\to\infty} \mathbb{P}\Big(L(\widehat{\theta}_n) - L(\theta^*) \leqslant \frac{d + 2\log(1/\delta)}{n}\Big) \geqslant 1 - \delta\,, \tag{6}$$

with the convention that $L(\widehat{\theta}_n) - L(\theta^*) = +\infty$ if the dataset is linearly separated. Note that the convergence (5) holds only in the well-specified case, and that in the misspecified case the normalized excess risk $2n\{L(\widehat{\theta}_n) - L(\theta^*)\}$ converges to a different limiting distribution that depends on the distribution P of (X, Y); see [vdV98, Example 5.25 p. 55] and (for instance) the introductions of [OB21, MG22] for additional discussions on this point.

On the positive side, the high-probability guarantee (6) is sharp, in light of the convergence in distribution (5) of the excess risk. On the other hand, it should be noted that this guarantee is purely asymptotic: it holds as $n \to \infty$ while all other parameters of the problem are fixed. This does not allow one to handle the modern high-dimensional regime, where the dimension d may be large and possibly comparable to n. In addition, it does not state how large the sample size n should be (in terms of B, d, δ) for the asymptotic behavior (6) to occur—in particular, it provides no information on the sample size required for the existence of the MLE.

High-dimensional asymptotics. Several of the shortcomings of the classical asymptotic theory can be addressed by considering a different asymptotic framework, namely the "high-dimensional asymptotic regime", where $d, n \to \infty$ while d/n converges to a fixed constant. This framework has attracted significant interest in statistics over the last decade (see, e.g., [EK18b, Mon18] and references therein for a partial overview of this line of work). The interest of this framework is that it allows one to capture high-dimensional effects, since the dimension is no longer negligible compared to the sample size.

The question of existence of the MLE under high-dimensional asymptotics was addressed in the seminal work of Candès and Sur [CS20], extending a previous result of Cover [Cov65] in the "null" case where $\theta^* = 0$. Specifically, the main result of Candès and Sur [CS20, Theorems 1–2] can be stated as follows[1]: there exists a function $h : \mathbb{R}^+ \to (0, 1)$ such that the following holds. Fix $\beta \in \mathbb{R}^+$ and $\gamma \in (0, 1)$, and let $d = d_n \to \infty$ as $n \to \infty$, with $d/n \to \gamma$. If $X \sim \mathsf{N}(0, I_d)$ and $\mathbb{P}(Y = 1|X) = \sigma(\langle \theta^*, X\rangle)$, with $\theta^* = \theta^*_d \in \mathbb{R}^d$ such that $\|\theta^*\| = \beta$, and the dataset consists of n i.i.d. copies of (X, Y), then

$$\lim_{n\to\infty} \mathbb{P}(\text{MLE exists}) = \begin{cases} 0 & \text{if } \gamma > h(\beta) \\ 1 & \text{if } \gamma < h(\beta)\,. \end{cases} \tag{7}$$

In addition, the quantity $h(\beta)$ is defined as the infimum of the expectation of an explicit family random variables that depend on β (see eq. (2.4) in [CS20]), and the curve of the function h is plotted numerically in this paper.

The conditions (7) provide a precise characterization of the existence of the MLE under high-dimensional asymptotics, and in particular establish a sharp phase transition for this property, depending on the value of the aspect ratio $\gamma = \lim d/n$.

While this result conclusively answers the question of existence of the MLE in this asymptotic setting, it does not cover the general regime where the problem parameters may be of arbitrary order of magnitude relative to each other. Indeed, although this regime captures high-dimensional effects by allowing the dimension to grow with the sample size, it assumes the signal strength B to be fixed while $n \to \infty$. This excludes "strong signal" regimes, where the sample size may not be large enough relative to B for the asymptotic characterization (7) to provide an accurate approximation. As an example, a finite-sample condition of the form $n \gg \exp(B)$ would always be satisfied under high-dimensional asymptotics, and thus would not be visible from results framed in this setting. In addition, the characterization (7) is a qualitative zero-one law, stating that the considered probability converges to 0 or 1. However, one may wish for more precise information, namely sharp quantitative estimates on the probabilities.

Finally, the characterization (7) is specific to the case of a Gaussian design, and indeed one should expect the precise threshold for existence of the MLE to be sensitive to the distribution of the design (see [EK18a] for results in this spirit in the case of robust regression). One may therefore want to identify general conditions on the design distribution under which the MLE

[1] In fact, Theorem 1 in [CS20] deals with the case of logistic regression with an intercept, while Theorem 2 therein is concerned with logistic regression without intercept that we discuss here.

behaves in a similar way as for Gaussian design. Likewise, the characterization (7) holds in the well-specified case, which raises the question of existence of the MLE in the misspecified case.

These considerations motivate a finite-sample analysis that would allow one to handle general values of the problem parameters, and extend to more general situations. We would like however to clarify that the finite-sample results do not imply the asymptotic ones: indeed, the non-asymptotic characterizations we will obtain feature universal constant factors (and even in some cases logarithmic factors), while the asymptotic characterization (7) is precise down to the numerical constants. This loss in precision may be a price to pay for a non-asymptotic analysis in the general regime; on the positive side, it will allow us to obtain conditions that are easier to interpret. For these reasons, we view the finite-sample and asymptotic perspectives as complementary.

Non-asymptotic guarantees. We now discuss available non-asymptotic guarantees for the MLE in logistic regression from the literature, focusing on those that are most relevant to our setting. First, it follows from the results of [CLL20] (specifically, combining Theorems 1 and 8 therein) that there is a constant $c > 0$ such that the following holds: if $n \geqslant e^{cB} d$, then with probability $1 - 2e^{-d/c}$ the MLE $\widehat{\theta}_n$ exists and satisfies

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant \frac{e^{cB} d}{n} \, . \tag{8}$$

On the positive side, this result is fully explicit and features an optimal dependence on the sample size n and dimension d; the probability $1 - e^{-d/c}$ under which the $O_B(d/n)$ bound holds is also optimal, in light of the asymptotic results (5) and (6). On the other hand, the dependence on the signal strength B is exponential, which turns out to be highly suboptimal for a Gaussian design. In fact, the bound (8) holds in a more general setting, where the model may be misspecified and where the design is only assumed to be sub-Gaussian. As we will discuss below, some exponential dependence on the norm turns out to be unavoidable if one only assumes the design distribution to be sub-Gaussian.

Up until recently, the sharpest available non-asymptotic guarantees for the MLE in logistic regression with a Gaussian design were due to Ostrovskii and Bach [OB21]. Specifically, combining Theorem 4.2 in [OB21] with Proposition D.1 therein shows that there is a constant $c > 0$ such that the following holds: for $\delta \leqslant 1/2$, if $n \geqslant c \log^4(B) B^8 d \log(1/\delta)$, then with probability at least $1 - \delta$ the MLE exists and satisfies

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant \frac{B^3 d \log(1/\delta)}{n} \, . \tag{9}$$

Like the bound (8), this result features an optimal dependence on the dimension d and sample size n; and while the bound involves a deviation term $d \log(1/\delta)$ proportional to the dimension (which is suboptimal for small δ), it could be tightened to an additive deviation term $d + \log(1/\delta)$ with very minor changes to the proof of [OB21]. Importantly, this result significantly improves over the general bound (8) in the case of a Gaussian design, by replacing the exponential dependence on the norm B by a polynomial one. In addition, it is worth mentioning that the result of [OB21] holds in the general misspecified case, and that in this case it is actually the best available guarantee in the literature. This being said, as we will see below, the polynomial dependence on B in both the condition for existence of the MLE and in the risk bound can be improved. For instance, in the well-specified case, the risk bound (9) is larger than the asymptotic risk (6) by a factor of B^3, which suggests possible improvements.

Recently and while this manuscript was under preparation, two additional works [KvdG23, HM23] contributed significantly to the study of logistic regression with a Gaussian design, with an emphasis on the dependence on the signal strength B. Closest to our setting is the work of Kuchelmeister and van de Geer [KvdG23], who study the MLE for logistic regression under a Gaussian design, but assuming that the conditional distribution of Y given X follows a probit rather than a logit model. Despite real technical differences between the probit and logit models, this is qualitatively related to the well-specified logit model. With a natural notion of signal strength B in the probit model (the inverse of the parameter σ in their work), Theorem 2.1.1 in [KvdG23] states that: for some absolute constant c, if $n \geqslant cB(d \log n + \log(1/\delta))$, the MLE exists and satisfies

$$\left\| \frac{\widehat{\theta}_n}{\|\widehat{\theta}_n\|} - \frac{\theta^*}{\|\theta^*\|} \right\| \leqslant c\sqrt{\frac{d \log n + \log(1/\delta)}{Bn}}, \quad \left| \|\widehat{\theta}_n\| - \|\theta^*\| \right| \leqslant cB^{3/2}\sqrt{\frac{d \log n + \log(1/\delta)}{n}}. \tag{10}$$

While the bound (10) controls the estimation errors on the norm and direction of the parameter, we note that it can be equivalently restated in terms of excess logistic risk, as

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c' \, \frac{d \log n + \log(1/\delta)}{n} \tag{11}$$

for some constant $c' > 0$. This guarantee matches the asymptotic risk (6) up to an additional $\log n$ factor, and as we will show below the condition for existence of the MLE from [KvdG23] is also almost sharp up logarithmic factors. We also note that further results on linear separation in more general contexts have been obtained by Kuchelmeister [Kuc24].

Hsu and Mazumdar [HM23] consider the problem of estimating the parameter direction $\theta^*/\|\theta^*\|$ (which suffices for the task of classification, namely of predicting the most likely value of Y given X, as opposed to estimating conditional probabilities), again with an emphasis on the dependence on the signal strength B. Like [KvdG23] they consider the case of a Gaussian design, but assume that the data follows a logit model rather than a probit model. Notably, they consider different estimators than the MLE for logistic regression, in particular the minimizer of a classification error. They establish upper bounds on the estimation error of the same order as the first bound in (10), again with logarithmic factors in n. In addition, they establish minimax lower bounds on the estimation error of $\theta^*/\|\theta^*\|$, which show that the previous upper bound is sharp up to logarithmic factors. They also explicitly raise the question of whether or not the MLE achieves optimal upper bounds.

While these results constitute decisive advances, they leave some important questions. First, the guarantees feature additional logarithmic factors in the sample size, which are presumably suboptimal but seem hard to avoid in the analyses of [KvdG23] and [HM23], leaving a gap between upper and lower bounds. Although logarithmic factors are admittedly a mild form of suboptimality, logistic regression with a Gaussian design is arguably a basic enough problem to justify aiming for sharp results. Second and perhaps more importantly, these results are specific to the case of a Gaussian design and a well-specified model, which raises the question of the behavior of the MLE for more general design distributions or under a misspecified model.

1.3 Summary of contributions

We are now in position to provide a high-level overview of our main results; we refer to subsequent sections for precise statements and additional comments.

Gaussian design, well-specified model. First, in the case of a Gaussian design and a well-specified logit model, Theorem 1 provides optimal (up to absolute constants) guarantees for the existence and accuracy of the MLE. Specifically, there exists a universal constant c such that the following holds: for any $\delta \leqslant 1/2$, if $n \leqslant c^{-1}B(d + \log(1/\delta))$, then

$$\mathbb{P}(\text{MLE exists}) \leqslant 1 - \delta\,. \tag{12}$$

On the other hand, if $n \geqslant cB(d + \log(1/\delta))$, then with probability *at least* $1 - \delta$ the MLE exists and satisfies

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c\,\frac{d + \log(1/\delta)}{n}\,. \tag{13}$$

This removes a $\log n$ factor from the bound (11) deduced from the work of [KvdG23] in the case of a probit model, and answers in the affirmative (after translating this risk bound into a bound on the estimation error) a question from [HM23] on the optimality of the MLE.

In short, this result provides necessary and sufficient conditions on the sample size n (up to numerical constant factors) for the MLE to exist with high probability, and shows that in the regime where the MLE exists, it achieves non-asymptotically the same risk as predicted by the asymptotic behavior (6) for fixed B, d, δ and $n \to \infty$.

The previous result implies in particular that, if $n \gg Bd$, then the MLE exists with probability at least $1 - \exp\big(-\frac{n}{c'B}\big)$ for some constant c', and that this estimate is optimal. This provides a quantitative version of the convergence to 1 in the phase transition (7) from Candès and Sur [CS20]. On the other hand, in the regime where $n \ll Bd$, Theorem 1 only shows that the probability of existence of the MLE is bounded by a constant (say, 1/2), rather than converging to 0 as in the phase transition (7). We therefore complement Theorem 1 by a result on non-existence of the MLE (Theorem 2), which states that if $n \ll Bd/\kappa$ for some $\kappa \geqslant 1$, then

$$\mathbb{P}(\text{MLE exists}) \leqslant c \exp\big(-\max\big\{\kappa\sqrt{d}, \kappa^2 d/B^2\big\}/c\big) \tag{14}$$

for some constant c. This can be seen as a quantitative version of the convergence to 0 in the phase transition (7) from [CS20].

Regular design, well-specified model. The previous results are specific to the case of a Gaussian design, which can be seen as the most favorable case. This raises the following natural question: which properties of the Gaussian distribution are responsible for the previously described behavior of the MLE? Or equivalently, for which distributions of the design does the MLE behave (at least in the well-specified case) similarly as for a Gaussian design?

Perhaps a natural guess is that a light-tailed design distribution would lead to a similar behavior as a Gaussian design, and indeed this would be the case for linear regression. However, this is far from being the case for logistic regression: as previously alluded to, if the design distribution is only assumed to be sub-Gaussian (as in [CLL20]), then an exponential dependence on the norm is unavoidable.

In Section 2.2, we identify suitable assumptions on the design distribution leading to a near-Gaussian behavior. Aside from light tails (Assumption 1), the assumptions include a condition on the behavior of one-dimensional linear projections of the design near 0 (Assumption 2), which is related to standard margin conditions in the classification literature [MT99, Tsy04]. However, as shown in Proposition 1, another assumption is necessary to obtain a near-Gaussian behavior (in a suitable sense); this non-standard condition (Assumption 3) bears on *two-dimensional* linear projections of the design, rather than merely its one-dimensional marginals. By analogy with the standard (one-dimensional) margin condition, we refer to this condition as "two-dimensional margin assumption".

Under these regularity assumptions on the design but still in the well-specified case, Theorem 3 shows that the MLE behaves similarly as in the Gaussian case, up to poly-logarithmic factors in the norm B. Specifically, for some constant c (depending on the constants of the regularity conditions), if $n \geqslant c \log^4(B) B(d + \log(1/\delta))$, then with probability $1-\delta$ the MLE exists and satisfies

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c \log^4(B) \, \frac{d + \log(1/\delta)}{n} \, . \tag{15}$$

Regular design, misspecified model. Finally, we turn to the most general setting, where no assumption is made on the conditional distribution of Y given X; in particular, it is no longer assumed that it belongs to the logit model. This being said, as previously discussed it is still possible to define the minimizer θ^* of the logistic risk, and to consider the excess risk $L(\widehat{\theta}_n) - L(\theta^*)$ of the MLE, which corresponds to the empirical risk minimizer (ERM) under the logistic loss. This corresponds to the problem of Statistical Learning under logistic loss.

As discussed in Section 1.2, in many regimes of interest the best available guarantees for this problem in the literature are those from [OB21], namely the excess risk bound (9) of order $B^3 d \log(1/\delta)/n$, when the design is Gaussian but the model may be misspecified. Theorem 4 below improves these guarantees in the following way: it shows that if the design is regular (in the same sense as before) and $n \geqslant c \log^4(B)(Bd + B^2 \log(1/\delta))$, with probability at least $1-\delta$ the MLE exists and satisfies

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c \log^4(B) \frac{d + B\log(1/\delta)}{n} \, ; \tag{16}$$

here and as in (15), c is a constant that depends on the constants from the regularity conditions on X. For instance, for constant δ and $B \lesssim d$, this removes a factor of almost B^3 from the previous best guarantee (9) for statistical learning with logistic loss.

Our guarantees in the misspecified case feature a stronger dependence on the norm B than in the well-specified case; specifically, the "deviation terms" (those that depend on the failure probability δ) in both the condition for existence of the MLE and its excess risk bound are larger by a factor of B. This raises the question of whether this gap is essential or an artifact of the analysis. As it turns out, even for a Gaussian design, the guarantee (16) is best possible (up to $\mathrm{polylog}(B)$ factors) in the general misspecified setting, both in the condition for existence of the MLE and for its excess risk bound, as shown in Theorem 4.

Distributions satisfying the regularity assumptions. While identifying conditions on the design distribution under which the MLE behaves as in the Gaussian case may be interesting in its own right, in order for these general assumptions to constitute a genuine extension of the Gaussian case one must exhibit other meaningful examples of distributions satisfying them.

To illustrate these conditions, in Section 3 we consider two families of distributions, namely log-concave distributions and product measures. The log-concave case is overall similar to the Gaussian case (which it contains as a special case), in that the regularity conditions hold for any value of the parameter $\theta^* \in \mathbb{R}^d$—that is, for any parameter direction $u^* = \theta^*/\|\theta^*\|$ and signal strength $B = \max\{e, \|\theta^*\|\}$. In contrast, the case of product measures is more subtle, since the regularity conditions are highly sensitive to the parameter direction u^*. Indeed, we show that for designs with i.i.d. coordinates (a prototypical example being the *Bernoulli design*, with i.i.d. coordinates uniform over $\{-1, 1\}$), depending on the parameter direction the MLE may behave as in the Gaussian case either only for trivial (constant) signal strength $B = O(1)$, or up to a large signal strength $B = O(\sqrt{d})$.

1.4 Additional related work

We now survey additional relevant prior work on logistic regression, beyond the results discussed in Section 1.2.

Logistic regression as convex statistical learning. Logistic regression is a special case of convex statistical learning (or convex stochastic optimization), allowing to leverage guarantees from this setting. For instance, a standard uniform convergence argument using the Lipschitz property of logistic loss implies an excess risk bound of order $B\sqrt{d/n}$ for ERM over a ball of radius $O(B)$. This upper bound exhibits a slow convergence rate of $n^{-1/2}$ as $n \to \infty$, as opposed to the actual asymptotic rate of n^{-1}.

In order to improve over the slow rate, a strengthening of mere convexity is needed, in the form of additional assumptions on the curvature of the loss or risk. A common notion of curvature in optimization is strong convexity [BV04, Bub15, Bac24] of the loss, however the logistic loss is not strongly convex with respect to the regression parameter θ as it only varies in one direction. A more appropriate notion of curvature is "exponential concavity" (exp-concavity), which originates from online learning [Vov98, CBL06]. Using this property of logistic loss, it is shown in [PZ23] (see also [Meh17, DVW21]) that ERM for logistic regression constrained to a ball of radius $O(B)$ achieves an excess risk of at most $O(de^{cBR}/n)$, where $R > 0$ is such that $\|X\| \leqslant R$ almost surely. In the isotropic case where $\mathbb{E}[XX^\top] = I_d$, one has $R \geqslant \mathbb{E}[\|X\|^2]^{1/2} = \sqrt{d}$, hence the previous guarantee scales at best as $de^{cB\sqrt{d}}/n$, with an exponential dependence on the norm B and (square root of the) dimension d. This reflects the fact that logistic loss only possesses very weak deterministic curvature.

Another relevant property of logistic loss is (pseudo-)self-concordance (a bound on the third derivative of the loss in terms of the second derivative) which was put forward by [Bac10], and used to analyze logistic regression in a series of works [Bac10, Bac14, BM13, OB21], the sharpest results in this direction being those of Ostrovskii and Bach [OB21] discussed in Section 1.2.

Finally, a classical condition to obtain fast rates for ERM in Statistical Learning Theory is a bound on the variance of loss differences in terms of the excess risk [Mas07, Kol11], an assumption known as Bernstein condition [BM06]. General guarantees for ERM in statistical learning under convex and Lipschitz loss are obtained in [ACL19] using the Bernstein condition. These results are refined in the work [CLL20] by using a local version of this condition; we discussed the instantiation of their results to logistic regression in Section 1.2.

High-dimensional asymptotics. As discussed in Section 1.2, Candès and Sur [CS20] characterized the phase transition for existence of the MLE in the well-specified case and with a Gaussian design in the high-dimensional asymptotic regime where $d/n \to \gamma \in (0, 1)$ and $\beta = \|\theta^*\| \in \mathbb{R}^+$ is fixed. This result on existence is complemented in [SC19] by a result on the behavior of the MLE under the same assumptions and asymptotic regime; specifically, it is shown in this work that the joint distribution of the true and estimated coefficients converges to a certain distribution. These results have been extended, among others, to arbitrary covariance matrices of the design [ZSC22], to ridge-regularized logistic regression [SAH19], to more general binary models [TPT20], to multinomial logistic regression [TB24] and to missing data [VM24].

Worst-case design distributions, improper and robust estimators. Our focus in this work is to characterize the performance of the MLE in "regular" situations, namely when the design distribution satisfies some suitable conditions ensuring a near-Gaussian behavior. A rather different but complementary perspective consists in considering the performance of the MLE or other estimators for logistic regression under worst-case design distributions.

As one might expect, the performance of the MLE is considerably degraded for worst-case design distributions. In particular, a lower bound from [HKL14] for statistical learning with logistic loss implies that, when no assumption is made on the design asides from boundedness ($\|X\| \leqslant R$ almost surely), then the MLE or any "proper" estimator (that returns a conditional density belonging to the logistic model) can achieve no better expected excess risk (with respect to the ball of radius B) than $O(BR/\sqrt{n})$, as long as $n \leqslant e^{cBR}$. This exponential dependence on the parameter norm can be bypassed by resorting to "improper estimators", that is, estimators that return conditional densities that do not belong to the logistic model; these include Bayesian model averaging [KN05, FKL+18, QRZ24] or adjusted estimators that account for uncertainty using "virtual labels" [MG22, JGR20]. We also refer to [Vij21, vdHZCB23] for alternative procedures achieving sharp high-probability guarantees, albeit at a high computational cost.

A related direction is that of robust logistic regression. In [CLL20], high-probability risk bounds in logistic loss are established for estimators based on medians-of-means, when the design X may be heavy-tailed. In addition, estimators for logistic regression achieving near-optimal guarantees in Hellinger distance were proposed in [BC24]. From a statistical perspective, these estimators are more robust than the MLE; on the other hand, their computational cost appears to be exponential in the dimension, making them less directly usable in practice.

1.5 Outline and notation

Paper outline. This paper is organized as follows. Section 2 contains the precise statements of our main results, as well as the definition and discussion of the regularity assumptions we consider on the design distribution. In Section 3, we illustrate these assumptions by investigating to which extent they hold for three standard classes of design distributions. Section 4 describes the structure of the proofs of the main results from Section 2, including statements of the main lemmas. In particular, a convex localization argument reduces the proof of existence and risk bounds for the MLE to two components: an upper bound on the gradient of the empirical risk, and a uniform lower bound on the Hessian of the empirical risk in a neighborhood of the true parameter. Section 5 is devoted to the proofs of upper bounds on the empirical gradients, while lower bounds on the empirical Hessian are established in Section 6. Next, in Section 7 we provide the proof of Theorem 2 on linear separation (non-existence of the MLE) with high probability. In Section 8, we conclude the proofs of the main results of Section 2 by putting together the results of Sections 4.1, 5 and 6 and providing additional lower bounds. Finally, Section 9 contains the proofs of results from Section 3, namely regularity of log-concave and product measures. Appendices A and B gather technical facts on real random variables and polar coordinates, while Appendix C contains the proof of Proposition 1 on necessity of the two-dimensional margin condition.

Notation. Throughout the paper, the sample size will be denoted by n and the dimension by d. We let $\langle \cdot, \cdot \rangle$ denote the standard inner product on $\mathbb{R}^d$, and $\| \cdot \| = \| \cdot \|_2$ the associated Euclidean norm. We denote by B_2^d the unit Euclidean ball and by S^{d-1} the unit sphere in $\mathbb{R}^d$. For a positive semi-definite matrix A, we let $\| \cdot \|_A$ be the semi-norm induced by A, defined by $\|x\|_A^2 = \langle Ax, x \rangle = \|A^{1/2}x\|^2$ for $x \in \mathbb{R}^d$. The operator norm of a matrix A is denoted by $\|A\|_{\mathrm{op}}$. Given two $d \times d$ symmetric matrices A, B, we write $A \preccurlyeq B$ if $\langle Av, v \rangle \leqslant \langle Bv, v \rangle$ for every $v \in \mathbb{R}^d$.

If $f : \mathbb{R}^d \to \mathbb{R}^d$ is a twice continuously differentiable function, we let $\nabla f(x) \in \mathbb{R}^d$ and $\nabla^2 f(x) \in \mathbb{R}^{d \times d}$ respectively denote its gradient and Hessian at $x \in \mathbb{R}^d$. For $a, b \in \mathbb{R}$ we let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$, as well as $a_+ = \max(a, 0)$ and $a_- = \max(-a, 0)$.

Recall the notation introduced in Section 1.1: we let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. pairs having the same distribution P as a generic pair (X, Y), and let $\widehat{\theta}_n$ denote the MLE defined by (2). Throughout, we assume without loss of generality that the design X has an isotropic distribution, in the sense that $\mathbb{E}[XX^\top] = I_d$. We say that the logit model is *well-specified* if the conditional distribution of Y given X belongs to the model $\mathcal{P}_{\mathrm{logit}}$, that is if there exists $\theta^* \in \mathbb{R}^d$ such that $\mathbb{P}(Y = 1|X) = \sigma(\langle \theta^*, X \rangle)$; otherwise, the model is said to be misspecified.

2 Main results

In this section, we provide the precise statements of our main results on logistic regression, which we presented informally in Section 1.3.

This section is organized as follows. First, Section 2.1 contains the results in the case of a well-specified model and a Gaussian design. In Section 2.2, we introduce and discuss the regularity assumptions we consider on the design to generalize the Gaussian case. In Section 2.3 we extend the results of Section 2.1 to the case of a regular design, while still assuming that the model is well-specified. Finally, in Section 2.4 we consider the most general case, where the design is regular but no assumption is made on the conditional distribution of Y given X.

2.1 Well-specified model, Gaussian design

We start with the case of a well-specified model and a Gaussian design. Theorem 1 below provides a sharp condition (up to universal constant factors) on the sample size n for the MLE to exist with high-probability, as well as an optimal upper bound in deviation on its excess risk. Its proof can be found in Section 8.2.

Theorem 1. *Assume that the design $X \sim \mathsf{N}(0, I_d)$ is Gaussian and that the model is well-specified with parameter $\theta^* \in \mathbb{R}^d$, and let $B = \max\{e, \|\theta^*\|\}$. There exist universal constants $c_1, c_2, c_3 > 0$ such that, for any $t > 0$: if*

$$n \geqslant c_1 B(d + t)\,, \tag{17}$$

then, with probability $1 - e^{-t}$, the MLE $\widehat{\theta}_n$ exists and satisfies

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c_2 \frac{d + t}{n}\,. \tag{18}$$

Moreover, for any $d \geqslant 53$ and $t \geqslant 1$, if $n \leqslant c_3 B(d + t)$ then the MLE exists with probability at most $1 - e^{-t}$.

It follows from Theorem 1 that, up to numerical constants, the condition (17) is both necessary and sufficient for the MLE to exist with high probability, and that whenever this condition holds, the MLE admits the same risk guarantee as the asymptotic one (6) in the regime where B, d are fixed while $n \to \infty$, which is optimal in light of the convergence in distribution (5). In particular, the condition on n that ensures that the MLE exists also ensures that it achieves its asymptotic excess risk.

We note also that, using Lemma 3, the proof of Theorem 1 also provides guarantees for the estimation error of the direction and norm of θ^*: if $\|\theta^*\| \geqslant e$, we have for some universal constants $c_3, c_4 > 0$: if $n \geqslant c_3 B(d + t)$, then with probability at least $1 - e^{-t}$,

$$\left\| \frac{\widehat{\theta}_n}{\|\widehat{\theta}_n\|} - \frac{\theta^*}{\|\theta^*\|} \right\| \leqslant c_4 \sqrt{\frac{d + t}{Bn}}\,, \qquad \big| \|\widehat{\theta}_n\| - \|\theta^*\| \big| \leqslant c_4 \sqrt{\frac{B^3(d + t)}{n}}\,. \tag{19}$$

The proof of Theorem 1 can be found in Section 8.2 (combining results from Sections 5.2 and 6.3), while the scheme of proof is described in Section 4. In particular, a key structural result in the analysis is Theorem 6, which provides a sharp high-probability lower bound on the Hessian of the empirical risk $\widehat{H}_n(\theta) = \nabla^2 \widehat{L}_n(\theta)$, uniformly for θ belonging to a neighborhood of θ^* that is "as large as possible".

Let us now come back to the question of existence of the MLE; as noted in the introduction, non-existence of the MLE amounts to linear separation of the dataset. Theorem 1 implies in particular that if $n \geqslant 2C_1 Bd$, then the probability that the MLE exists is at least $1-\exp(-\frac{n}{2C_1 B})$, which is optimal by the last part of Theorem 1. This can be seen as a quantitative version of the convergence to 1 in the phase transition (7) for the existence of the MLE established by Candès and Sur [CS20]. On the other hand, if $n \ll Bd$, then Theorem 1 (with $t = 0$) only implies that the probability of existence of the MLE is bounded away from 1, rather than close to 0 as in the phase transition (7).

Theorem 2 below shows that if $n \ll Bd$, then the probability of existence of the MLE indeed approaches 0, at a rate exponential in the dimension. This can be seen as a quantitative version of the convergence to 0 in the phase transition (7) from [CS20]. Theorem 2 is proved in Section 7.

Theorem 2. *Let $d \geqslant 53$, and assume that $X \sim \mathsf{N}(0, I_d)$ and that the logistic model is well-specified. For every $\kappa \geqslant 1$, if $n \leqslant Bd/(23000\kappa)$ then*

$$\mathbb{P}(\text{MLE exists}) \leqslant \exp\big(-\max\big\{\kappa\sqrt{d}, \kappa^2 d/B^2\big\}\big) + 6e^{-d/21}\,. \tag{20}$$

Let us now discuss the interpretation of this result.

First, the question of non-existence of the MLE (that is, linear separation) is mainly of interest when $n \geqslant d$, since for $n < d$ the dataset is linearly separated, as the points $X_1, \dots, X_n$ do not span $\mathbb{R}^d$. We are therefore interested in the regime $d \ll n \ll Bd$, where linear separation no longer occurs deterministically because of the dimension, but instead with high probability due to the fact that the signal is strong ($B \gg 1$). Specifically, in this regime, a strong signal B entails that a large fraction of labels Y_i will be of the same sign as the predictions $\langle\theta^*, X_i\rangle$, which effectively constrains the directions of the vectors $Y_i X_i$, making it easier to find a $\theta \neq 0$ such that $\langle\theta, Y_i X_i\rangle \geqslant 0$ for every $i = 1, \dots, n$. Interestingly, if $n \gg B$ (while $n \ll Bd$), then the true parameter θ^* will typically not satisfy this property; instead, linear separation will be achieved by another (random) parameter θ, thanks to the flexibility due to the large dimension d. As such, in the regime $\max\{B, d\} \ll n \ll Bd$, linear separation holds with high probability owing to the *combination* of the "signal strength" and "dimension" effects, rather than one of the two taken individually.

We now comment on the quantitative bound (20). Theorem 2 implies that if n is small compared to the threshold of order Bd, then the probability of existence of the MLE is smaller than $\exp(-c\max\{\sqrt{d}, d/B^2\})$ for some constant c. In addition, the parameter $\kappa \geqslant 1$ quantifies how small the sample size n is relative to the critical threshold of Bd, and the smaller the sample size (that is, the larger κ is), the smaller the bound (20). In particular, in the regime where $n \asymp d$ and $B \gg 1$ (so that $\kappa \asymp B$), Theorem 2 shows that the probability of existence of the MLE is smaller than $\exp(-cd)$ for some constant c.

The proof of Theorem 2, which builds on the approach of Candès and Sur [CS20], can be found in Section 7. Specifically, following [CS20], the starting point of the proof is to reformulate the property of linear separation into the property that a certain random cone Λ in $\mathbb{R}^n$ has a non-trivial intersection with an independent uniform random subspace. Now, it follows from the work [ALMT14] that the probability of such an event depends on the dimension of the random subspace, and on a certain geometric parameter of the cone Λ called "statistical dimension". In order to control the probability of existence of the MLE, one must therefore combine two steps: (i) conditionally on the cone Λ, apply a phase transition result showing that the probability that a random subspace does not intersect Λ is small; (ii) in order to apply the previous result to the random cone Λ, control of the statistical dimension of Λ with high probability.

For the first point, Candès and Sur use a phase transition result from [ALMT14]. For the second point, they establish that the statistical dimension of the random cone Λ converges in probability to a deterministic value as $n, d \to \infty$ while $d/n \to \gamma$, for fixed $\beta = \|\theta^*\|$. To show this, they first relate the statistical dimension to (a family of) averages of i.i.d. random variables, and then establish uniform convergence of the averages to the corresponding expectations.

While these arguments suffice to establish the 0-1 law (7) in this asymptotic regime, several refinements are required in order to obtain the quantitative bound of Theorem 2. First, a more precise phase transition result [ALMT14, Theorem 6.1] must be used in order to finely capture the dependence on the statistical dimension of Λ. Second and more importantly, one must establish a refined high-probability control on the statistical dimension of the random cone. This requires a high-probability bound on the sum of i.i.d. random variables that (as shown in [CS20]) controls this dimension. We achieve this by first obtaining a tight control on the moments of the individual summands, and then applying a sharp estimate of Latała [Lat97] on moments of sums of independent random variables.

2.2 Regularity assumptions

In this section, we introduce formally the setup that we called "regular" in the introduction and present our two generalizations of Theorem 1, first to the case of a non-Gaussian design (but still assuming a well-specified model) and then to the most general case of a regular design with a misspecified model. Examples of settings satisfying these regularity assumptions are given and discussed in Section 3.

The first assumption on the design is standard and states that the design X is light-tailed; we refer to Definition 5 in Appendix A for the definition of the ψ_1-norm.

Assumption 1. The random vector X is K*-sub-exponential* for some $K \geqslant e$, in the sense that $\|\langle v, X\rangle\|_{\psi_1} \leqslant K$ for every $v \in S^{d-1}$.

The second assumption is also standard in the literature on supervised classification. It states that the design X does not put too much mass close to the separation hyperplane. It is related to the *margin assumption* that allows to derive fast rates of convergence, see [MT99, Tsy04, AT07].

Assumption 2. Let $u^* \in S^{d-1}$ and $\eta \in (0, 1]$. For some $c \geqslant 1$, one has for every $t \geqslant \eta$ that

$$\mathbb{P}\big(|\langle u^*, X\rangle| \leqslant t\big) \leqslant ct\,. \tag{21}$$

The third assumption on the other hand is new to the best of our knowledge. It is an assumption on the two-dimensional marginals of the design X that we call *two-dimensional margin condition.*

Assumption 3. Let $u^* \in S^{d-1}$, $\eta \in [0, 1/e]$ and $c \geqslant 1$. For every $v \in S^{d-1}$ such that $\langle u^*, v\rangle \geqslant 0$, one has

$$\mathbb{P}\big(|\langle u^*, X\rangle| \leqslant c\eta,\ |\langle v, X\rangle| \geqslant c^{-1} \max\{\eta, \|u^* - v\|\}\big) \geqslant \eta/c\,. \tag{22}$$

Remark 1. Using that

$$\|u^* - v\|/\sqrt{2} \leqslant \sqrt{1 - \langle u^*, v\rangle^2} = \|u^* - v\|\sqrt{(1 + \langle u^*, v\rangle)/2} \leqslant \|u^* - v\|$$

if $\langle u^*, v\rangle \geqslant 0$, another way of stating Assumption 3 is that for every $v \in S^{d-1}$, one has

$$\mathbb{P}\Big(|\langle u^*, X\rangle| \leqslant c\eta,\ |\langle v, X\rangle| \geqslant c^{-1} \max\Big\{\eta, \sqrt{1 - \langle u^*, v\rangle^2}\Big\}\Big) \geqslant \eta/c\,. \tag{23}$$

This only changes the value of the parameter c from (22) by a factor $\sqrt{2}$. This equivalent formulation turns out to be more convenient in some situations.

Let us now discuss this new assumption. Recall that the purpose here is to provide a setting more general than the Gaussian one where the MLE behaves as in the Gaussian case. We argue here that Assumption 3 is necessary for this task. To see why, note first that, when the design

$X = G$ is a standard Gaussian vector, the Hessian $H_G(\theta^*) = \nabla^2 L(\theta^*)$ is, by (36) and (37), within constant factors of the matrix

$$H = \frac{1}{B^3} u^* u^{*\top} + \frac{1}{B}(I_d - u^* u^{*\top}), \tag{24}$$

where $B = \max\{\|\theta^*\|, e\}$. This Hessian is also the Fisher information matrix of the statistical model at θ^*. This implies that, when the design is Gaussian and the model well-specified, the MLE converges in distribution as $n \to \infty$:

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \to \mathsf{N}(0, H_G(\theta^*)^{-1}).$$

For more general design X and still in the well-specified case, the Fisher information matrix is still equal to the Hessian $H_X(\theta^*)$ (but computed using the distribution of X rather than the Gaussian one) and the MLE converges, as $n \to \infty$, to

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \to \mathsf{N}(0, H_X(\theta^*)^{-1}).$$

Therefore, for the MLE to behave as well as in the Gaussian case, at least asymptotically, it is necessary that the Hessian matrix $H_X(\theta^*)$ is at least as large as in the Gaussian case $H_G(\theta^*)$, meaning that it should satisfy an inequality of the form $H_X(\theta) \succcurlyeq CH$ for an absolute constant C. We will show that Assumptions 1, 2 and 3 are sufficient to prove that this inequality holds, and even that the Hessian $H_X(\theta)$ is locally equivalent to H, see Lemma 28. Actually, the main results in Sections 2.3 and 2.4 show that the MLE indeed behaves within $\log B$ factors as in the Gaussian setting in this general framework.

Moreover, the following result establishes that the new Assumption 3 is necessary to bound the Hessian $H_X(\theta^*)$ in the sense that if Assumptions 1 and 2 hold and $H_X(\theta^*) \succcurlyeq CH$, then Assumption 3 must also hold.

Proposition 1. *Let X be a random vector satisfying Assumption 1 with parameter K and Assumption 2 with parameters $u^* \in S^{d-1}$, $\eta = B^{-1} \leqslant e^{-1}$ and $c \geqslant 1$. If there exists $C_0 \geqslant e$ such that $H_X(\theta^*) \succcurlyeq C_0^{-1} H$, then Assumption 3 holds with parameters $u^* = \theta^*/\|\theta^*\|$, $\eta = 1/B$, and*

$$c' = \max\left(c_0 \log B, 2\sqrt{c_0 C_0 c \log B}, 144 K^2 \log^2(C_0 K B)\right),$$

where $c_0 = 3 + \log(4C_0)$.

The proof of this result can be found in Appendix C.

We can now formulate the definition of "regular distributions" that we use throughout.

Definition 1. Let $u^* \in S^{d-1}$, $\eta \in (0, e^{-1}]$ and $c \geqslant 1$. A random vector X in $\mathbb{R}^d$ is said to have an (u^*, η, c)-*regular* distribution if it is isotropic (that is, $\mathbb{E}[XX^\top] = I_d$) and satisfies Assumptions 2 and 3 with parameters u^*, η, c.

2.3 Well-specified model, regular design

We can now state our main result on the performance of the MLE in the case of a regular design and a well-specified model, whose proof can be found in Section 8.3.

Theorem 3. *Assume that the model is well-specified, with unknown parameter $\theta^* = \|\theta^*\| u^*$ where $u^* \in S^{d-1}$ and let $B = \max\{e, \|\theta^*\|\}$. Assume that X satisfies Assumptions 1, 2 and 3 with parameters $K \geqslant e$, u^*, $\eta = B^{-1}$ and c. There exist constants c_1, c_2 that depend only on c, K such that, if*

$$n \geqslant c_1 B \log^4(B)(d + t),$$

then with probability at least $1 - e^{-t}$, the MLE $\widehat{\theta}_n$ exists and satisfies

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c_2 \log^4(B) \frac{d + t}{n}. \tag{25}$$

The guarantees of Theorem 3 almost match (up to poly-logarithmic factors in B) those of Theorem 1 in the Gaussian case, which are optimal as discussed above. In fact, one can almost recover (again up to $\log^4(B)$ factors) the guarantees of Theorem 1 from this result, since one can show that the Gaussian design satisfies the regularity assumptions for all u^*, η and with c, K being universal constants.

2.4 Misspecified model, regular design

We now turn to the general case where the logit model may be misspecified. In this setting, the conditional distribution of Y given X is no longer determined by θ^*, conversely θ^* is now a function of the joint distribution of (X, Y), as is the case in Statistical Learning. We define θ^* as the minimizer of the population risk L (see (3)), namely

$$\theta^* \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} L(\theta)\,, \quad L(\theta) = \mathbb{E}\big[\log\big(1 + e^{-Y\langle \theta, X\rangle}\big)\big]. \tag{26}$$

By the discussion in the introduction, θ^* exists whenever the distribution of (X, Y) is not linearly separated, meaning that there is no $\theta \neq 0$ such that $Y\langle \theta, X\rangle \geqslant 0$ almost surely—which we assume in this section. In addition, θ^* is unique since we assume that $\mathbb{E}[XX^\top] = I_d$, which ensures strict convexity of L. Theorem 4 below is proved in Section 8.4.

Theorem 4. *Suppose that X satisfies Assumptions 1, 2 and 3 with parameters $K \geqslant e$, u^*, $\eta = B^{-1}$ and c, and that $\theta^* = \|\theta^*\| u^*$. Let $B = \max\{e, \|\theta^*\|\}$. There exist constants c_1, c_2 that depend only on c, K such that for any $t > 0$, if*

$$n \geqslant c_1 B \log^4(B)(d + Bt)\,,$$

then with probability at least $1 - e^{-t}$, the MLE $\widehat{\theta}_n$ exists and satisfies

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c_2 \log^4(B)\frac{d + Bt}{n}\,. \tag{27}$$

Moreover, for any $B \geqslant e$, there exists a distribution of (X, Y) with $X \sim \mathsf{N}(0, I_d)$ and $\|\theta^*\| = B$ such that if $n \leqslant c_3 B(d + Bt)$ (for some universal constant c_3), then

$$\mathbb{P}(\text{MLE exists}) \leqslant 1 - e^{-t}\,. \tag{28}$$

In addition, for the same distribution,

$$\liminf_{n\to\infty} \mathbb{P}\Big(L(\widehat{\theta}_n) - L(\theta^*) \geqslant c_3 \frac{d + Bt}{n}\Big) \geqslant e^{-t}\,. \tag{29}$$

Theorem 4 improves the previous best guarantees for the MLE in logistic regression in the general misspecified case. As discussed in Section 1.2, these are from [CLL20] for a sub-Gaussian design, and [OB21] for a Gaussian design. The guarantees in [CLL20] in the sub-Gaussian case feature an exponential dependence on B. The guarantees in [OB21] in the Gaussian case (a special case of regular design), which are actually the previous best guarantees for the MLE in a misspecified setting[2], feature a polynomial dependence on B but a stronger one than in Theorem 4: the condition for existence of the MLE writes (ignoring polylog(B) factors) $n \gtrsim B^8 dt$, and the risk is bounded by $B^3 dt/n$.

[2]Technically speaking, the guarantees in [OB21] (obtained by combining Theorem 4.2 with Proposition D.1) are stated in the well-specified case. However, they can be extended to the misspecified case through a very minor modification (renormalizing gradients by the Hesssian of the risk instead of their covariance matrix). Likewise, the deviation terms in $d \cdot t$ in these results can be tightened to $d + t$ with no changes to the analysis.

It should be noted that both the sample size needed for the MLE to exist and the bound on its excess risk in Theorem 4 exhibit a stronger dependence on B compared to the well-specified case. As shown by (28) and (29), this stronger dependence on B is in fact necessary in the misspecified case. This shows that the non-asymptotic guarantees of Theorem 4 for the existence and the excess risk of the MLE are sharp, up to polylogarithmic factors in B. It should be pointed out that the degradation only affects an additive term that does not multiply the dimension d, hence as long as $Bt = O(d)$ (a regime that covers many situations of interest), the guarantees in the misspecified case actually match those of the well-specified case.

3 Examples of regular design distributions

In the previous section, we introduced certain regularity assumptions (Assumptions 1, 2 and 3) on the distribution of the design X, which we argued were essentially necessary and sufficient to obtain the same results as in the Gaussian case. In this section, we provide examples of distributions that satisfy these assumptions.

The three examples we consider are: sub-exponential distributions when the signal strength is of constant order (Section 3.1), log-concave distributions (Section 3.2), and product measures (Section 3.3).

We recall that the regularity assumptions introduced in Section 2.2 depend on both a direction $u^* \in S^{d-1}$ and a scale parameter $\eta \in (0, e^{-1}]$. When applied to logistic regression, these correspond respectively to the parameter direction $u^* = \theta^*/\|\theta^*\|$ and inverse signal strength $\eta = 1/B = 1/\max(\|\theta^*\|, e)$. In particular, the stronger the signal, the finer the scale η at which the regularity assumptions should hold for our guarantees of Section 2 to apply.

3.1 Regularity at constant scales

First, we note that the regularity assumptions at a lower-bounded scale η (corresponding to a bounded signal strength) are automatically satisfied when the design is sub-exponential.

Proposition 2. *Let X be an isotropic and K-sub-exponential random vector (Assumption 1). Then X is $(u^*, \eta, c_{K,\eta})$-regular for any $u^* \in S^{d-1}$ and $\eta \in (0, e^{-1}]$, where*

$$c_{K,\eta} = \max\left\{\frac{2K\log(2K)}{\eta}, 2K^4\right\}.$$

The content of Proposition 2 (proved in Section 9.1) is that the regularity assumptions are general enough to include all sub-exponential distributions, with the caveat that the involved constant c depends on the scale η. However, it should be noted that the bounds in Theorems 3 and 4 depend exponentially on c, leading to an exponential dependence on the signal strength $B = \eta^{-1}$. For this reason, Proposition 2 is mainly relevant in the case of constant signal strength.

3.2 Regularity of log-concave distributions

The issue of the general reduction from sub-exponential to regular is that it ultimately leads to a poor (exponential) dependence on the signal strength in the guarantees of Section 2. As we shall see in Section 3.3, this exponential dependence is necessary in general, hence in order to obtain similar guarantees as for a Gaussian design, one must strengthen the assumptions on the design beyond merely sub-exponential tails.

A natural class of probability measures that contains Gaussian measures, and often exhibit similar properties, is the class of *log-concave* distributions on $\mathbb{R}^d$. Specifically, recall that the distribution P_X on $\mathbb{R}^d$ is log-concave (see e.g. [SW14]) if, for all Borel sets $S, T \subset \mathbb{R}^d$ and $\lambda \in (0,1)$ such that $\lambda S + (1-\lambda)T = \{\lambda s + (1-\lambda)t : s \in S, t \in T\}$ is measurable,

$$P_X(\lambda S + (1-\lambda)T) \geqslant P_X(S)^\lambda P_X(T)^{1-\lambda}.$$

We are interested in the case where X is centered and isotropic, in which case it is log-concave if and only if it admits a density on $\mathbb{R}^d$ of the form $\exp(-\phi)$, for some convex function $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$.

The following result shows that centered isotropic and log-concave distributions are regular in all directions and at all scales.

Proposition 3. *Assume that X has a centered isotropic (that is, $\mathbb{E}[X] = 0$ and $\mathbb{E}[XX^\top] = I_d$) and log-concave distribution on $\mathbb{R}^d$. Then X is c-sub-exponential and (u^*, η, c)-regular with a universal constant c, for every direction $u^* \in S^{d-1}$ and every scale $\eta \in (0, e^{-1}]$.*

The proof of Proposition 3 is provided in Section 9.2. The fact that log-concave distributions are regular (with universal constants) mainly comes from a key stability property: the distributions of their lower-dimensional linear projections are also log-concave [SW14], which is applied here to two-dimensional projections. In addition, low-dimensional, centered and isotropic distributions admit a density that is upper and lower-bounded around the origin. Hence, at small scales they are essentially equivalent to the Lebesgue (or Gaussian) measure, which admits a "product" or "independence" property for orthogonal linear projections that implies regularity.

3.3 Regularity for i.i.d. coordinates

Besides log-concave measures, another class of distributions that tend to behave similarly to Gaussian distributions in many high-dimensional contexts is that of product measures, that is, distributions of random vectors with independent coordinates. In this section, we therefore consider the question of regularity of product measures, which turns out to be much more subtle than in the log-concave case.

Specifically, in this section we consider the class of random vectors with i.i.d. sub-exponential coordinates:

Assumption 4. The random vector $X = (X_1, \dots, X_d)$ is such that: $X_1, \dots, X_d$ are i.i.d., with $\mathbb{E}[X_j] = 0$, $\mathbb{E}[X_j^2] = 1$ and $\|X_j\|_{\psi_1} \leqslant K$ (for some $K \geqslant e$) for $j = 1, \dots, d$.

It is a simple fact (see Lemma 33 in Section 9.3 below) that such a random vector is $4K$-sub-exponential. Hence, the main question is whether Assumptions 2 and 3 are satisfied.

A concrete example which illustrates the main issues is the Bernoulli design $X = (X_1, \dots, X_d)$, whose coordinates are i.i.d. random signs, namely $\mathbb{P}(X_j = 1) = \mathbb{P}(X_j = -1) = 1/2$ for $1 \leqslant j \leqslant d$ (that is, X is uniform on the discrete hypercube $\{-1, 1\}^d$). This design satisfies Assumption 4; in fact, its tails are even lighter than sub-exponential, since its coordinates are bounded and it is a sub-Gaussian random vector. This design is similar to the Gaussian design in many ways; for instance, it possesses strong concentration properties.

Despite these facts, the behavior of the MLE under a Bernoulli design can be drastically different from the case of a Gaussian design. Indeed, as noted below, an exponential dependence on the signal strength is necessary for the MLE to exist. This contrasts with the linear dependence on B in the Gaussian case (Theorem 1). As an aside, the example below shows that for a sub-Gaussian design, an exponential dependence on the norm is unavoidable in general, a fact we alluded to previously. In what follows, we denote by $(e_1, \dots, e_d)$ the canonical basis of $\mathbb{R}^d$.

Fact 1. *Let $X = (X_1, \dots, X_d)$ be a Bernoulli design, and let Y given X follow the logit model with parameter $\theta^* = Be_1$ for some $B \geqslant e$. Given an i.i.d. sample of size $n \geqslant 1$ from the same distribution as (X, Y), if $n \leqslant 0.1 \exp(B)$ then $\mathbb{P}(\text{MLE exists}) \leqslant 0.1$.*

Proof. First, since the model is well-specified, one readily verifies that

$$\mathbb{P}(Y\langle \theta^*, X\rangle \leqslant 0) = \mathbb{E}\big[\mathbb{P}(Y\langle \theta^*, X\rangle \leqslant 0|X)\big] = \mathbb{E}[\sigma(-|\langle \theta^*, X\rangle|)] \leqslant \mathbb{E}[\exp(-|\langle \theta^*, X\rangle|)]\,.$$

Now, note that $|\langle \theta^*, X\rangle| = B|X_1| = B$ since $X_1 = \pm 1$, and in particular $|\langle \theta^*, X\rangle| \geqslant B$. Thus, the above formula shows that $\mathbb{P}(Y\langle \theta^*, X\rangle \leqslant 0) \leqslant \exp(-B)$. Now, let $Z = YX$ and define similarly $Z_1, \dots, Z_n$ from the i.i.d. sample. If the MLE exists, then in particular θ^* does not linearly separate the dataset, hence there exists $1 \leqslant i \leqslant n$ such that $\langle \theta^*, Z_i\rangle \leqslant 0$. By a union bound, the probability of this event is lower than $n \exp(-B) \leqslant 0.1$ by assumption on n. □

This exponential dependence on the norm B comes from the fact that X is not regular at small scales in the direction $u^* = e_1$. Indeed, the random variable $\langle e_1, X\rangle = X_1$ is a random sign, which puts no mass in the neighborhood $(-1, 1)$ of 0, therefore violating Assumption 3 for small η and constant c. This illustrates the fact that the existence of the MLE is sensitive to the behavior of linear marginals of X around the origin, and not merely to the tails of X. Hence, the "discrete" nature of the Bernoulli design X (supported on a finite set) can lead to a very different behavior from the Gaussian case.

Although the previous example shows very different behaviors between the Gaussian and Bernoulli designs, one should keep in mind that it concerns a very specific direction $u^* = (1, 0, \dots, 0)$, which is a coordinate vector. This worst-case direction is highly "sparse"; this contrasts with a typical vector on the sphere, which is "dense" or "delocalized" in the sense that most of its coordinates are small, namely of order $O(1/\sqrt{d})$. One may expect that for such vectors, the behavior of the MLE is markedly different than for a sparse direction.

In order to capture this effect, we now consider the "densest" direction $u^* = (1/\sqrt{d}, \dots, 1/\sqrt{d})$, all of whose coefficients are small. Our aim is to characterize the smallest scale $\eta = \eta_d^*$ for which a design X with i.i.d. coordinates satisfies the regularity assumptions (Definition 1) at scale η in this direction u^*. In particular, if one could show that $\eta_d^* \to 0$ as $d \to \infty$, then this would establish sensitivity of the behavior of the MLE to the structure of the parameter direction u^*.

We start with Assumption 2 on the one-dimensional marginal $\langle u^*, X\rangle = \frac{1}{\sqrt{d}} \sum_{j=1}^d X_j$. Under Assumption 4, this random variable is a normalized sum of i.i.d. random variables. It then follows from the Berry-Esseen inequality that its distribution approaches the standard Gaussian distribution, down to a scale of order $1/\sqrt{d}$. This implies the following:

Lemma 1. *Let X satisfy Assumption 4. Then, for every $u \in S^{d-1}$ such that $\|u\|_3 \leqslant K^{-1}$ and any $t \in [K^3 \|u\|_3^3, 1]$, one has*

$$\frac{t}{4} \leqslant \mathbb{P}\big(|\langle u, X\rangle| \leqslant t\big) \leqslant t\,. \tag{30}$$

In particular, if $d \geqslant K^6$ and $u^* = (1/\sqrt{d}, \dots, 1/\sqrt{d})$, then Assumption 2 holds with $\eta = K^3/\sqrt{d}$ and $c = 1$.

Lemma 1 (whose proof is provided in Section 9.3) shows that the one-dimensional marginal $\langle u^*, X\rangle$ exhibits the "right" behavior down to a scale $\eta \asymp 1/\sqrt{d}$.

However, as discussed in Section 2.2 (see Proposition 1), Assumption 2 on the one-dimensional marginal $\langle u^*, X\rangle$ does not suffice to establish a near-Gaussian behavior of the MLE; indeed, for this task one must establish Assumption 3 on two-dimensional marginals $(\langle u^*, X\rangle, \langle v, X\rangle)$ for every $v \in S^{d-1}$. In order to simplify the discussion, let us consider the special case where $v \in S^{d-1}$ is orthogonal to u^*. In this case, Assumption 3 is of the form

$$\mathbb{P}\Big(|\langle u^*, X\rangle| \leqslant c\eta,\, |\langle v, X\rangle| \geqslant \frac{1}{c}\Big) \geqslant \frac{\eta}{c} \tag{31}$$

for some constant c. In the case where $X \sim \mathsf{N}(0, I_d)$ is Gaussian, condition (31) immediately follows from the fact that $\langle u^*, X\rangle$ and $\langle v, X\rangle$ are independent if $\langle u^*, v\rangle = 0$. However, this property is highly specific to the Gaussian case, and does not extend to the more general case of product measures.

By analogy with the proof of Assumption 2, a natural attempt to establish condition (31) is to resort to Gaussian approximation. Specifically, by applying a two-dimensional Berry-Esseen inequality to the random vector $(\langle u^*, X\rangle, \langle v, X\rangle) = \sum_{j=1}^d X_j \omega_j$ with $\omega_j = (u_j^*, v_j) = (1/\sqrt{d}, v_j)$ (such that $\sum_{j=1}^d \omega_j \omega_j^\top = I_2$) and proceeding as in Lemma 1, one can show that condition (31) holds down to $\eta \asymp \sum_{j=1}^d \|\omega_j\|_2^3 \asymp \max\{\|u^*\|_3^3, \|v\|_3^3\} \asymp \max\{1/\sqrt{d}, \|v\|_3^3\}$. This approach ensures that (31) holds for small η whenever v is sufficiently diffuse that $\|v\|_3^3$ is small. Unfortunately, condition (31) must hold for *every* $v \in S^{d-1}$ such that $\langle u^*, v\rangle = 0$, and in particular for non-diffuse vectors v such that $\|v\|_3^3 \asymp 1$ (for instance $v = (1/\sqrt{2}, -1/\sqrt{2}, 0, \dots, 0)$). For such vectors $v \in S^{d-1}$, Gaussian approximation gives vacuous guarantees.

As it happens, an entirely different argument (based on "approximate separation of supports") can be used to handle the case of "sparse" vectors, which—when suitably combined with Gaussian approximation—allows one to establish regularity at a non-trivial scale $\eta_d \asymp d^{-1/4} \to 0$. In order to convey the idea of this argument, and to illustrate how the $d^{-1/4}$ scaling naturally arises from this approach, we provide a high-level overview of the argument at the end of Section 9.3. Since the estimate on η_d obtained with this approach is sub-optimal and is improved in Lemma 2 below, we only provide a sketch of proof that omits significant technical details.

The argument we just alluded to leads to a scale of $d^{-1/4}$ for the two-dimensional margin assumption, which is larger than the scale of $d^{-1/2}$ obtained in Lemma 1 for one-dimensional marginals. This naturally raises the question of whether the $d^{-1/4}$ scale can be improved to $d^{-1/2}$ by a refined analysis. Lemma 2 below shows that this is indeed the case:

Lemma 2. *Let $X = (X_1, \dots, X_d)$ have i.i.d. coordinates, with $\mathbb{E}[X_1] = 0$, $\mathbb{E}[X_1^2] = 1$ and $\mathbb{E}[X_1^8] \leqslant \kappa^8$ for some $\kappa \geqslant 1$. Assume that $d \geqslant 2025\kappa^6$, define $u^* = (1/\sqrt{d}, \dots, 1/\sqrt{d})$ and let $\eta \in [45\kappa^3/\sqrt{d}, 1]$. Then, for every $v \in S^{d-1}$ such that $\langle u^*, v\rangle \geqslant 0$, one has*

$$\mathbb{P}\Big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X\rangle| \geqslant 0.2\max\{\eta, \|u^* - v\|_2\}\Big) \geqslant \frac{\eta}{70\,000\,\kappa^4}\,. \tag{32}$$

In particular, if X satisfies Assumption 4, then Assumption 3 holds for any $\eta \in [18K^3/\sqrt{d}, e^{-1}]$ with $c = 21\,000$.

Lemma 2 is a somewhat delicate result, so before discussing its implications we first explain the main idea behind its proof. The detailed proof may be found in Section 9.3.

We need to show that, conditionally on the fact that $|\langle u^*, X\rangle| \leqslant \eta$, the variable $\langle v, X\rangle$ fluctuates on a scale of order at least $\max\{\eta, \|u^* - v\|\}$. Since $\langle v, X\rangle = \langle u^*, v\rangle\langle u^*, X\rangle + \sqrt{1 - \langle u^*, v\rangle^2}\langle w, X\rangle$ with $\langle u^*, w\rangle = 0$, this means roughly speaking that the variables $\langle u^*, X\rangle$ and $\langle w, X\rangle$ behave as if they were independent. Of course, the main difficulty is that these variables are not in fact independent, except in the very special case where the vectors u^* and w have disjoint supports. In addition, Gaussian approximation on the vector $(\langle u^*, X\rangle, \langle w, X\rangle)$ fails in general since $w \in S^{d-1}$ is arbitrary.

We therefore need to show that $\langle v, X\rangle$ exhibits some variability under the event that $\langle u^*, X\rangle$ is small, in the absence of independence properties. The main idea to achieve this is to "perturb" the vector $X = (X_1, \dots, X_d)$ by randomly permuting its coordinates. Specifically, given a permutation $\sigma \in \mathfrak{S}_d$ of $\{1, \dots, d\}$, we let $X^\sigma = (X_{\sigma(1)}, \dots, X_{\sigma(d)})$. We introduce an additional source of randomness (besides X) by taking σ to be random, drawn uniformly over the symmetric group $\mathfrak{S}_d$, and independent of X. These transformations are useful thanks to the following properties:

1. The vector X^σ has the same distribution as X for a fixed σ, and thus also for random σ;
2. Permutations preserve $\langle u^*, X\rangle$, as $\langle u^*, X^\sigma\rangle = \frac{1}{\sqrt{d}} \sum_{j=1}^d X_{\sigma(j)} = \frac{1}{\sqrt{d}} \sum_{j=1}^d X_j = \langle u^*, X\rangle$;
3. Conditionally on X (for most values of X), the quantity $\langle v, X^\sigma\rangle = \sum_{j=1}^d v_j X_{\sigma(j)}$ fluctuates on the desired scale of $\max\{\eta, \|u^* - v\|\}$, as the random permutation σ varies.

Since the first claim (exchangeability) follows immediately from Assumption 4, the main step is to justify the third claim. We establish it by applying the Paley-Zygmund inequality, which reduces the task to lower-bounding one moment of $\langle v, X^\sigma\rangle$ (conditionally on X and with respect to random σ), and to upper-bounding a higher-order moment, ideally to conclude that they are of the same order of magnitude. In addition, one may explicitly evaluate the moments of even integer order, as this reduces to computations over symmetric polynomials in $X_1, \dots, X_d$. After suitable simplifications (exploiting that $\sum_{j=1}^d w_j = \sqrt{d}\langle u^*, w\rangle = 0$), we can show that this is indeed the case, provided that $X = (X_1, \dots, X_d)$ satisfies some symmetric conditions that do hold with high probability. We refer to Section 9.3 for more details on this proof.

We can now gather the conclusions of Lemmas 1 and 2 into the following statement, which is the main result of the present section.

Proposition 4. *Let $X = (X_1, \dots, X_d)$ satisfy Assumption 4, set $u^* = (1/\sqrt{d}, \dots, 1/\sqrt{d})$ and assume that $d \geqslant K^6$. Then X is $4K$-sub-exponential and (u^*, η, c)-regular with $c = 21\,000$ for any $\eta \in [18K^3/\sqrt{d}, e^{-1}]$.*

It then follows from Theorem 3 that, if $\theta^* = (B/\sqrt{d}, \dots, B/\sqrt{d})$, then the MLE behaves in a similar way as if the design was Gaussian as long as $B = O(\sqrt{d})$. Hence, in this direction, the "discrete" nature of the design has no impact, even for a moderately strong signal.

It is natural to ask if the sufficient condition $B = O(\sqrt{d})$ is also necessary to exhibit a Gaussian-like behavior. The following simple example shows that this is indeed the case.

Fact 2. *Let d be an odd integer, $X = (X_1, \dots, X_d)$ a Bernoulli design, and let Y given X follow the logit model with parameter $\theta^* = (B/\sqrt{d}, \dots, B/\sqrt{d})$ for some $B \geqslant \sqrt{d}$. Given an i.i.d. sample of size $n \geqslant 1$ from this distribution, if $n \leqslant 0.1 \exp(B/\sqrt{d})$ then $\mathbb{P}(\text{MLE exists}) \leqslant 0.1$.*

Proof. The proof is the same as that of Fact 1, except that the condition $|\langle\theta^*, X\rangle| \geqslant B$ therein is now replaced by $|\langle\theta^*, X\rangle| \geqslant B/\sqrt{d}$. Indeed, one has $|\langle\theta^*, X\rangle| = B|\sum_{j=1}^d X_j|/\sqrt{d} \geqslant B/\sqrt{d}$ since $\sum_{j=1}^d X_j$ is an odd integer. □

In other words, if $B \gg \sqrt{d}$ then some exponential dependence on B is again necessary for the MLE to exist. In particular, the regularity scale of $\eta \asymp 1/\sqrt{d}$ is indeed optimal for the Bernoulli design in the direction $u_d^* = (1/\sqrt{d}, \dots, 1/\sqrt{d})$.

Now, since $u_d^* = (1/\sqrt{d}, \dots, 1/\sqrt{d})$ is the most "well-spread" vector in S^{d-1}, it is perhaps tempting to conjecture that it is the "best" direction from the perspective of logistic regression, that is, the one with the smallest regularity scale η. If this were indeed the case, then for a "typical" direction $u^* \in S^{d-1}$ one would expect a regularity scale of $1/\sqrt{d}$ at best.

Interestingly, this is *not* the case, at least for the one-dimensional Assumption 2. It turns out that, for a "typical" direction $u^* \in S^{d-1}$, Assumption 2 is satisfied down to a smaller scale, of order $1/d$ instead of $1/\sqrt{d}$. This follows from a remarkable result of Klartag and Sodin [KS12], which states that for a typical direction $u = (u_1, \dots, u_d) \in S^{d-1}$, the distribution of the linear combination $\langle u, X \rangle = \sum_{j=1}^d u_j X_j$ approaches the Gaussian distribution at a rate of $1/d$, which is faster than the $1/\sqrt{d}$ rate for the normalized sum $\frac{1}{\sqrt{d}} \sum_{j=1}^d X_j$. We discuss the nature of this improvement and raise related open questions in Section 9.4.

4 Proof scheme and main lemmas

In this section, we describe the general scheme of proof that we use to establish Theorems 1, 3 and 4, as well as the main lemmas in the analysis.

4.1 Convex localization

We start with the lemma that is used to both establish existence of, and obtain risk bounds for, the MLE. It is based on a simple convex localization argument, which is purely deterministic. This reduction is general: the only properties that it uses, besides those explicitly stated in Lemma 3, are that $\widehat{L}_n, L$ are twice continuously differentiable, that $\widehat{L}_n$ is convex and that θ^* is a global minimizer of L.

Lemma 3. *Assume that there exists a positive-definite matrix $H \in \mathbb{R}^{d \times d}$ and real numbers $r_0, c_0, c_1, \nu > 0$ such that the following conditions hold:*

- $\|\nabla \widehat{L}_n(\theta^*)\|_{H^{-1}} \leqslant \nu$*;*
- *For every $\theta \in \mathbb{R}^d$ such that $\|\theta - \theta^*\|_H \leqslant r_0$, one has $\nabla^2 \widehat{L}_n(\theta) \succcurlyeq c_0 H$;*
- *For every $\theta \in \mathbb{R}^d$ such that $\|\theta - \theta^*\|_H \leqslant r_0$, one has $\nabla^2 L(\theta) \preccurlyeq c_1 H$.*

If $\nu < c_0 r_0 / 2$, then the empirical risk $\widehat{L}_n$ admits a unique global minimizer $\widehat{\theta}_n$, which satisfies

$$\big\|\widehat{\theta}_n - \theta^*\big\|_H \leqslant \frac{2\nu}{c_0} \qquad \text{and} \qquad L(\widehat{\theta}_n) - L(\theta^*) \leqslant \frac{2 c_1 \nu^2}{c_0^2}\,. \tag{33}$$

If in addition $\nu < c_0 r_0 / 4$, then for any $\widetilde{\theta}_n \in \mathbb{R}^d$ such that $\widehat{L}_n(\widetilde{\theta}_n) - \widehat{L}_n(\widehat{\theta}_n) < c_0 r_0^2 / 4$, one has

$$L(\widetilde{\theta}_n) - L(\theta^*) \leqslant \frac{c_1}{2} \big\|\widetilde{\theta}_n - \theta^*\big\|_H^2 \leqslant \max\left\{ \frac{8 c_1 \nu^2}{c_0^2}, \frac{2c_1}{c_0} \Big[\widehat{L}_n(\widetilde{\theta}_n) - \widehat{L}_n(\widehat{\theta}_n) \Big] \right\}. \tag{34}$$

Lemma 3 (proved in Section 8.1) reduces the proof of existence and risk bounds for the MLE (or approximate minimizers $\widetilde{\theta}_n$ of the empirical risk $\widehat{L}_n$ thanks to (34)) to two main components:

- a high-probability upper bound on the H^{-1}-norm $\|\nabla \widehat{L}_n(\theta^*)\|_{H^{-1}}$ of the empirical gradient at θ^*;
- a high-probability lower bound $\nabla^2 \widehat{L}_n(\theta) \succcurlyeq c_0 H$ on the Hessian of the empirical risk at θ, uniformly over all $\theta \in \Theta = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant r_0\}$.

The risk bound is then given by (33), while the condition for existence of $\widehat{\theta}_n$ is that $\nu < c_0 r_0 / 2$. In particular, the smaller $\nu = \nu_n$ and the larger r_0, the weaker the condition for existence of $\widehat{\theta}_n$.

Although the matrix H (and the corresponding parameters c_0, c_1, r_0, ν) from Lemma 3 can in principle be arbitrary, in order to obtain tight guarantees, a natural choice is to take H to be equivalent up to constant factors to $\nabla^2 L(\theta^*)$, the Hessian of the risk at θ^*, which coincides in the well-specified case with the Fisher information.

Indeed, in order to obtain sharp bounds we would like c_0, c_1 to be of constant order, and indeed in the Gaussian case these will be universal constants. Now by assumption one has $\nabla^2 L(\theta) \preccurlyeq c_1 H$ for all $\theta \in \Theta = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant r_0\}$, while $\nabla^2 \widehat{L}_n(\theta) \succcurlyeq c_0 H$ for any $\theta \in \Theta$. Now for large n, by the law of numbers $\nabla^2 \widehat{L}_n(\theta)$ should be close to its expectation $\mathbb{E}[\nabla^2 \widehat{L}_n(\theta)] = \nabla^2 L(\theta)$, so for the latter condition to hold with high probability, one should also have $\nabla^2 L(\theta) \succcurlyeq c_0 H$ for all $\theta \in \Theta$. This implies that H is equivalent to $\nabla^2 L(\theta^*)$, namely $c_0 H \preccurlyeq \nabla^2 L(\theta^*) \preccurlyeq c_1 H$. In addition, this constrains the domain Θ (namely the parameter r_0), which must be contained in the set

$$\Theta' = \left\{\theta \in \mathbb{R}^d : c_2^{-1} \nabla^2 L(\theta^*) \preccurlyeq \nabla^2 L(\theta) \preccurlyeq c_2 \nabla^2 L(\theta^*)\right\} \tag{35}$$

with $c_2 = c_1/c_0$.

It follows from these considerations that, in order to apply Lemma 3 effectively, a first step is to understand the behavior of the Hessian $\nabla^2 L(\theta)$ for $\theta \in \mathbb{R}^d$—both to set the matrix $H \approx \nabla^2 L(\theta^*)$, and to identify the largest possible region (35) where the conditions of Lemma 3 could be expected to hold.

By rotation-invariance of the Gaussian distribution, when $X \sim \mathsf{N}(0, I_d)$ the Hessian $\nabla^2 L(\theta)$ commutes with any linear isometry of $\mathbb{R}^d$ that fixes θ, and is therefore of the form

$$\nabla^2 L(\theta) = c_0(\|\theta\|) u u^\top + c_1(\|\theta\|)(I_d - u u^\top)\,, \tag{36}$$

where $u = \theta/\|\theta\|$ (for $\theta \neq 0$), and letting $G \sim \mathsf{N}(0,1)$ we have for $\beta \in \mathbb{R}^+$:

$$c_0(\beta) = \mathbb{E}[\sigma'(\beta G) G^2], \quad c_1(\beta) = \mathbb{E}[\sigma'(\beta G)]\,.$$

In addition, one may verify (see Lemma 25 in Section 8.1) that for some numerical constants c_0', c_0'', c_1', c_1'':

$$\frac{c_0'}{(\beta+1)^3} \leqslant c_0(\beta) \leqslant \frac{c_0''}{(\beta+1)^3}\,, \qquad \frac{c_1'}{\beta+1} \leqslant c_1(\beta) \leqslant \frac{c_1''}{\beta+1}\,. \tag{37}$$

We will therefore set H to be the matrix

$$H = \frac{1}{B^3} u^* u^{*\top} + \frac{1}{B}(I_d - u^* u^{*\top}), \qquad u^* = \frac{\theta^*}{\|\theta^*\|} \in S^{d-1}, \qquad B = \max(e, \|\theta^*\|)\,, \tag{38}$$

so that $c_0 H \preccurlyeq \nabla^2 L(\theta^*) \preccurlyeq c_1 H$ for some absolute constants c_0, c_1 for a Gaussian design.

In addition, it can be deduced from this characterization of $\nabla^2 L(\theta)$ that the region (35) (for large B and constant c_2) where the Hessian is equivalent to $\nabla^2 L(\theta^*)$ coincides up to constants with an ellipsoid of the form $\{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant r_0\}$, where $r_0 \asymp 1/\sqrt{B}$.

4.2 Upper bounds on the empirical gradient

We now consider the first ingredient in the application of Lemma 3, namely high-probability upper bounds on the H^{-1}-norm of the empirical gradient:

$$\|\nabla \widehat{L}_n(\theta^*)\|_{H^{-1}} = \left\| \frac{1}{n} \sum_{i=1}^n \nabla \ell(\theta^*, (X_i, Y_i)) \right\|_{H^{-1}} = \left\| \frac{1}{n} \sum_{i=1}^n \sigma(-Y_i \langle \theta^*, X_i \rangle) H^{-1/2} X_i \right\|. \tag{39}$$

We describe below our guarantees in the following three cases: (i) Gaussian design, well-specified model, (ii) regular design, well-specified model and (iii) regular design, misspecified model. (We note in passing that in order to control the gradient, we only require Assumptions 1 and 2.)

We start with the first case. A natural approach (which is essentially that of [OB21]) is to use that $\sigma_i = \sigma(-Y_i \langle \theta^*, X_i \rangle) \leqslant 1$ and that X_i is sub-Gaussian for each i, to deduce that the individual summands in (39) are H^{-1}-sub-Gaussian. By standard deviation bounds for sub-Gaussian vectors, this implies that for some constant $c > 0$, with probability at least $1 - e^{-t}$,

$$\|\nabla \widehat{L}_n(\theta^*)\|_{H^{-1}} \leqslant c\sqrt{\frac{\mathrm{Tr}(H^{-1}) + \|H^{-1}\|_{\mathrm{op}} t}{n}} \leqslant c\sqrt{\frac{Bd + B^3(t+1)}{n}}\,.$$

Unfortunately, this bound features a suboptimal dependence on the norm B. In order to improve it, the key observation is the following: if $Y_i\langle\theta^*, X_i\rangle \geqslant 0$, then the sigmoid σ_i is bounded as

$$\sigma_i = \sigma(-Y_i\langle\theta^*, X_i\rangle) = \sigma(-|\langle\theta^*, X_i\rangle|) \leqslant \exp(-|\langle\theta^*, X_i\rangle|)\,,$$

which is very small if $|\langle\theta^*, X_i\rangle|$ is large. On the other hand, if $Y_i\langle\theta^*, X_i\rangle < 0$, then the sigmoid is no longer small (specifically, $\frac{1}{2} \leqslant \sigma_i \leqslant 1$). However, *this configuration is highly unlikely if* $|\langle\theta^*, X_i\rangle|$ *is large*: indeed, using that the model is well-specified, one has

$$\mathbb{P}(Y_i\langle\theta^*, X_i\rangle < 0 | X_i) = \sigma(-|\langle\theta^*, X_i\rangle|) \leqslant \exp(-|\langle\theta^*, X_i\rangle|)\,. \tag{40}$$

Hence, the only remaining situation where σ_i may not be small is when $|\langle\theta^*, X_i\rangle|$ is upper-bounded; but since $\langle\theta^*, X_i\rangle \sim \mathsf{N}(0, \|\theta^*\|^2)$, the probability that $|\langle\theta^*, X_i\rangle| \lesssim 1$ is of order $1/B$, which is small when B is large.

From a technical standpoint, the considerations above allow us to obtain improved upper bounds (compared to those obtained by bounding $|\sigma_i| \leqslant 1$) on the moments of the random variables $\langle v, H^{-1/2}\nabla\ell(\theta^*, (X_i, Y_i))\rangle$ for $v \in S^{d-1}$, whose supremum is precisely the norm (39). Specifically, these random variables can be shown to satisfy the *sub-gamma* property [BLM13, §2.4], which we recall in Definition 6. Using a deviation bound for sub-gamma random vectors (Lemma 5), we deduce the following result, proved in Section 5.2:

Proposition 5. *Assume that X is Gaussian and the model is well-specified. Let H be the matrix defined in* (38)*. For any $t > 0$, if $n \geqslant 4B(d\log 5 + t)$ then with probability at least $1 - 2e^{-t}$,*

$$\big\|\nabla \widehat{L}_n(\theta^*)\big\|_{H^{-1}} \leqslant 27\sqrt{\frac{d+t}{n}}\,.$$

We now turn to the more general case of a regular design, but still assuming a well-specified model. Here the guarantees are quite similar to the Gaussian case, and the high-level argument sketched above remains valid. However, two important properties of the Gaussian distribution that we used in the proof of Proposition 5 no longer hold for general regular distributions: (1) linear marginals $\langle u, X\rangle$ and $\langle v, X\rangle$ in orthogonal directions $u, v \in S^{d-1}$ are independent, and (2) the distribution of $\langle u^*, X\rangle$ admits a bounded (by $\frac{1}{\sqrt{2\pi}}$) density. The lack of independence is handled by using that X is sub-exponential (leading to an additional $\log B$ factor); while to get around the lack of bounded density, we decompose the relevant expectations (that define the moments of the gradient) over a geometric grid of scales. Using these arguments to again show that gradients admit sub-gamma moments, we obtain the following bound, proved in Section 5.3.

Proposition 6. *Assume that X satisfies Assumptions 1 and 2 with parameters K such that $K\log B \geqslant 4$, u^*, $\eta = B^{-1}$ and $c \geqslant 1$, and that the model is well-specified. For any $t > 0$, if $n \geqslant B(d+t)$ then with probability at least $1 - 2e^{-t}$,*

$$\big\|\nabla \widehat{L}_n(\theta^*)\big\|_{H^{-1}} \leqslant c' \log B\sqrt{\frac{d+t}{n}}\,,$$

where $c' > 0$ is a constant that depends only on K and c.

We now conclude with the most general case we consider, where the design is regular but the model is no longer assumed to be well-specified. The fact that the model may be misspecified induces a significant change: the key bound (40) on the conditional probability of misclassification no longer holds. Given that if $Y_i\langle\theta^*, X_i\rangle < 0$, then $\sigma_i = \sigma(-Y_i\langle\theta^*, X_i\rangle) \in [1/2, 1]$ is of constant order, and that no bound on $\mathbb{P}(Y_i\langle\theta^*, X_i\rangle < 0|X_i)$ is available, it might be tempting to simply bound $|\sigma_i| \leqslant 1$, which as discussed above leads to a bound of order $\sqrt{(Bd + B^3t)/n}$.

As it happens, this bound is suboptimal and can be improved even in the misspecified case. The reason for this is that, if the parameter $\theta^* = \arg\min_{\theta\in\mathbb{R}^d} L(\theta)$ has a large norm B, then the (unconditional) probability $\mathbb{P}(Y\langle\theta^*, X\rangle < 0)$ of misclassification of θ^* must be small. The key result that expresses this intuition is Lemma 7, which shows that the probability of misclassification $\mathbb{P}(Y\langle\theta^*, X\rangle < 0)$ and the first moment $\mathbb{E}[|\langle u^*, X\rangle|\mathbf{1}(Y\langle\theta^*, X\rangle < 0)]$ are bounded in the general misspecified case in a similar way as in the well-specified case. This allows one to refine the naive bound of $\sqrt{(Bd + B^3t)/n}$ into a near-optimal bound of $\log(B)\sqrt{(d + Bt)/n}$.

Proposition 7. *Assume that X satisfies Assumptions 1 and 2 with parameters K and (u^*, B^{-1}, c), but not that the model is well-specified. For any $t > 0$, if $n \geqslant B(d + Bt)$, then with probability at least $1 - 2e^{-t}$,*

$$\big\|\nabla\widehat{L}_n(\theta^*)\big\|_{H^{-1}} \leqslant c' \log(B)\sqrt{\frac{d + Bt}{n}}\,,$$

where $c' > 0$ is a constant that depends only on K and c.

The proof of Proposition 7 may be found in Section 5.4.

4.3 Lower bounds on empirical Hessian matrices

We now turn to the second component of the proof scheme of Lemma 3, namely a high-probability lower bound on the Hessian of the empirical risk:

$$\widehat{H}_n(\theta) = \nabla^2\widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sigma'(\langle\theta, X_i\rangle)X_iX_i^\top\,, \tag{41}$$

where $\sigma'(s) = \{(1 + e^s)(1 + e^{-s})\}^{-1}$ for $s \in \mathbb{R}$, uniformly for θ in a neighborhood of θ^* that is as large as possible. Specifically, it follows from the discussion of Section 4.1 that an "ideal" guarantee would be of the form: for n large enough (depending on B, d, t),

$$\mathbb{P}\Big(\forall\theta \in \Theta,\ \widehat{H}_n(\theta) \succcurlyeq c_0 H\Big) \geqslant 1 - e^{-t}\,, \quad \text{where} \quad \Theta = \Big\{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant \frac{c_1}{\sqrt{B}}\Big\} \tag{42}$$

for some constants c_0, c_1 that should not depend (or weakly depend) on n, B, d, where the matrix H is defined in (38).

As is clear from the expression (41), the empirical Hessian matrix $\widehat{H}_n(\theta)$ only depends on $X_1, \dots, X_n$ and not on the labels $Y_1, \dots, Y_n$. Hence, the behavior of $\widehat{H}_n(\theta)$ depends on the distribution of X but not on the conditional distribution of Y given X. As such, there is no distinction between the well-specified and misspecified cases, and we only have to consider two cases: Gaussian design and regular design. While the Gaussian design is a special case of regular design, we consider it separately because in this case we obtain sharper guarantees, involving universal constants rather than poly-logarithmic factors in B.

We start with the general case of a regular design, because its analysis is actually simpler than that of the Gaussian case. Theorem 5 below provide an almost optimal uniform lower bound on the empirical Hessian of the form (42), up to logarithmic factors in B. We note in passing that this control on the Hessian only requires Assumptions 1 and 3, while Assumption 2 was used in the control of the gradient discussed in Section 4.2.

Theorem 5. *Let X be a random vector satisfying Assumptions 1 and 3 with parameter $K \geqslant e$, $u^* = \theta^*/\|\theta^*\|$, $\eta = 1/B$ and $c \geqslant 1$. There exist constants $c_1, c_2, c_3 > 0$ that depend only on c and K for which the following holds: for any $t > 0$, if*

$$n \geqslant c_1 B\big(\log(B)d + t\big)$$

then with probability at least $1 - e^{-t}$,

$$\widehat{H}_n(\theta) \succcurlyeq c_2 H \quad \text{for every } \theta \in \mathbb{R}^d \text{ such that } \quad \|\theta - \theta^*\|_H \leqslant \frac{c_3}{\log(B)\sqrt{B}} .$$

Theorem 5 is proved in Section 6.1, and we discuss here the main ideas of the proof. The key observation is that a certain property of the dataset implies the desired behavior (42). Specifically, the first step is to notice that if $X_1, \dots, X_n$ satisfy, for some constants c_0, c_1,

$$\inf_{\substack{u,v \in S^{d-1} \\ \|u-u^*\| \leqslant c_0/B,\, \langle u^*, v\rangle \geqslant 0}} \left[\sum_{i=1}^n \mathbf{1}\left\{ |\langle u, X_i\rangle| \leqslant \frac{c_1}{B}\, ; \, |\langle v, X_i\rangle| \geqslant \frac{\max\{B^{-1}, \|u^* - v\|\}}{c_1} \right\} \right] \geqslant \frac{n}{2c_1 B}, \quad (43)$$

then $\widehat{H}_n(\theta) \succcurlyeq c_2 H$ for every $\theta \in \mathbb{R}^d$ such that $\|\theta - \theta^*\|_H \leqslant c_3/\sqrt{B}$, for some constants c_2, c_3 that depend on c_0, c_1. This follows from properties of the function σ' and the structure of H.

It then remains to establish that condition (43) holds with high probability over the random draw of $X_1, \dots, X_n$. To achieve this, observe first that condition (43) is essentially a variant of Assumption 3, with two differences: (i) it holds for any $u \in S^{d-1}$ such that $\|u - u^*\| \leqslant c_0/B$, rather than just for $u = u^*$, and (ii) it holds for the random sample $X_1, \dots, X_n$, rather than for the distribution P_X.

Condition (43) (or rather, a slightly weaker version with additional $\log B$ factors) is thus established in two steps. First, we show that Assumption 3 on P_X extends to all directions $u \in S^{d-1}$ such that $\|u - u^*\| \leqslant c_3/(B \log B)$. Second, we show that this condition on P_X is stable under random sampling with high probability, by using that the class of events in (43) is a Vapnik-Chervonenkis (VC) class with VC dimension at most $O(d)$, and then applying a uniform lower bound on the empirical frequencies of a VC class of sets.

Theorem 5 applies to the general regular case, and in particular to the special case of a Gaussian design. (That the Gaussian design satisfies Assumption 3 may be verified using independence of orthogonal linear marginals, or alternatively follows from Proposition 3.) In fact, the identification of condition (43) as a structural property implying the "right" behavior of the empirical Hessian in the Gaussian case is what motivates the definition of Assumption 3.

At the same time, it should be noted that the guarantees of Theorem 5 feature additional $\log B$ factors, compared to "ideal" guarantees that would lead to Theorem 1 in the Gaussian case. In particular, the (sufficient) condition on the sample size n from Theorem 5 is stronger by a $\log B$ factor than the necessary condition presented at the end of Theorem 1.

We address this suboptimality of Theorem 5 in the Gaussian case in Theorem 6 below, which provides an optimal uniform lower bound on empirical Hessian matrices.

Theorem 6. *Assume that $X \sim \mathsf{N}(0, I_d)$. For any $t > 0$, if $n \geqslant 320000B(d + t)$ then with probability at least $1 - 2e^{-t}$,*

$$\widehat{H}_n(\theta) \succcurlyeq \frac{1}{1000} H \quad \text{for every } \theta \in \mathbb{R}^d \text{ such that } \quad \|\theta - \theta^*\|_H \leqslant \frac{1}{100\sqrt{B}} .$$

The proof of Theorem 6 is provided in Section 6.3. The proof of this sharp result in the Gaussian case happens to be significantly more delicate than that of the more general, but less precise, Theorem 5. The reason for this is that the techniques used to establish Theorem 5 (specifically, the use of Vapnik-Chervonenkis arguments) inherently lead to additional logarithmic factors, hence the proof of Theorem 6 requires a fundamentally different approach.

In order to obtain optimal results in the Gaussian case, we rely instead on the so-called PAC-Bayes method, which involves controlling a "smoothed" version of the process of interest. The use of this technique in non-asymptotic statistics was pioneered by Catoni and co-authors [AC11, Cat16], and has found several applications to the non-asymptotic study of random matrices [Oli16, Mou22, Zhi24]. In the logistic regression setting we consider, the presence of nonlinear terms (due to the sigmoid σ') in the empirical Hessian (41) is an additional source of difficulty, which requires new technical ideas. In particular, instead of applying the PAC-Bayes method to the process of interest, and later controlling the difference between the smoothed version of the process and the process itself, we apply it to an auxiliary process whose smoothed version is (a bound on) the process of interest. In addition, the smoothing distributions we employ differ from the isotropic Gaussian distributions that have been used in previous works, in two ways: first, they exhibit an anisotropic structure, and second, one of their component is far from being Gaussian. We refer to Section 6.3 for more details on this point.

5 Proofs of upper bounds on the empirical gradient

The concentration of the empirical gradient derives from the fact that the vectors $H^{-1/2}\nabla\widehat{L}_n(\theta^*)$ are sub-gamma, using a classical result recalled in Section 5.1. To prove this, we first use the definition of H. Let $v \in S^{d-1}$ and let $w \in S^{d-1}$ be such that $\langle u^*, w\rangle = 0$ and $v - \langle u^*, v\rangle u^* = \|v - \langle u^*, v\rangle u^*\| w$, then

$$\langle v, H^{-1/2}\nabla\widehat{L}_n(\theta^*)\rangle \leqslant B^{3/2}|\langle u^*, \nabla\widehat{L}_n(\theta^*)\rangle| + B^{1/2}|\langle w, \nabla\widehat{L}_n(\theta^*)\rangle| \,. \tag{44}$$

Then, for any $v \in S^{d-1}$, we have by definition of the empirical gradient

$$\langle v, \nabla\widehat{L}_n(\theta^*)\rangle = \frac{1}{n}\sum_{i=1}^{n}\langle v, \nabla\ell(\theta^*, Z_i)\rangle \,.$$

Thus, by Lemma 5, it is sufficient to prove that the variables $\langle v, \nabla\ell(\theta^*, Z_i)\rangle$ are sub-gamma. These random variables are centered, thus, by Point 4 in Lemma 35, this property can be obtained by proving a proper upper bound on the moments of these random variables. This upper bound will be proved using the following lemma.

Lemma 4. *Let $Z = (X, Y)$ denote a random variable taking value in $\mathbb{R}^d \times \{-1, 1\}$. Let $u^* = \theta^*/\|\theta^*\|$ and let $p \geqslant 2$. For any $v \in S^{d-1}$,*

$$\mathbb{E}[|\langle v, \nabla\ell(\theta^*, Z)\rangle|^p] \leqslant \mathbb{E}\big[\big(\exp(-|\langle\theta^*, X\rangle|) + \mathbf{1}\{Y\langle\theta^*, X\rangle < 0\}\big)|\langle v, X\rangle|^p\big] \,. \tag{45}$$

Moreover, when the model is well-specified, we have

$$\mathbb{E}[|\langle v, \nabla\ell(\theta^*, Z)\rangle|^p] \leqslant 2\mathbb{E}\big[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p\big] \,. \tag{46}$$

Proof. As $\nabla\ell(\theta^*, Z) = -Y\sigma(-Y\langle\theta^*, X\rangle)X$, for any $v \in S^{d-1}$

$$|\langle v, \nabla\ell(\theta^*, Z)\rangle| = \sigma(-Y\langle\theta^*, X\rangle)|\langle v, X\rangle| \,.$$

Moreover, as $\sigma(-|x|) \leqslant \exp(-|x|)$ and $\sigma(x) \leqslant 1$,

$$\sigma(-Y\langle\theta^*, X\rangle) \leqslant \exp(-|\langle\theta^*, X\rangle|) + \mathbf{1}\{Y\langle\theta^*, X\rangle < 0\} \,.$$

This proves (45) since, for any $p \geqslant 1$, we have $\sigma^p \leqslant \sigma$.

For (46), when the model is well-specified, we have

$$\mathbb{E}[\mathbf{1}\{Y\langle\theta^*, X\rangle < 0\}\,|X] = \sigma(-|\langle\theta^*, X\rangle|) \leqslant \exp(-|\langle\theta^*, X\rangle|)\,. \tag{47}$$

Hence, (46) follows by plugging this bound into (45). □

5.1 Concentration of sub-gamma random vectors

The concentration of empirical gradients are based on the following classical concentration inequality for sub-gamma random vectors.

Lemma 5. *Let $Z_1, \dots, Z_n$ denote independent random vectors and let V denote a linear subspace of $\mathbb{R}^d$. Assume that, for any $v \in S^{d-1} \cap V$, $\langle v, Z_i\rangle$ is (ν^2, K) sub-gamma (see Definition 6). Then, for any $t > 0$,*

$$\mathbb{P}\Big(\sup_{v\in V\cap S^{d-1}} \frac{1}{n}\sum_{i=1}^n \langle v, Z_i\rangle > 2\nu\sqrt{\frac{2(d\log 5 + t)}{n}} + 2K\frac{d\log 5 + t}{n}\Big) \leqslant \exp(-t)\,.$$

Proof. By Point 3 in Lemma 35, the random variables

$$\frac{1}{n}\sum_{i=1}^n \langle v, Z_i\rangle, \quad v \in V\cap S^{d-1},$$

are $(\nu^2/n, K/n)$ sub-gamma. Thus, by Bernstein's inequality, recalled in point 2 of Lemma 35, for any $v \in V \cap S^{d-1}$ and any $t > 0$, it holds

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \langle v, Z_i\rangle > \nu\sqrt{\frac{2t}{n}} + K\frac{t}{n}\Big) \leqslant \exp(-t)\,. \tag{48}$$

To make this bound uniform and conclude the proof, we use an ε-net argument. Let N denote a maximal set of $1/2$-separated points in the unit ball B_V for the Euclidean norm in V. Then each point in B_V is at distance at most $1/2$ of a point in N, so, for every $v \in B_V$, there exists $v' \in N$ such that $\|v - v'\| \leqslant 1/2$. Therefore, for any $v \in B_V$,

$$\frac{1}{n}\sum_{i=1}^n \langle v, Z_i\rangle = \frac{1}{n}\sum_{i=1}^n \langle v', Z_i\rangle + \frac{1}{n}\sum_{i=1}^n \langle v - v', Z_i\rangle \leqslant \max_{v'\in N}\frac{1}{n}\sum_{i=1}^n \langle v', Z_i\rangle + \frac{1}{2}\sup_{v\in B_V}\frac{1}{n}\sum_{i=1}^n \langle v, Z_i\rangle\,.$$

As this holds for any $v \in B_V$, it shows that

$$\sup_{v\in B_V}\frac{1}{n}\sum_{i=1}^n \langle v, Z_i\rangle \leqslant 2\max_{v'\in N}\frac{1}{n}\sum_{i=1}^n \langle v', Z_i\rangle\,.$$

Therefore, using a union bound, for any $t > 0$,

$$\mathbb{P}\Big(\sup_{v\in B_V}\frac{1}{n}\sum_{i=1}^n \langle v, Z_i\rangle > 2\nu\sqrt{\frac{2t}{n}} + 2K\frac{t}{n}\Big) \leqslant \mathbb{P}\Big(\max_{v\in N}\frac{1}{n}\sum_{i=1}^n \langle v, Z_i\rangle > \nu\sqrt{\frac{2t}{n}} + K\frac{t}{n}\Big) \leqslant |N|\exp(-t)\,.$$

Now by [Ver18, Lemma 4.2.13], we have $|N| \leqslant 5^d$, therefore the last inequality applied with $t' = d\log 5 + t$ shows the result. □

5.2 Proof of Proposition 5 (Gaussian design, well-specified model)

By (46), we have to bound the random variables $\mathbb{E}[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p]$ in the case where the design X is Gaussian. We start with the simplest case where $\|\theta^*\| < e$, so $B = e$. In this case, we use that $\exp(-|\langle\theta^*, X\rangle|) \leqslant 1$ to say that, for any $v \in S^{d-1}$, we have

$$\mathbb{E}[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p] \leqslant \mathbb{E}\big[|\langle v, X\rangle|^p\big] = \frac{\sqrt{2}^p}{\sqrt{\pi}}\Gamma\left(\frac{p+1}{2}\right) \leqslant \frac{p!}{\sqrt{\pi}}\,. \tag{49}$$

By Point 4 in Lemma 35, the variable $\langle v, \nabla\ell(\theta^*, Z_i)\rangle$ is thus $(2/\sqrt{\pi}, 1)$ sub-gamma. Hence, by Lemma 5, for any $t > 0$, with probability larger than $1 - \exp(-t)$

$$\forall v \in S^{d-1}, \qquad \langle v, \nabla\widehat{L}_n(\theta^*)\rangle \leqslant 2\left(\frac{2}{(\pi)^{1/4}}\sqrt{\frac{d\log 5 + t}{n}} + \frac{d\log 5 + t}{n}\right) \leqslant 6\sqrt{\frac{d+t}{n}}\,, \tag{50}$$

where the last inequality holds as $n \geqslant 4(d\log 5 + t)$. Taking the supremum over $v \in S^{d-1}$ and using that $\|\nabla\widehat{L}_n(\theta^*)\|_{H^{-1}} \leqslant e^{3/2}\|\nabla\widehat{L}_n(\theta^*)\|$ and $6e^{3/2} \leqslant 27$ gives the desired bound. Assume now that $\|\theta^*\| \geqslant e$, so $\|\theta^*\| = B$. For any $k \geqslant 0$, using that the density of $\langle u^*, X\rangle \sim \mathsf{N}(0, 1)$ is upper-bounded by $1/\sqrt{2\pi}$, we get

$$\begin{aligned}\mathbb{E}\big[\exp(-|\langle\theta^*, X\rangle|)|\langle u^*, X\rangle|^k\big] &= \mathbb{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle u^*, X\rangle|^k\big] \\ &\leqslant \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}|x|^k\exp(-B|x|)\mathrm{d}x = \sqrt{\frac{2}{\pi}}\frac{k!}{B^{k+1}}\,.\end{aligned} \tag{51}$$

Thus, (51) and (46) show that

$$\mathbb{E}[|\langle u^*, \nabla\ell(\theta^*, Z)\rangle|^p] \leqslant \frac{2\sqrt{2}}{B^3\sqrt{\pi}}\frac{p!}{B^{p-2}}\,.$$

This shows, by Bernstein's inequality recalled in (48) that, for any $t > 0$, with probability larger than $1 - \exp(-t)$

$$|\langle u^*, \nabla\widehat{L}_n(\theta^*)\rangle| \leqslant \frac{1}{B^{3/2}}\sqrt{\frac{t}{n}}\left(2\left(\frac{8}{\pi}\right)^{1/4} + \sqrt{\frac{Bt}{n}}\right) \leqslant \frac{3}{B^{3/2}}\sqrt{\frac{t}{n}}\,, \tag{52}$$

where the last inequality holds because $n \geqslant 4Bt$. Now let $w \in S^{d-1}$ such that $\langle w, u^*\rangle = 0$, the Gaussian random variables $\langle\theta^*, X\rangle$ and $\langle w, X\rangle$ are independent, so

$$\mathbb{E}[\exp(-|\langle\theta^*, X\rangle|)|\langle w, X\rangle|^p] = \mathbb{E}[\exp(-|\langle\theta^*, X\rangle|)]\mathbb{E}[|\langle w, X\rangle|^p]\,.$$

We bound the first term in the right-hand side with (51) with $k = 0$ and the second one with (49) to get

$$\mathbb{E}[\exp(-|\langle\theta^*, X\rangle|)|\langle w, X\rangle|^p] \leqslant \frac{2}{\pi}\frac{p!}{B}\,.$$

By Point 4 in Lemma 35, the vectors $\langle v, \nabla\ell(\theta^*, Z_i)\rangle$ are $(8/(\pi B), 1)$ sub-gamma. Hence, by Lemma 5, for any $t > 0$, with probability larger than $1 - \exp(-t)$, simultaneously for any $w \in S^{d-1}$ such that $\langle w, u^*\rangle = 0$,

$$\langle w, \nabla\widehat{L}_n(\theta^*)\rangle \leqslant 8\sqrt{\frac{d\log 5 + t}{\pi Bn}} + \frac{2(d\log 5 + t)}{n} \leqslant 6\sqrt{\frac{d+t}{Bn}}\,. \tag{53}$$

Plugging (52) and (53) into (44) concludes the proof of the second part of the proposition.

5.3 Proof of Proposition 6 (regular design, well-specified model)

We now prove Proposition 6, which is a deviation bound on the empirical gradient when the model is still well-specified, but the design is no longer Gaussian and instead satisfies Assumptions 1 and 2 with parameters u^*, $\eta = B^{-1}$ and $c \geqslant 1$.

Since the model is well-specified, by (46), we have to bound $\mathbb{E}[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p]$ when the design X is regular to prove the result.

Bounds on moments for regular designs. We start with the following bound on moments.

Lemma 6. *Let $\theta^* \in \mathbb{R}^d \setminus \{0\}$, $u^* = \theta^*/\|\theta^*\|$ and $B = \max(e, \|\theta^*\|)$. Suppose that X satisfies Assumptions 1 with parameter $K > 0$ and 2 with parameters $\eta = 1/B$ and $c \geqslant 1$. Then, for any $p \geqslant 0$, for any $v \in S^{d-1}$,*

$$\mathbb{E}\big[\exp(-|\langle\theta^*, X\rangle|)|\langle u^*, X\rangle|^p\big] \leqslant \frac{9c}{B}\left(\frac{K\log(B)}{2B}\right)^p p!\,. \tag{54}$$

$$\mathbb{E}\big[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p\big] \leqslant \frac{5ec}{B}\left(\frac{K\log(B)}{2}\right)^p p!\,. \tag{55}$$

Proof of Lemma 6. By Assumption 1, $\|\langle v, X\rangle\|_{\psi_1} \leqslant K$, so by Definition 5,

$$\mathbb{E}[|\langle v, X\rangle|^p] \leqslant \frac{K^p}{(2e)^p}p^p \leqslant \left(\frac{K}{2}\right)^p p!\,.$$

This proves the result in the case where $\|\theta^*\| \leqslant e$. Therefore, in the remaining of the proof, we assume that $\|\theta^*\| > e$, so $B = \|\theta^*\| > e$. Since X satisfies Assumption 2, for any $b > 0$

$$\begin{aligned}
\mathbb{E}[\exp(-b|\langle u^*, X\rangle|)] &\leqslant \mathbb{P}\left(|\langle u^*, X\rangle| \leqslant \frac{1}{B}\right) + \sum_{k\geqslant 0}\exp\left(-\frac{b2^k}{B}\right)\mathbb{P}\left(|\langle u^*, X\rangle| \leqslant \frac{2^{k+1}}{B}\right) \\
&\leqslant \frac{c}{B}\left(1 + \sum_{k\geqslant 0}2^{k+1}\exp\left(-\frac{b2^k}{B}\right)\right) \\
&\leqslant \frac{c}{B}\left(1 + 4\int_{1/2}^{+\infty}\exp\left(-\frac{bt}{B}\right)\mathrm{d}t\right) \leqslant \frac{c}{B}\left(1 + \frac{4B}{b}\right).
\end{aligned}$$

This yields

$$\mathbb{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle u^*, X\rangle|^p\big] \leqslant \sup_{t>0}\{t^p e^{-Bt/2}\}\mathbb{E}\left[\exp\left(-\frac{B}{2}|\langle u^*, X\rangle|\right)\right] \leqslant \left(\frac{2}{B}\right)^p\frac{9c}{B}p!\,.$$

This proves (54). For (55), Hölder's inequality implies that for any $\nu \in (0,1)$,

$$\begin{aligned}
\mathbb{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle v, X\rangle|^p\big] &\leqslant \mathbb{E}[|\langle v, X\rangle|^{p/\nu}]^\nu\,\mathbb{E}\big[\exp\big(-B|\langle u^*, X\rangle|\big)\big]^{1-\nu} \\
&\leqslant \left(\frac{K}{2\nu}\right)^p p!\left(\frac{5c}{B}\right)^{1-\nu}.
\end{aligned}$$

Letting $\nu = 1/\log(B)$,

$$\mathbb{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle v, X\rangle|^p\big] \leqslant \frac{5ec}{B}\left(\frac{K\log(B)}{2}\right)^p p!\,.$$

This concludes the proof of (55). □

Conclusion of the proof. By (54) and Bernstein's inequality, we deduce that, for any $t > 0$, with probability larger than $1 - 2e^{-t}$,

$$|\langle u^*, \nabla \widehat{L}_n(\theta^*)\rangle| \leqslant \frac{K \log B}{B^{3/2}} \sqrt{\frac{t}{n}} \left(\sqrt{\frac{9c}{2}} + \sqrt{\frac{Bt}{4n}} \right) \leqslant 3.2 \frac{K \log B}{B^{3/2}} \sqrt{\frac{ct}{n}}, \tag{56}$$

where the last inequality follows from $n \geqslant 4Bt$.

From (55) and Lemma 5, with probability larger than $1 - e^{-t}$, for any $w \in S^{d-1}$ such that $\langle u^*, w\rangle = 0$,

$$\langle w, \nabla \widehat{L}_n(\theta^*)\rangle \leqslant \frac{5K \log B}{\sqrt{B}} \sqrt{\frac{d+t}{n}} \left(\sqrt{c} + \sqrt{\frac{B(d \log 5 + t)}{4n}} \right) \leqslant \frac{6K \log B}{\sqrt{B}} \sqrt{\frac{c(d+t)}{n}}.$$

Plugging this upper bound and (56) into (44) concludes the proof of the proposition.

5.4 Proof of Proposition 7 (regular design, misspecified model)

We now turn to the proof of Proposition 7, which provides a deviation bound on the empirical gradient in the misspecified case. Specifically, X satisfies Assumptions 1 and 2 and the model might not be well-specified. The parameter θ^* is defined using the joint distribution of $Z = (X, Y)$ by

$$\theta^* = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} L(\theta), \quad \text{where} \quad L(\theta) = \mathbb{E}[\ell(\theta, Z)].$$

Recall that, from Lemma 4, the main task is to bound, for all $p \geqslant 2$, the expectations

$$\mathbb{E}\big[(\exp(-|\langle \theta^*, X\rangle|) + \mathbf{1}\{Y\langle \theta^*, X\rangle < 0\})|\langle v, X\rangle|^p\big].$$

The first expectation is bounded using Lemma 6. Therefore, we focus in this proof on the second expectation

$$\mathbb{E}\big[\mathbf{1}\{Y\langle \theta^*, X\rangle < 0\}\,|\langle v, X\rangle|^p\big].$$

The main problem here is that (47) does not hold and has to be extended. When the model might not be well-specified, the control of this last expectation is slightly worse than the one provided in Lemma 6 for the first expectation, yielding the extra $\sqrt{B}$ in front of $\sqrt{t}$.

Bounds on the first moments. In this section, we bound the expectation in the case where $p = 0$ and $p = 1$. This is a key step toward the bound for $p = 2$ and then for general p.

Lemma 7. *Let $\theta^* \in \mathbb{R}^d$ such that $\|\theta^*\| \geqslant e$ and let $u^* = \theta^*/\|\theta^*\|$. Suppose that X satisfies Assumption 2 with parameters (u^*, B^{-1}, c). Then,*

$$\mathbb{E}\big[|\langle \theta^*, X\rangle|\mathbf{1}(Y\langle \theta^*, X\rangle < 0)\big] \leqslant \frac{6c}{B^2}; \tag{57}$$

$$\mathbb{P}(Y\langle \theta^*, X\rangle < 0) \leqslant \frac{3.21c}{B}. \tag{58}$$

Remark 2. Notice that the second bound also holds when $\|\theta^*\| \leqslant e$ as it is trivial then and the first one also becomes trivial (and therefore holds) in this case as soon as $c \geqslant e^2/6 \approx 1.23$.

Proof. Since L is minimized in θ^*, one has $\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=1} L(t\theta^*) = 0$. Hence

$$\begin{aligned} 0 &= \mathbb{E}\big[Y\langle \theta^*, X\rangle \sigma(-Y\langle \theta^*, X\rangle)\big] \\ &= \mathbb{E}\Big[|\langle \theta^*, X\rangle|\sigma\big(-|\langle \theta^*, X\rangle|\big)\mathbf{1}(Y\langle \theta^*, X\rangle \geqslant 0) - |\langle \theta^*, X\rangle|\sigma\big(|\langle \theta^*, X\rangle|\big)\mathbf{1}(Y\langle \theta^*, X\rangle < 0)\Big]. \end{aligned}$$

Now, using that $\sigma(t) = 1 - \sigma(-t)$, we obtain:

$$\begin{aligned}0 &= \mathbb{E}\Big[|\langle\theta^*, X\rangle|\Big\{\sigma(-|\langle\theta^*, X\rangle|)\big[1 - \mathbf{1}(Y\langle\theta^*, X\rangle < 0)\big] - \big[1 - \sigma(-|\langle\theta^*, X\rangle|)\big]\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\Big\}\Big] \\ &= \mathbb{E}\Big[|\langle\theta^*, X\rangle|\big\{\sigma(-|\langle\theta^*, X\rangle|) - \mathbf{1}(Y\langle\theta^*, X\rangle < 0)\big\}\Big],\end{aligned}$$

which writes

$$\mathbb{E}\big[|\langle\theta^*, X\rangle|\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\big] = \mathbb{E}\big[|\langle\theta^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big]. \tag{59}$$

By (54) applied with $p = 1$, this shows (57). Moreover, as $1 \leqslant |\langle\theta^*, X\rangle| + \mathbf{1}(|\langle\theta^*, X\rangle| \leqslant 1)$, we have

$$\begin{aligned}\mathbb{P}(Y\langle\theta^*, X\rangle < 0) &\leqslant \mathbb{E}\big[|\langle\theta^*, X\rangle|\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\big] + \mathbb{E}\big[\mathbf{1}(|\langle\theta^*, X\rangle| \leqslant 1)\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\big] \\ &\leqslant \mathbb{E}\big[|\langle\theta^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big] + \mathbb{P}(|\langle\theta^*, X\rangle| \leqslant 1).\end{aligned}$$

Bounding the first term with (57) and the second using Assumption 2 concludes the proof. □

Bounds on the second moments. In this paragraph, we bound the expectation of interest for $p = 2$. We deduce an upper bound on the variance of the gradients.

Lemma 8. *Let $\theta^* \in \mathbb{R}^d$ of direction $u^* \in S^{d-1}$. Suppose that X satisfies Assumptions 1 and 2 with parameters $K \geqslant e$, (u^*, B^{-1}, c). Then, for any $v \in S^{d-1}$,*

$$\begin{aligned}\mathbb{E}\big[\mathbf{1}\{Y\langle\theta^*, X\rangle < 0\}\langle u^*, X\rangle^2\big] &\leqslant \frac{6cK\log(KB^2)}{B^2}. \\ \mathbb{E}\big[\mathbf{1}\{Y\langle\theta^*, X\rangle < 0\}\langle v, X\rangle^2\big] &\leqslant \frac{\max\{4e^2, 3.21cK^2\log^2 B\}}{4eB}.\end{aligned}$$

Proof. We start with the second inequality. As $\mathbb{E}[\langle v, X\rangle^2] = 1$, the left-hand side is smaller than 1 while the right-hand side is at least 1 if $B = e$. Therefore, we can assume that $\|\theta^*\| = B > e$.

By (58), $\mathbb{P}(Y\langle\theta^*, X\rangle < 0) \leqslant \min\{1, 3.21c/B\}$. Hence, by the first part of Lemma 9,

$$\mathbb{E}\big[\mathbf{1}\{Y\langle\theta^*, X\rangle < 0\}\langle v, X\rangle^2\big] \leqslant \frac{3.21cK^2\log^2(B)}{4eB},$$

which shows the second inequality since $B \geqslant e$.

For the first inequality, when $\|\theta^*\| < e$, the upper bound is larger than 1 while

$$\mathbb{E}\big[\mathbf{1}\{Y\langle\theta^*, X\rangle < 0\}\langle u^*, X\rangle^2\big] \leqslant \mathbb{E}[\langle u^*, X\rangle^2] = 1,$$

so the inequality holds in this case. Hence, we may assume that $\|\theta^*\| \geqslant e$. In this case, by (57),

$$\mathbb{E}\big[|\langle u^*, X\rangle|\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\big] \leqslant \frac{6c}{B^2}.$$

Thus, applying the second part of Lemma 9 below with $U = |\langle u^*, X\rangle|\mathbf{1}(Y\langle\theta^*, X\rangle < 0)$ and $V = |\langle u^*, X\rangle|$, we get

$$\mathbb{E}\big[\langle u^*, X\rangle^2\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\big] = \mathbb{E}[UV] \leqslant \frac{6c}{B^2}K\log\left(e \vee \frac{KB^2}{6c}\right) \leqslant \frac{6cK\log(KB^2)}{B^2}. \qquad \square$$

Note that, together with Lemma 6 and (45), Lemma 8 shows that, for any $v \in S^{d-1}$,

$$\mathbb{E}[\langle u^*, \nabla\ell(\theta^*, Z)\rangle^2] \leqslant \frac{7.1cK\log(KB^2)}{B^2}. \tag{60}$$

$$\mathbb{E}[\langle v, \nabla\ell(\theta^*, Z)\rangle^2] \leqslant \frac{cK^2\log^2 B}{B}. \tag{61}$$

Conclusion of the proof. The concentration of the gradients $\langle v, \nabla \widehat{L}_n(\theta^*)\rangle$ now follows from general facts on sub-exponential random variables recalled in Lemma 35. Recall that, for any $v \in S^{d-1}$,

$$\langle v, \nabla \widehat{L}_n(\theta^*)\rangle = \frac{1}{n}\sum_{i=1}^n \langle v, \nabla \ell(\theta^*, Z_i)\rangle$$

The random variables $\langle u^*, \nabla \ell(\theta^*, Z_i)\rangle$ are centered, their variance is by (60) bounded from above by $7.1cK\log(KB^2)/B^2 \leqslant C^2K^2(\log B)/B^2 = \nu^2$, where $C^2 = 14.2c(\log K)/K$. Moreover, as

$$|\langle u^*, \nabla \ell(\theta^*, Z)\rangle| = \sigma(-Y\langle \theta^*, X\rangle)|\langle u^*, X\rangle| \leqslant |\langle u^*, X\rangle|\,,$$

they also satisfy by Assumption 1, $\|\langle u^*, \nabla \ell(\theta^*, Z)\rangle\|_{\psi_1} \leqslant K$. Hence, by Point 6 in Lemma 35, they are (ν^2, K') sub-gamma, with

$$\nu^2 = \frac{C^2K^2\log(B)}{B^2}\,, \qquad K' = \max(e\nu, K)\log\left(\frac{B\max(e\nu, K)}{CK}\right) \leqslant c'K\log(B)\,, \tag{62}$$

where c' denote a function of c and K whose value may change from line to line. Therefore, by Lemma 35, for any $t > 0$, with probability larger than $1 - 2\exp(-t)$,

$$|\langle u^*, \nabla \widehat{L}_n(\theta^*)\rangle| \leqslant \frac{c'K}{B}\sqrt{\frac{t}{n}}\left(\sqrt{\log B} + B\log B\sqrt{\frac{t}{n}}\right) \leqslant c'\frac{K\log B}{B}\sqrt{\frac{t}{n}}\,. \tag{63}$$

In the last inequality, we used that $n \geqslant 4B^2t$.

Now, for any $v \in S^{d-1}$, the random variables $\langle v, \nabla \ell(\theta^*, Z_i)\rangle$ are centered, with variance bounded from above by $cK^2\log^2(B)/B$ from (61). Moreover, as

$$|\langle v, \nabla \ell(\theta^*, Z)\rangle| = \sigma(-Y\langle \theta^*, X\rangle)|\langle v, X\rangle| \leqslant |\langle v, X\rangle|\,,$$

they also satisfy by Assumption 1, $\|\langle v, \nabla \ell(\theta^*, Z)\rangle\|_{\psi_1} \leqslant K$. Hence, by Point 6 in Lemma 35, they are (ν^2, K') sub-gamma, with

$$\nu^2 = \frac{cK^2\log(B)^2}{B}\,, \qquad K' = \max(e\nu, K)\log\left(\frac{\sqrt{B}\max(e\nu, K)}{\sqrt{c}K\log(B)}\right) \leqslant c'K\log(B)\,.$$

Therefore, by Lemma 5, for any $t > 0$ and any $w \in S^{d-1}$ such that $\langle w, u^*\rangle = 0$,

$$\langle w, \nabla \widehat{L}_n(\theta^*)\rangle \leqslant \frac{c'\log B}{\sqrt{B}}\sqrt{\frac{d+t}{n}}\left(1 + \sqrt{\frac{B(d+t)}{n}}\right) \leqslant c'\frac{\log B}{\sqrt{B}}\sqrt{\frac{d+t}{n}}\,,$$

where the last inequality holds since $n \geqslant B(d + Bt)$. Plugging this upper bound and (63) into (44) concludes the proof of the proposition.

We conclude this section with the following lemma that was used in the proof of Lemma 8.

Lemma 9. *Let U, V be nonnegative real random variables such that $\mathbb{E}[U] \leqslant \varepsilon$ and $\|V\|_{\psi_1} \leqslant K$ for some $\varepsilon, K > 0$.*

1. *If $U \leqslant 1$ almost surely and $\varepsilon \leqslant 1$, then $\mathbb{E}[UV^2] \leqslant \varepsilon \cdot \frac{K^2\log^2(e\vee\varepsilon^{-1})}{4e}$.*
2. *If $\|U\|_{\psi_1} \leqslant K$, then $\mathbb{E}[UV] \leqslant \varepsilon K\log(e \vee K/\varepsilon)$.*

Proof. We start with the first point. Using Hölder's inequality, for any $p > 1$ we have (using that $u^{p/(p-1)} \leqslant u$ for $u \in [0, 1]$)

$$\mathbb{E}\big[UV^2\big] \leqslant \mathbb{E}\big[|V|^{2p}\big]^{1/p} \mathbb{E}\big[U^{p/(p-1)}\big]^{1-1/p} \leqslant \|V\|_{2p}^2 \mathbb{E}[U]^{1-1/p}$$
$$\leqslant \Big(\frac{Kp}{2e}\Big)^2 \varepsilon^{1-1/p} = \frac{K^2\varepsilon}{4e^2} \varepsilon^{-1/p} p^2 .$$

Now, letting $p' \to p = \max(1, \log(1/\varepsilon)) \geqslant 1$, we obtain

$$\mathbb{E}[UV^2] \leqslant \frac{K^2\varepsilon}{4e^2} \cdot e \cdot \max(1, \log^2(1/\varepsilon)) = \frac{K^2}{4e} \cdot \varepsilon \log^2(e \vee \varepsilon^{-1}) .$$

We now prove the second inequality. For any $p > 1$, letting $q = p/(p-1)$ we have

$$\mathbb{E}[UV] \leqslant \mathbb{E}[V^p]^{1/p}\mathbb{E}[U^q]^{1/q} \leqslant \frac{Kp}{2e}\mathbb{E}[U^q]^{1/q} , \tag{64}$$

where the second inequality comes from the fact that $\|V\|_{\psi_1} \leqslant K$. Next, for any $r > 1$, write $q = 1 - \frac{1}{r} + \frac{q'}{r}$ with $q' = 1 + r(q-1) > q$. We also have by Hölder's inequality

$$\mathbb{E}[U^q] = \mathbb{E}[U^{1-1/r}(U^{q'})^{1/r}] \leqslant \mathbb{E}[U]^{1-1/r} \|U\|_{q'}^{q'/r} \leqslant \varepsilon^{1-1/r} \Big(\frac{Kq'}{2e}\Big)^{q'/r}$$
$$= \varepsilon^{1-1/r} \Big(\frac{K[1 + r(q-1)]}{2e}\Big)^{q-1+1/r} = \varepsilon^q \Big(\frac{K[1 + r(q-1)]}{2e\varepsilon}\Big)^{q-1+1/r} ,$$

where we used that $\mathbb{E}[U] \leqslant \varepsilon$ and $\|U\|_{\psi_1} \leqslant K$. Hence, using that $q - 1 = 1/(p-1)$, letting $r = p - 1$ (assuming $p > 2$) so that $qr = p$, we obtain

$$\mathbb{E}[U^q]^{1/q} \leqslant \varepsilon \Big(\frac{K[1 + r(q-1)]}{2e\varepsilon}\Big)^{1-1/q+1/(qr)} = \varepsilon \Big(\frac{K}{e\varepsilon}\Big)^{2/p} .$$

Plugging this inequality into (64) and letting $p \to 2\log(e \vee K/\varepsilon) \geqslant 2$, so that $\lim(K/\varepsilon)^{2/p} \leqslant e$, we get

$$\mathbb{E}[UV] \leqslant \frac{K \cdot 2\log(e \vee K/\varepsilon)}{2e} \cdot \varepsilon e = \varepsilon K \log(e \vee K/\varepsilon) ,$$

which establishes the second point. □

6 Proofs of lower bounds on empirical Hessian matrices

This section is devoted to the proofs of the uniform lower bound on empirical Hessian matrices stated in Section 4.3. Specifically, Sections 6.1 and 6.2 contain the proof of Theorem 5 (in the regular case), while Sections 6.3 and 6.4 contain the proof of Theorem 6 (in the Gaussian case).

6.1 Proof of Theorem 5 (regular design)

In this section, we prove Theorem 5, namely the uniform lower bound on empirical Hessian matrices in the case of a regular design. Specifically, we assume that X satisfies Assumptions 1 and 3 with parameters $K \geqslant e$, $u^* = \theta^*/\|\theta^*\|$, $\eta = 1/B$ and $c \geqslant 1$.

Fix $v \in S^{d-1}$ and $\theta \in \Theta$, we want to bound from below

$$\langle \widehat{H}_n(\theta)v, v\rangle = \frac{1}{n}\sum_{i=1}^n \sigma'(\langle \theta, X_i\rangle)\langle v, X_i\rangle^2 .$$

The function $\sigma'(x) = \exp(x)/(1+\exp(x))^2$ is even, non negative, non increasing on $[0,+\infty)$. Therefore, for any $m, M > 0$,

$$\langle \widehat{H}_n(\theta)v, v\rangle \geqslant \frac{\sigma'(m(1+r)B)M^2}{n} \sum_{i=1}^n \mathbf{1}\{|\langle u, X_i\rangle| \leqslant m,\ |\langle v, X_i\rangle| \geqslant M\}\ , \tag{65}$$

where we also used that, as $\|\theta-\theta^*\|_H \leqslant r/\sqrt{B}$, $\|\theta\| \leqslant (1+r)B$ by Lemma 17. It remains to bound from below the empirical process $n^{-1}\sum_{i=1}^n \mathbf{1}\big\{\big|\langle u, X_i\rangle\big| \leqslant m,\ |\langle v, X_i\rangle| \geqslant M\big\}$ uniformly over $\theta \in \Theta$ and $v \in S^{d-1}$, for a proper choice of m and M. We want to apply Lemma 10. For this, we have to estimate $\mathbb{P}(|\langle u, X_i\rangle| \leqslant m,\ |\langle v, X_i\rangle| \geqslant M)$ for each $\theta \in \Theta$ and $v \in S^{d-1}$. If $\|\theta^*\| \leqslant e$, $B = e$ so $\eta = 1/e$, so Proposition 2 shows that Assumption 3 is satisfied with constant $\max\{2eK\log(2K), 2K^4\} = 2K^4$. Therefore,

$$\mathbb{P}\left(\big|\langle u, X\rangle\big| \leqslant \frac{2K^4}{B}\,;\, \big|\langle v, X\rangle\big| \geqslant \frac{\max\{1/B, \|u^*-v\|\}}{2K^4}\right) \geqslant \frac{1}{2K^4B}\,.$$

When $\|\theta^*\| \geqslant e$, the third point of Lemma 17 implies that for every $\theta \in \Theta$,

$$\|u-u^*\| \leqslant \frac{\sqrt{2}}{[K\log(c(c+1)B) - r]}\frac{r}{B} \leqslant \frac{2r}{KB\log(c(c+1)B)}\,.$$

By Lemma 11, this implies that for all $\theta \in \Theta$ and $v \in S^{d-1}$, one has for all $t \geqslant 1/B$

$$\mathbb{P}\left(\big|\langle u, X\rangle\big| \leqslant \frac{c+1}{B}\,;\, \big|\langle v, X\rangle\big| \geqslant \frac{\max\{1/B, \|u^*-v\|\}}{c+1}\right) \geqslant \frac{1}{(c+1)B}\,.$$

This suggests to choose $m = \gamma/B$, $M = \max(1/B, \|u^*-v\|)/\gamma$ in (65), where $\gamma = c+1$ if $\|\theta^*\| \geqslant e$ and $\gamma = 2K^4$ if $\|\theta^*\| < e$. With this choice, we have, for all $\theta \in \Theta$ and all $v \in S^{d-1}$,

$$\mathbb{P}\left(\big|\langle u, X\rangle\big| \leqslant \frac{\gamma}{B}\,;\, \big|\langle v, X\rangle\big| \geqslant \frac{\max\{1/B, \|u^*-v\|\}}{\gamma}\right) \geqslant \frac{1}{\gamma B}\,. \tag{66}$$

The next step to apply Lemma 10 is to bound the VC dimension of the class of sets of the form $\big\{x : |\langle u, x\rangle| \leqslant m,\ |\langle v, x\rangle| \geqslant M\big\}$ for any $u, v \in S^{d-1}$ and any $m, M > 0$. For this, remark that each of these sets is the union of two intersections of 3 half-spaces. The class of all half-spaces in $\mathbb{R}^d$ has VC dimension d [DGL96, Theorem 13.8]. Therefore, by [vdVW09, Theorem 1.1], the class of all intersections of 3 half-spaces is bounded from above by $6.9\log(12)d$, and therefore, by the same result, the VC dimension of the class of all unions of 2 intersections of 3 half spaces is bounded from above by

$$4.6\log(8) \times 6.9\log(12)d \leqslant 165d\,.$$

Hence, Lemma 10 applies with $p = 1/(\gamma B)$ and VC dimension $165d$. It shows that, whenever

$$n \geqslant \max\Big\{270000\log\big(730000\gamma B\big)d, 80\gamma Bt\Big\}\,,$$

with probability at least $1-e^{-t}$, one has simultaneously for all $\theta \in \Theta$ and $v \in S^{d-1}$,

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}\big\{\big|\langle u, X_i\rangle\big| \leqslant m,\ \langle v, X_i\rangle \geqslant M\big\} \geqslant \frac{1}{2\gamma B}\,.$$

Plugging this estimate into (65) shows that, on the same event,

$$\begin{aligned}\langle \widehat{H}_n(\theta)v, v\rangle^2 &\geqslant \frac{\sigma'(\gamma(1+r))}{2\gamma^2 B} \max\left\{\frac{1}{B}, \|u^* - v\|\right\}^2 \\ &\geqslant \frac{\sigma'(\gamma(1+r))}{4\gamma^2}\left(\frac{\langle u^*, v\rangle^2}{B^3} + \frac{1-\langle u^*, v\rangle^2}{B}\right) = \frac{\sigma'(\gamma(1+r))}{4\gamma^2}\langle Hv, v\rangle\,.\end{aligned}$$

In addition, for all real x, $\sigma'(x) \geqslant \frac{e^{-|x|}}{2}\mathbf{1}(|x| \geqslant 1)$. One can also check that $x^2\exp(\alpha x) \leqslant \exp((\alpha + 2/e)x)$ for every $x, \alpha \geqslant 1$. The result then follows by applying this with $x = \gamma$ and $\alpha = 1 + r$. The condition on r ensures that $1 + r + 2/e \leqslant 2$.

6.2 Technical lemmas for the proof of Theorem 5

This section gathers the technical tools that we used in the previous proof. We used the following VC-type inequality (see e.g. [BLM13, Example 3.10] for the definition of the VC dimension), whose proof is recalled for the sake of completeness and to justify our numerical constants.

Lemma 10. *Let $X_1, \dots, X_n$ be i.i.d. random variables taking values in $\mathcal{X}$ with common distribution P, and let $\mathcal{A}$ be a collection of subsets of $\mathcal{X}$ with VC dimension at most $d \geqslant 1$. Let $p \in (0,1)$, and assume that $P(A) \geqslant p$ for any $A \in \mathcal{A}$. If*

$$n \geqslant \frac{1600\log(4400/p)\,d}{p}\,, \tag{67}$$

then with probability at least $1 - e^{-np/80}$, one has

$$\inf_{A\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \in A) \geqslant \frac{p}{2}\,. \tag{68}$$

Proof. We first modify the class $\mathcal{A}$ to ensure that all events have a probability equal to p, rather than larger than p. For $i = 1, \dots, n$, we let $X_i' = (X_i, U_i)$, where $U_1, \dots, U_n$ are i.i.d. random variables uniformly distributed on $[0,1]$ and independent from $X_1, \dots, X_n$, and denote by P' the common distribution of the i.i.d. variables $X_1', \dots, X_n'$. In addition, we define the class $\mathcal{A}'$ of subsets of $\mathcal{X} \times [0,1]$ by

$$\mathcal{A}' = \left\{A \times \left[0, \frac{p}{P(A)}\right] : A \in \mathcal{A}\right\}.$$

Note that, for any $A' = A \times [0, p/P(A)] \in \mathcal{A}'$, one has $P'(A') = P(A) \times \frac{p}{P(A)} = p$. In addition,

$$\inf_{A'\in\mathcal{A}'} \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i' \in A') = \inf_{A\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \in A,\, U_i \leqslant p/P(A)) \leqslant \inf_{A\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \in A)\,.$$

In light of this inequality, in order to show (68), it suffices to show that $Z \leqslant p/2$, where

$$Z = p - \inf_{A'\in\mathcal{A}'} \frac{1}{n}\sum_{i=1}^n \mathbf{1}\big(X_i' \in A'\big) = \sup_{A'\in\mathcal{A}'} \frac{1}{n}\sum_{i=1}^n \big\{p - \mathbf{1}\big(X_i' \in A'\big)\big\}\,.$$

Now by Talagrand's inequality [Bou02, Theorem 2.3], as $\mathrm{Var}(\mathbf{1}(X_i' \in A')) = p(1-p) \leqslant p$ for any i and $A' \in \mathcal{A}'$, for any $t \geqslant 0$, with probability at least $1 - e^{-t}$ one has

$$Z \leqslant 2\left(\mathbb{E}[Z] + \sqrt{\frac{pt}{n}} + \frac{t}{n}\right). \tag{69}$$

Hence, if $\mathbb{E}[Z] \leqslant p/8$, we get that with probability at least $1-e^{-np/80}$,

$$Z \leqslant 2\Big(\frac{p}{8}+\sqrt{\frac{p}{n}\cdot\frac{np}{80}}+\frac{np}{80n}\Big)=\frac{p}{2}\Big(\frac{1}{2}+\frac{1}{\sqrt{5}}+\frac{1}{20}\Big)<\frac{p}{2}\,.$$

Hence, it suffices to show that $\mathbb{E}[Z] \leqslant p/8$. First, by symmetrization (e.g. [Kol11, Theorem 2.1]), one has

$$\mathbb{E}[Z] \leqslant \frac{2}{n}\,\mathbb{E}\Big[\sup_{A'\in\mathcal{A}'}\sum_{i=1}^{n}\varepsilon_i\mathbf{1}(X_i'\in A')\Big]\,, \tag{70}$$

where $\varepsilon_1,\dots,\varepsilon_n$ are i.i.d. random variables with $\mathbb{P}(\varepsilon_i=\pm 1)=1/2$. Next, it follows from Hoeffding's lemma [BLM13, §2.3], together with the maximal inequality from [BLM13, §2.5], that

$$\mathbb{E}\Big[\sup_{A'\in\mathcal{A}'}\sum_{i=1}^{n}\varepsilon_i\mathbf{1}(X_i'\in A')\Big|X_1,\dots,X_n\Big] \leqslant \sqrt{2\Big(\sup_{A'\in\mathcal{A}'}\sum_{i=1}^{n}\mathbf{1}(X_i'\in A')\Big)\log S_n(\mathcal{A}')}\,,$$

where $S_n(\mathcal{A}')=\max_{x_1',\dots,x_n'}|\{(\mathbf{1}(x_i'\in A'))_{1\leqslant i\leqslant n}:A'\in\mathcal{A}'\}|$ denotes the n-th shattering number of $\mathcal{A}'$. Let $\mathcal{B}=\{[0,t]:t\in[0,1]\}$. By definition of $\mathcal{A}'$, one has $S_n(\mathcal{A}')\leqslant S_n(\mathcal{A})S_n(\mathcal{B})\leqslant(n+1)S_n(\mathcal{A})$. In addition, since $\mathcal{A}$ is a VC class with VC dimension at most d and $n\geqslant d+1$, Sauer's lemma (e.g., [vH14, Lemma 7.12]) implies that $S_n(\mathcal{A})\leqslant(en/d)^d$. Plugging this bound into the above and applying Jensen's inequality gives:

$$\begin{aligned}
R_n=\mathbb{E}\Big[\sup_{A'\in\mathcal{A}'}\sum_{i=1}^{n}\varepsilon_i\mathbf{1}(X_i'\in A')\Big]&=\mathbb{E}\Big[\mathbb{E}\Big[\sup_{A'\in\mathcal{A}'}\sum_{i=1}^{n}\varepsilon_i\mathbf{1}(X_i'\in A')\Big|X_1,\dots,X_n\Big]\Big]\\
&\leqslant\sqrt{2\,\mathbb{E}\Big[\sup_{A'\in\mathcal{A}'}\sum_{i=1}^{n}\mathbf{1}(X_i'\in A')\Big]\log\Big[(n+1)\Big(\frac{en}{d}\Big)^d\Big]}\,.
\end{aligned}\tag{71}$$

On the other hand, since $\mathbb{P}(X_i'\in A')=p$ for every $i=1,\dots,n$, another application of the symmetrization inequality gives:

$$\begin{aligned}
\mathbb{E}\Big[\sup_{A'\in\mathcal{A}'}\sum_{i=1}^{n}\mathbf{1}(X_i'\in A')\Big]&=np+\mathbb{E}\Big[\sup_{A'\in\mathcal{A}'}\sum_{i=1}^{n}\big\{\mathbf{1}(X_i'\in A')-\mathbb{P}(X_i'\in A')\big\}\Big]\\
&\leqslant np+2\,\mathbb{E}\Big[\sup_{A'\in\mathcal{A}'}\sum_{i=1}^{n}\varepsilon_i\mathbf{1}(X_i'\in A')\Big]=np+2R_n\,.
\end{aligned}\tag{72}$$

Plugging (72) into (71) and using that $n+1\leqslant en\leqslant(en/d)^d$, denoting $H_n=d\log(en/d)$ we get:

$$R_n^2\leqslant 4(np+2R_n)H_n\,,$$

which after solving for this second-order inequality in R_n gives

$$R_n\leqslant 4H_n+2\sqrt{4H_n^2+npH_n}\,.$$

Hence, recalling from (70) that $\mathbb{E}[Z]\leqslant 2R_n/n$, we get whenever $H_n\leqslant\varepsilon np$ with $\varepsilon=1/1160$:

$$\mathbb{E}[Z]\leqslant\frac{8H_n}{n}+\frac{4\sqrt{4H_n^2+npH_n}}{n}\leqslant\Big(8\varepsilon+4\sqrt{4\varepsilon^2+\varepsilon}\Big)p<\frac{p}{8}\,,$$

which is precisely what we aimed to show. It thus remains to show that $H_n\leqslant np/1160$, namely

$$\frac{d\log(en/d)}{n}\leqslant\frac{p}{1160}\,.$$

But this follows from the assumption (67), together with the basic fact that if $u,v\geqslant 1$ satisfy $u\geqslant(1+e^{-1})v\log\big((e+1)v\big)$, then $\log(eu)/u\leqslant 1/v$ (applied to $u=n/d$ and $v=1160/p$). $\square$

To apply Lemma 10 to the sets $\{x \in \mathbb{R}^d : |\langle u, x\rangle| \leqslant m, |\langle v, x\rangle| \geqslant M\}$, we had to lower-bound the probability of these events. This bound can be deduced from the fact that the lower bound on these event provided for $u = u^*$ by Assumption 3 can be extended to all u in a neighborhood of u^* as shown in the following result.

Lemma 11. *Suppose that Assumptions 1 and 3 hold with respective parameters $K \geqslant e$, $u^* \in S^{d-1}$, $c \geqslant 1$ and $\eta \in (0, 1)$. Then for all $u, v \in S^{d-1}$ such that*

$$\|u - u^*\| \leqslant \frac{2\eta}{K \log(c(c+1)/\eta)} \quad \textit{and} \quad \langle u^*, v\rangle \geqslant 0\,, \tag{73}$$

one has

$$\mathbb{P}\Big(|\langle u, X\rangle| \leqslant (c+1)\eta\,;\, |\langle v, X\rangle| \geqslant \frac{\max\{\eta, \|u^* - v\|\}}{c+1}\Big) \geqslant \frac{\eta}{c+1}\,.$$

Proof. Let $u, v \in S^{d-1}$ satisfy (73). The triangle inequality

$$|\langle u, X\rangle| \leqslant |\langle u^*, X\rangle| + |\langle u - u^*, X\rangle|$$

implies that

$$\begin{aligned} &\mathbb{P}\Big(|\langle u, X\rangle| \leqslant (c+1)\eta\,;\, |\langle v, X\rangle| \geqslant \frac{\max\{\eta, \|u^* - v\|\}}{c}\Big) \\ &\geqslant \mathbb{P}\Big(|\langle u^*, X\rangle| \leqslant c\eta\,;\, |\langle v, X\rangle| \geqslant \frac{\max\{\eta, \|u^* - v\|\}}{c}\Big) - \mathbb{P}\big(|\langle u - u^*, X\rangle| > \eta\big)\,. \end{aligned} \tag{74}$$

Next, on the one hand, Assumption 3 asserts that

$$\mathbb{P}\Big(|\langle u^*, X\rangle| \leqslant c\eta\,;\, |\langle v, X\rangle| \geqslant \frac{\max\{\eta, \|u^* - v\|\}}{c}\Big) \geqslant \frac{\eta}{c}\,,$$

and on the other hand, Assumption 1 together with Point 1 in Lemma 35 implies that

$$\mathbb{P}\left(\big|\langle u - u^*, X\rangle\big| > \eta\right) \leqslant \exp\left(-\frac{2\eta}{K\|u - u^*\|}\right) \leqslant \exp\left(-\log\left(c(c+1)/\eta\right)\right) \leqslant \frac{\eta}{c(c+1)}\,.$$

Plugging the previous two inequalities into (74) concludes the proof, since $\frac{\eta}{c} - \frac{\eta}{c(c+1)} = \frac{\eta}{c+1}$. □

6.3 Proof of Theorem 6 (Gaussian design)

In this section, we let

$$\Theta = \Big\{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant \frac{1}{100\sqrt{B}}\Big\}\,.$$

All along the section, we denote by $C > 0$ absolute constants, whose value may change from line to line. This section is devoted to the proof of Theorem 6.

We have to prove that the smallest eigenvalue of the matrices $H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}$ is bounded from below or, equivalently, that

$$\forall v \in S^{d-1}, \forall \theta \in \Theta, \qquad \langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle \geqslant 1/C\,.$$

For this, we use the PAC-Bayes inequality, see e.g. [Cat07]. This inequality involves the Kullback-Leibler divergence between two probability measures. Recall that, if μ and ν denote two probability measures on a same space Ω such that ν is dominated by μ, the Kullback-Leibler divergence from ν to μ is defined by

$$D(\nu\|\mu) = \int_\Omega \log\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\Big)\mathrm{d}\nu\,. \tag{75}$$

Lemma 12 (PAC-Bayes inequality). *Let $(E, \mathcal{E}, \pi)$ denote a probability space and $Z = (Z(\omega))_{\omega \in E}$ a measurable real process indexed by $\omega \in E$. Let $Z_1, \ldots, Z_n$ be independent copies of the process Z. Let also $\lambda > 0$ be such that $\mathbb{E}\exp(\lambda Z(\omega)) < \infty$ for every $\omega \in \Omega$. For any $t > 0$, with probability at least $1 - e^{-t}$, simultaneously for every probability measure ρ on E dominated by π,*

$$\frac{1}{n}\sum_{i=1}^{n}\int_E Z_i(\omega)\rho(\mathrm{d}\omega) \leqslant \frac{1}{\lambda}\int_E \log\left(\mathbb{E}\big[e^{\lambda Z(\omega)}\big]\right)\rho(\mathrm{d}\omega) + \frac{D(\rho\|\pi) + t}{\lambda n}\,. \tag{76}$$

We will refer to the distribution π as the "prior", and to the smoothing distributions ρ as the "posteriors". We will apply the PAC-Bayes inequality with index set $E = \Theta' \times S^{d-1}$, where

$$\Theta' = \left\{\theta' \in \mathbb{R}^d : \|\theta' - \theta^*\|_H \leqslant \frac{1}{10\sqrt{B}}\right\} \supset \Theta\,,$$

and process $Z_i(\omega) = Z_i(\theta', v')$ defined by

$$Z_i(\theta', v') = -\mathbf{1}\Big\{|\langle\theta', X_i\rangle| \leqslant 1; \|X_i\| \leqslant 2\sqrt{d}\Big\}\left\langle H^{-1/2}v', X_i\right\rangle^2. \tag{77}$$

The prior π and posteriors $\rho_{\theta,v} = \rho_\theta \otimes \rho_v$ on E will be defined during the proof.

An interesting fact is that the PAC-Bayes inequality applied to this process will show a uniform lower bound on the Hessian, hence, a lower bound on the process $\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle$ with respect to both θ and v. To see why, we start by bounding from bellow

$$\begin{aligned}\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle &= \frac{1}{n}\sum_{i=1}^{n}\sigma'(\langle\theta, X_i\rangle)\langle H^{-1/2}v, X_i\rangle^2 \\ &\geqslant \frac{1}{n}\sum_{i=1}^{n}\sigma'(\langle\theta, X_i\rangle)\mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\langle H^{-1/2}v, X_i\rangle^2\end{aligned}$$

In the next paragraph, we will associate to each $\theta \in \Theta$ a probability measure ρ_θ on Θ' and prove a key smoothing lemma (Lemma 13), which allows us to further bound from below:

$$\begin{aligned}\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle &\geqslant \frac{1}{n}\sum_{i=1}^{n}\frac{1}{15}\int_{\mathbb{R}^d}\mathbf{1}\Big\{|\langle\theta', X_i\rangle| \leqslant 1, \|X_i\| \leqslant 2\sqrt{d}\Big\}\rho_\theta(\mathrm{d}\theta')\langle H^{-1/2}v, X_i\rangle^2 \\ &= \frac{1}{15}\int_{\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n} -Z_i(\theta', v)\rho_\theta(\mathrm{d}\theta')\,.\end{aligned} \tag{78}$$

Then, in the following paragraphs, we compute all quantities appearing in the PAC-Bayes inequality. We bound first the Laplace transform, then the smoothed process in the left-hand side of (76) and finally the Kullback-Leibler divergence. A last paragraph gathers all the results and optimizes the choice of λ to conclude the proof.

Remark 3. The expected value of the minorizing process appearing in (78), namely $\mathbb{E}[-Z_i(\widetilde{\theta}, v)]$, is equal to $\langle H^{-1/2}\widetilde{H}(\theta')H^{-1/2}v, v\rangle$, where the matrix $\widetilde{H}(\theta)$ is defined by

$$\widetilde{H}(\theta) = \mathbb{E}\left[\mathbf{1}\Big\{|\langle\theta, X\rangle| \leqslant 1; \|X\| \leqslant 2\sqrt{d}\Big\}XX^\top\right]. \tag{79}$$

It is proved in Lemma 18 that $\widetilde{H}(\theta) \succcurlyeq 0.05 \cdot H_\theta$, where

$$H_\theta = \frac{1}{B^3}uu^\top + \frac{1}{B}(I_d - uu^\top)\,. \tag{80}$$

Then, in Lemma 19, it is proved that $H_\theta \succcurlyeq 0.97H$. This means that $\widetilde{H}(\theta) \succcurlyeq 0.04H$, so $H^{-1/2}\widetilde{H}(\theta)H^{-1/2} \succcurlyeq 0.04I_d$ and therefore, for any $v \in S^{d-1}$,

$$\mathbb{E}\big[-Z_i(\theta, v)\big] = \langle H^{-1/2}\widetilde{H}(\theta)H^{-1/2}v, v\rangle \geqslant 0.04\,. \tag{81}$$

The posteriors ρ_θ. Let us first define the posterior ρ_θ. For any $\theta \in \Theta$, let $u = \theta/\|\theta\|$.

Definition 2. For any $\theta \in \Theta$, we let ρ_θ denote the distribution of $\theta' = U\theta + Z$, where

(i) U, Z are independent;

(ii) U is uniform over $[0.99, 1.01]$;

(iii) the distribution of Z is the conditional distribution of $Z' \sim \mathsf{N}(0, (I_d - uu^\top)/(2 \cdot 100^2 \cdot d))$ given that $\|Z'\| \leqslant 1/100$.

The motivation behind the choice of the posterior (or smoothing distribution) ρ_θ in Definition 2 is twofold. On the one hand, it is sufficiently spread out that, every $\theta \in \Theta$, the Kullback-Leibler divergence between ρ_θ and a suitably chosen prior is controlled: Lemma 16 below show that it is at most of order d, with no dependence on B. At the same time, it is sufficiently localized around θ (in particular, along the direction of θ itself) that smoothing an indicator with respect to this distribution provides a lower bound on the sigmoid, as shown in Lemma 13 below.

We note in passing that such a lower bound would not hold if instead of being uniform on $[0.99, 1.01]$, the variable U in Definition 2 was Gaussian (say, if $U \sim \mathsf{N}(1, 0.01^2)$), as in this case the smoothed indicator would be of order $(1 + |\langle \theta, x\rangle|)^{-1}$, which is much larger than $\sigma'(|\langle \theta, x\rangle|)$. This contrasts with prior applications of the PAC-Bayes method to the study of random matrices [Cat16, Oli16, Mou22, Zhi24], which essentially rely on Gaussian or truncated Gaussian posteriors.

Lemma 13. *For every $\theta \in \Theta$, the measure ρ_θ is supported on Θ'. In addition, for every $x \in \mathbb{R}^d$ such that $\|x\| \leqslant 2\sqrt{d}$, one has*

$$\sigma'(\langle \theta, x\rangle) \geqslant \frac{1}{15} \int_{\mathbb{R}^d} \mathbf{1}\{|\langle \theta', x\rangle| \leqslant 1\} \rho_\theta(\mathrm{d}\theta') \,. \tag{82}$$

Proof. We start with the first claim. Let $\theta \in \Theta$ and $\theta' = U\theta + Z \sim \rho_\theta$, with U, Z distributed as in Definition 2. By Lemma 19, as $\|\theta - \theta^*\|_H \leqslant 1/100\sqrt{B}$, if $\theta' \sim \rho_\theta$ then

$$\|\theta' - \theta\|_H \leqslant \frac{1}{0.97} \|\theta' - \theta\|_{H_\theta} = \frac{1}{0.97} \sqrt{\frac{(U-1)^2 \|\theta\|^2}{B^3} + \frac{\|Z\|^2}{B}} \,.$$

Now $|U - 1| \leqslant 0.01$, $\|\theta\|/B \leqslant 1.01$ and $\|Z\| \leqslant 1/100$ a.s., so by Lemma 17, as $\theta \in \Theta$,

$$\|\theta' - \theta\|_H \leqslant \frac{0.015}{\sqrt{B}} \qquad \text{and} \qquad \|\theta' - \theta^*\|_H \leqslant \frac{0.025}{\sqrt{B}} \,. \tag{83}$$

We now prove inequality (82). Let $x \in \mathbb{R}^d$ be such that $\|x\| \leqslant 2\sqrt{d}$. We have

$$\int_{\mathbb{R}^d} \mathbf{1}\{|\langle \theta', x\rangle| \leqslant 1\} \rho_\theta(\mathrm{d}\theta') = \mathbb{E}\big[\mathbb{P}\left(|U\langle \theta, x\rangle + \langle Z, x\rangle| \leqslant 1 | U\right)\big] \,.$$

If $Z' \sim \mathsf{N}(0, (I_d - uu^\top)/(2 \cdot 100^2 \cdot d))$, we have, as $\mathbb{P}(\|Z'\| \leqslant 1/100) \geqslant 1 - 100^2 \mathbb{E}\|Z'\|^2 \geqslant 3/4$,

$$\mathbb{P}\left(|U\langle \theta, x\rangle + \langle Z, x\rangle| \leqslant 1 | U\right) \leqslant \frac{4}{3} \mathbb{P}\left(|U\langle \theta, x\rangle + \langle Z', x\rangle| \leqslant 1 | U\right) \,.$$

Now, if g is a standard Gaussian random variable, for any $a \in \mathbb{R}$, $b > 0$ and $\sigma > 0$

$$\mathbb{P}(|\sigma g - a| \leqslant 1) \leqslant 2\mathbb{P}(g > (|a| - 1)_+/\sigma) \leqslant \exp\left(- \frac{(|a| - 1)_+^2}{2\sigma^2} \right) \leqslant C \exp(-b|a|) \,,$$

with $C = \exp\left(\frac{\sigma^2 b^2}{2} + b\right)$.

We apply this result with

$$\sigma^2 = \mathrm{Var}(\langle g', x\rangle) \leqslant \frac{\|x\|^2}{2\cdot(100)^2\cdot d} \leqslant \frac{1}{5000}, \qquad b = \frac{1}{U} \leqslant \frac{1}{0.99}, \qquad a = U\langle\theta, x\rangle\,.$$

This shows that

$$\mathbb{P}\left(|U\langle\theta, x\rangle + \langle Z, x\rangle| \leqslant 1|U\right) \leqslant 3.7\exp\big(-|\langle\theta, x\rangle|\big) \leqslant 15\sigma'(\langle\theta, x\rangle)\,.$$

This proves the lower bound (82). □

Upper bound on the Laplace transform. In this section, we prove the following upper bound on the Laplace transfom of $Z(\theta', v')$, for $\theta' \in \Theta'$ and $v' \in S^{d-1}$. It controls the Laplace transform in the PAC-Bayes inequality for any posterior distribution supported on Θ'.

Lemma 14. *For any $\theta' \in \mathbb{R}^d$ such that $\|\theta' - \theta^*\|_H \leqslant 1/10\sqrt{B}$ and any $v' \in S^{d-1}$, we have*

$$\log \mathbb{E}\exp(\lambda Z(\theta', v')) \leqslant -0.04\lambda + 2.2\lambda^2 B\,.$$

Proof. Let $\lambda > 0$, θ' such that $\|\theta' - \theta^*\|_H \leqslant 1/10\sqrt{B}$ and $v' \in S^{d-1}$. As, for any $s > 0$, $e^{-s} \leqslant 1 - s + s^2/2$, one has

$$\mathbb{E}\exp(\lambda Z_i(\theta', v')) \leqslant 1 + \lambda\mathbb{E}Z_i(\theta', v') + \frac{\lambda^2}{2}\mathbb{E}Z_i(\theta', v')^2.$$

By (81), one has $\mathbb{E}Z_i(\theta', v') \leqslant -0.04$. Moreover, by Lemma 19,

$$\mathbb{E}Z_i(\theta', v')^2 \leqslant \mathbb{E}\big[\mathbf{1}\big\{|\langle\theta', X_i\rangle| \leqslant 1\big\}\,\langle H^{-1/2}v', X_i\rangle^4\big] \leqslant 1.7\mathbb{E}\big[\mathbf{1}\big\{|\langle\theta', X_i\rangle| \leqslant 1\big\}\,\langle H_{\theta'}^{-1/2}v', X_i\rangle^4\big]\,.$$

We have, for $u' = \theta'/\|\theta'\|$,

$$H_{\theta'}^{-1/2}v' = \langle u', v'\rangle B^{3/2}u' + \sqrt{B}(v' - \langle v', u'\rangle u')\,.$$

As $\langle u', X_i\rangle$ and $\langle v' - \langle v', u'\rangle u', X_i\rangle$ are independent Gaussian random variables, we have

$$\mathbb{E}Z_i(\theta', v')^2 \leqslant 1.7\mathbb{E}\big[\mathbf{1}\big\{|\langle\theta', X_i\rangle| \leqslant 1\big\}\,(B^6\langle u', X_i\rangle^4 + 6B^4\langle u', X_i\rangle^2 + B^2)\big]\,.$$

Next, using that $|\langle\theta', X\rangle| = \|\theta'\|\cdot|\langle u', X\rangle|$ and that the density of $\langle u', X_i\rangle \sim \mathsf{N}(0,1)$ is upper-bounded by $1/\sqrt{2\pi}$, we bound for $k \in \{0, 1, 2\}$:

$$\mathbb{E}\big[\mathbf{1}\big\{|\langle\theta', X_i\rangle| \leqslant 1\big\}\,\langle u', X_i\rangle^{2k}\big] \leqslant \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathbf{1}\big\{|x| \leqslant \|\theta'\|^{-1}\big\}\,x^{2k}\mathrm{d}x = \sqrt{\frac{2}{\pi}}\frac{1}{(2k+1)\|\theta'\|^{2k+1}}\,.$$

This yields $\mathbb{E}Z_i(\theta', v')^2 \leqslant 4.4B$ and thus, the result. □

Bounding the linear term. Let us first define the posterior ρ_v on S^{d-1}.

Definition 3. Let $\varepsilon \in (0,1)$. For any $v \in S^{d-1}$, let $\rho_v = \mathcal{U}\big(\mathsf{C}(v,\varepsilon)\big)$ denote the uniform distribution on the spherical cap of width ε around v, that is

$$\mathsf{C}(v,\varepsilon) = \Big\{v' \in S^{d-1} : \langle v, v'\rangle \geqslant 0,\, |\sin(v, v')| = \sqrt{1 - \langle v, v'\rangle^2} \leqslant \varepsilon\Big\}\,. \tag{84}$$

In this paragraph, we prove the following lower bound on the linear term.

Lemma 15. *Let $\rho_{\theta,v} = \rho_\theta \otimes \rho_v$ denote the posterior distribution defined as the product of the posterior ρ_θ of Definition 2 and the posterior ρ_v of Definition 84. Then,*

$$\frac{1}{n}\sum_{i=1}^n \int_{\Theta'\times S^{d-1}} Z_i(\theta', v')\rho_{\theta,v}(\mathrm{d}\theta', dv') \geqslant -15\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle - 12000\varepsilon^2 BR\,, \tag{85}$$

where $R = \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \exp\big(-0.99|\langle\theta, X_i\rangle|\big)\mathbf{1}\big(\|X_i\| \leqslant 2\sqrt{d}\big)$.

Proof. Recall that, by Lemma 13,

$$-\frac{1}{n}\sum_{i=1}^n \sigma'(\langle\theta, X_i\rangle)\mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\,\langle X_i, H^{-1/2}v'\rangle^2 \leqslant \frac{1}{15n}\sum_{i=1}^n \int_{\Theta'} Z_i(\theta', v')\rho_\theta(\mathrm{d}\theta')\,.$$

Therefore, our main task is to bound from above

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \int_{S^{d-1}} &\sigma'(\langle\theta, X_i\rangle)\mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\,\langle X_i, H^{-1/2}v'\rangle^2\rho_v(\mathrm{d}v')\\ &= \int_{\mathsf{C}(v,\varepsilon)} \langle H^{-1/2}\Big(\frac{1}{n}\sum_{i=1}^n \sigma'(\langle\theta, X_i\rangle)\mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\,X_iX_i^\top\Big)H^{-1/2}v', v'\rangle\rho_v(\mathrm{d}v')\,.\end{aligned}$$

Using the computations from [Mou22, Eqs. (42) and (43)] and Fact 6, we obtain

$$\begin{aligned}\int_{\mathsf{C}(v,\varepsilon)} &\langle H^{-1/2}\overline{H}_n(\theta)H^{-1/2}v', v'\rangle\rho_v(\mathrm{d}v')\\ &\leqslant \langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle + \frac{2\varepsilon^2}{d-1}\mathrm{Tr}\left(H^{-1/2}\overline{H}_n(\theta)H^{-1/2}\right)\,,\end{aligned} \tag{86}$$

where $\overline{H}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \sigma'(\langle\theta, X_i\rangle)\mathbf{1}\{\|X_i\| \leqslant 2\sqrt{d}\}X_iX_i^\top \preccurlyeq \widehat{H}_n(\theta)$.

Thus, (85) is proved if we show that

$$\mathrm{Tr}\left(H^{-1/2}\overline{H}_n(\theta)H^{-1/2}\right) \leqslant 3000\frac{Bd}{n}\sum_{i=1}^n \exp\big(-0.99|\langle\theta, X_i\rangle|\big)\mathbf{1}\big(\|X_i\| \leqslant 2\sqrt{d}\big)\,. \tag{87}$$

First, by Lemma 19, for all $\theta\in\Theta$, $H^{-1} \preccurlyeq 1.03H_\theta^{-1}$, so

$$\mathrm{Tr}\left(H^{-1/2}\overline{H}_n(\theta)H^{-1/2}\right) \leqslant 1.03\mathrm{Tr}\left(H_\theta^{-1/2}\overline{H}_n(\theta)H_\theta^{-1/2}\right)\,.$$

Now, as $\|X_i - \langle u, X_i\rangle u\|^2 \leqslant \|X_i\|^2$,

$$\begin{aligned}&\mathrm{Tr}\big(H^{-1/2}\overline{H}_n(\theta)H^{-1/2}\big)\\ &\quad= \frac{1.03}{n}\sum_{i=1}^n \sigma'\big(\langle\theta, X_i\rangle\big)\mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\,\big(B^3\langle u, X_i\rangle^2 + B\|X_i - \langle u, X_i\rangle u\|^2\big)\\ &\quad\leqslant \frac{1.03B}{n}\sum_{i=1}^n \sigma'\big(\langle\theta, X_i\rangle\big)\mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\,\big(B^2\langle u, X_i\rangle^2 + 4d\big)\,.\end{aligned}$$

Now, we use the inequalities $\sigma'(t) \leqslant e^{-|t|}$, for all $t \geqslant 0$, $t^2e^{-t} \leqslant (200/e)^2e^{-0.99t}$ and, by Lemma 17, $\|\theta\| \geqslant 0.99B$ to get

$$B^2\langle u, X_i\rangle^2\sigma'\big(\langle\theta, X_i\rangle\big) \leqslant 5550\exp\big(-0.99|\langle\theta, X_i\rangle|\big)\,.$$

This shows (87) with as $d \geqslant 2$ and thus (85). □

Upper bounds on the Kullback-Leibler divergence. In this section, we define the prior distribution π and bound from above the Kullback-Leibler divergence term $D(\rho_{\theta,v}\|\pi)$, where $\rho_{\theta,v}$ was defined in Lemma 15.

Let us start with the definition of the prior π. For any $\mu \in \mathbb{R}^d$, $\Sigma \succcurlyeq 0$ and any measurable subset $S \subset \mathbb{R}^d$, let $\mathsf{TN}(\mu, \Sigma, S)$ denote the Gaussian distribution $\mathsf{N}(\mu, \Sigma)$ conditioned on S, that is the distribution with density

$$\mathrm{d}\nu = \frac{\mathbf{1}_S}{\gamma(S)}\,\mathrm{d}\gamma\,, \tag{88}$$

where γ is the Gaussian distribution $\mathsf{N}(\mu, \Sigma)$.

Definition 4. The prior distribution π on $\Theta \times S^{d-1}$ is the product $\pi = \pi_\Theta \otimes \pi_S$, where π_S is the uniform distribution on the unit sphere and $\pi_\Theta = \mathsf{TN}(\theta^*, \Gamma, \Theta')$, for

$$\Gamma = \frac{1}{100^2}\left(B^2 u^* u^{*\top} + \frac{1}{2d}(I_d - u^* u^{*\top})\right).$$

The purpose of this section is to show the following upper bound on the Kullback divergence between $\rho_{\theta,v}$ and π.

Lemma 16. *Let $\theta \in \Theta$ and $v \in S^{d-1}$. Let $\rho_{\theta,v}$ denote the prior defined in Lemma 15 and π denote the prior distribution of Definition 4. Then,*

$$D(\rho_{\theta,v}\|\pi) \leqslant \left(6.5 + \log\left(1 + \frac{2}{\varepsilon}\right)\right)d\,. \tag{89}$$

Proof. Since the prior and all posterior distributions are product measures, the divergence writes

$$D(\rho_{\theta,v}\|\pi) = D(\rho_v\|\pi_S) + D(\rho_\theta\|\pi_\Theta)\,.$$

On one hand, we have

$$D(\rho_v\|\pi_S) = \int_{S^{d-1}} \log\left(\frac{\mathrm{d}\rho_v}{\mathrm{d}\pi_S}\right)\mathrm{d}\rho_v = \log\left(\frac{\mathrm{Vol}_{d-1}(S^{d-1})}{\mathrm{Vol}_{d-1}(\mathsf{C}(v,\varepsilon))}\right).$$

By [Mou22, §4.4] and Fact 6, this yields

$$D(\rho_v\|\pi_S) \leqslant d\log\left(1 + \frac{2}{\varepsilon}\right). \tag{90}$$

It remains to bound $D(\rho_\theta\|\pi_\Theta)$, which is more delicate. We first define an intermediate distribution $\tilde{\rho}_\theta$ and show, see (93), that

$$D(\rho_\theta\|\pi_\Theta) \leqslant 1.5(\log(1.5) + D(\tilde{\rho}_\theta\|\pi_\Theta))\,.$$

Then, we bound this last divergence. The intermediate distribution $\tilde{\rho}_\theta = \mathsf{TN}(\theta, \Gamma_\theta, \mathcal{E}_\theta)$ is the Gaussian distribution $\mathsf{N}(\theta, \Gamma_\theta)$ conditioned on the ellipsoid $\mathcal{E}_\theta = \{\theta_0 : \|\theta_0 - \theta\|_H \leqslant \frac{0.02}{\sqrt{B}}\}$, chosen such that, by Lemma 13,

$$\mathrm{Supp}\,(\rho_\theta) \subset \mathrm{Supp}\,(\tilde{\rho}_\theta) \subset \Theta' = \mathrm{Supp}\,(\pi_\Theta)\,.$$

For any $\theta \in \Theta$, the covariance Γ_θ is defined as

$$\Gamma_\theta = \frac{1}{100^2}\left(B^2 u u^\top + \frac{1}{2d}(I_d - u u^\top)\right), \qquad u = \frac{\theta}{\|\theta\|}\,.$$

Before we bound the Kullback-Leibler divergences $D(\rho_\theta\|\pi_\Theta)$, we check the following facts.

1. the density of ρ_θ satisfies $\frac{\mathrm{d}\rho_\theta}{\mathrm{d}\widetilde{\rho}_\theta} \leqslant 1.5$,
2. the normalizing factors satisfy $\gamma_\theta(\mathcal{E}_\theta) \geqslant 0.5$ and $\gamma(\mathcal{E}_\theta) \geqslant 0.3$, where $\gamma_\theta = \mathsf{N}(\theta, \Gamma_\theta)$ and $\gamma = \mathsf{N}(\theta^*, \Gamma)$.

Let us briefly check these facts. Start with point 1. We have on one hand that the density f_θ of ρ_θ satisfies, for every $\theta_0 = tu + z$,

$$f_\theta(\theta_0) \leqslant \frac{1}{p0.02\|\theta\|}\Big(\frac{(100)^2 d}{\pi}\Big)^{\frac{d-1}{2}} \exp\Big(-\frac{d\|z\|^2}{100^2}\Big)\mathbf{1}\Big(t/\|\theta\| \in [0.99, 1.01]; \|z\| \leqslant 1/100\Big)$$

and on the other hand $\widetilde{\rho}_\theta$ has density given for every $\theta_0 = tu + z$ by

$$\tilde{f}_\theta(\theta_0) = \frac{1}{\gamma_\theta(\mathcal{E}_\theta)} \cdot \frac{e^{-\frac{(t-\|\theta\|)^2}{2(B/100)^2}}}{B/100} \cdot \frac{(2\cdot(100)^2 d)^{\frac{d-1}{2}}}{(2\pi)^{d/2}} \exp\big(-(100)^2 d\|z\|^2\big)\mathbf{1}(\theta_0 \in \mathcal{E}_\theta)\,.$$

For any t such that $|t/\|\theta\| - 1| \leqslant 1/100$ and $\theta \in \Theta$ so, by Lemma 17, $\|\theta\|/B \in [0.99, 1.01]$, we deduce that

$$\frac{f_\theta(\theta_0)}{\tilde{f}_\theta(\theta_0)} \leqslant \frac{1}{2p}\frac{B}{\|\theta\|}\sqrt{\frac{\pi}{2}}\gamma_\theta(\mathcal{E}_\theta)\exp\Big(\frac{(t-\|\theta\|)^2}{2(B/100)^2}\Big)\mathbf{1}\left(\frac{t}{\|\theta\|} \in [0.99, 1.01]; \theta_0 \in \mathcal{E}_\theta\right) \leqslant 1.5\cdot\mathbf{1}(\theta_0 \in \mathcal{E}_\theta)\,. \tag{91}$$

Let us move to point 2. Fix $\theta \in \Theta$ so by Lemma 17, $\|u - u^*\| \leqslant 1/50B$. We have, if $N \sim \mathsf{N}(\theta, \Gamma_\theta)$, by Chebychev's inequality,

$$1 - \gamma_\theta(\mathcal{E}_\theta) = \mathbb{P}\left(\|N - \theta\|_H > \frac{0.02}{\sqrt{B}}\right) \leqslant \frac{B\mathbb{E}[\|N-\theta\|_H^2]}{0.02^2}\,.$$

Besides,

$$\begin{aligned}\mathbb{E}\|N-\theta\|_H^2 = \mathrm{Tr}\big(H^{1/2}\Gamma_\theta H^{1/2}\big) &= \frac{1}{100^2}\bigg(\Big(B^2 - \frac{1}{2d}\Big)\|H^{1/2}u\|^2 + \frac{1}{2d}\mathrm{Tr}(H)\bigg)\\ &\leqslant \frac{1}{100^2}\bigg(\Big(B^2 - \frac{1}{2d}\Big)\Big(\frac{1}{B^3} + \frac{\|u-u^*\|^2}{2B}\Big) + \frac{1}{2d}\Big(\frac{1}{B^3} + \frac{d-1}{B}\Big)\bigg) \leqslant \frac{2}{100^2 B}\,.\end{aligned}$$

This shows the first lower bound. For the second one, we proceed similarly: Let $N \sim \mathsf{N}(\theta^*, \Gamma)$ so, by Chebychev's inequality,

$$1 - \gamma(\mathcal{E}_\theta) = \mathbb{P}\left(\|N - \theta\|_H > \frac{0.02}{\sqrt{B}}\right) \leqslant \frac{B\mathbb{E}[\|N-\theta\|_H^2]}{0.02^2}\,.$$

Besides, as $\theta \in \Theta$,

$$\begin{aligned}\mathbb{E}\|N-\theta\|_H^2 &= \|\theta-\theta^*\|_H^2 + \mathrm{Tr}\big(H^{1/2}\Gamma H^{1/2}\big)\\ &\leqslant \frac{1}{100^2 B} + \frac{1}{100^2}\bigg(\Big(B^2 - \frac{1}{2d}\Big)\|H^{1/2}u^*\|^2 + \frac{1}{2d}\mathrm{Tr}(H)\bigg)\\ &= \frac{1}{100^2 B}\left(2 + \frac{d-1}{2d}\right) \leqslant \frac{2.5}{100^2 B}\,.\end{aligned} \tag{92}$$

This concludes the proof of Point 2.

We are now in position to bound the Kullback-Leibler divergence $D(\rho_\theta\|\pi_\Theta)$. By point 1, we have

$$D(\rho_\theta\|\pi_\Theta) = \int_{\mathcal{E}_\theta}\log\left(\frac{\mathrm{d}\rho_\theta}{\mathrm{d}\pi_\Theta}\right)\mathrm{d}\rho_\theta \leqslant \int_{\mathcal{E}_\theta}\log\left(\frac{1.5\mathrm{d}\widetilde{\rho}_\theta}{\mathrm{d}\pi_\Theta}\right)1.5\mathrm{d}\widetilde{\rho}_\theta = 1.5(\log 1.5 + D(\widetilde{\rho}_\theta\|\pi_\Theta))\,. \tag{93}$$

Now, denote $\gamma_\theta = \mathsf{N}(\theta, \Gamma_\theta)$ and $\gamma = \mathsf{N}(\theta^*, \Gamma)$ so $\widetilde{\rho}_\theta$ and π_Θ are the restrictions of γ_θ and γ. We have

$$D(\widetilde{\rho}_\theta \| \pi_\Theta) = \int_{\mathcal{E}_\theta} \frac{\mathrm{d}\gamma_\theta}{\gamma_\theta(\mathcal{E}_\theta)} \log\left(\frac{\mathrm{d}\gamma_\theta/\gamma_\theta(\mathcal{E}_\theta)}{\mathrm{d}\gamma/\gamma(\mathcal{E}_\theta)}\right) + \log\left(\frac{\gamma(\Theta')}{\gamma(\mathcal{E}_\theta)}\right).$$

Using Lemma 20 and Point 2 to bound the first term in the left-hand-side and Point 2 for the second, we get,

$$D(\rho_\theta \| \pi_\Theta) \leqslant 1.5\log(5) + 3D(\gamma_\theta \| \gamma)\,. \tag{94}$$

Finally, we compute the divergence from γ_θ to γ. Recall that, as $\det(\Gamma) = \det(\Gamma_\theta)$, it is equal to

$$D(\gamma_\theta \| \gamma) = \frac{1}{2}\big(\mathrm{Tr}(\Gamma^{-1/2}\Gamma_\theta\Gamma^{-1/2}) + \|\theta - \theta^*\|^2_{\Gamma^{-1}} - d\big)\,.$$

As $\Gamma^{-1} \preccurlyeq 2(100)^2 dBH$, we have on one side, by (92),

$$\mathrm{Tr}(\Gamma^{-1/2}\Gamma_\theta\Gamma^{-1/2}) \leqslant 2(100)^2 dB\,\mathrm{Tr}(H^{1/2}\Gamma_\theta H^{1/2}) \leqslant 3d\,,$$

and, on the other side,

$$\|\theta - \theta^*\|^2_{\Gamma^{-1}} \leqslant 2(100)^2 dB\|\theta - \theta^*\|^2_H \leqslant 2d\,.$$

Thus, $D(\gamma_\theta \| \gamma) \leqslant 2d$ and, by (94),

$$D(\rho_\theta \| \pi_\Theta) \leqslant 6.5d\,. \tag{95}$$

Combining this inequality with (90) concludes the proof. □

Conclusion. We apply PAC-Bayes inequality of Lemma 12 to the variables $Z_i(\theta_0, v')$ defined in (77). We use the upper bound on the Laplace transform given in Lemma 14, the one on the Kullback-Leibler divergence given in Lemma 16 and the lower bound on the linear term obtained in Lemma 15 to obtain that, for all $t > 0$, with probability at least $1 - e^{-t}$, simultaneously for every $\theta \in \Theta$ and $v \in S^{d-1}$,

$$15\big\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\big\rangle \geqslant 0.04 - 2.2\lambda B - \frac{\big(6.5 + \log\big(1 + 2\varepsilon^{-1}\big)\big)d + t}{\lambda n} - 12000\varepsilon^2 BR\,, \tag{96}$$

with $R = \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \exp\big(-0.99|\langle\theta, X_i\rangle|\big)\mathbf{1}\big(\|X_i\| \leqslant 2\sqrt{d}\big)$. As $\theta \in \Theta$, by Lemma 17, $\|\theta\| \geqslant 0.99B$ and $\sqrt{1 - \langle u, u^*\rangle^2} \leqslant 1/50B$, so $\{\theta/\|\theta\| : \theta \in \Theta\} \subset \mathsf{C}(u^*, 1/50B)$. Therefore, if $u = \theta/\|\theta\|$,

$$R \leqslant \sup_{u\in\mathsf{C}(u^*,1/50B)} \frac{1}{n}\sum_{i=1}^n \exp\big(-(0.99)^2 B|\langle u, X_i\rangle|\big)\mathbf{1}\big(\|X_i\| \leqslant 2\sqrt{d}\big)\,.$$

The right-hand side can be bounded using Lemma 21, we have, for any $t > 0$ such that $n \geqslant 1.1B(d + t)$, with probability larger than $1 - e^{-t}$,

$$\sup_{u\in\mathsf{C}(u^*,1/50B)} \frac{1}{n}\sum_{i=1}^n \exp\big(-(0.99)^2 B\,|\langle u, X_i\rangle|\big)\mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\} \leqslant \frac{4}{B}\,.$$

Plugging this bound into (96) shows that, for any $\lambda > 0$, with probability $1 - 2e^{-t}$

$$15\big\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\big\rangle \geqslant 0.04 - 2.2\lambda B - \frac{\big(6.5 + \log\big(1 + 2\varepsilon^{-1}\big)\big)d + t}{\lambda n} - 48000\varepsilon^2. \tag{97}$$

Choosing $\varepsilon = 1/2200$ so $48000\varepsilon^2 \leqslant 0.01$ and $6.5 + \log(1 + 2\varepsilon^{-1}) \leqslant 17.2$ shows that

$$15\big\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\big\rangle \geqslant 0.03 - 2.2\lambda B - \frac{17.2d + t}{\lambda n}\,. \tag{98}$$

Finally, we choose $\lambda = \sqrt{8(d+t)/n}$ to get

$$15\big\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\big\rangle \geqslant 0.03 - 2\sqrt{\frac{8(d+t)}{n}}\,.$$

For $n \geqslant 320000(d+t)$, this last lower bound is larger than 0.02, which concludes the proof.

6.4 Technical lemmas for the proof of Theorem 6

This section gathers technical tools used repeatedly in the proofs.

Lemma 17. *Let $\theta^* \in \mathbb{R}^d$ be such that $B = \|\theta^*\| > 1$, let $H = B^{-3}u_*u_*^\top + B^{-1}(I_d - u_*u_*^\top)$ and let $r \in (0,1)$. Then, for every $\theta \in \mathbb{R}^d$ such that $\|\theta - \theta^*\|_H \leqslant r/\sqrt{B}$, $u^* = \theta^*/\|\theta^*\|$ and $u = \theta/\|\theta\|$,*

1. $(1-r)B \leqslant \|\theta\| \leqslant (1+r)B$,
2. $\frac{\|\theta - \langle u^*,\theta\rangle u^*\|}{\|\theta\|} = \|u - \langle u^*, u\rangle u^*\| \leqslant \frac{r}{(1-r)B}$,
3. $\|u - u^*\| \leqslant \frac{\sqrt{2}r}{(1-r)B}$.

Proof. The constraint $\|\theta - \theta^*\|_H \leqslant r/\sqrt{B}$ can be written

$$\frac{(\langle\theta, u^*\rangle - B)^2}{B^3} + \frac{\|\theta - \langle\theta, u^*\rangle u^*\|^2}{B} \leqslant \frac{r^2}{B}\,. \tag{99}$$

For the upper bound in the first point, remark first that $\theta = (1+r)\theta^*$ satisfies (99) and $\|(1+r)\theta^*\| = (1+r)B$. Let now θ be such that $\|\theta - \theta^*\|_H \leqslant r/\sqrt{B}$ and let $\|\theta - \langle\theta, u^*\rangle u^*\| = \alpha$. By (99), $(\langle\theta, u^*\rangle - B)^2 \leqslant B(r^2 - \alpha^2)$. Therefore, as $B > 1$,

$$\begin{aligned}\|\theta\|^2 = \langle\theta, u^*\rangle^2 + \|\theta - \langle\theta, u^*\rangle u^*\|^2 &\leqslant B^2(1 + \sqrt{r^2 - \alpha^2})^2 + \alpha^2 \\ &\leqslant B^2(1 + 2\sqrt{r^2 - \alpha^2} + r^2) \leqslant B^2(1+r)^2 = \|(1+r)\theta^*\|^2\,.\end{aligned}$$

The lower bound is obtained using the same arguments.

The second point follows from the remark that by (99), we have $\|\theta - \langle\theta, u^*\rangle u^*\| \leqslant r$ and from the first point $\|\theta\| \geqslant (1-r)B$.

For the last point, we first remark that, as $|\langle\theta, u^*\rangle - B| < rB$, we have $\langle u, u^*\rangle > 0$. Then, we write

$$\|\theta - \langle\theta, u^*\rangle u^*\|^2 = \|\theta\|^2\|u - \langle u, u^*\rangle u^*\|^2 = \|\theta\|^2(1 - \langle u, u^*\rangle^2)\,.$$

Therefore, by (99),

$$\|u - u^*\|^2 = 2(1 - \langle u, u^*\rangle) \leqslant 2(1 - \langle u, u^*\rangle^2) \leqslant \frac{2r^2}{\|\theta\|^2}\,.$$

The proof is concluded by Point 1. □

The second lemma proves a lower bound on the expectation of $Z_i(\theta, v)$.

Lemma 18. *For any $\theta \in \mathbb{R}^d$, let*

$$\widetilde{H}(\theta) = \mathbb{E}\left[\mathbf{1}\left\{|\langle\theta, X\rangle| \leqslant 1;\ \|X\| \leqslant 2\sqrt{d}\right\} XX^\top\right].$$

For any θ such that $\|\theta - \theta^*\|_H \leqslant 1/10\sqrt{B}$, we have

$$\widetilde{H}(\theta) \succcurlyeq 0.05 \cdot H_\theta$$

Proof. Let $u = \theta/\|\theta\|$ and $v \in S^{d-1}$. We want to show that

$$\langle \widetilde{H}(\theta)v, v\rangle = \mathbb{E}\left[\mathbf{1}\left\{|\langle\theta, X\rangle| \leqslant 1;\ \|X\| \leqslant 2\sqrt{d}\right\} \langle v, X\rangle^2\right] \geqslant 0.05\langle H_\theta v, v\rangle\,.$$

write $v = \langle v, u\rangle u + (v - \langle v, u\rangle u)$. As $\langle\theta, X\rangle$ is independent of $\langle v - \langle v, u\rangle u, X\rangle$ and $\langle v - \langle v, u\rangle u, X\rangle \sim \mathsf{N}(0, 1 - \langle v, u\rangle^2)$, we have

$$\begin{aligned}\langle \widetilde{H}(\theta)v, v\rangle = \langle u, v\rangle^2 \mathbb{E}\big[\mathbf{1}\big\{|\langle\theta, X\rangle| \leqslant 1;\ \|X\| \leqslant 2\sqrt{d}\big\}\langle u, X\rangle^2\big] \\ + (1 - \langle u, v\rangle^2)\mathbb{E}[\mathbf{1}\left\{|\langle\theta, X\rangle| \leqslant 1;\ \|X\| \leqslant 2\sqrt{d}\right\}]\,. \end{aligned} \tag{100}$$

It remains to bound from below both expectations in the right-hand side term. Let us start with the second one, we have $\|X\|^2 = \langle u, X\rangle^2 + \|X - \langle u, X\rangle u\|^2$, where $X - \langle u, X\rangle u$ is a Gaussian vector independent from $\langle u, X\rangle$. Thus, if $\|X - \langle u, X\rangle u\|^2 \leqslant 4d - 1/\|\theta\|^2$ and $|\langle u, X\rangle| \leqslant 1/\|\theta\|$, then $|\langle\theta, X\rangle| \leqslant 1;\ \|X\| \leqslant 2\sqrt{d}$, so

$$\mathbb{E}\left[\mathbf{1}\left\{|\langle\theta, X\rangle| \leqslant 1;\ \|X\| \leqslant 2\sqrt{d}\right\}\right] \geqslant \mathbb{P}(|\langle u, X\rangle| \leqslant 1/\|\theta\|)\mathbb{P}(\|X - \langle u, X\rangle u\|^2 \leqslant 4d - 1/\|\theta\|^2)\,.$$

By Lemma 17, as $\|\theta - \theta^*\|_H \leqslant 1/10\sqrt{B}$,

$$\frac{1}{2} \leqslant 0.9 \cdot B \leqslant \|\theta\| \leqslant 1.1 \cdot B\,. \tag{101}$$

By Markov's inequality, we have thus

$$\mathbb{P}(\|X - \langle u, X\rangle u\|^2 \leqslant 4d - 1/\|\theta\|^2) \geqslant 1 - \frac{d-1}{4d - 1/\|\theta\|^2} \geqslant \frac{3}{4}\,.$$

As $x \mapsto x\exp(-x^2)$ is non-decreasing on $[0, 1/2]$,

$$\mathbb{P}(|\langle u, X\rangle| \leqslant 1/\|\theta\|) \geqslant \frac{2}{\|\theta\|}\frac{\exp(-\|\theta\|^2/2)}{\sqrt{2\pi}} \geqslant \frac{0.07}{B}\,.$$

Thus

$$\mathbb{E}[\mathbf{1}\left\{|\langle\theta, X\rangle| \leqslant 1;\ \|X\| \leqslant 2\sqrt{d}\right\}] \geqslant \frac{0.05}{B}\,.$$

We use the same arguments to bound the first expectation in the right hand side of (100), we get

$$\mathbb{E}\big[\mathbf{1}\big(|\langle\theta, X\rangle| \leqslant 1;\ \|X\| \leqslant 2\sqrt{d}\big)\langle u, X\rangle^2\big] \geqslant \frac{3}{4}\mathbb{E}\big[\mathbf{1}\{|\langle u, X\rangle| \leqslant 1/\|\theta\|\}\,\langle u, X\rangle^2\big]\,.$$

We have

$$\begin{aligned}\mathbb{E}[\mathbf{1}\{|\langle u, X\rangle| \leqslant 1/\|\theta\|\}\,\langle u, X\rangle^2] &\geqslant \frac{\exp(-1/2\|\theta\|^2)}{\sqrt{2\pi}}\int_{-1/\|\theta\|}^{1/\|\theta\|} x^2 \mathrm{d}x \\ &= \sqrt{\frac{2}{\pi}}\frac{\exp(-1/2\|\theta\|^2)}{3\|\theta\|^3} \geqslant \frac{0.18}{B^3}\,.\end{aligned}$$

Thus

$$\mathbb{E}\big[\mathbf{1}\big(|\langle\theta,X\rangle|\leqslant 1;\ \|X\|\leqslant 2\sqrt{d}\big)\langle u,X\rangle^2\big]\geqslant\frac{0.1}{B^3}\,.$$

Plugging these bounds into (100) yields

$$\langle\widetilde{H}(\theta)v,v\rangle\geqslant 0.05\bigg(\frac{2\langle u,v\rangle^2}{B^3}+\frac{(1-\langle u,v\rangle^2)}{B}\bigg)\geqslant 0.05\langle H_\theta v,v\rangle\,.$$ □

Lemma 19. *Let $r\in[0,1/10]$. For every $\theta\in\mathbb{R}^d$ such that $\|\theta-\theta^*\|_H\leqslant r/\sqrt{B}$, we have*

$$(1-2.35r)H\preccurlyeq H_\theta\preccurlyeq(1+2.35r)H\,.$$

Proof. Let $v\in S^{d-1}$, $u=\theta/\|\theta\|$, $u^*=\theta^*/\|\theta^*\|$, we want to compare

$$\langle Hv,v\rangle=\frac{1}{B^3}\langle u^*,v\rangle^2+\frac{1}{B}(1-\langle u^*,v\rangle^2),\qquad\langle H_\theta v,v\rangle=\frac{1}{B^3}\langle u,v\rangle^2+\frac{1}{B}(1-\langle u,v\rangle^2)\,.$$

We have

$$|\langle v,u\rangle|\leqslant|\langle v,u^*\rangle|+\|u-\langle u,u^*\rangle u^*\|\|v-\langle v,u^*\rangle u^*\|\,,$$
$$v-\langle u,v\rangle u=(v-\langle u^*,v\rangle u^*)-\langle v-\langle u^*,v\rangle u^*,u\rangle u+\langle u^*,v\rangle(u^*-\langle u,u^*\rangle u)\,.$$

By Lemma 17, $\|u-\langle u,u^*\rangle u^*\|\leqslant\frac{r}{(1-r)B}$. Using Cauchy-Schwarz inequality and $(a+b)^2\leqslant(1+r)a^2+(1+r^{-1})b^2$, we deduce

$$\langle u,v\rangle^2\leqslant(1+r)\bigg(\langle v,u^*\rangle^2+\frac{r}{(1-r)^2B^2}(1-\langle v,u^*\rangle^2)\bigg)\,,$$
$$1-\langle u,v\rangle^2\leqslant(1+r)\bigg((1-\langle v,u^*\rangle^2)+\frac{r}{(1-r)^2B^2}\langle v,u^*\rangle^2\bigg)\,.$$

Hence,

$$\begin{aligned}\langle H_\theta v,v\rangle&\leqslant(1+r)\bigg(\frac{1+r(1-r)^{-2}}{B^3}\langle u,v\rangle^2+\frac{1+r(1-r)^{-2}B^{-4}}{B}(1-\langle u,v\rangle^2)\bigg)\\&\leqslant(1+r)(1+r(1-r)^{-2})\langle Hv,v\rangle\leqslant(1+2.35r)\langle Hv,v\rangle\,,\end{aligned}$$

where the last inequality holds as $r\leqslant 0.1$. The lower bound is obtained using similar arguments. □

Lemma 20. *Let P,Q be probability measures and A an event such that $\mathsf{P}(A)>0$. One has*

$$D(\mathsf{P}_{|A}\|\mathsf{Q}_{|A})\leqslant\frac{1}{\mathsf{P}(A)}D(\mathsf{P}\|\mathsf{Q})\,.$$

Proof. Without loss of generality, let us assume that P and Q have densities p and q respectively, with respect to a common dominating measure μ (e.g. $\mathsf{P}+\mathsf{Q}$). Let also $p_{|A}$ and $q_{|A}$ denote their conditional densities. One has

$$D(\mathsf{P}\|\mathsf{Q})=\int_A p\log\Big(\frac{p}{q}\Big)\mathrm{d}\mu+\int_{A^c}p\log\Big(\frac{p}{q}\Big)\mathrm{d}\mu\,.\tag{102}$$

By symmetry we do the computations on the event A.

$$
\begin{aligned}
\int_A p \log\Big(\frac{p}{q}\Big)\mathrm{d}\mu &= \mathsf{P}(A)\int_A \frac{p}{\mathsf{P}(A)}\log\Big(\frac{p/\mathsf{P}(A)}{q/\mathsf{Q}(A)}\cdot\frac{\mathsf{P}(A)}{\mathsf{Q}(A)}\Big) \\
&= \mathsf{P}(A)\int_A p_{|A}\log\Big(\frac{p_{|A}}{q_{|A}}\Big) + \mathsf{P}(A)\log\Big(\frac{\mathsf{P}(A)}{\mathsf{Q}(A)}\Big) \\
&= \mathsf{P}(A)D(\mathsf{P}_{|A}\|\mathsf{Q}_{|A}) + \mathsf{P}(A)\log\Big(\frac{\mathsf{P}(A)}{\mathsf{Q}(A)}\Big)\,.
\end{aligned}
$$

Hence, by symmetry,

$$
D(\mathsf{P}\|\mathsf{Q}) = \mathsf{P}(A)D(\mathsf{P}_{|A}\|\mathsf{Q}_{|A}) + \mathsf{P}(A^c)D(\mathsf{P}_{|A^c}\|\mathsf{Q}_{|A^c}) + D(\mathsf{P}(A)\|\mathsf{Q}(A))\,,
$$

where $D(\mathsf{P}(A)\|\mathsf{Q}(A))$ denotes the divergence between Bernoulli distributions with parameters $\mathsf{P}(A), \mathsf{Q}(A)$. The last two terms being non-negative, the claim is proved. □

Lemma 21. *Let $h \in (0,1)$. Then, for any $t > 0$ such that $n \geqslant 1.1B(d+t)$, with probability larger than $1-e^{-t}$,*

$$
\sup_{u\in\mathsf{C}(u^*,1/10B)} \frac{1}{n}\sum_{i=1}^n \exp\big(-hB\,|\langle u, X_i\rangle|\big)\mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\} \leqslant \frac{3.6}{hB}\,.
$$

Proof. For any $u \in S^{d-1}$, let

$$
Z_i(u) = \exp\big(-hB\,|\langle u, X_i\rangle|\big)\mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\,.
$$

To bound $\sup_{u\in\mathsf{C}(u^*,1/10B)} \frac{1}{n}\sum_{i=1}^n Z_i(u)$, we apply the PAC-Bayes inequality (recalled in Lemma 12), and for this, we bound all terms appearing in this inequality.

We apply this inequality with the collection of posteriors $(\rho_u)_{u\in\mathsf{C}(u^*,1/10B)}$ and the prior π defined as follows: For every $u \in \mathsf{C}(u^*, 1/10B)$, ρ_u is the uniform distribution over $\mathsf{C}(u, 1/10B)$ and π is the uniform distribution over $\mathsf{C}(u^*, \sqrt{2}/5B)$, chosen such that, for any $u \in \mathsf{C}(u^*, 1/10B)$, the support of ρ_u is included into the one of π.

Bounds on the Kullback-Leibler divergence. We prove in this paragraph that

$$
D(\rho_u\|\pi) \leqslant 1.1\cdot d\,. \tag{103}
$$

We have directly:

$$
D(\rho_u\|\pi) = \log\Big(\frac{\mathrm{Vol}_{d-1}(\mathsf{C}(u^*,\sqrt{2}/5B))}{\mathrm{Vol}_{d-1}(\mathsf{C}(u,1/10B))}\Big)\,,
$$

where Vol_{d-1} denote the Lebesgue measure on S^{d-1}. To compute these volumes, we let U denote a random variable uniformly distributed on the sphere and $u \in S^{d-1}$. It is a standard fact that $\langle u, U\rangle$ has density given by

$$
f(s) = c_d\left(1-s^2\right)^{\frac{d-3}{2}}\mathbf{1}(-1 \leqslant s \leqslant 1)\,,
$$

where c_d is a normalizing constant. Therefore, for any $\varepsilon \in (0,1)$,

$$
\mathrm{Vol}_{d-1}(\mathsf{C}(u,\varepsilon)) = \mathbb{P}\big(\langle U, u\rangle > \sqrt{1-\varepsilon^2}\big) = c_d\int_{\sqrt{1-\varepsilon^2}}^1 \left(1-s^2\right)^{\frac{d-3}{2}}\mathrm{d}s = \int_0^{\varepsilon^2}\frac{t^{(d-3)/2}}{\sqrt{1-t}}\mathrm{d}t\,.
$$

Hence,

$$\frac{2c_d\varepsilon^{d-1}}{d-1} \leqslant \mathrm{Vol}_{d-1}(\mathsf{C}(u,\varepsilon)) \leqslant \frac{1}{\sqrt{1-\varepsilon^2}}\frac{2c_d\varepsilon^{d-1}}{d-1}\,.$$

Therefore,

$$D(\rho_u\|\pi) \leqslant (d-1)\log\big(2\sqrt{2}\big) + \frac{1}{2}\log\left(\frac{1}{1-\sqrt{2}/5B}\right).$$

As $\frac{1}{2}\log\big(\frac{1}{1-\sqrt{2}/5B}\big) \leqslant \log\big(2\sqrt{2}\big)$, further bounding numerical constants gives the bound (90).

Bounds on the Laplace transform. In this paragraph, we prove that, for any $\lambda \in [0,1]$,

$$\log \mathbb{E}[\exp(\lambda Z_i(u))] \leqslant \frac{1}{hB}\big(\lambda+\lambda^2\big)\,. \tag{104}$$

Indeed, the random variable $Z_i(u) \leqslant \exp(-hB|\langle u, X\rangle|) \leqslant 1$ a.s., so, for any $\lambda \in [0,1]$,

$$\mathbb{E}[\exp(\lambda Z_i(u))] = 1 + \lambda\mathbb{E}[Z_i(u)] + \lambda^2\mathbb{E}[Z_i(u)^2]\,.$$

The moments of $Z_i(u)$ can then be bounded as follows:

$$\mathbb{E}[Z_i(u)] \leqslant \mathbb{E}[\exp(-hB|\langle u, X\rangle|)] = 2\int_0^{+\infty}\exp\left(-hBx-\frac{x^2}{2}\right)\frac{\mathrm{d}x}{\sqrt{2\pi}} \leqslant \frac{1}{hB}\,,$$

and, similarly, $\mathbb{E}[Z_i(u)^2] \leqslant \frac{1}{\sqrt{2\pi}}\frac{1}{hB}$. The conclusion then follows by using $\log(1+s) \leqslant s$ valid for any $s > -1$.

Lower bound on the linear term. In this paragraph, we show that

$$\int_{\mathsf{C}(u,1/10B)} Z_i(u_0)\rho_u(\mathrm{d}u_0) \geqslant 0.85 \cdot \mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\exp\Big(-hB|\langle u, X_i\rangle|\Big)\,. \tag{105}$$

We have by Jensen's inequality,

$$\begin{aligned}\int_{\mathsf{C}(u,1/10B)} Z_i(u_0)\rho_u(\mathrm{d}u_0) &= \mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\int_{\mathsf{C}(u,1/10B)}\exp(-hB|\langle u_0, X_i\rangle|)\rho_u(\mathrm{d}u_0)\\ &\geqslant \mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\exp\left(-hB\int_{\mathsf{C}(u,1/10B)}|\langle u_0, X_i\rangle|\rho_u(\mathrm{d}u_0)\right)\\ &\geqslant \mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\exp\left(-hB\left(\int_{\mathsf{C}(u,1/10B)}\langle u_0, X_i\rangle^2\rho_u(\mathrm{d}u_0)\right)^{1/2}\right).\end{aligned}$$

This proves the result as, by [Mou22, Eq (42) and (43)] and Fact 6,

$$\int_{\mathsf{C}(u,1/10B)}\langle u_0, X_i\rangle^2\rho_u(\mathrm{d}u_0) \leqslant \langle u, X_i\rangle^2 + \frac{2}{100(d-1)B}\|X_i\|^2\,.$$

Conclusion of the proof. By the PAC-Bayes inequality, we have thus, for any $t > 0$, with probability at least $1-e^{-t}$, simultaneously for any $u \in \mathsf{C}(u^*, 1/10B)$ and any $\lambda \in [0,1]$,

$$\begin{aligned}\frac{0.85}{n}\sum_{i=1}^n \mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\}\exp\big(-hB|\langle u, X_i\rangle|\big) &\leqslant \frac{1}{hB}\big(1+\lambda\big) + \frac{1.1\cdot d + t}{\lambda n}\\ &\leqslant \frac{1}{hB}\big(1+\lambda\big) + \frac{1}{\lambda B}\,.\end{aligned}$$

We conclude by taking $\lambda = 1$. □

7 Linear separation: Proof of Theorem 2

In this section, we provide the proof of Theorem 2 on linear separation for small sample sizes. In particular, Theorem 2 can be seen as a non-asymptotic analogue of (one direction in) the result of [CS20].

Throughout this section, we assume that the design is isotropic Gaussian and that the model is well-specified. Specifically, given a dimension $d \geqslant 1$, a parameter $\theta^* \in \mathbb{R}^d$ with norm $\beta = \|\theta^*\|$ and a sample size $n \geqslant d$, the dataset consists of n i.i.d. random pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ with $X_i \sim \mathsf{N}(0, I_d)$ and $\mathbb{P}(Y_i = 1|X_i) = \sigma(\langle \theta^*, X_i \rangle)$. Note that if $n \geqslant d$, then almost surely $X_1, \dots, X_n$ span $\mathbb{R}^d$, hence (by the discussion in the introduction) the MLE exists if and only if the dataset is not linearly separated. In addition, by rotation invariance of the standard Gaussian distribution in $\mathbb{R}^d$, the probability of linear separation (non-existence of the MLE) only depends on θ^* through its norm β.

We start by recalling the result of Candès and Sur [CS20] in the proportional asymptotic regime. In this setting, we consider a sequence of parameters $(d_n, \theta_n^*)_{n \geqslant 1}$ with $d_n/n \to \gamma \in (0, 1)$ and $\beta_n = \|\theta_n^*\| \to \beta \in \mathbb{R}^+$.

Theorem 7 ([CS20], Theorem 2.2)**.** *In the setting described above, one has*

$$\mathbb{P}(\text{MLE exists}) \underset{n\to\infty}{\longrightarrow} \begin{cases} 1 \ \textit{if}\ \gamma < h(\beta) \\ 0 \ \textit{if}\ \gamma > h(\beta)\,, \end{cases} \tag{106}$$

where the function $h : \mathbb{R}^+ \to [0, 1]$ is defined as follows. For $\beta \in \mathbb{R}^+$, let (X', Y'_β) be a random pair in $\mathbb{R} \times \{-1, 1\}$, with $X' \sim \mathsf{N}(0, 1)$ and $\mathbb{P}(Y'_\beta = 1|X') = \sigma(\beta X')$, and let $V_\beta = Y'_\beta X'$. In addition, let $Z \sim \mathsf{N}(0, 1)$ be independent of V_β. Then,

$$h(\beta) = \min_{t \in \mathbb{R}} \mathbb{E}\big[(tV_\beta - Z)_+^2\big]\,. \tag{107}$$

7.1 Proof of Theorem 2

The proof of Theorem 2 relies on the approximate kinematic formula from conic geometry recalled hereafter. We first recall the definition of the statistical dimension of a cone $\mathbf{C}$ in $\mathbb{R}^n$. It is defined as $\delta(\mathbf{C}) = \mathbb{E}\|\Pi_{\mathbf{C}} \boldsymbol{Z}\|^2$ where $\Pi_{\mathbf{C}}$ is the Euclidean projection on $\mathbf{C}$ and $\boldsymbol{Z} \sim \mathsf{N}(0, I_n)$.

Lemma 22 (Approximate kinematic formula, Theorem 7.1 in [ALMT14])**.** *Let $\mathcal{L}$ be a random subspace of $\mathbb{R}^n$ drawn uniformly from all subspaces of dimension k and let $\mathbf{C} \subset \mathbb{R}^n$ be a cone. For all $t > 0$, if*

$$n - k \leqslant \delta(\mathbf{C}) - t\,, \tag{108}$$

then

$$\mathbb{P}\left(\mathbf{C} \cap \mathcal{L} \neq \{0\}\right) \geqslant 1 - 4 \exp\left(-\frac{t^2/8}{\min\{\delta(\mathbf{C}), n - \delta(\mathbf{C})\} + t}\right).$$

We can now proceed with the proof of Theorem 2. First, if $n \leqslant d$, a simple induction shows that almost surely, the points $X_1, \dots, X_n$ are linearly independent in $\mathbb{R}^d$. Hence, there exists $\theta \in \mathbb{R}^d$ such that for $i = 1, \dots, n$, one has $\langle \theta, X_i \rangle = Y_i$ and thus $Y_i \langle \theta, X_i \rangle = Y_i^2 = 1 > 0$. Thus $\inf_{\theta' \in \mathbb{R}^d} \widehat{L}_n(\theta') = 0$, but $\widehat{L}_n > 0$ on $\mathbb{R}^d$ and thus $\widehat{L}_n$ admits no global minimizer in $\mathbb{R}^d$. We thus assume from now on that $n > d$. Since $n \leqslant Bd/23000$, this implies that $\|\theta^*\| = B > e$.

First, following [CS20], we express the probability that the MLE does not exist as the probability that some random cone non-trivially intersects a random subspace of dimension $d-1$ in $\mathbb{R}^n$. Let $\{(X_i, Y_i)\}_{1 \leqslant i \leqslant n}$ denote the dataset, where all the X_i's are independently drawn from $\mathsf{N}(0, I_d)$. Using the rotational invariance of the standard Gaussian distribution, we can assume without loss of generality that for every $i \in \{1, \ldots, n\}$, $\mathbb{P}(Y_i = 1|X_i) = \sigma(BX_i^1)$ where X_i^j denotes the j-th coordinate of X_i for every $j \in \{1, \ldots, d\}$. Below we let $U_i = X_i^1$ and $V_i = Y_i U_i$ for all $i \in \{1, \ldots, n\}$. Let also $\boldsymbol{V} = (V_1, \ldots, V_n) \in \mathbb{R}^n$ and $\Lambda = \mathbb{R}\boldsymbol{V} + \mathbb{R}_+^n$, which is a random cone in $\mathbb{R}^n$. The proof of Theorem 2 relies on the following observation.

Lemma 23. *Let $\mathcal{L}$ be a random subspace drawn uniformly from all subspaces of dimension $d-1$ in $\mathbb{R}^n$. Then*

$$\mathbb{P}(\text{MLE does not exist}) \geqslant \mathbb{P}(\Lambda \cap \mathcal{L} \neq \{0\}) .$$

The proof of this result is postponed to the end of the section and is a straightforward adaptation of [CS20, Proposition 2] to the case where the model does not include an intercept.

In view of this characterization, we want to apply Lemma 22 to the cone Λ but we cannot do it in a straightforward way, as this cone is random. We therefore show that the sufficient condition (108) regarding the statistical dimension of Λ is satisfied with high probability. Hereafter we denote by $E = \{\Lambda \cap \mathcal{L} \neq \{0\}\}$, and for every $t \geqslant 0$, we define the event

$$A_t = \{n - d + 1 \leqslant \delta(\Lambda) - t\} . \tag{109}$$

Our main task in this proof is to show that

$$\mathbb{P}(A_{\alpha d}) \geqslant 1 - \exp\big(-\max\big\{\kappa\sqrt{d}, \kappa^2 d/B^2\big\}\big) - 2e^{-\tau d} \tag{110}$$

for some $\tau \in (0, 1/2)$ and $\alpha \in (1/2, 1)$. We now establish (110) with explicit constants. Conditionning on $\boldsymbol{V}$, one has $\delta(\Lambda) = n - \mathbb{E}_{\boldsymbol{Z}}\big[\mathrm{dist}(\boldsymbol{Z}, \Lambda)^2 | \boldsymbol{V}\big]$, where $\mathbb{E}_{\boldsymbol{Z}}$ denotes the expectation with respect to $\boldsymbol{Z}$. Throughout the rest of this proof, we let

$$F(\boldsymbol{V}) = \mathbb{E}_{\boldsymbol{Z}}\big[\mathrm{dist}(\boldsymbol{Z}, \Lambda)^2 | \boldsymbol{V}\big] = \mathbb{E}_{\boldsymbol{Z}}\left[\min_{\lambda \in \mathbb{R}} \sum_{i=1}^n (\lambda V_i - Z_i)_+^2 \middle| \boldsymbol{V}\right] .$$

This way, we will prove (110) by showing that $F(\boldsymbol{V}) \ll d$ with high probability. It is reasonable to believe that this is true in the regime of interest where $n \leqslant Bd/(C_0\kappa)$. Indeed, we note that

$$\mathbb{E}[F(\boldsymbol{V})] \leqslant \min_{\lambda \in \mathbb{R}} \mathbb{E}\left[\sum_{i=1}^n (\lambda V_i - Z_i)_+^2\right] = nh(B) ,$$

where h is the phase transition function (107) from [CS20]. In addition, one can show that $h(B) \lesssim 1/B$ (we will not need this exact claim, hence we will not prove it, although it could be deduced from the analysis below). Thus $\mathbb{E}[F(\boldsymbol{V})] \lesssim n/B \lesssim d$.

Hereafter, we let $\psi : s \in \mathbb{R} \mapsto \mathbb{E}\big[(s - Z)_+^2\big]$ with $Z \sim \mathsf{N}(0, 1)$ and start by bounding

$$F(\boldsymbol{V}) = \mathbb{E}\left[\min_{\lambda \in \mathbb{R}} \sum_{i=1}^n (\lambda V_i - Z_i)_+^2 \middle| \boldsymbol{V}\right] \leqslant \min_{\lambda \in \mathbb{R}} \sum_{i=1}^n \mathbb{E}\big[(\lambda V_i - Z_i)_+^2 | V_i\big] = \min_{\lambda \in \mathbb{R}} \sum_{i=1}^n \psi(\lambda V_i) .$$

By Fact 4, for all $\lambda \geqslant 0$,

$$\psi(-\lambda V_i) \leqslant e^{-\lambda^2 U_i^2/2} + \mathbf{1}(Y_i U_i \leqslant 0) + U_i^2 \mathbf{1}(Y_i U_i \leqslant 0) . \tag{111}$$

We thus define for all $i \in \{1, \ldots, n\}$ and $\lambda \in \mathbb{R}$ the variables

$$\zeta_{i,\lambda} = e^{-\lambda^2 U_i^2/2} , \quad \varepsilon_i = \mathbf{1}(Y_i U_i \leqslant 0) , \quad \psi_i = U_i^2 \mathbf{1}(Y_i U_i \leqslant 0) , \tag{112}$$

so that we can further bound

$$F(\boldsymbol{V}) \leqslant \min_{\lambda>0}\bigg\{\sum_{i=1}^{n}\zeta_{i,\lambda}+\sum_{i=1}^{n}\varepsilon_i+\lambda^2\sum_{i=1}^{n}\psi_i\bigg\}\,. \tag{113}$$

We now separately bound from above the three sums and then optimize the resulting bound over λ. We use Bernstein's inequality to bound the first two sums involving the $\zeta_{i,\lambda}$ and ε_i, but bounding the sum of the ψ_i's is a more subtle task, for which we resort to Latała's bound on the moments of sums of independent variables [Lat97] to control the moments of $\sum_{i=1}^{n}\psi_i$. Let us start with the first sum. For every $i \in \{1,\dots,n\}$, every $\lambda>0$ and $k\in\{1,2\}$,

$$\mathbb{E}[\zeta_{i,\lambda}^k]=\mathbb{E}\Big[\exp\Big(-\frac{k\lambda^2U_i^2}{2}\Big)\Big]=\int_{\mathbb{R}}\frac{e^{-(k\lambda^2+1)u^2/2}}{\sqrt{2\pi}}\mathrm{d}u=\frac{1}{\sqrt{k\lambda^2+1}}\leqslant\frac{1}{\lambda}\,.$$

Since in addition $\zeta_{i,\lambda}\leqslant 1$ almost surely, by Lemma 36 and the second and third points of Lemma 35, for all $t\geqslant 0$, with probability larger than $1-e^{-t}$,

$$\sum_{i=1}^{n}\zeta_{i,\lambda}\leqslant\frac{n}{\lambda}+\sqrt{\frac{2nt}{\lambda}}+3t\,. \tag{114}$$

Regarding the second sum, inequality (51) shows that for every i, $\mathbb{E}[\varepsilon_i]=\mathbb{E}[\exp(-B|U_i|)]\leqslant B^{-1}$. Since $\varepsilon_i^2=\varepsilon_i$ and $\varepsilon_i\leqslant 1$, the same argument as before shows that for all $t\geqslant 0$, it holds with probability larger than $1-e^{-t}$ that

$$\sum_{i=1}^{n}\varepsilon_i\leqslant\frac{n}{B}+\sqrt{\frac{2nt}{B}}+3t\,. \tag{115}$$

Finally, we turn to the control of the last sum, for which we use Latała's bound, recalled hereafter.

Lemma 24 ([Lat97], Corollary 1)**.** *Let $\xi,\xi_1,\dots,\xi_n$ be i.i.d. nonnegative random variables. Then for any $p\geqslant 1$,*

$$\bigg\|\sum_{i=1}^{n}\xi_i\bigg\|_p\leqslant 2e^2\sup\Big\{\frac{p}{s}\Big(\frac{n}{p}\Big)^{1/s}\|\xi\|_s : 1\vee\frac{p}{n}\leqslant s\leqslant p\Big\}\,.$$

From now on, we let $S_n=\psi_1+\cdots+\psi_n$ and $p\in[1,n]$. By Markov's inequality, $\mathbb{P}(S_n\leqslant e\|S_n\|_p)\geqslant 1-e^{-p}$, hence we want to bound $\|S_n\|_p$ from above by some factor of d, with p as large as possible. We are thus led to bound the individual L^s norms $\|\psi_i\|_s$ and then optimize over $s\in[1,p]$.

Bound on individual moments. Regarding the bound on $\|\psi_i\|_s$, the result is obtained by taking advantage of either the fact that U_i^2 is sub-exponential (by neglecting the indicator) or by conditioning on U_i, which allows to use an exponential moment inequality. Let us formalize this. Let U,Y,ψ denote random variables having the same distribution as U_i,Y_i,ψ_i. On the one hand, $\psi\leqslant U^2$, so for all $s\geqslant 1$,

$$\mathbb{E}[\psi^s]\leqslant\mathbb{E}[|U|^{2s}]=\frac{2^s}{\sqrt{2\pi}}\Gamma\Big(s+\frac{1}{2}\Big)\,.$$

Hence, using [OLBC10, Eq. (5.6.1)] and simplifying we obtain

$$\|\psi\|_s=\mathbb{E}[\psi^{2s}]^{1/s}\leqslant(3/e)s\,. \tag{116}$$

On the other hand, we use the fact that conditionally on U, $\{YU \leqslant 0\}$ happens with exponentially small probability. More precisely, we write

$$\begin{aligned}\mathbb{E}[\psi^s] &= \mathbb{E}\big[|U|^{2s}\mathbb{E}[\mathbf{1}(YU \leqslant 0)|U]\big] = \mathbb{E}\big[|U|^{2s}\sigma(-B|U|)\big] \\ &\leqslant \mathbb{E}\big[|U|^{2s}\exp(-B|U|)\big] \leqslant \sqrt{\frac{2}{\pi}}\frac{\Gamma(2s+1)}{B^{2s+1}}\,.\end{aligned}$$

We then bound in a similar way $\Gamma(2s+1)^{1/s}$ and thus, combining the previous two bounds we deduce that

$$\|\psi\|_s \leqslant \frac{9}{e^2}\min\Big\{\frac{s^2}{B^2B^{1/s}}, s\Big\}\,. \tag{117}$$

Upper bound on the supremum. Using the control on the moments of the ψ_i's (117), it follows from Latała's inequality (Lemma 24) that

$$\|S_n\|_p \leqslant \frac{18}{B^2}\sup\Big\{\min\Big\{ps\Big(\frac{n}{pB}\Big)^{1/s},\, B^2p\Big(\frac{n}{p}\Big)^{1/s}\Big\};\, 1 \leqslant s \leqslant p\Big\}\,. \tag{118}$$

We now proceed with a bound on the supremum in the right-hand side by a function of p, n and B. For this technical step, we define for every $s > 0$

$$G(s) = \min\Big\{ps\Big(\frac{n}{pB}\Big)^{1/s},\, B^2p\Big(\frac{n}{p}\Big)^{1/s}\Big\},\quad G_1(s) = ps\Big(\frac{n}{pB}\Big)^{1/s},\quad G_2(s) = B^2p\Big(\frac{n}{p}\Big)^{1/s},$$

and then bound $M(p) = \sup_{1\leqslant s\leqslant p} G(s)$ for every $p \geqslant 1$. We first note that G_2 decreases on $(0, +\infty)$ and that $G_1(s) \leqslant G_2(s)$ for every $s \leqslant B^2$. Let also

$$g(s) = \log\Big(s\Big(\frac{n}{pB}\Big)^{1/s}\Big) = \frac{1}{s}\log\Big(\frac{n}{pB}\Big) + \log s\,.$$

Then

$$g'(s) = \frac{1}{s}\Big(1 - \frac{1}{s}\log\Big(\frac{n}{pB}\Big)\Big)\,.$$

Now let $s_1 = \log(\frac{n}{pB})$. We first deal with the case where $s_1 \geqslant \min\{p, B^2\}$. In this configuration, G_1 decreases on $[1, s_1]$, hence G decreases on $[1, p]$ and therefore $M(p) = G_1(1) = n/B$. From now on we assume that $s_1 < \min\{p, B^2\}$.

If $s_1 \leqslant 1$, $g'(s) > 0$ for every $s > 1$, hence G_1 increases on $[1, +\infty)$. Since G_2 decreases on $(0, +\infty)$, we deduce that the supremum is attained at either the value s_c where G_1 and G_2 coincide or at p, depending on whether p is smaller or larger than s_c, hence $M(p) = \min\{G_1(p), G(s_c)\}$. In addition s_c is solution to $s = B^{2+1/s}$, therefore it (i) does not depend on p and (ii) is slightly larger than B^2 (more precisely, $G_1(B^2) \leqslant G_2(B^2)$ but these quantities only differ by a multiplicative constant). Consequently $G(s_c) = G_2(s_c) \leqslant G_2(B^2)$. Using the fact that $p^{1/p} = e^{\log p/p} \geqslant 1$ (since $p \geqslant 1$), we deduce that in this configuration,

$$\|S_n\|_p \leqslant \frac{18}{B^2}\min\Big\{p^2\Big(\frac{n}{B}\Big)^{1/p},\, B^2p\Big(\frac{n}{p}\Big)^{1/B^2}\Big\}\,.$$

Now, if $s_1 > 1$, G decreases on $[1, s_1]$ and increases on $[s_1, \min\{s_c, p\}]$, then decreases again on $[\min\{s_c, p\}, p]$ as it coincides with G_2 on this last segment. Hence the only difference with the previous case is that the supremum might be attained at $s = 1$.

Putting everything together, we conclude that

$$\|S_n\|_p \leqslant \frac{18}{B^2}\max\Big\{\frac{n}{B}, \min\Big\{p^2\Big(\frac{n}{B}\Big)^{1/p},\, B^2p\Big(\frac{n}{p}\Big)^{1/B^2}\Big\}\Big\}\,. \tag{119}$$

High-probability upper bound on $\sum_{i=1}^n \psi_i$. Using the bound on the moments of S_n established above, we apply Markov's inequality to derive a high probability bound for S_n. To that end, we let p be as large as possible under the constraint that $\|S_n\|_p$ does not exceed $O(\kappa^2 d)$, namely $\|S_n\|_p \leqslant L_0\kappa^2 d/B^2$, where L_0 only depends on C_0. We prove that this is achieved by taking

$$p = \max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}\,. \tag{120}$$

To do so, we first show that if $p = \kappa\sqrt{d}$, then

$$p^2\Big(\frac{n}{B}\Big)^{1/p} \leqslant \kappa^2 d \exp\Big(\frac{2}{e\sqrt{C_0}}\Big)\,. \tag{121}$$

Using that $n/B \leqslant d/(C_0\kappa)$,

$$p^2\Big(\frac{n}{B}\Big)^{1/p} \leqslant \kappa^2 d\Big(\frac{d}{C_0\kappa}\Big)^{1/(\kappa\sqrt{d})} = \kappa^2 d \exp\left(\frac{\log\big(d/(C_0\kappa)\big)}{\kappa\sqrt{d}}\right).$$

The function $h: t \mapsto \log(t)/\sqrt{t}$ reaches its maximum at $t = e^2$ and thus satisfies $h(t) \leqslant 2/e$ for all $t > 0$. Hence

$$\frac{\log\big(d/(C_0\kappa)\big)}{\kappa\sqrt{d}} \leqslant \frac{2}{e\kappa^{3/2}\sqrt{C_0}}\,, \tag{122}$$

from which (121) follows, since $\kappa \geqslant 1$.

Similarly, if $p = \kappa^2 d/B^2$, using that $n/d \leqslant B/(C_0\kappa)$, one has

$$B^2 p\Big(\frac{n}{p}\Big)^{1/B^2} \leqslant \kappa^2 d\bigg(\frac{1}{C_0}\Big(\frac{B}{\kappa}\Big)^3\bigg)^{1/B^2}.$$

Using the same argument as before, we find that

$$\bigg(\frac{1}{C_0}\Big(\frac{B}{\kappa}\Big)^3\bigg)^{1/B^2} \leqslant \exp\left(\frac{3}{2eC_0^{2/3}}\right).$$

It is clear that $\exp(3/\big(2eC_0^{2/3}\big)) \geqslant \exp(2/(e\sqrt{C_0}))$, hence, with $p = \max\big\{\kappa\sqrt{d}, \kappa^2 d/B^2\big\}$

$$\min\left\{p^2\Big(\frac{n}{B}\Big)^{1/p}, B^2 p\Big(\frac{n}{p}\Big)^{1/B^2}\right\} \leqslant L_0\kappa^2 d\,, \quad L_0 = \exp\left(\frac{3}{2eC_0^{2/3}}\right).$$

Plugging this in (119) and using again the assumption $n/B \leqslant d/(C_0\kappa)$, we deduce that for $p = \max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}$,

$$\|S_n\|_p \leqslant \frac{18}{B^2}\max\Big\{\frac{d}{C_0\kappa}, L_0\kappa^2 d\Big\} = 18L_0\frac{\kappa^2 d}{B^2}\,. \tag{123}$$

High-probability bound on the statistical dimension. We now apply Markov's inequality using the moment bound (123). This yields

$$\mathbb{P}\Big(S_n \leqslant \frac{18eL_0\kappa^2 d}{B^2}\Big) \geqslant 1 - \exp\big(-\max\big\{\kappa\sqrt{d}, \kappa^2 d/B^2\big\}\big)\,.$$

Finally, we combine this with (114) and (115), showing that for every $\lambda > 0$ and every $t \geqslant 0$, it holds with probability larger than $1 - e^{-\max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}} - 2e^{-t}$ that

$$\sum_{i=1}^n \psi(-\lambda V_i) \leqslant \frac{n}{B} + \sqrt{\frac{2nt}{B}} + \frac{n}{\lambda} + \sqrt{\frac{2nt}{\lambda}} + 6t + \frac{L\lambda^2\kappa^2}{B^2}d\,, \quad L = 18eL_0\,.$$

We then set $t = \tau d$ for some $\tau \in (0,1)$. As $n \leqslant Bd/(C_0\kappa)$, the above rewrites

$$\sum_{i=1}^n \psi(-\lambda V_i) \leqslant \frac{d}{C_0\kappa} + \sqrt{\frac{2\tau}{C_0\kappa}}d + \frac{Bd}{C_0\kappa\lambda} + \sqrt{\frac{2B\tau}{C_0\kappa\lambda}}d + 6\tau d + \frac{L\lambda^2\kappa^2}{B^2}d\,.$$

Then, by the arithmetic mean–geometric mean inequality,

$$\sum_{i=1}^n \psi(-\lambda V_i) \leqslant \left[\frac{2}{C_0\kappa} + 7\tau + \frac{2B}{C_0\kappa\lambda} + \frac{L\kappa^2\lambda^2}{B^2}\right]d\,. \tag{124}$$

We now optimize the terms depending on λ and write, using the arithmetic mean–geometric mean inequality

$$\frac{2B}{C_0\kappa\lambda} + \frac{L\kappa^2\lambda^2}{B^2} = \frac{2}{3}\cdot\frac{3B}{C_0\kappa\lambda} + \frac{1}{3}\cdot\frac{3L\kappa^2\lambda^2}{B^2} \geqslant \left(\frac{3B}{C_0\kappa\lambda}\right)^{2/3}\left(\frac{3L\kappa^2\lambda^2}{B^2}\right)^{1/3}.$$

The last inequality is an equality if $\lambda = \lambda^*$, the value such that $\frac{3B}{C_0\kappa\lambda} = \frac{3L\kappa^2\lambda^2}{B^2}$. Simplifying the constants yields

$$\inf_{\lambda>0}\left\{\frac{2B}{C_0\kappa\lambda} + \frac{L\kappa^2\lambda^2}{B^2}\right\} = 3\cdot(18e)^{1/3}\frac{\exp\left(\frac{1}{2eC_0^{2/3}}\right)}{C_0^{2/3}} \leqslant \frac{13.2}{C_0^{2/3}}\,.$$

We finally plug this in (124) and obtain that

$$\mathbb{P}\left(\sum_{i=1}^n \psi(-\lambda^* V_i) \leqslant \left[7\tau + \frac{2}{C_0} + \frac{13.2}{C_0^{2/3}}\right]d\right) \geqslant 1 - e^{-\max\{\kappa\sqrt{d},\kappa^2 d/B^2\}} - 2e^{-\tau d}\,.$$

We choose C_0 large enough so that $2C_0^{-1} \leqslant 0.8C_0^{-2/3}$. Then the above rewrites in particular

$$\mathbb{P}\big(F(\boldsymbol{V}) \leqslant \big[7\tau + 14C_0^{-2/3}\big]d\big) \geqslant 1 - e^{-\max\{\kappa\sqrt{d},\kappa^2 d/B^2\}} - 2e^{-\tau d}\,.$$

Given $\alpha \in (0,1)$, for any $\tau \in (0,1)$ and $C_0 \geqslant 1$ such that

$$\big(7\tau + 14C_0^{-2/3}\big)d \leqslant d - 1 - \alpha d \tag{125}$$

it holds on the last event that $F(\boldsymbol{V}) \leqslant d - 1 - \alpha d$. Equivalently, recalling that $\delta(\Lambda) = n - F(\boldsymbol{V})$, this rewrites

$$\mathbb{P}(A_{\alpha d}) = \mathbb{P}\big(n - d + 1 \leqslant \delta(\Lambda) - \alpha d\big) \geqslant 1 - e^{-\max\{\kappa\sqrt{d},\kappa^2 d/B^2\}} - 2e^{-\tau d}\,, \tag{126}$$

provided that α, τ and C_0 satisfy (125). We have proved (110).

Conclusion of the proof. The final step of the proof consists in applying the kinematic formula conditionally on the event where the statistical dimension of Λ is well-behaved. Let $\mathcal{L}$ be a random subspace drawn uniformly from all subspaces of dimension $d-1$ in $\mathbb{R}^n$ and let E denote the event $\{\Lambda \cap \mathcal{L} \neq \{0\}\}$ (the event where linear separation occurs). By Lemma 22, on the event $A_{\alpha d}$,

$$\mathbb{P}(E|\boldsymbol{V}) \geqslant 1 - 4\exp\left(-\frac{(\alpha d)^2}{8(\min\{\delta(\Lambda), n-\delta(\Lambda)\} + \alpha d)}\right) \geqslant 1 - 4e^{-\alpha^2 d/8}\,. \tag{127}$$

The last inequality stems from the fact that on $A_{\alpha d}$, it also holds that $\min\{F(\boldsymbol{V}), n - F(\boldsymbol{V})\} \leqslant d - 1 - \alpha d$. We thus showed that, given $\alpha \in (0,1)$, for any τ and C_0 satisfying (125),

$$\mathbb{P}(A_{\alpha d}) \geqslant \mathbb{P}\big(F(\boldsymbol{V}) \leqslant \big[7\tau + 14C_0^{-2/3}\big]d\big) \geqslant 1 - e^{-\max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}} - 2e^{-\tau d}.$$

To conclude the proof, we bound from below the probability of $\mathcal{L}$ intersecting Λ in a non trivial way by following the final steps of the proof of Theorem 1 in [CS20]. Using (127), one has

$$\begin{aligned}\mathbf{1}(A_{\alpha d}) \leqslant \mathbf{1}\big(\mathbb{P}(E|\boldsymbol{V}) \geqslant 1 - 4e^{-\alpha^2 d/8}\big) &= \mathbf{1}\big(\mathbb{P}(E|\boldsymbol{V}) + 4e^{-\alpha^2 d/8} \geqslant 1\big) \\ &\leqslant \mathbb{P}(E|\boldsymbol{V}) + 4e^{-\alpha^2 d/8}.\end{aligned}$$

Taking expectation with respect to $\boldsymbol{V}$, this implies that

$$\mathbb{P}(E) \geqslant \mathbb{P}(A_{\alpha d}) - 4e^{-\alpha^2 d/8} \geqslant 1 - e^{-\max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}} - 2e^{-\tau d} - 4e^{-\alpha^2 d/8}.$$

We will thus choose $\tau \geqslant \alpha^2/8$ so that $e^{-\tau d} \leqslant e^{-\alpha^2 d/8}$, under the constraint (125). This constraint rewrites

$$\tau \leqslant \frac{1}{7}\left(1 - \frac{1}{d} - \alpha - \frac{14}{C_0^{2/3}}\right).$$

Since $d \geqslant 50$, if $C_0 \geqslant 22630$, then

$$\frac{1-\alpha}{8} \leqslant \frac{1}{7}\left(1 - \frac{1}{d} - \alpha - \frac{14}{C_0^{2/3}}\right).$$

To optimize the constants, we let α be such that $\alpha^2 = 1 - \alpha$ (denoted by α_* and choose $\tau = \alpha^2/8 = (1-\alpha)/8$. Hence the above rewrites

$$\mathbb{P}(E) \geqslant 1 - e^{-\max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}} - 6e^{-\alpha_*^2 d/8}.$$

We finally bound $\alpha_*^2/8 = (1-\alpha_*)/8 = (1 - (\sqrt{5}-1)/2)/8 \geqslant 1/21$.

7.2 Remaining proofs and additional results

Proof of Lemma 23. By definition, there exists a separating hyperplane if there is some $\theta \in \mathbb{R}^d \setminus \{0\}$ such that for all $i \in \{1, \dots, n\}$,

$$Y_i\langle\theta, X_i\rangle \geqslant 0. \tag{128}$$

From now on, for $1 \leqslant j \leqslant d$, we let $\boldsymbol{X}^j$ denote the n-dimensional vector $(X_1^j, \dots, X_n^j)$ whose entries are all j-th coordinates of $X_1, \dots, X_n$. For every i, the random vectors (Y_i, X_i^1) and $(X_i^2, \dots, X_i^d)$ are independent (and the latter has a symmetric distribution), hence the vectors $(Y_iX_i^1, Y_iX_i^2, \dots, Y_iX_i^d)$ and $(Y_iX_i^1, X_i^2, \dots, X_i^d)$ have the same distribution. Therefore,

$$\mathbb{P}\big(\exists\theta \in \mathbb{R}^d \setminus \{0\}, \forall i, Y_i\langle\theta, X_i\rangle \geqslant 0\big) = \mathbb{P}\left(\exists\theta \in \mathbb{R}^d \setminus \{0\}, \theta^1\boldsymbol{V} + \sum_{j=2}^d \theta^j\boldsymbol{X}^j \in \mathbb{R}_+^n\right). \tag{129}$$

Now, let $\mathcal{L} = \mathrm{span}\{\boldsymbol{X}^2, \dots, \boldsymbol{X}^d\}$. Since $\boldsymbol{X}^{(2)}, \dots, \boldsymbol{X}^{(d)}$ are i.i.d. random vectors with distribution $\mathsf{N}(0, I_n)$, the distribution of $\mathcal{L}$ is rotation-invariant and thus uniform over $(d-1)$-dimensional subspaces of $\mathbb{R}^n$. Also, $\mathcal{L}$ is independent from $\Lambda = \mathbb{R}\boldsymbol{V} + \mathbb{R}_+^n$, and if $\Lambda \cap \mathcal{L} \neq \{0\}$, then there exists $\theta^1 \in \mathbb{R}$ and $(\theta^2, \dots, \theta^d) \in \mathbb{R}^{d-1}$, as well as $w \in \mathbb{R}_+^n$ such that $-\theta^1\boldsymbol{V} + w = \sum_{j=2}^d \theta^j\boldsymbol{X}^j$, thus $\theta^1\boldsymbol{V} + \sum_{j=2}^d \theta^j\boldsymbol{X}^j \in \mathbb{R}_+^n$. Combining this fact with (129) concludes the proof. □

Fact 3. *Let $p \in (0,1/2)$ and $u^* \in S^{d-1}$ be such that $\mathbb{P}(Y\langle u^*, X\rangle < 0) \leqslant p$. For any $t > 0$, if $n \leqslant t/(2p)$, then with probability at least e^{-t} the dataset $(X_1, Y_1), \dots, (X_n, Y_n)$ of i.i.d. copies of (X, Y) is linearly separated.*

Proof. We have

$$\mathbb{P}\big(\forall i \leqslant n,\ Y_i\langle u^*, X_i\rangle \geqslant 0\big) = \big(1 - \mathbb{P}(Y\langle u^*, X\rangle < 0)\big)^n \geqslant (1-p)^n = \exp\big(n\log(1-p)\big)\,.$$

By concavity, for all $x \in [0, 1/2]$, $\log(1-x) \geqslant -2\log(2)x$. Thus, since $n \leqslant t/(2p) \leqslant t/(2\log(2)p)$, one has

$$\mathbb{P}\big(\forall i \leqslant n,\ Y_i\langle \theta^*, X_i\rangle > 0\big) \geqslant \exp\big(-2np\big) \geqslant \exp(-t)\,.$$

□

Fact 4. *Let $\psi(s) = \mathbb{E}\big[(s - Z)_+^2\big]$ for every $s \in \mathbb{R}$, with $Z \sim \mathsf{N}(0,1)$. Then*

$$\psi(s) \leqslant \frac{e^{-s^2/2}}{2}\mathbf{1}(s < 0) + (s^2+1)\mathbf{1}(s \geqslant 0)\,.$$

Proof. Using the symmetry of Z and the fact that $(-x)_+ = x_-$, we have, for every real s,

$$\psi(-s) = \mathbb{E}(-s - Z)_+^2 = \mathbb{E}(-(s+Z))_+^2 = \mathbb{E}(s+Z)_-^2 = \mathbb{E}(s-Z)_-^2\,.$$

Also, observing that for all $x \in \mathbb{R}$, $x^2 = x_+^2 + x_-^2$, we have

$$\psi(-s) + \psi(s) = \mathbb{E}(s-Z)_-^2 + \mathbb{E}(s-Z)_+^2 = \mathbb{E}(s-Z)^2 = s^2 + 1\,. \tag{130}$$

We start with the case where $s < 0$. In this case, denoting by g the density of $\mathsf{N}(0,1)$, one has

$$\begin{aligned}\psi(s) &= \mathbb{E}(s-Z)_+^2 = \mathbb{E}\left[(s-Z)^2\mathbf{1}\{s - Z > 0\}\right] = \int_{-\infty}^{s}(s-z)^2 g(z)\mathrm{d}z = \int_0^{+\infty} z^2 g(s-z)\mathrm{d}z \\ &= \int_0^{+\infty} z^2 \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{s^2 - 2sz + z^2}{2}\Big)\mathrm{d}z = e^{-s^2/2}\int_0^{+\infty} z^2 e^{sz} g(z)\mathrm{d}z \\ &= e^{-s^2/2}\,\mathbb{E}\left[Z^2 e^{sZ}\mathbf{1}\{Z > 0\}\right]\,.\end{aligned}$$

Note that, since $s < 0$, $e^{sZ}\mathbf{1}\{Z > 0\} \leqslant \mathbf{1}\{Z > 0\}$, which implies that

$$\mathbb{E}\left[Z^2 e^{-sZ}\mathbf{1}\{Z < 0\}\right] \leqslant \mathbb{E}\left[Z^2\mathbf{1}\{Z < 0\}\right] = \frac{1}{2}\,,$$

which proves the first part of the result. Regarding the case where $s \geqslant 0$, we deduce from (130) that $\psi(s) = s^2 + 1 - \psi(-s)$, and, from the previous point, that $0 \leqslant \psi(-s) \leqslant 1/2$. □

8 Proofs of the main results

This section gathers the results of Sections 5 and 6 to establish the upper bounds on the excess risk of the MLE thanks to Lemma 3.

8.1 Preliminaries: convex localization and Hessian

We start with the proof of the localization lemma (Lemma 3).

Proof of Lemma 3. Let r be arbitrary such that $2\nu/c_0 < r < r_0$, which exists since $r_0 > 2\nu/c_0$ by assumption. For any $\theta \in \mathbb{R}^d$ such that $\|\theta - \theta^*\|_H = r$, a Taylor expansion of order 2 shows that

$$\begin{aligned}
\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) &= \langle \nabla \widehat{L}_n(\theta^*), \theta - \theta^* \rangle + \int_0^1 (1-t) \big\langle \nabla^2 \widehat{L}_n\big((1-t)\theta^* + t\theta\big)(\theta - \theta^*), \theta - \theta^* \big\rangle \mathrm{d}t \\
&\geqslant - \|\nabla \widehat{L}_n(\theta^*)\|_{H^{-1}} \|\theta - \theta^*\|_H + \frac{c_0}{2} \|\theta - \theta^*\|_H^2 \qquad (131) \\
&\geqslant -\nu r + c_0 r^2/2 > 0 \,, \qquad (132)
\end{aligned}$$

where inequality (131) comes from the fact that $\nabla^2 \widehat{L}_n((1-t)\theta^* + t\theta) \succcurlyeq c_0 H$ by assumption, and (132) from the condition $r > 2\nu/c_0$. Now, for any $\theta' \in \mathbb{R}^d$ such that $r' = \|\theta' - \theta^*\|_H \geqslant r$, the parameter $\theta = (1-t)\theta^* + t\theta'$ with $t = r/r' \in (0,1]$ satisfies $\|\theta - \theta^*\|_H = r$, hence by the preceding and by convexity of $\widehat{L}_n$ one has

$$(1-t)\widehat{L}_n(\theta^*) + t\widehat{L}_n(\theta') \geqslant \widehat{L}_n\big((1-t)\theta^* + t\theta'\big) = \widehat{L}_n(\theta) > \widehat{L}_n(\theta^*) \,,$$

which simplifies to $\widehat{L}_n(\theta') > \widehat{L}_n(\theta^*)$. Hence $\inf_{\mathbb{R}^d} \widehat{L}_n = \inf_{\theta : \|\theta - \theta^*\|_H \leqslant r} \widehat{L}_n(\theta)$, and the latter infimum is attained by compactness and continuity of $\widehat{L}_n$. Since in addition $\widehat{L}_n$ is strictly convex on the set $\{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant r_0\}$ due to the second assumption, the function $\widehat{L}_n$ admits a unique global minimizer $\widehat{\theta}_n \in \mathbb{R}^d$, such that $\|\widehat{\theta}_n - \theta^*\|_H \leqslant r$. Since this holds for every $r \in (2\nu/c_0, r_0)$, we deduce that $\|\widehat{\theta}_n - \theta^*\|_H \leqslant 2\nu/c_0$.

The excess risk bound (33) then follows from the fact that $L(\theta) - L(\theta^*) \leqslant \frac{c_1}{2}\|\theta - \theta^*\|_H^2$ for any θ with $\|\theta - \theta^*\|_H \leqslant r_0$, since $\nabla L(\theta^*) = 0$ and $\nabla^2 L \preccurlyeq c_1 H$ over this domain.

To prove the second point, let $\varepsilon = \widehat{L}_n(\widetilde{\theta}_n) - \widehat{L}_n(\widehat{\theta}_n)$ and r be such that $\max\{4\nu/c_0, 2\sqrt{\varepsilon/c_0}\} < r < r_0$. For any θ such that $\|\theta - \theta^*\|_H = r$, proceeding as before (and using that $\widehat{L}_n(\widehat{\theta}_n) \leqslant \widehat{L}_n(\theta^*)$) we get

$$\widehat{L}_n(\theta) - \widehat{L}_n(\widehat{\theta}_n) \geqslant \widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) \geqslant -\nu r + c_0 r^2/2 \geqslant c_0 r^2/4 > \varepsilon \,,$$

where the last two inequalities follow from the conditions on r. By the same convexity argument as before, this implies that $\widehat{L}_n(\theta) - \widehat{L}_n(\widehat{\theta}_n) > \varepsilon$ for any θ such that $\|\theta - \theta^*\|_H \geqslant r$, hence $\|\widetilde{\theta}_n - \theta^*\|_H < r$. Letting $r \to \max\{4\nu/c_0, 2\sqrt{\varepsilon/c_0}\}$ and using that $L(\widetilde{\theta}_n) - L(\theta^*) \leqslant \frac{c_1}{2}\|\widetilde{\theta}_n - \theta^*\|_H^2$ concludes the proof. □

We now turn to the structure of the Hessian $\nabla^2 L(\theta^*)$ in the case of a Gaussian design, which is given by (36). It then remains to justify the estimates (37) on the components $c_0(\cdot), c_1(\cdot)$ of the Hessian. Lemma 25 below shows that

$$\frac{2\sqrt{2}}{3e^4\sqrt{\pi}} \min\Big(1, \frac{1}{\beta^3}\Big) \leqslant c_0(\beta) \leqslant 2\sqrt{\frac{2}{\pi}} \min\Big(1, \frac{1}{\beta^3}\Big) \,; \qquad (133)$$

$$\frac{1}{2e^4}\sqrt{\frac{2}{\pi}} \min\Big(1, \frac{1}{\beta}\Big) \leqslant c_1(\beta) \leqslant \sqrt{2} \min\Big(1, \frac{1}{\beta}\Big) \,. \qquad (134)$$

Lemma 25. *Let $G \sim \mathsf{N}(0,1)$. For any $\beta > 0$ and integer $k \geqslant 0$,*

$$\sqrt{\frac{2}{\pi}} \frac{2^{k+1}}{k+1} \min\Big(\frac{1}{4e^4\beta^{k+1}}, \frac{\sigma'(2)}{e^2}\Big) \leqslant \mathbb{E}[\sigma'(\beta G)|G|^k] \leqslant \sqrt{\frac{2}{\pi}} \min\Big(\Gamma\Big(\frac{k+1}{2}\Big), \frac{k!}{\beta^{k+1}}\Big) \,.$$

Proof. We have

$$\mathbb{E}[\sigma'(\beta G)|G|^k] = \sqrt{\frac{2}{\pi}} \int_0^{+\infty} x^k \sigma'(\beta x) \exp\Big(-\frac{x^2}{2}\Big) \mathrm{d}x \,.$$

For the upper bound, we use that, for any x, $\sigma'(x) \leqslant \exp(-|x|) \leqslant 1$ and $\exp(-x^2/2) \leqslant 1$ to get

$$\mathbb{E}[\sigma'(\beta G)|G|^k] \leqslant \sqrt{\frac{2}{\pi}} \min\left(\int_0^{+\infty} x^k \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x, \int_0^{+\infty} x^k \exp\left(-\beta x\right)\mathrm{d}x\right).$$

Computing the integrals yields the upper bound.

For the lower bound, as the function we integrate is nonnegative and $\sigma'(x) \geqslant \exp(-x)/4$, we have

$$\begin{aligned}
\mathbb{E}[\sigma'(\beta G)|G|^k] &\geqslant \sqrt{\frac{2}{\pi}} \int_0^2 x^k \sigma'(\beta x) \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x \\
&\geqslant \sqrt{\frac{2}{\pi}} \max\left(\frac{1}{4e^2}\int_0^2 x^k \exp(-\beta x)\mathrm{d}x, \sigma'(2\beta)\int_0^2 x^k \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x\right) \\
&= \sqrt{\frac{2}{\pi}} \max\left(\frac{1}{4e^2\beta^{k+1}}\int_0^2 x^k \exp(-x)\mathrm{d}x, \sigma'(2\beta)\int_0^2 x^k \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x\right) \\
&\geqslant \sqrt{\frac{2}{\pi}} \frac{2^{k+1}}{k+1} \max\left(\frac{1}{4e^4\beta^{k+1}}, \frac{\sigma'(2\beta)}{e^2}\right).
\end{aligned}$$

To get the lower bound, we use the first bound when $\beta > 1$ and the second one when $\beta \leqslant 1$. □

8.2 Proof of Theorem 1

By Proposition 5, since $n \geqslant 4B(d\log 5 + t)$, with probability larger than $1 - 2e^{-t}$,

$$\left\|\nabla \widehat{L}_n(\theta^*)\right\|_{H^{-1}} \leqslant 27\sqrt{\frac{d+t}{n}},$$

Moreover, let

$$\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant \frac{1}{100\sqrt{B}}\right\},$$

Theorem 6 ensures that, as $n \geqslant 320000B(d+t)$, then with probability at least $1-2e^{-t}$, simultaneously for all $\theta \in \Theta$, $\widehat{H}_n(\theta) \succcurlyeq \frac{1}{1000}H$. Therefore, by Lemma 3, as soon as $n \geqslant (5400000)^2 B(d+t)$,

$$27\sqrt{\frac{d+t}{n}} < \frac{1}{2000} \cdot \frac{1}{100\sqrt{B}},$$

so, with probability at least $1 - 5e^{-t}$,

$$\|\widehat{\theta}_n - \theta^*\|_H \leqslant 54000\sqrt{\frac{d+t}{n}}.$$

By Lemmas 27 and 3, we also have on the same event

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant 420 \cdot (54)^2 \cdot 10^6 \frac{d+t}{n}.$$

Regarding the necessity of the sample size condition, we combine Theorem 2 with Fact 3 which in the case of a well-specified model shows that the condition $n \gtrsim Bt$ is also necessary. Indeed, as the model is well-specified, $\mathbb{P}(Y\langle\theta^*, X\rangle < 0) = \mathbb{E}\sigma(-|\langle\theta^*, X\rangle|)$. In addition, for all real t, $\sigma(-|t|) \leqslant \min\{1/2, e^{-|t|}\}$, hence

$$\mathbb{P}(Y\langle\theta^*, X\rangle < 0) \leqslant \frac{1}{\max\{2, \|\theta^*\|\}} \leqslant \frac{e}{2B}.$$

We use Fact 3 with $p = e/(2B)$ to conclude that if $n \leqslant Bt/e$, then

$$\mathbb{P}(\{\text{MLE does not exist}\}) \geqslant \exp(-t)\,. \tag{135}$$

With these results at hand, one has that whenever

$$n \leqslant \frac{B}{2}\Big(\frac{d}{C_0} + \frac{t}{e}\Big) \leqslant \max\Big\{\frac{Bd}{C_0}, \frac{Bt}{e}\Big\}\,, \tag{136}$$

either $n \leqslant Bt/e$ or $n \leqslant Bd/C_0$. The former is already dealt with by (135). The latter, by Theorem 2 (with parameter $\kappa = 1$), implies that

$$\mathbb{P}(\{\text{MLE does not exist}\}) \geqslant 1 - \exp\Big(-\max\Big\{\sqrt{d}, \frac{d}{B^2}\Big\}\Big) - 6e^{-d/21}\,. \tag{137}$$

As $d \geqslant 53$ and $t \geqslant 1$, one has $6e^{-d/21} \leqslant 1/2$ and $1/2 - e^{-\sqrt{d}} \geqslant e^{-t}$. Hence, taking the minimum of the two lower bounds (135) and (137) shows that $\mathbb{P}(\{\text{MLE does not exist}\}) \geqslant \exp(-t)$ and concludes the proof of Theorem 1.

8.3 Proof of Theorem 3

By Proposition 6, we have, for any $n \geqslant B(d+t)$, for any $t > 0$, with probability at least $1-3e^{-t}$,

$$\big\|\nabla \widehat{L}_n(\theta^*)\big\|_{H^{-1}} \leqslant c' \log B \sqrt{\frac{d+t}{n}}\,.$$

Moreover, by Theorem 5, if

$$n \geqslant c_1 B(\log(B)d + t)\ ,$$

then, with probability $1 - \exp(-t)$,

$$\widehat{H}_n(\theta) \succcurlyeq c_2 H, , \qquad \forall \|\theta - \theta^*\|_H \leqslant \frac{c_3}{\log(B)\sqrt{B}}\ .$$

By Lemma 3, it follows that, if

$$n \geqslant 4\Big(\frac{c'}{c_2 c_3}\Big)^2 (\log B)^2 B(d+t)\,,$$

with probability at least $1 - 4e^{-t}$,

$$\|\widehat{\theta}_n - \theta^*\|_H \leqslant \frac{2c'(\log B)^2}{c_0}\sqrt{\frac{d+t}{n}}\,. \tag{138}$$

By Lemmas 28 and 3, on the same event, there exists a function c' of c and K only such that

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c'(\log B)^4 \frac{d+t}{n}\,.$$

8.4 Proof of Theorem 4

By Proposition 7, if $n \geqslant B(d + Bt)$, then with probability larger than $1 - 3e^{-t}$,

$$\big\|\nabla \widehat{L}_n(\theta^*)\big\|_{H^{-1}} \leqslant c' \log(B)\sqrt{\frac{d+Bt}{n}}\,.$$

Moreover, by Theorem 5, if

$$n \geqslant c_1 B(\log(B)d + t) \ ,$$

then, with probability $1 - \exp(-t)$,

$$\widehat{H}_n(\theta) \succcurlyeq c_2 H \qquad \text{for every } \theta \text{ such that } \|\theta - \theta^*\|_H \leqslant \frac{c_3}{\log(B)\sqrt{B}} \, .$$

By Lemma 3, it follows that, if

$$n \geqslant 4\bigg(\frac{c'}{c_2 c_3}\bigg)^2 (\log B)^2 B(d + Bt)\,, \tag{139}$$

with probability at least $1 - 4e^{-t}$,

$$\|\widehat{\theta}_n - \theta^*\|_H \leqslant \frac{2c'(\log B)^2}{c_0}\sqrt{\frac{d + Bt}{n}}\,. \tag{140}$$

By Lemmas 28 and 3, on the same event, there exists a function c' of c and K only such that

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c'(\log B)^4 \frac{d + Bt}{n}\,.$$

Regarding the necessity of the sample size condition, the fact that the condition $n \gtrsim Bd$ is necessary comes from the well-specified case, which is a particular case of the current setting. Regarding the necessity of the extra B factor in the sample size condition, consider the following distribution of (X, Y): X is a standard Gaussian vector and the conditional distribution of Y given X is such that $\mathbb{P}(Y\langle u^*, X\rangle < 0|X)$ is constant (see (141)). The first point of Lemma 26 shows that for this distribution, $\mathbb{P}(Y\langle \theta^*, X\rangle < 0) \leqslant 1/B^2$. It then follows from Fact 3 that if $n \leqslant B^2 t/2$,

$$\mathbb{P}(\text{MLE exists}) \leqslant e^{-t}\,.$$

The conclusion follows from the same argument as in the proof of Theorem 3 in the previous section, see (136) and after.

We now turn to the optimality of our bound on the excess risk (29). It is known from asymptotic theory (see e.g. [vdV98, Example 5.25 p. 55]) that in the misspecified case,

$$\sqrt{n} H(\theta^*)^{1/2}(\widehat{\theta}_n - \theta^*) \xrightarrow[n\to\infty]{(\mathrm{d})} \mathsf{N}(0, \Gamma)\,, \qquad \Gamma = H(\theta^*)^{-1/2} G H(\theta^*)^{-1/2}\,,$$

where $H(\theta^*) = \nabla^2 L(\theta^*)$ is the population Hessian and $G = \mathbb{E}[\nabla\ell(\theta^*)\nabla\ell(\theta^*)^\top]$ is the the covariance of the gradient at θ^*. Hence, the rescaled excess risk $2n(L(\widehat{\theta}_n) - L(\theta^*))$ converges in distribution to $\|\xi\|^2$ where $\xi \sim \mathsf{N}(0, \Gamma)$. The argument showing the optimality of our result is twofold. First, in the case where the model is well-specified, $\Gamma = I_d$, so $\mathrm{Tr}(\Gamma) = d$. Second, the argument regarding the necessity of the deviation term builds upon the same conditional distribution that explains the necessity of $B^2 t$ in the sample size condition that we described above (*i.e.* $X \sim \mathsf{N}(0, I_d)$ and $Y|X$ is given by (141)). Indeed, by Lemma 27, $\Gamma \succcurlyeq C_1^{-1} H^{-1/2} G H^{-1/2}$, with $C_1 = 2\sqrt{2/\pi}$, so by the second point of Lemma 26, for this particular distribution, it holds that

$$\|\Gamma\|_{\mathrm{op}} \geqslant \frac{B}{8C_1} \geqslant \frac{B}{13}\,.$$

In addition, by standard concentration arguments, one can find an absolute constant c_1 such that on one hand, the median of the distribution $\chi^2(d)$ is at least $c_1 d$; and on the other hand, if $v \in S^{d-1}$ denotes an eigenvector of Γ associated to its largest eigenvalue,

$$\mathbb{P}(\|\xi\|^2 \geqslant c_1\|\Gamma\|_{\mathrm{op}} t) \geqslant \mathbb{P}(\langle v, \xi\rangle^2 \geqslant c_1 \|\Gamma\|_{\mathrm{op}} t) \geqslant e^{-t}\,.$$

This concludes the proof of Theorem 4.

Worst misspecified case. In this paragraph we provide an example of conditional distribution of Y given X which accounts for the extra factors in the sample size and the risk bound on the MLE. Such a distribution is obtained by taking X to be a standard Gaussian and $Y|X$ such that the event where Y differs from the sign of $\langle \theta^*, X\rangle$ has a constant probability and is independent of X.

Lemma 26. *Let $X \sim \mathsf{N}(0, I_d)$, $u^* \in S^{d-1}$ and $p \in (0, e^{-2}/2)$. Let Y be such that*

$$\mathbb{P}(Y\langle u^*, X\rangle < 0|X) = p\,. \tag{141}$$

Then

1. *the signal strength $B = \max\{e, \|\theta^*\|\}$ is related to the probability of misclassification by*
$$\frac{1}{2B^2} \leqslant \mathbb{P}(Y\langle u^*, X\rangle < 0) \leqslant \frac{1}{B^2}\,.$$
2. *The covariance of the gradient $G = \mathbb{E}[\nabla\ell(\theta^*, Z)\nabla\ell(\theta^*, Z)^\top]$ satisfies*
$$\|H^{-1/2}GH^{-1/2}\|_{\mathrm{op}} \geqslant \frac{B}{8}\,.$$

Proof. Recall that since the model is misspecified, θ^* is defined as the unique minimizer of $L(\theta)$ (uniqueness follows from the strict convexity of L). We first note that for any isometry Q such that $Qu^* = u^*$, it holds for all $\theta \in \mathbb{R}^d$ that

$$L(Q\theta) = L(\theta)\,. \tag{142}$$

This stems from the fact that the distribution of X is invariant under any isometry and the distribution of Y given X is invariant under any isometry that preserves u^*. This holds in particular at the point θ^*. Hence $Q\theta^* = \theta^*$ and, letting $Q = 2u^*u^{*\top} - I_d$, this shows that $\theta^* \in \mathbb{R}u^*$. We show in addition that $\theta^* \in \mathbb{R}_+u^*$, namely

$$\theta^* = \|\theta^*\|u^*\,. \tag{143}$$

This amounts to showing that $L(-\|\theta^*\|u^*) > L(\|\theta^*\|u^*)$ which we do next. Let $\phi(t) = \log(1+e^t)$ denote the logistic loss and write

$$\begin{aligned}
L(-\|\theta^*\|u^*) &= \mathbb{E}[\phi(Y\|\theta^*\|\langle u^*, X\rangle)] \\
&= (1-p)\mathbb{E}[\phi(\|\theta^*\||\langle u^*, X\rangle|)] + p\mathbb{E}[\phi(-\|\theta^*\||\langle u^*, X\rangle|)] \\
&> p\mathbb{E}[\phi(\|\theta^*\||\langle u^*, X\rangle|)] + (1-p)\mathbb{E}[\phi(-\|\theta^*\||\langle u^*, X\rangle|)] \\
&= \mathbb{E}[\phi(-Y\|\theta^*\|\langle u^*, X\rangle)] = L(\|\theta^*\|u^*)\,.
\end{aligned}$$

This proves (143).

It remains to show that $B = \|\theta^*\| \geqslant e$ and that $B \asymp p^{-1/2}$. In view of (141), $\mathbf{1}(Y\langle\theta^*, X\rangle < 0)$ does not depend on X. Hence (59) rewrites

$$p\mathbb{E}\big[|\langle u^*, X\rangle|\big] = \mathbb{E}\big[|\langle u^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big]\,.$$

In addition, since $\mathbb{E}|\langle u^*, X\rangle| = \sqrt{2/\pi}$, one has

$$p = \sqrt{\frac{\pi}{2}}\,\mathbb{E}\big[|\langle u^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big]\,. \tag{144}$$

On one hand, $\sigma(-t) \geqslant e^{-t}/2$ for all $t \geqslant 0$. Using that $\|\theta^*\| \leqslant B$, it follows from Lemma 29 that

$$\mathbb{E}\big[|\langle u^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big] \geqslant \frac{1}{2}\mathbb{E}\big[|\langle u^*, X\rangle|\exp(-B|\langle u^*, X\rangle|)\big] \geqslant \frac{1}{\sqrt{2\pi}B^2}\,. \tag{145}$$

Hence, using (144) and since $p \leqslant e^{-2}/2$ we deduce that

$$B \geqslant \frac{1}{\sqrt{2p}} \geqslant e\,.$$

The lower bound of the first point of the lemma is therefore a straightforward consequence of (145) and (144) and we have

$$B = \|\theta^*\| \geqslant e \quad \text{and} \quad p \geqslant \frac{1}{2B^2}. \tag{146}$$

We now prove the upper bound of the first point, which is a consequence of the exponential moment bound (51), since $B = \|\theta^*\| \geqslant e$. Using that $\sigma(-t) \leqslant e^{-t}$ for all $t \geqslant 0$, we deduce

$$\mathbb{E}\big[|\langle u^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big] \leqslant \sqrt{\frac{2}{\pi}}\cdot\frac{1}{B^2}\,.$$

We plug this in (144) to get that $p \leqslant 1/B^2$, which is the desired upper bound.

We now prove the second point. As $\sigma(t) \geqslant 1/2$ for every $t \geqslant 0$,

$$\langle Gu^*, u^*\rangle = \mathbb{E}\langle u^*, \nabla\ell(\theta^*, Z)\rangle^2 = \mathbb{E}\big[\sigma(-Y\langle\theta^*, X\rangle)^2\langle u^*, X\rangle^2\big] \geqslant \frac{1}{4}\mathbb{E}\big[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\langle u^*, X\rangle^2\big]\,.$$

The distribution of $Y|X$ is designed so that $\mathbb{E}[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)|X]$ is actually not a function of X, but constant and equal to p. More precisely,

$$\begin{aligned}\mathbb{E}\big[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\langle u^*, X\rangle^2\big] &= \mathbb{E}\big[\mathbb{E}\big[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\langle u^*, X\rangle^2|X\big]\big] \\ &= \mathbb{E}\big[\langle u^*, X\rangle^2\mathbb{E}\big[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)|X\big]\big] \\ &= p\,\mathbb{E}\langle u^*, X\rangle^2 = p\,.\end{aligned}$$

Therefore

$$\langle Gu^*, u^*\rangle \geqslant \frac{p}{4} \geqslant \frac{1}{8B^2}\,.$$

Finally, since $H^{-1/2}u^* = B^{3/2}u^*$, it follows that $\langle H^{-1/2}GH^{-1/2}u^*, u^*\rangle \geqslant B/8$. □

8.5 Technical tools

In the previous proofs, we used the following lemmas linking the Hessians $\nabla^2 L(\theta) = H(\theta) = \mathbb{E}[\sigma'(\langle\theta, X\rangle)XX^\top]$ to H to conclude the proof.

Lemma 27. *Let $\theta \in \mathbb{R}^d \setminus \{0\}$ denote a vector such that $\|\theta - \theta^*\|_H \leqslant 1/10\sqrt{B}$, let $u = \theta/\|\theta\|$ and let X denote a standard Gaussian vector. Then,*

$$\frac{1}{500}H \preccurlyeq H(\theta) \preccurlyeq 420H\,.$$

Proof. We write that, for any $v \in S^{d-1}$

$$\langle H_\theta^{-1/2}H(\theta)H_\theta^{-1/2}v, v\rangle = B^3\langle u, v\rangle^2\mathbb{E}[\sigma'(\langle\theta, X\rangle)\langle u, X\rangle^2] + B(1 - \langle u, v\rangle^2)\mathbb{E}[\sigma'(\langle\theta, X\rangle)]\,.$$

We start with the upper bound. By Lemma 25,

$$\mathbb{E}[\sigma'(\langle\theta, X\rangle)|\langle u, X\rangle|^k] \leqslant \sqrt{\frac{2}{\pi}} \min\left(\Gamma\left(\frac{k+1}{2}\right), \frac{\Gamma(k+1)}{\|\theta\|^{k+1}}\right).$$

Then, if $B = e$, we use the first bound to get

$$\langle H_\theta^{-1/2} H(\theta) H_\theta^{-1/2} v, v\rangle \leqslant \sqrt{2}e^3 .$$

This proves that $H(\theta) \preccurlyeq \sqrt{2}e^3 H_\theta$. As $H_\theta \preccurlyeq e^{-1} I_d \preccurlyeq e^2 H$, this proves the result when $B = e$.

If $B > e$, we have by Lemma 17, $\|\theta\| \geqslant 0.9 \cdot B$, so the second bound on the moments gives

$$\langle H_\theta^{-1/2} H(\theta) H_\theta^{-1/2} v, v\rangle \leqslant \sqrt{\frac{2}{\pi}} \frac{2}{0.9^3} .$$

This proves that $H(\theta) \preccurlyeq 2.2 \cdot H_\theta$ and this proves the result in the case $B > e$ since by Lemma 19 we also have $H_\theta \preccurlyeq 1.3 \cdot H$.

We now turn to the lower bound. By Lemma 25,

$$\mathbb{E}[\sigma'(\langle\theta, X\rangle)|\langle u, X\rangle|^k] \geqslant \sqrt{\frac{2}{\pi}} \frac{2^{k+1}}{k+1} \min\left(\frac{1}{4e^4\|\theta\|^{k+1}}, \frac{\sigma'(2)}{e^2}\right).$$

Then, if $B = e$, we use the second bound to get

$$\langle H_\theta^{-1/2} H(\theta) H_\theta^{-1/2} v, v\rangle \geqslant 0.02 .$$

This proves that $H(\theta) \succcurlyeq c_2 H_\theta$ and as $H_\theta \succcurlyeq e^{-3} I_d \succcurlyeq e^{-2} H$, this proves the result when $B = e$.

If $B > e$, we have by Lemma 17, $\|\theta\| \geqslant 0.9 \cdot B$, so the first bound on the moments gives

$$\langle H_\theta^{-1/2} H(\theta) H_\theta^{-1/2} v, v\rangle \geqslant 0.0027 .$$

This proves the result in the case $B > e$ since by Lemma 19 we also have $H_\theta \succcurlyeq 0.76 \cdot H$. $\square$

Lemma 28. *Let $\theta \in \mathbb{R}^d \setminus \{0\}$ be such that $\|\theta - \theta^*\|_H \leqslant 1/10\sqrt{B}$ and let $u = \theta/\|\theta\|$. Suppose that X satisfies Assumptions 1 with parameter $K > 0$ and 2 with parameters $\eta = 1/B$ and $c \geqslant 1$. Then, there exists c' depending on c and K such that*

$$\frac{1}{c'} H \preccurlyeq H(\theta) \preccurlyeq c' H .$$

Proof. We start with the proof of the upper bound. Let $v \in S^{d-1}$ and let $w \in S^{d-1}$ denote a vector such that $\langle u, w\rangle = 0$ and $v - \langle u, v\rangle u = \sqrt{1 - \langle u, v\rangle^2} w$. As $\sigma'(x) \leqslant \exp(-|x|)$, we have

$$\begin{aligned}\langle H_\theta^{-1/2} H(\theta) H_\theta^{-1/2} v, v\rangle &= B^3 \langle u, v\rangle^2 \mathbb{E}[\exp(-|\langle\theta, X\rangle|)\langle u, X\rangle^2] \\ &\quad + B(1 - \langle u, v\rangle^2)\mathbb{E}[\exp(-|\langle\theta, X\rangle|)\langle w, X\rangle^2] .\end{aligned}$$

If $B = e$, it follows from $\sigma'(x) \leqslant 1$ that $H(\theta) \preccurlyeq e^3 H_\theta$ and thus $H(\theta) \preccurlyeq e^5 H$ since $H_\theta \preccurlyeq e^2 H$ in this case.

If $B > e$, we have by Lemma 17, $\|\theta\| \geqslant (1 - r)B$. Thus, by Lemma 6, it follows that

$$\langle H_\theta^{-1/2} H(\theta) H_\theta^{-1/2} v, v\rangle \leqslant \frac{3c}{1 - r} (K \log((1 + r)B))^2 .$$

This proves that $H(\theta) \preccurlyeq \frac{3c}{1-r}(K\log((1+r)B))^2 H_\theta$ and this proves the result in the case $B > e$ since by Lemma 19 we also have $H_\theta \preccurlyeq (1+2.35r)H$.

We now turn to the lower bound. Let $v \in S^{d-1}$, we have

$$\langle H(\theta)v, v\rangle = \mathbb{E}[\sigma'(\langle \theta, X\rangle)\langle v, X\rangle^2] .$$

The function $\sigma'(x) = \exp(x)/(1+\exp(x))^2$ is even, non negative, non increasing on $[0, +\infty)$. Therefore, for any $m, M > 0$,

$$\langle H(\theta)v, v\rangle \geqslant \sigma'(m(1+r)B)M^2\mathbb{P}\big(|\langle u, X\rangle| \leqslant m,\ |\langle v, X\rangle| \geqslant M\big) , \tag{147}$$

where we also used that, as $\|\theta - \theta^*\| \leqslant r/\sqrt{B}$, $\|\theta\| \leqslant (1+r)B$ by Lemma 17.

If $\|\theta^*\| \leqslant e$, $B = e$, so Proposition 2 shows that Assumptions 2 holds with $c = e$ and Assumption 3 is satisfied with constant $\max\{2eK\log(2K), 2K^4\} = 2K^4$. Therefore,

$$\mathbb{P}\left(\big|\langle u, X\rangle\big| \leqslant \frac{2K^4}{B}\,;\, \big|\langle v, X\rangle\big| \geqslant \frac{\max\{1/B, \|u^* - v\|\}}{2K^4}\right) \geqslant \frac{1}{2K^4 B} .$$

Hence, choosing $m = 2K^4/B$ and $M = \max\{1/B, \|u^* - v\|\}/2K^4$ in (147), we get that

$$\langle H(\theta)v, v\rangle \geqslant \frac{\sigma'((1+r)2K^4)}{8K^{12}}\frac{1}{B}\max\left\{\frac{1}{B^2}, \|u^* - v\|^2\right\} \geqslant \frac{\sigma'((1+r)2K^4)}{16K^{12}}\langle Hv, v\rangle .$$

When $B > e$, the third point of Lemma 17 implies that for every $\theta \in \Theta$,

$$\|u - u^*\| \leqslant \frac{\sqrt{2}}{[K\log(c(c+1)B) - 1]}\frac{r}{B} \leqslant \frac{2r}{KB\log(c(c+1)B)} .$$

By Lemma 11, this implies that for all $\theta \in \Theta$ and $v \in S^{d-1}$, one has for all $t \geqslant 1/B$

$$\mathbb{P}\left(\big|\langle u, X\rangle\big| \leqslant \frac{c+1}{B}\,;\, \big|\langle v, X\rangle\big| \geqslant \frac{\max\{1/B, \|u^* - v\|\}}{c+1}\right) \geqslant \frac{1}{(c+1)B} .$$

Hence, choosing $m = (c+1)/B$, $M = \max(1/B, \|u^* - v\|)/(c+1)$ in (65), we get that

$$\langle H(\theta)v, v\rangle \geqslant \frac{\sigma'((1+r)(1+c))}{(1+c)^3}\frac{1}{B}\max\left\{\frac{1}{B^2}, \|u^* - v\|^2\right\} \geqslant \frac{\sigma'((1+r)(1+c))}{2(1+c)^3}\langle Hv, v\rangle .$$

$\square$

Lemma 29. *Let $N \sim \mathsf{N}(0,1)$ and $B \geqslant e$. Then*

$$\mathbb{E}\big[|N|\exp(-B|N|)\big] \geqslant \frac{1}{\sqrt{2\pi}B^2} .$$

Proof. Let $g(t) = e^{-t^2/2}/\sqrt{2\pi}$ denote the standard real Gaussian density. First, by symmetry,

$$\mathbb{E}\big[|N|\exp(-B|N|)\big] = \sqrt{\frac{2}{\pi}}\int_0^{+\infty} te^{-Bt}e^{-t^2/2}\mathrm{d}t . \tag{148}$$

Then we proceed with an expansion of the integral.

$$\begin{aligned}\int_0^{+\infty} te^{-Bt}e^{-t^2/2}\mathrm{d}t &= e^{B^2/2}\int_0^{+\infty} te^{-\frac{(t+B)^2}{2}}\mathrm{d}t = e^{B^2/2}\int_B^{+\infty}(x-B)e^{-x^2/2}\mathrm{d}x \\ &= e^{B^2/2}\Big(\int_B^{+\infty} xe^{-x^2/2}\mathrm{d}x - B\int_B^{+\infty} e^{-x^2/2}\mathrm{d}x\Big) \\ &= 1 - Be^{B^2/2}\int_B^{+\infty} e^{-t^2/2}\mathrm{d}t .\end{aligned}$$

Now we use twice the formula

$$\int_x^{+\infty} \frac{1}{t^k} e^{-t^2/2}\mathrm{d}t = \frac{e^{-x^2/2}}{x^{k+1}} - (k+1)\int_x^{+\infty} \frac{1}{t^{k+2}} e^{-t^2/2}\mathrm{d}t\,,$$

which holds for all $x > 0$ and all $k \geqslant 0$ by a simple integration by parts. This yields

$$e^{B^2/2}\int_B^{+\infty} e^{-t^2/2}\mathrm{d}t = \frac{1}{B} - \frac{1}{B^3} + \frac{3}{B^5} - 15\int_B^{+\infty} \frac{1}{t^6} e^{-t^2/2}\mathrm{d}t \leqslant \frac{1}{B} - \frac{1}{B^3} + \frac{3}{B^5}\,.$$

Then

$$\int_0^{+\infty} te^{-Bt} g(t)\mathrm{d}t = 1 - Be^{B^2/2}\int_B^{+\infty} e^{-t^2/2}\mathrm{d}t \geqslant \frac{1}{B^2} - \frac{3}{B^4} = \frac{1}{B^2}\Big(1 - \frac{3}{B^2}\Big)\,.$$

Combining with (148) proves the claim. Finally, as $B \geqslant e$, $3/B^2 \leqslant 3/e^2 \leqslant 1/2$ and the result follows. □

9 Proofs of results from Section 3

9.1 Proof of Proposition 2 (regularity at constant scales)

First, note that Assumption 2 holds with $c = \eta^{-1} < c(K,\eta)$, since $\mathbb{P}(|\langle u^*, X\rangle| \leqslant t) \leqslant 1 \leqslant \eta^{-1}\cdot t$ for any $t \geqslant \eta$. We now show that Assumption 3 also holds for $c = c(K,\eta)$. We start by writing, for any $v \in S^{d-1}$ such that $\langle u^*, v\rangle \geqslant 0$ and $s, t > 0$,

$$\mathbb{P}\big(|\langle u^*, X\rangle| \leqslant s,\, |\langle v, X\rangle| \geqslant t\big) \geqslant \mathbb{P}(|\langle v, X\rangle| \geqslant t) - \mathbb{P}(|\langle u^*, X\rangle| > s)\,.$$

In order to lower bound the first term above, we apply the Paley-Zygmund inequality (162) to $Z = \langle v, X\rangle^2$ (with $\mathbb{E}[Z] = 1$), which gives

$$\mathbb{P}\Big(|\langle v, X\rangle| \geqslant \frac{1}{\sqrt{2}}\Big) = \mathbb{P}\Big(\langle v, X\rangle^2 \geqslant \frac{1}{2}\mathbb{E}[\langle v, X\rangle^2]\Big) \geqslant \frac{1}{4}\frac{\mathbb{E}[\langle v, X\rangle^2]^2}{\mathbb{E}[\langle v, X\rangle^4]} \geqslant \frac{1}{4K^4}\,,$$

where the last inequality follows from the fact that $\|\langle v, X\rangle\|_{\psi_1} \leqslant K$, which by Definition 5 implies that $\|\langle v, X\rangle\|_4 \leqslant 4K/(2e) \leqslant K$. In addition, since $\|\langle u^*, X\rangle\|_{\psi_1} \leqslant K$, Lemma 35 implies that

$$\mathbb{P}\big(|\langle u^*, X\rangle| > 2K\log(2K)\big) \leqslant e^{-2\times 2\log(2K)} = \frac{1}{16K^4}\,. \tag{149}$$

Combining the previous inequalities and using that $\|u^* - v\| \leqslant \sqrt{2}$ and $\eta \leqslant e^{-1}$, we obtain that

$$\begin{aligned}&\mathbb{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{2K\log(2K)}{\eta}\cdot\eta,\, |\langle v, X\rangle| \geqslant \frac{\max\{\|u^* - v\|, \eta\}}{2}\Big) \geqslant \frac{1}{4K^4} - \frac{1}{16K^4}\\ &= \frac{3}{16K^4} \geqslant \frac{3e\eta}{16K^4} \geqslant \frac{\eta}{2K^4}\,,\end{aligned}$$

which shows that Assumption 3 holds with $c = c_{K,\eta}$ given by (149).

9.2 Proof of Proposition 3 (regularity of log-concave distributions)

In this section, we show that centered isotropic log-concave distributions satisfy Assumptions 1, 2 and 3, in every direction $u^* \in S^{d-1}$ and at any scale $\eta \in (0, e^{-1})$.

First, it is a standard fact that log-concave measures are sub-exponential.

Lemma 30. *For every isotropic log-concave random vector X in $\mathbb{R}^d$, one has $\|\langle v, X\rangle\|_{\psi_1} \leqslant \sqrt{2}e$ for every $v \in S^{d-1}$.*

Proof. Corollary 5.7 in [GNT14] with $q = 2$ shows that for all $v \in S^{d-1}$ and $p \geqslant 1$,

$$\|\langle v, X\rangle\|_p \leqslant \frac{(p!)^{1/p}}{2^{1/2}} \|\langle v, X\rangle\|_2 \leqslant \frac{p}{\sqrt{2}}.$$

Hence $\langle v, X\rangle$ is sub-exponential with $\|\langle v, X\rangle\|_{\psi_1} \leqslant \sqrt{2}e$. □

Lemma 31. *Let X be an isotropic random vector in $\mathbb{R}^d$ with log-concave distribution. Then for all $u \in S^{d-1}$ and all $t > 0$,*

$$\mathbb{P}\big(|\langle u, X\rangle| \leqslant t\big) \leqslant 2t. \tag{150}$$

In other words, X satisfies Assumption 2 with constant $c_1 = 2$, for all $u \in S^{d-1}$ and $\eta > 0$.

Proof. The random variable $\langle u, X\rangle$ is log-concave since the random vector X is, and additionally $\mathbb{E}[\langle u, X\rangle] = 0$ and $\mathbb{E}[\langle u, X\rangle^2] = 1$. It then follows from [BL19, Proposition B.2] that $\langle u, X\rangle$ admits a density f_u that is upper-bounded by 1 on $\mathbb{R}$, which proves (150). □

We now show that the two-dimensional margin condition is satisfied at all scales and in every direction. Note that the case of constant scales follows from the sub-exponential tails, by Proposition 2. For small scales, the proof uses the fact that centered and isotropic low-dimensional log-concave densities are lower-bounded around the origin.

Fact 5. *There exist absolute constants $\varepsilon, c_2 \leqslant 1$ such that for any isotropic and centered density f on $\mathbb{R}^2$ and $z \in [-\varepsilon, \varepsilon]^2$, one has $f(z) \geqslant c_2$.*

Another way of saying this is that, with the same notation, f is bounded from below by a constant factor of the uniform density on $[-\varepsilon, \varepsilon]^2$.

Lemma 32. *There exist universal constants $C \geqslant 1$ and $\varepsilon \in (0, 1)$ such that for any centered and isotropic log-concave random vector X in $\mathbb{R}^d$, for all $\eta \in (0, \varepsilon]$ and $u \in S^{d-1}$, for all $v \in S^{d-1}$,*

$$\mathbb{P}\Big(|\langle u, X\rangle| \leqslant \eta\,;\, |\langle v, X\rangle| \geqslant \frac{1}{C} \max\big\{\eta, \|u - v\|\big\}\Big) \geqslant \frac{\eta}{C}. \tag{151}$$

Proof. Let $u, v \in S^{d-1}$ be such that $\langle u, v\rangle \geqslant 0$. Using Remark 1, we work with the quantity $\sqrt{1 - \langle u, v\rangle^2}$ rather than $\|u - v\|$. Based on Fact 5, we start by reducing the problem at hand to uniform distributions. We start by writing $\langle u, v\rangle = \cos\phi$ for some $\phi \in [0, \pi/2]$, so that $\sqrt{1 - \langle u, v\rangle^2} = \sin\phi$ and we define (if $v \neq u$)

$$w = \frac{v - \langle u, v\rangle u}{\sqrt{1 - \langle u, v\rangle^2}} = \frac{v - \cos(\phi)\, u}{\sin\phi} \in S^{d-1}.$$

This way, $\langle u, w\rangle = 0$ and in particular, $(\langle u, X\rangle, \langle w, X\rangle)$ is a centered and isotropic log-concave random vector in $\mathbb{R}^2$, whose density will be denoted by f_0 throughout the rest of this proof. By Fact 5, it holds that

$$f_0(s, t) \geqslant c_2 \mathbf{1}(|s| \leqslant \varepsilon, |t| \leqslant \varepsilon),$$

for some absolute constant ε and all $(s, t) \in \mathbb{R}^2$. In particular, letting U, W be i.i.d. uniform variables on $[-\varepsilon, \varepsilon]$, with joint density

$$g_0(s, t) = \frac{1}{4\varepsilon^2} \mathbf{1}(|s| \leqslant \varepsilon, |t| \leqslant \varepsilon),$$

the previous inequality can be rewritten as

$$f_0 \geqslant 4\varepsilon^2 c_2 g_0\,. \tag{152}$$

Now, let f denote the density of $(\langle u, X\rangle, \langle v, X\rangle)$ and let $\eta \in (0, \varepsilon]$. As we seek to establish (151), our goal is to bound from below (the integral of) f as

$$\int_{|s|\leqslant \eta} \int_{|t|\geqslant \frac{\max\{\eta, \sin\phi\}}{C}} f(s,t)\mathrm{d}s\mathrm{d}t \geqslant \frac{\eta}{C}\,, \tag{153}$$

for some $C \geqslant 1$ that may depend on c_2 and ε. As $\langle v, X\rangle = \cos(\phi)\langle u, X\rangle + \sin(\phi)\langle w, X\rangle$, we let $V = \cos(\phi)U + \sin(\phi)W$ and denote by g the joint density of (U, V). Then, since f is obtained from f_0 by the same change of variables as g is obtained from g_0, it is enough to prove that g satisfies (153). We now do so.

First, if u is close to v (in a sense measured by the scale η, *i.e.* that $\sin\phi < \eta$), the two-dimensional condition essentially reduces to a one-dimensional property. More precisely, if $\eta > \sin\phi$, then for every $c \geqslant 1$, one has

$$\mathbb{P}\left(|U| \leqslant \eta\,;\, |V| \geqslant \frac{\eta}{c}\right) \geqslant \mathbb{P}(|U| \leqslant \eta) - \mathbb{P}\left(|V| < \frac{\eta}{c}\right)$$

On one hand, as $\eta \leqslant \varepsilon$, $\mathbb{P}(|U| \leqslant \eta) = \eta/\varepsilon$. On the other hand, regarding the second term, using the fact that U is independent of W and symmetric, we find by conditioning on W that

$$\begin{aligned}\mathbb{P}\Big(|V| \leqslant \frac{\eta}{c}\Big|W\Big) &= \mathbb{P}\Big(\big|U\cos\phi - W\sin\phi\big| \leqslant \frac{\eta}{c}\Big|W\Big)\\ &= \mathbb{P}\Big(\Big|U - W\frac{\sin\phi}{\cos\phi}\Big| \leqslant \frac{\eta}{c\cos\phi}\Big|W\Big)\\ &\leqslant \mathbb{P}\Big(\Big|U - W\frac{\sin\phi}{\cos\phi}\Big| \leqslant 1.1\frac{\eta}{c}\Big|W\Big)\,,\end{aligned} \tag{154}$$

where the last line uses that $\sin\phi \leqslant \eta \leqslant e^{-1}$ hence $\cos\phi \geqslant \sqrt{1 - e^{-2}}$, and a bound on the numerical constant. Then, recalling that $U \sim \mathcal{U}([-\varepsilon, \varepsilon])$ it holds for all $t \in \mathbb{R}$ and $r \geqslant 0$ that

$$\mathbb{P}(|U - t| \leqslant r) \leqslant \frac{r}{\varepsilon}\,.$$

Thus

$$\mathbb{P}\Big(\Big|U - W\frac{\sin\phi}{\cos\phi}\Big| \leqslant 1.1\frac{\eta}{c}\Big|W\Big) \leqslant 1.1\frac{\eta}{c\varepsilon}\,.$$

It then follows that as soon as $c \geqslant 2.2$,

$$\mathbb{P}\left(|U| \leqslant \eta\,;\, |V| \geqslant \frac{\eta}{c}\right) \geqslant \frac{\eta}{\varepsilon} - 1.1\frac{\eta}{c\varepsilon} \geqslant \frac{\eta}{2\varepsilon}\,. \tag{155}$$

We now turn our attention to the other regime, where $\eta \leqslant \sin\phi$. We now rely on the following. Recalling that $W \sim \mathcal{U}([-\varepsilon, \varepsilon])$ it holds for all $t \in \mathbb{R}$ that

$$\mathbb{P}\Big(|W - t| \geqslant \frac{\varepsilon}{2}\Big) \geqslant \frac{1}{2}\,.$$

This implies (since W is independent of U and symmetric) that as soon as $c \geqslant 2/\varepsilon$,

$$\begin{aligned}\mathbb{P}\Big(|V| \geqslant \frac{\sin\phi}{c}\Big|U\Big) &= \mathbb{P}\Big(\big|W\sin\phi - U\cos\phi\big| \geqslant \frac{\sin\phi}{c}\Big|U\Big)\\ &= \mathbb{P}\Big(\Big|W - U\frac{\cos\phi}{\sin\phi}\Big| \geqslant \frac{1}{c}\Big|U\Big)\\ &\geqslant \mathbb{P}\Big(\Big|W - U\frac{\cos\phi}{\sin\phi}\Big| \geqslant \frac{\varepsilon}{2}\Big|U\Big) \geqslant \frac{1}{2}\,.\end{aligned}$$

It follows that

$$\mathbb{P}\Big(|U| \leqslant \eta\,;\, |V| \geqslant \frac{\sin\phi}{c}\Big) = \mathbb{E}\Big[\mathbf{1}\big(|U| \leqslant \eta\big)\mathbb{P}\Big(|V| \geqslant \frac{\sin\phi}{c}\Big|U\Big)\Big]$$
$$\geqslant \frac{1}{2}\mathbb{P}\big(|U| \leqslant \eta\big) = \frac{\eta}{2\varepsilon}\,. \tag{156}$$

Combining (155) and (156) shows that

$$\mathbb{P}\Big(|U| \leqslant \eta\,;\, |V| \geqslant \frac{\max\{\eta, \sin\phi\}}{c}\Big) \geqslant \frac{\eta}{2\varepsilon}\,, \tag{157}$$

meaning that g satisfies (153). Using (152) and applying the same change of variables to g_0 and f_0, we conclude that

$$\mathbb{P}\Big(|\langle u, X\rangle| \leqslant \eta\,;\, |\langle v, X\rangle| \geqslant \frac{\varepsilon}{2}\max\big\{\eta, \sqrt{1-\langle u, v\rangle^2}\big\}\Big) \geqslant 2\varepsilon c_2 \eta\,. \qquad \square$$

9.3 Proof of Proposition 4 (regularity for i.i.d. coordinates)

This section contains the proofs of the results from Section 3.3. Specifically, we show that random vectors X with i.i.d. sub-exponential coordinates (Assumption 4) satisfy Assumptions 1, 2 and 3 down to a scale $\eta \asymp 1/\sqrt{d}$ in the "diffuse" direction $u^* = (1/\sqrt{d}, \dots, 1/\sqrt{d})$.

Assumption 1

We first recall the standard fact that a random vector with independent sub-exponential coordinates is itself sub-exponential.

Lemma 33. *If $X_1, \dots, X_d$ are independent centered real random variables with $\|X_j\|_{\psi_1} \leqslant K$ for $1 \leqslant j \leqslant d$, for every $v = (v_j)_{1\leqslant j\leqslant d} \in S^{d-1}$, letting $X = (X_j)_{1\leqslant j\leqslant d}$ one has $\|\langle v, X\rangle\|_{\psi_1} \leqslant 4K$.*

Proof. By the sixth point of Lemma 35, X_j is $(K^2/2, K/2)$-sub-gamma for every j. By independence and the third point of the same lemma, $\langle v, X\rangle = \sum_{j=1}^d v_j X_j$ is sub-gamma, with parameters $K^2/2 \cdot \sum_{j=1}^d v_j^2 = K^2/2$ and $K/2 \cdot \max_{1\leqslant j\leqslant d} |v_j| \leqslant K/2$. Since the same also holds for $-\langle v, X\rangle = \langle -v, X\rangle$, the fifth point of Lemma 35 implies that $\|\langle v, X\rangle\|_{\psi_1} \leqslant 2\sqrt[3]{2e}\max(K/\sqrt{2}, 2 \cdot K/2) = 2\sqrt[3]{2e}K \leqslant 4K$. $\square$

Assumption 2: proof of Lemma 1

The second condition on one-dimensional marginals holds because, if $u \in S^{d-1}$ is sufficiently "diffuse", then the distribution of $\langle u, X\rangle$ is close to that of a standard Gaussian variable. This fact follows from the Berry-Esseen theorem (see, e.g., [Fel68]); we will use the version with small numerical constants from [She10, Tyu12].

Lemma 34 ([Tyu12], Theorem 1)**.** *Let $Z_1, \dots, Z_d$ be independent centered random variables with $\sum_{j=1}^d \mathbb{E}[Z_j^2] = 1$. Let $Z = \sum_{j=1}^d Z_j$ and $G \sim \mathsf{N}(0, 1)$. Then, for every $t \in \mathbb{R}$, one has*

$$\big|\mathbb{P}(Z \leqslant t) - \mathbb{P}(G \leqslant t)\big| \leqslant 0.56 \cdot \sum_{j=1}^d \mathbb{E}\big[|Z_j|^3\big]\,. \tag{158}$$

We now proceed with the proof of Lemma 1, which states that Assumption 2 holds.

Proof of Lemma 1. First, applying Lemma 34 to $-Z_1, \dots, -Z_d$ and $-s$ gives a similar bound as (158) for $\mathbb{P}(Z < s)$. After taking differences we deduce that, under the assumptions of Lemma 34, for every $s, t \in \mathbb{R}$ with $s \leqslant t$, one has

$$\big|\mathbb{P}(s \leqslant Z \leqslant t) - \mathbb{P}(s \leqslant G \leqslant t)\big| \leqslant 1.12 \cdot \sum_{j=1}^{d} \mathbb{E}\big[|Z_j|^3\big] . \tag{159}$$

We apply this inequality to $t \in [K^3 \|u\|_3^3, 1]$, $s = -t$ and $Z_j = u_j X_j$, so that $\mathbb{E}[Z_j] = 0$, $\sum_{j=1}^d \mathbb{E}[Z_j^2] = \sum_{j=1}^d u_j^2 = 1$, and $\mathbb{E}[|Z_j|^3] = |u_j|^3 \|X_j\|_3^3 \leqslant |u_j|^3 (3K/2e)^3$. As $Z = \langle u, X\rangle$,

$$\big|\mathbb{P}(|\langle u, X\rangle| \leqslant t) - \mathbb{P}(|G| \leqslant t)\big| \leqslant 1.12 \cdot \sum_{j=1}^{d} \Big(\frac{3K}{2e}\Big)^3 |u_j|^3 \leqslant \frac{K^3}{5} \|u\|_3^3 , \tag{160}$$

where we used that $1.12 \times (\frac{3}{2e})^3 \leqslant 1/5$. Now, since the density of G is between $e^{-1/2}/\sqrt{2\pi}$ and $1/\sqrt{2\pi}$ on $[-1, 1]$, one has

$$\frac{2t}{\sqrt{2\pi e}} \leqslant \mathbb{P}(|G| \leqslant t) \leqslant \frac{2t}{\sqrt{2\pi}} .$$

Plugging these inequalities into (160) and using that $K^3 \|u\|_3^3 \leqslant t$ gives

$$\Big(\frac{2}{\sqrt{2\pi e}} - \frac{1}{5}\Big) t \leqslant \mathbb{P}(|\langle u, X\rangle| \leqslant t) \leqslant \Big(\sqrt{\frac{2}{\pi}} + \frac{1}{5}\Big) t ,$$

which implies (30) by further bounding the numerical constants. □

Assumption 3: proof of Lemma 2

We now establish the two-dimensional margin condition.

Proof of Lemma 2. As discussed in Section 3.3, the idea of the proof is to perturb the vector X by a random permutation of its coordinates, and use the fact that such transformations do not affect the distribution of X nor the value of $\langle u^*, X\rangle$, but induce some variability in the quantity $\langle v, X\rangle$.

Perturbation by random permutations. Let σ be a permutation of $\{1, \dots, d\}$. For $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, we let $x^\sigma = (x_{\sigma(1)}, \dots, x_{\sigma(d)})$ denote the vector obtained by permuting the coordinates of x by σ. First, since $X_1, \dots, X_d$ are i.i.d., the vector X^σ has the same distribution as X. In addition, one has

$$\langle u^*, X^\sigma\rangle = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} X_{\sigma(i)} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} X_i = \langle u^*, X\rangle .$$

It follows that, for any $v \in S^{d-1}$ and $s > 0$,

$$\mathbb{P}\big(|\langle u^*, X\rangle| \leqslant \eta, \, |\langle v, X\rangle| \geqslant s\big) = \mathbb{P}\big(|\langle u^*, X\rangle| \leqslant \eta, \, |\langle v, X^\sigma\rangle| \geqslant s\big) .$$

From now on, we let σ denote a random permutation, drawn uniformly from the set $\mathfrak{S}_d$ of all permutations of $\{1, \dots, d\}$ and independent of X. We let $\mathbb{P}_\sigma$ and $\mathbb{E}_\sigma$ respectively denote the probability and expectation with respect to σ, conditionally on X. From the equality above applied to any $\sigma' \in \mathfrak{S}_d$, one has

$$\begin{aligned}\mathbb{P}\big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X\rangle| \geqslant s\big) &= \frac{1}{d!}\sum_{\sigma'\in\mathfrak{S}_d} \mathbb{P}\big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X^{\sigma'}\rangle| \geqslant s\big) \\ &= \mathbb{E}\big[\mathbb{P}\big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X^{\sigma}\rangle| \geqslant s \,|\, \sigma\big)\big] \\ &= \mathbb{E}\big[\mathbb{P}_\sigma\big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X^{\sigma}\rangle| \geqslant s\big)\big] \\ &= \mathbb{E}\big[\mathbf{1}\{|\langle u^*, X\rangle| \leqslant \eta\}\mathbb{P}_\sigma\big(|\langle v, X^{\sigma}\rangle| \geqslant s\big)\big]\,. \end{aligned} \tag{161}$$

Hence, in order to lower bound the left-hand side of (161), it suffices to lower bound $\mathbb{P}_\sigma\big(|\langle v, X^\sigma\rangle| \geqslant s\big)$ when X satisfies $|\langle u^*, X\rangle| \leqslant \eta$ (we will actually require additional symmetric conditions on X, but we omit them here for simplicity). In other words, we need to show that for such values of X, the fraction of permutations $\sigma \in \mathfrak{S}_d$ such that $|\langle v, X^\sigma\rangle| \geqslant s$ is lower-bounded.

We will achieve this by resorting to the Paley-Zygmund inequality (e.g., [Tal21, eq. (6.15) p. 181]), which asserts that for any non-negative random variable Z with $0 < \mathbb{E}[Z^2] < +\infty$, one has

$$\mathbb{P}\Big(Z \geqslant \frac{1}{2}\mathbb{E}[Z]\Big) \geqslant \frac{1}{4}\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}\,. \tag{162}$$

Applying this inequality to the random variable $Z = \langle v, X^\sigma\rangle^2$ conditionally on X gives

$$\mathbb{P}_\sigma\Big(|\langle v, X^\sigma\rangle| \geqslant \frac{1}{\sqrt{2}}\mathbb{E}_\sigma[\langle v, X^\sigma\rangle^2]^{1/2}\Big) \geqslant \frac{1}{4}\frac{\mathbb{E}_\sigma[\langle v, X^\sigma\rangle^2]^2}{\mathbb{E}_\sigma[\langle v, X^\sigma\rangle^4]}\,. \tag{163}$$

We are therefore led to bound $\mathbb{E}_\sigma[\langle v, X^\sigma\rangle^2]^{1/2}$ from below and $\mathbb{E}_\sigma[\langle v, X^\sigma\rangle^4]^{1/4}$ from above, ideally to conclude that these two quantities are both of the order of the value from Lemma 2. The advantage of this approach is that it reduces to evaluating expectations of polynomials of the variables $X_{\sigma(i)}, 1 \leqslant i \leqslant d$ under the uniform distribution on $\mathfrak{S}_d$, which can be computed exactly.

Lower bound on the second moment. Denote for $p \in \mathbb{N}$,

$$\phi = \phi(X) = \langle u^*, X\rangle = \frac{1}{\sqrt{d}}\sum_{i=1}^d X_i\,, \qquad \mu_p = \mu_p(X) = \frac{1}{d}\sum_{i=1}^d X_i^p\,.$$

In particular, one has $|\mu_p| \leqslant \mu_4^{p/4}$ for $1 \leqslant p \leqslant 4$. In the following, we assume that X satisfies $\mu_2(X) \geqslant 1/2$ and $|\phi(X)| \leqslant \eta \leqslant 1$.

Now, for any $v \in S^{d-1}$,

$$\mathbb{E}_\sigma\big[\langle v, X^\sigma\rangle^2\big] = \mathbb{E}_\sigma\Bigg[\Big(\sum_{i=1}^d v_i X_{\sigma(i)}\Big)^2\Bigg] = \sum_{1\leqslant i,j\leqslant d} v_i v_j \mathbb{E}_\sigma[X_{\sigma(i)}X_{\sigma(j)}]\,.$$

For $i = j$, since $\sigma(i)$ is uniformly distributed on $\{1, \dots, d\}$ one has

$$\mathbb{E}_\sigma[X_{\sigma(i)}X_{\sigma(j)}] = \mathbb{E}_\sigma[X_{\sigma(i)}^2] = \frac{1}{d}\sum_{k=1}^d X_k^2 = \mu_2\,.$$

On the other hand, if $i \neq j$, then $(\sigma(i), \sigma(j))$ is distributed uniformly on pairs (k, l) such that $k \neq l$, thus

$$\mathbb{E}_\sigma[X_{\sigma(i)}X_{\sigma(j)}] = \frac{1}{d(d-1)}\sum_{k\neq l} X_k X_l = \frac{1}{d(d-1)}\Bigg\{\Big(\sum_{k=1}^d X_k\Big)^2 - \sum_{k=1}^d X_k^2\Bigg\} = \frac{\phi^2 - \mu_2}{d-1}.$$

Combining the previous two equations, we get for any i, j that

$$\mathbb{E}_\sigma[X_{\sigma(i)}X_{\sigma(j)}] = \frac{\phi^2 - \mu_2}{d-1} + \Big(\mu_2 + \frac{\mu_2 - \phi^2}{d-1}\Big)\mathbf{1}(i = j)\,.$$

Hence,

$$\begin{aligned}
\mathbb{E}_\sigma\big[\langle v, X^\sigma\rangle^2\big] &= \frac{\phi^2 - \mu_2}{d-1}\Bigg(\sum_{i=1}^d v_i\Bigg)^2 + \Big(\mu_2 + \frac{\mu_2 - \phi^2}{d-1}\Big)\sum_{i=1}^d v_i^2 \\
&= (\phi^2 - \mu_2)\frac{d}{d-1}\langle u^*, v\rangle^2 + \mu_2 + \frac{\mu_2 - \phi^2}{d-1} \\
&= \frac{d}{d-1}\Big[\mu_2(1 - \langle u^*, v\rangle^2) + \phi^2\langle u^*, v\rangle^2 - \frac{\phi^2}{d}\Big] \\
&\geqslant \mu_2(1 - \langle u^*, v\rangle^2) + \phi^2\langle u^*, v\rangle^2 - \frac{\phi^2}{d}\,.
\end{aligned}$$

Recalling that $\mu_2 \geqslant 1/2$, that $|\phi| \leqslant \eta \leqslant 1$ and $d \geqslant 2025$, then either $\langle u^*, v\rangle^2 \geqslant 1/4$ and

$$\mu_2(1 - \langle u^*, v\rangle^2) + \phi^2\langle u^*, v\rangle^2 - \frac{\phi^2}{d} \geqslant \frac{1 - \langle u^*, v\rangle^2}{2} + 0.97\langle u^*, v\rangle^2\phi^2\,,$$

or $\langle u^*, v\rangle^2 < 1/4$ and then

$$\mu_2(1 - \langle u^*, v\rangle^2) + \phi^2\langle u^*, v\rangle^2 - \frac{\phi^2}{d} \geqslant \frac{3}{8} - \frac{1}{2025} \geqslant 0.37\big[1 - \langle u^*, v\rangle^2 + \langle u^*, v\rangle^2\phi^2\big]\,.$$

Combining the previous inequalities, we get in all cases that

$$\mathbb{E}_\sigma\big[\langle v, X^\sigma\rangle^2\big] \geqslant 0.37\big[1 - \langle u^*, v\rangle^2 + \langle u^*, v\rangle^2\phi^2\big]\,. \tag{164}$$

Upper bound on the fourth moment. We now turn to the control the conditional fourth moment. Let $v \in S^{d-1}$ such that $\langle u^*, v\rangle \geqslant 0$; we may write $v = \sqrt{1-\alpha^2}u^* + \alpha w$ where $\alpha = \sqrt{1 - \langle u^*, v\rangle^2}$ and $w \in S^{d-1}$ is such that $\langle u^*, w\rangle = 0$. We then have

$$\begin{aligned}
\mathbb{E}_\sigma\big[\langle v, X^\sigma\rangle^4\big] &= \mathbb{E}_\sigma\Big[\big(\alpha\langle w, X^\sigma\rangle + \sqrt{1-\alpha^2}\langle u, X^\sigma\rangle\big)^4\Big] \\
&\leqslant 8\,\mathbb{E}_\sigma\big[\alpha^4\langle w, X^\sigma\rangle^4 + (1-\alpha^2)^2\langle u, X^\sigma\rangle^4\big] \\
&= 8\big\{(1 - \langle u^*, v\rangle^2)^2\,\mathbb{E}_\sigma\big[\langle w, X^\sigma\rangle^4\big] + \langle u^*, v\rangle^4\phi^4\big\}\,.
\end{aligned} \tag{165}$$

In light of (165), it remains to show that $\mathbb{E}_\sigma[\langle w, X^\sigma\rangle^4] \lesssim_\kappa 1$.

We start by writing:

$$\mathbb{E}_\sigma\big[\langle w, X^\sigma\rangle^4\big] = \sum_{1\leqslant i,j,k,l\leqslant d} w_i w_j w_k w_l \mathbb{E}[X_{\sigma(i)}X_{\sigma(j)}X_{\sigma(k)}X_{\sigma(l)}]\,. \tag{166}$$

We abbreviate "pairwise distinct" (indices in $\{1, \ldots, d\}$) by p.d., and denote for i, j, k, l p.d.,

$$\begin{aligned}
\alpha_4 &= \mathbb{E}_\sigma[X_{\sigma(i)}^4] \\
\alpha_{31} &= \mathbb{E}_\sigma[X_{\sigma(i)}^3X_{\sigma(j)}] \\
\alpha_{22} &= \mathbb{E}_\sigma[X_{\sigma(i)}^2X_{\sigma(j)}^2] \\
\alpha_{211} &= \mathbb{E}_\sigma[X_{\sigma(i)}^2X_{\sigma(j)}X_{\sigma(k)}] \\
\alpha_{1111} &= \mathbb{E}_\sigma[X_{\sigma(i)}X_{\sigma(j)}X_{\sigma(k)}X_{\sigma(l)}]\,;
\end{aligned}$$

these quantities are independent of i,j,k,l p.d. since σ is distributed uniformly on the symmetric group, hence $(\sigma(i),\sigma(j),\sigma(k),\sigma(l))$ is distributed uniformly on the set of p.d. indices. Hence, collecting the terms in the right-hand side of (166) depending on the distinct indices, we obtain

$$\mathbb{E}_\sigma\left[\langle w, X^\sigma\rangle^4\right] = \Big(\sum_i w_i^4\Big)\alpha_4 + 4\Big(\sum_{i,j \text{ p.d.}} w_i^3 w_j\Big)\alpha_{31} + 3\Big(\sum_{i,j \text{ p.d.}} w_i^2 w_j^2\Big)\alpha_{22} + \\ + 6\Big(\sum_{i,j,k \text{ p.d.}} w_i^2 w_j w_k\Big)\alpha_{211} + \Big(\sum_{i,j,k,l \text{ p.d.}} w_i w_j w_k w_l\Big)\alpha_{1111}\,. \quad (167)$$

We control the sum in (167) by separately controlling the $\alpha_\cdot$ terms (that depend on X) and their coefficients depending on w. The control of the former terms is simple, as we simply bound all these terms by the empirical fourth moment μ_4: for every $1 \leqslant r \leqslant 4$ and $\iota_1 \geqslant \ldots \geqslant \iota_r \geqslant 1$ such that $\iota_1 + \cdots + \iota_r = 4$, we have

$$|\alpha_{\iota_1,\ldots,\iota_r}| \leqslant \mu_4\,. \quad (168)$$

To show (168), first note that since $\sigma(1)$ is uniformly distributed in $\{1,\ldots,d\}$, we have

$$\alpha_4 = \mathbb{E}_\sigma[X_{\sigma(1)}^4] = \frac{1}{d}\sum_{i=1}^d X_i^4 = \mu_4\,.$$

Now for $\iota_1,\ldots,\iota_r$ as above, Hölder's inequality (with $\iota_1/4+\cdots+\iota_r/4 = 1$) implies that

$$|\alpha_{\iota_1,\ldots,\iota_r}| \leqslant \mathbb{E}_\sigma[|X_{\sigma(1)}|^{\iota_1}\cdots|X_{\sigma(r)}|^{\iota_r}] = \mathbb{E}_\sigma[(X_{\sigma(1)}^4)^{\iota_1/4}\cdots(X_{\sigma(r)}^4)^{\iota_r/4}] \\ \leqslant \mathbb{E}_\sigma[X_{\sigma(1)}^4]^{\iota_1/4}\cdots\mathbb{E}_\sigma[X_{\sigma(r)}^4]^{\iota_r/4} = \mu_4\,.$$

We now turn to the control of the coefficients in (167) that depend on w. Although one could in principle use the same method as above, namely Hölder's inequality combined with the fact that $\|w\|_4^4 \leqslant \|w\|_2^4 = 1$, this would result in a highly suboptimal bound in $O(d^2)$. In order to improve this bound, we exploit the additional information that w is orthogonal to u^*, namely

$$0 = \langle u^*, w\rangle = \frac{1}{\sqrt{d}}\sum_{i=1}^d w_i\,,$$

so that $\sum_i w_i = 0$. We will therefore decompose the sums in (167) by making the quantities $\sum_i w_i = 0$ and $\sum_i w_i^2 = 1$ appear.

For the first term, we have

$$0 \leqslant \sum_i w_i^4 \leqslant \sum_i w_i^2 = 1\,.$$

For the second term, we write

$$\sum_{i,j \text{ p.d.}} w_i^3 w_j = \Big(\sum_i w_i^3\Big)\Big(\sum_j w_j\Big) - \sum_i w_i^4 = -\sum_i w_i^4 \in [-1, 0]\,.$$

For the third term,

$$\sum_{i,j \text{ p.d.}} w_i^2 w_j^2 = \Big(\sum_i w_i^2\Big)^2 - \sum_i w_i^4 = 1 - \sum_i w_i^4 \in [0,1]\,.$$

For the fourth term, by distinguishing the different possible configurations of $i,j,k \in \{1,\ldots,d\}$,

$$\sum_{i,j,k \text{ p.d.}} w_i^2 w_j w_k = \Big(\sum_i w_i^2\Big)\Big(\sum_j w_j\Big)\Big(\sum_k w_k\Big) - \sum_i w_i^4 - \sum_{i,j \text{ p.d.}} w_i^2 w_j^2 - 2\sum_{i,j \text{ p.d.}} w_i^3 w_j\,. \quad (169)$$

Plugging the previous identities in (169), we obtain

$$\sum_{i,j,k\,\text{p.d.}} w_i^2 w_j w_k = -\sum_i w_i^4 - \Big(1 - \sum_i w_i^4\Big) - 2\Big(-\sum_i w_i^4\Big)$$
$$= 2\sum_i w_i^4 - 1 \in [-1,1]\,.$$

Finally, for the fifth term, we write (collecting the terms similarly to (167))

$$\sum_{i,j,k,l\,\text{p.d.}} w_i w_j w_k w_l = \Big(\sum_i w_i\Big)\Big(\sum_j w_j\Big)\Big(\sum_k w_k\Big)\Big(\sum_l w_l\Big) - \sum_i w_i^4 - \\ - 4\sum_{i,j\,\text{p.d.}} w_i^3 w_j - 3\sum_{i,j\,\text{p.d.}} w_i^2 w_j^2 - 6\sum_{i,j,k\,\text{p.d.}} w_i^2 w_j w_k\,. \tag{170}$$

Using the identities for the previous four terms, equation (170) becomes

$$\sum_{i,j,k,l\,\text{p.d.}} w_i w_j w_k w_l = -\sum_i w_i^4 - 4\Big(-\sum_i w_i^4\Big) - 3\Big(1 - \sum_i w_i^4\Big) - 6\Big(2\sum_i w_i^4 - 1\Big)$$
$$= -6\sum_i w_i^4 + 3 \in [-3,3]\,.$$

Finally, injecting the previous bounds into the decomposition (167), we obtain

$$\mathbb{E}_\sigma\big[\langle w, X^\sigma\rangle^4\big] \leqslant \Big|\sum_i w_i^4\Big|\cdot|\alpha_4| + 4\Big|\sum_{i,j\,\text{p.d.}} w_i^3 w_j\Big|\cdot|\alpha_{31}| + 3\Big|\sum_{i,j\,\text{p.d.}} w_i^2 w_j^2\Big|\cdot|\alpha_{22}| + \\ + 6\Big|\sum_{i,j,k\,\text{p.d.}} w_i^2 w_j w_k\Big|\cdot|\alpha_{211}| + \Big|\sum_{i,j,k,l\,\text{p.d.}} w_i w_j w_k w_l\Big|\cdot|\alpha_{1111}| \\ \leqslant \big(1 + 4\times 1 + 3\times 1 + 6\times 1 + 3\big)\mu_4 = 17\,\mu_4\,,$$

which combined with (165) gives

$$\mathbb{E}_\sigma\big[\langle v, X^\sigma\rangle^4\big] \leqslant 8\big\{17(1-\langle u^*, v\rangle^2)^2\mu_4 + \langle u^*, v\rangle^4\phi^4\big\}\,. \tag{171}$$

Symmetric condition. So far, we have established a lower bound on the second moment $\mathbb{E}_\sigma[\langle v, X^\sigma\rangle^2]$ and an upper bound on the fourth moment $\mathbb{E}_\sigma[\langle v, X^\sigma\rangle^4]$, both over the random permutation σ and conditionally on X. These upper and lower bounds are of the desired order whenever X satisfies the following three conditions: $\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta$, $\mu_2(X) \geqslant 1/2$, and $\mu_4(X) = O_\kappa(1)$. We are therefore reduced to lower-bounding the probability that X simultaneously satisfies those three conditions, which are symmetric in the coordinates of X.

We thus establish a lower bound on $\mathbb{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta,\ \mu_2(X) \geqslant 1/2,\ \mu_4(X) \leqslant 2\kappa^4\big)$. We start by writing

$$\mathbb{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta,\ \mu_2(X) \geqslant 1/2,\ \mu_4(X) \leqslant 2\kappa^4\big) \\ = \mathbb{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta\big) - \mathbb{P}\big(\{|\langle u^*, X\rangle| \leqslant \eta\} \cap \{\mu_2(X) < 1/2 \text{ or } \mu_4(X) > 2\kappa^4\}\big) \\ \geqslant \mathbb{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta\big) - \mathbb{P}\big(\mu_2(X) < 1/2\big) - \mathbb{P}\big(\mu_4(X) > 2\kappa^4\big)\,. \tag{172}$$

Now, applying the Berry-Esseen inequality (Lemma 34) and proceeding as in the proof of Lemma 1, using that $\mathbb{E}[(|X_i|/\sqrt{d})^3] \leqslant \mathbb{E}[|X_i|^8]^{3/8}/d^{3/2} \leqslant \kappa^3/d^{3/2}$, we get

$$\mathbb{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta\big) \geqslant \frac{\eta}{\sqrt{2\pi e}} - \frac{2.24\kappa^3}{\sqrt{d}} \geqslant 0.24\,\eta - \frac{2.25\kappa^3}{\sqrt{d}}\,. \tag{173}$$

We now upper bound $\mathbb{P}(\mu_4(X) > 2\kappa^4)$. Applying Chebyshev's inequality to $\sum_{i=1}^d X_i^4$ gives, for any $t > 0$,

$$\mathbb{P}\Big(\Big|\frac{1}{d}\sum_{i=1}^d X_i^4 - \mathbb{E}[X_1^4]\Big| > t\Big) \leqslant \frac{\mathbb{E}[X_1^8]}{d \cdot t^2} \leqslant \frac{\kappa^8}{d \cdot t^2}\,.$$

In particular, taking $t = \kappa^4$, applying the triangle inequality and using that $\mathbb{E}[X_1^4] \leqslant \mathbb{E}[X_1^8]^{1/2} \leqslant \kappa^4$ by assumption, we get

$$\mathbb{P}\big(\mu_4(X) > 2\kappa^4\big) \leqslant 1/d\,. \tag{174}$$

Likewise, Chebyshev's inequality implies that

$$\mathbb{P}\big(\mu_2(X) < 1/2\big) \leqslant \mathbb{P}\Big(\Big|\frac{1}{d}\sum_{i=1}^d X_i^2 - 1\Big| > \frac{1}{2}\Big) \leqslant \frac{4\,\mathbb{E}[X_1^4]}{d} \leqslant \frac{4\kappa^4}{d}\,. \tag{175}$$

Plugging inequalities (173), (174) and (175) into (172) gives

$$\mathbb{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta,\, \mu_2(X) \geqslant 1/2,\, \mu_4(X) \leqslant 2\kappa^4\big) \geqslant 0.24\,\eta - \frac{2.25\kappa^3}{\sqrt{d}} - \frac{1}{d} - \frac{4\kappa^4}{d}\,,$$

which is larger than $0.12\,\eta$ whenever $\eta \geqslant \max(45\kappa^3/\sqrt{d}, 80\kappa^4/d)$. Now since $d \geqslant 2025\kappa^6$ by assumption, one has $\sqrt{d} \geqslant 45\kappa^3 \geqslant 45\kappa$ and thus $80\kappa^4/d \leqslant 80\kappa^3/(45\sqrt{d}) < 45\kappa^3/\sqrt{d}$; therefore, the previous condition reduces to $\eta \geqslant 45\kappa^3/\sqrt{d}$, which is satisfied by assumption.

Putting things together. We now conclude the proof. Define the event E by

$$E = \big\{\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta,\, \mu_2(X) \geqslant 1/2,\, \mu_4(X) \leqslant 2\kappa^4\big\}\,,$$

so that $\mathbb{P}(E) \geqslant 0.12\eta$ by the above. In addition, it follows respectively from (164) and (171) that, under the event E,

$$\mathbb{E}_\sigma\big[\langle v, X^\sigma\rangle^2\big] \geqslant 0.37\big[1 - \langle u^*, v\rangle^2 + \langle u^*, v\rangle^2(\eta/2)^2\big]\,;$$
$$\mathbb{E}_\sigma\big[\langle v, X^\sigma\rangle^4\big] \leqslant 8\big\{34\kappa^4(1 - \langle u^*, v\rangle^2)^2 + \langle u^*, v\rangle^4\eta^4\big\}\,.$$

Plugging these upper and lower bounds into (163) gives:

$$\begin{aligned}
&\mathbb{P}_\sigma\Big(|\langle v, X^\sigma\rangle| \geqslant \frac{0.6}{\sqrt{2}}\big[1 - \langle u^*, v\rangle^2 + \langle u^*, v\rangle^2\eta^2/4\big]^{1/2}\Big)\\
&\geqslant \frac{1}{4}\frac{0.37^2\big[1 - \langle u^*, v\rangle^2 + \langle u^*, v\rangle^2\eta^2/4\big]^2}{8\big[34\kappa^4(1 - \langle u^*, v\rangle^2)^2 + \langle u^*, v\rangle^4\eta^4\big]}\\
&\geqslant \frac{0.37^2}{32}\frac{(1 - \langle u^*, v\rangle^2)^2 + \langle u^*, v\rangle^4\phi^4/16}{34\kappa^4(1 - \langle u^*, v\rangle^2)^2 + \langle u^*, v\rangle^4\eta^4} \geqslant \frac{1}{8000\kappa^4}\,.
\end{aligned}$$

Now, let $s = \frac{0.6}{\sqrt{2}}[1 - \langle u^*, v\rangle^2 + \langle u^*, v\rangle^2\eta^2/4]^{1/2}$. From (161) and the above, we obtain

$$\begin{aligned}
\mathbb{P}\big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X\rangle| \geqslant s\big) &= \mathbb{E}\big[\mathbf{1}\{|\langle u^*, X\rangle| \leqslant \eta\}\mathbb{P}_\sigma\big(|\langle v, X^\sigma\rangle| \geqslant s\big)\big]\\
&\geqslant \mathbb{E}\big[\mathbf{1}_E \cdot \mathbb{P}_\sigma\big(|\langle v, X^\sigma\rangle| \geqslant s\big)\big]\\
&\geqslant \frac{\mathbb{P}(E)}{8000\kappa^4} \geqslant \frac{0.12\eta}{8000\kappa^4} \geqslant \frac{\eta}{70\,000\kappa^4}\,.
\end{aligned}$$

To conclude, note that

$$s = \frac{0.6}{\sqrt{2}} \big[1 - \langle u^*, v\rangle^2 + \langle u^*, v\rangle^2 \eta^2/4\big]^{1/2} \geqslant \frac{0.6}{\sqrt{2}} \max\big\{1 - \langle u^*, v\rangle^2, \eta^2/4\big\}^{1/2},$$

and that by Lemma 37, if $\langle u^*, v\rangle \geqslant 0$ then $\sqrt{1 - \langle u^*, v\rangle^2} \geqslant \|u^* - v\|/\sqrt{2}$; the numerical constant in Lemma 2 is obtained by lower-bounding $0.6/(2\sqrt{2}) > 0.2$.

Finally, the last part of Lemma 2 follows from (32), since under Assumption 4 one has $\mathbb{E}[X_1^4]^{1/4} \leqslant \kappa = \frac{4}{2e}\|X_1\|_{\psi_1} \leqslant \frac{2}{e}K$, which gives the desired claims by substituting for κ and bounding the numerical constants. □

Sketch of the argument to obtain the $d^{-1/4}$ scaling

We now provide an (incomplete) high-level sketch of the argument alluded to in Section 3.3, that leads to a nontrivial guarantee by combining Gaussian approximation with approximate separation of supports.

The main idea is that an arbitrary vector $v \in S^{d-1}$ either admits a "dense" sub-vector $v_I = (v_i)_{i\in I}$ (for some $I \subset \{1, \dots, d\}$) with lower-bounded ℓ^2 norm, or a "sparse" sub-vector v_I with lower-bounded ℓ^2 norm. In the first case one may resort to Gaussian approximation, and in the second case one may argue that the supports of the vectors u^* and v are "almost separated". In addition, in both cases we use the fact that the random vectors $(\langle u_I^*, X_I\rangle, \langle v_I, X_I\rangle)$ and $(\langle u_{I^c}^*, X_{I^c}\rangle, \langle v_{I^c}, X_{I^c}\rangle)$ are independent for any subset $I \subset \{1, \dots, d\}$ (since they depend on disjoint subsets of the independent variables $(X_j)_{1\leqslant j\leqslant d}$).

Specifically, let $v \in S^{d-1}$ be arbitrary. Without loss of generality one may assume that $|v_1| \geqslant \ldots \geqslant |v_d|$. Define $k = \min\{1 \leqslant k \leqslant d : \sum_{j=1}^k v_j^2 \geqslant 0.01\}$ and let $I = \{1, \dots, k\}$. In particular, one has $\sum_{j=1}^k v_j^2 \geqslant 0.01$ and $k \leqslant 0.01d$, and either $k = 1$ or $\sum_{j>k} v_j^2 > 0.98$.

On the one hand, if $k > 1$, we have $\sum_{j>k} |v_j|^3 \leqslant |v_k| \sum_{j>k} v_j^2 \leqslant |v_k| \leqslant 1/\sqrt{k}$, since $kv_k^2 \leqslant \sum_{j=1}^k v_j^2 \leqslant 1$. Combining this with the fact that $\sum_{j>k} v_j^2 > 0.98$, that $\sum_{j>k} (u_j^*)^2 = (d-k)/d > 0.99$ and $|\sum_{j>k} u_j^* v_j| = |\langle u^*, v\rangle - \sum_{j=1}^k u_j^* v_j| = |\sum_{j=1}^k u_j^* v_j| \leqslant \sqrt{(\sum_{j=1}^k (u_j^*)^2)(\sum_{j=1}^k v_j^2)} \leqslant \sqrt{k/d} \leqslant 0.1$, applying the Berry-Esseen Gaussian approximation bound on $(\langle u_{I^c}^*, X_{I^c}\rangle, \langle v_{I^c}, X_{I^c}\rangle)$ and using independence with the remaining variables, one may show that condition (31) holds with $\eta \asymp 1/\sqrt{k}$.

On the other hand, regardless of the value of $k \leqslant 0.01d$, one has $\sqrt{\sum_{j=1}^k (u_j^*)^2} = \sqrt{k/d}$ while $\sum_{j=1}^k v_j^2 \geqslant 0.01$. In other words, a constant fraction of the "energy" of the vector v is supported in I, while if $k \ll d$ only a small fraction of the energy of u^* is supported on I. This "approximate separation" of the supports of u^*, v implies that $\sum_{j=1}^k u_j^* X_j$ is very small, while $\sum_{j=1}^k v_j X_j$ fluctuates on a constant scale. By using (one-dimensional) Gaussian approximation on $\sum_{j>k} u_j^* X_j$, conditioning and independence with $\sum_{j=1}^k u_j^* X_j, \sum_{j=1}^k v_j X_j$, and the fact that $|\sum_{j=1}^k u_j^* X_j| \lesssim \sqrt{k/d}$ with high probability, one may show that condition (31) holds with $\eta \asymp \sqrt{k/d}$.

Taking the best of the two guarantees above (depending on the value of $k = k(v)$), condition (31) holds down to $\eta \asymp \min(\sqrt{k/d}, 1/\sqrt{k}) \leqslant d^{-1/4}$ for any $v \in S^{d-1}$.

9.4 Improved regularity scales in generic directions?

We now discuss the phenomenon alluded to in Section 3.3, namely that Assumption 2 holds down to a scale of $1/d$ in "typical" directions $u^* \in S^{d-1}$. This is a consequence of the following result of Klartag and Sodin [KS12], which states that for a "typical" vector $u = (u_1, \dots, u_d) \in$

S^{d-1}, if $X = (X_1, \dots, X_d)$ has i.i.d. coordinates then the distribution of the linear combination $\langle u, X\rangle = \sum_{j=1}^d u_j X_j$ approaches the Gaussian distribution at a rate of order $1/d$. This rate is faster than the usual $1/\sqrt{d}$ rate from the Berry-Esseen theorem for the usual normalized sum $\langle u_d^*, X\rangle = \frac{1}{\sqrt{d}} \sum_{j=1}^d X_j$.

Theorem (Theorem 1.1 in [KS12])**.** *There exists a constant $c \geqslant 1$ such that the following holds. Let $\varepsilon \in (0, 1/2)$ and $d \geqslant 1$. Assume that $X = (X_1, \dots, X_d)$ has independent coordinates, with $\mathbb{E}[X_j] = 0$ and $\mathbb{E}[X_j^2] = 1$ for $j = 1, \dots, d$ and with finite fourth moment. Let*

$$\kappa = \left(\frac{1}{d} \sum_{j=1}^d \mathbb{E}[X_j^4] \right)^{1/4}.$$

Then, there is a subset $A_\varepsilon \subset S^{d-1}$ with $\mu_{d-1}(A_\varepsilon) \geqslant 1 - \varepsilon$ (where μ_{d-1} stands for the uniform probability measure on S^{d-1}) such that, for any $u \in A_\varepsilon$, one has

$$\sup_{a,b \in \mathbb{R},\, a \leqslant b} \left| \mathbb{P}(a \leqslant \langle u, X\rangle \leqslant b) - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-s^2/2} \mathrm{d}s \right| \leqslant \frac{c \log^2(1/\varepsilon) \kappa^4}{d}. \tag{176}$$

This immediately implies that there exists a subset A of S^{d-1} with $\mu_{d-1}(A) \geqslant 1 - 1/d \to_{d \to \infty} 1$ such that, for any $u \in A$, the margin probability $\mathbb{P}(|\langle u, X\rangle| \leqslant t)$ is of order t as long as $t \gtrsim \log^2(d)/d$ (hence, Assumption 2 holds at least down to $\eta \asymp \log^2(d)/d$).

The reason why a "generic" direction $u \in S^{d-1}$ leads to a faster rate of Gaussian approximation (and therefore a smaller scale η for Assumption 2) than $u_d^* = (1/\sqrt{d}, \dots, 1/\sqrt{d})$ is the following. For the parameter u_d^*, the rate of Gaussian approximation of order $1/\sqrt{d}$ cannot be improved due to an arithmetic obstruction: if $X_1, \dots, X_d$ are i.i.d. Bernoulli, the quantity $\langle u_d^*, X\rangle = \frac{1}{\sqrt{d}} \sum_{j=1}^d X_j$ takes values in the lattice $\mathbb{Z}/\sqrt{d}$. This is due to the strong additive structure of u_d^*, all of whose coefficients are equal: hence, there are many cancellations in the sum $\langle u_d^*, X\rangle = \frac{1}{\sqrt{d}} \sum_{j=1}^d X_j$, as any two opposite signs X_j, X_k cancel out. This means that many different values of the vector X lead to the same value of $\langle u_d^*, X\rangle$. However, this arithmetic obstruction vanishes for a "generic" direction $u = (u_1, \dots, u_d) \in S^{d-1}$, which is much less structured (for instance, all ratios u_j/u_k with $j \neq k$ are irrational numbers with probability 1).

These results suggest that, for a generic parameter direction $u^* \in S^{d-1}$, the regularity conditions (Definition 1) may hold at a scale $\eta_d \ll 1/\sqrt{d}$. However, we do not know how to prove this for the *two-dimensional* margin Assumption 3. Indeed, as previously discussed, a key difficulty is that the property (23) must be established for *every* direction $v \in S^{d-1}$, including those v for which Gaussian approximation fails. In addition, it is not clear how to extend our arguments in Lemma 2 from the case of $u^* = u_d^*$ to a generic $u^* \in S^{d-1}$ lacking additive structure, while incorporating the $1/d$ improvement of [KS12] in this case. We therefore leave this question as an open problem:

Problem 1. Does there exist a sequence $(\eta_d)_{d \geqslant 1}$ with $\sqrt{d} \cdot \eta_d \to 0$ as $d \to \infty$ such that the following holds? Let $X = (X_1, \dots, X_d)$ be a random vector with i.i.d. sub-exponential coordinates (Assumption 4 with $K \lesssim 1$), for instance a Bernoulli design. There exists a subset $A_d \subset S^{d-1}$ with $\mu_{d-1}(A_d) \to 1$ as $d \to \infty$, such that for every $u^* \in S^{d-1}$, the distribution X satisfies Assumption 3 with parameter u^*, η_d and $c \lesssim 1$.

In addition, does $\eta_d = 1/d$ satisfy this property? And what is the smallest order of magnitude of η_d such that this property holds?

In short, Problem 1 asks about the regularity scale of product measures (such as the Bernoulli design) in "typical" directions. By Theorem 3 and Proposition 1, this amounts to investigating the values of the parameter norm (for typical parameter directions) for which the MLE for logistic regression behaves as in the case of a Gaussian design.

A Tail conditions on real random variables

In this section, we gather some definitions and basic properties regarding tails of real valued random variables. These are well-known that are simply recalled here to fix the constants. We start with the definition of the sub-exponential and sub-Gaussian norms:

Definition 5 (ψ_α-norm)**.** Let $\alpha > 0$. If X is a real random variable, its *ψ_α-norm* is defined as

$$\|X\|_{\psi_\alpha} = \sup_{p \geqslant 2} \left[\frac{2^{1/\alpha} e \|X\|_p}{p^{1/\alpha}} \right] \in [0, +\infty] \,, \tag{177}$$

where the supremum is taken over all real values of $p \geqslant 2$. We say that X is *sub-exponential* if $\|X\|_{\psi_1} < +\infty$, and *sub-Gaussian* if $\|X\|_{\psi_2} < +\infty$.

We mostly consider the cases $\alpha = 1$ and $\alpha = 2$. We refer to [Ver18, §2.5 and §2.7] for equivalent definitions of the ψ_1 and ψ_2-norms.

Note that the normalization in the definition (177) ensures that (i) if $\mathbb{E}[X^2] = 1$, then $\|X\|_{\psi_\alpha} \geqslant e$ and (ii) if $\alpha \leqslant \alpha'$, then $\|X\|_{\psi_\alpha} \leqslant \|X\|_{\psi_{\alpha'}}$. In addition, one has $\|X + X'\|_{\psi_\alpha} \leqslant \|X\|_{\psi_\alpha} + \|X'\|_{\psi_\alpha}$ for every real valued random variables X, X' and every parameter $\alpha > 0$.

In order to obtain sharp guarantees, we need the additional notion of *sub-gamma* random variables [BLM13, §2.4].

Definition 6 (Sub-gamma random variables)**.** Let X be a real valued random variable and $\sigma, K > 0$. We say that X is *(σ^2, K)-sub-gamma* if for every $\lambda \in [0, 1/K)$ one has

$$\mathbb{E} \exp(\lambda X) \leqslant \exp\left(\frac{\sigma^2 \lambda^2}{2(1 - \lambda K)} \right) . \tag{178}$$

Recall that X is said to be centered if $\mathbb{E}[X] = 0$. The basic properties of sub-gamma and sub-exponential variables are gathered in the following lemma:

Lemma 35. *Let X be a real random variable and $\sigma, K > 0$.*

1. *If $\|X\|_{\psi_\alpha} \leqslant K$, then for every $t \geqslant 1$ one has*

$$\mathbb{P}\big(|X| \geqslant K t^{1/\alpha}\big) \leqslant e^{-2t} . \tag{179}$$

2. *If X is (σ^2, K)-sub-gamma, then for every $t \geqslant 0$, one has*

$$\mathbb{P}(X \geqslant \sigma\sqrt{2t} + Kt) \leqslant e^{-t} . \tag{180}$$

3. *If $X_1, \dots, X_n$ are independent random variables such that X_i is (σ_i^2, K_i)-sub-gamma (with $\sigma_i, K_i > 0$) for every $i = 1, \dots, n$, then $X_1 + \cdots + X_n$ is $(\sigma_1^2 + \cdots + \sigma_n^2, \max(K_1, \dots, K_n))$-sub-gamma. Also, if X is (σ^2, K)-sub-gamma and $\alpha \geqslant 0$, then αX is $(\alpha^2\sigma^2, \alpha K)$-sub-gamma.*

4. *If X is centered and satisfies for every integer $p \geqslant 2$ that*

$$\mathbb{E}[|X|^p] \leqslant \sigma^2 K^{p-2} p!/2 \,, \tag{181}$$

then X is (σ^2, K)-sub-gamma.

5. *If X and $-X$ are (σ^2, K)-sub-gamma, then* $\mathrm{Var}(X) \leqslant \sigma^2$ *and* $\|X\|_{\psi_1} \leqslant 2\sqrt[3]{2e} \max(\sigma, 2K)$.

6. *If X is centered,* $\mathrm{Var}(X) \leqslant \sigma^2$ *and* $\|X\|_{\psi_1} \leqslant K$ *(where* $K \geqslant e\sigma$*), then X is* $(\sigma^2, K\log(K/\sigma))$*-sub-gamma. In addition, X is* $(K^2/2, K/2)$*-sub-gamma.*

In particular, it follows from the last two points of Lemma 35 that if X is centered and $K \geqslant e\sigma$, the property that X (and $-X$) is (σ^2, K)-sub-gamma is closely related to the conditions $\mathrm{Var}(X) \leqslant \sigma^2$ and $\|X\|_{\psi_1} \leqslant K$. The sub-gamma condition is however slightly stronger, and allows one to gain a factor of order $\log(K/\sigma)$. We actually use this improvement in order to avoid additional $\log B$ factors in the setting of Theorem 1.

Proof. For the first point, for any $p \geqslant 2$, Markov's inequality implies that

$$\mathbb{P}(|X| \geqslant e\|X\|_p) = \mathbb{P}\big(|X|^p \geqslant e^p\|X\|_p^p\big) \leqslant \frac{\mathbb{E}[|X|^p]}{e^p\|X\|_p^p} = e^{-p} \,.$$

Letting $p = 2t$ and bounding $e\|X\|_p \leqslant \|X\|_{\psi_\alpha}(p/2)^{1/\alpha} \leqslant Kt^{1/\alpha}$ concludes.

The second point is established in [BLM13, p. 29] using the Chernoff method, namely bounding $\mathbb{P}(X \geqslant t) \leqslant e^{-\lambda t}\mathbb{E}e^{\lambda X}$ and optimizing over $\lambda \geqslant 0$.

The third point follows from the definition and the fact that, by independence,

$$\mathbb{E}\big[e^{\lambda(X_1+\cdots+X_n)}\big] = \mathbb{E}[e^{\lambda X_1}]\cdots\mathbb{E}[e^{\lambda X_n}] \,.$$

We now turn to the fourth point. For every $\lambda \in [0, 1/K)$, one has

$$\begin{aligned}\mathbb{E}e^{\lambda X} &\leqslant 1 + \lambda\mathbb{E}[X] + \sum_{p\geqslant 2}\frac{\lambda^p\,\mathbb{E}[|X|^p]}{p!} \leqslant 1 + \frac{\sigma^2\lambda^2}{2}\sum_{p\geqslant 2}\frac{\lambda^{p-2}K^{p-2}p!}{p!}\\ &= 1 + \frac{\sigma^2\lambda^2}{2(1-\lambda K)} \leqslant \exp\left(\frac{\sigma^2\lambda^2}{2(1-\lambda K)}\right).\end{aligned}$$

For the fifth point, we first note that, as $\mathbb{E}[e^{|X|/(2K)}] \leqslant \mathbb{E}[e^{X/(2K)}] + \mathbb{E}[e^{-X/(2K)}] < \infty$, by dominated convergence the function $\phi : \lambda \mapsto \log\mathbb{E}[e^{\lambda X}]$ is well-defined and twice continuously differentiable over $(-1/(2K), 1/(2K))$, with $\phi(0) = 0$, $\phi'(0) = \mathbb{E}[X]$ and $\phi''(0) = \mathrm{Var}(X)$. Hence, $\phi(\lambda) = \mathbb{E}[X]\lambda + \mathrm{Var}(X)\lambda^2/2 + o(\lambda^2)$ as $\lambda \to 0$, and by assumption one has $\phi(\lambda) \leqslant \frac{\sigma^2\lambda^2}{2(1-\lambda K)} = \sigma^2\lambda^2/2 + o(\lambda^2)$, hence $\mathbb{E}[X] = 0$ and $\mathrm{Var}(X) \leqslant \sigma^2$. Next, in order to bound $\|X\|_{\psi_1}$, we apply the sub-gamma condition (178) to $\lambda = 1/(\sigma \vee 2K)$, which gives:

$$\mathbb{E}\left[\exp\left(\frac{|X|}{\sigma\vee 2K}\right)\mathbf{1}(X \geqslant 0)\right] \leqslant \mathbb{E}\left[\exp\left(\frac{X}{\sigma\vee 2K}\right)\right] \leqslant \mathbb{E}\left[\exp\left(\frac{\sigma^2/\sigma^2}{2(1-K/(2K))}\right)\right] = e \,.$$

Applying the same inequality to $-X$ and summing gives:

$$\mathbb{E}\left[\exp\left(\frac{|X|}{\sigma\vee 2K}\right)\right] \leqslant 2e \,.$$

Now, a simple analysis of function shows that $e^u - eu \geqslant 0$ for any $u \geqslant 0$, hence (applying this to u/p) $\left(\frac{eu}{p}\right)^p \leqslant e^u$. Hence, for any $p \geqslant 3$, one has

$$\mathbb{E}\left[\left(\frac{e|X|}{p(\sigma\vee 2K)}\right)^p\right] \leqslant \mathbb{E}\left[\exp\left(\frac{|X|}{\sigma\vee 2K}\right)\right] \leqslant 2e \,,$$

so that $2e\|X\|_p/p \leqslant 2(2e)^{1/p}(\sigma\vee 2K) \leqslant 2\sqrt[3]{2e}(\sigma\vee 2K)$, which proves the desired bound since we also have $2e\|X\|_2/2 \leqslant e\sigma \leqslant 2\sqrt[3]{2e}\sigma$.

Let us now establish the sixth point. For every $p > 2$, one has for $r > 1$, using Hölder's inequality:

$$\begin{aligned}
\mathbb{E}[|X|^p] &= \mathbb{E}[|X|^{2(1-1/r)}|X|^{p-2+2/r}] \\
&\leqslant \mathbb{E}[X^2]^{1-1/r}\mathbb{E}[X^{(p-2)r+2}]^{1/r} \\
&\leqslant \sigma^{2-2/r}\|X\|_{(p-2)r+2}^{[(p-2)r+2]/r} \\
&\leqslant \sigma^{2-2/r}\Big[\frac{[(p-2)r+2]K}{2e}\Big]^{p-2+2/r} \\
&= \sigma^2\Big(\frac{Kr}{2}\Big)^{p-2}\Big(\frac{r}{2}\Big)^{2/r}\Big(\frac{K}{\sigma}\Big)^{2/r}\Big(\frac{p-2+2/r}{e}\Big)^{p-2+2/r}.
\end{aligned}$$

Now let $r/2 = \log(K/\sigma) \geqslant 1$, so that $(K/\sigma)^{2/r} = e$. A direct analysis shows that the function $u \mapsto (u/e)^u$ increases on $[1, +\infty)$, and since $r/2 \geqslant 1$ one has $1 \leqslant p-2 \leqslant p-2+2/r \leqslant p-1$. Hence, for any integer $p \geqslant 3$,

$$\Big(\frac{p-2+2/r}{e}\Big)^{p-2+2/r} \leqslant \Big(\frac{p-1}{e}\Big)^{p-1} \leqslant (2\pi(p-1))^{-1/2}(p-1)! \leqslant p!/(6\sqrt{\pi}),$$

where we used the standard Stirling-type inequalities

$$\sqrt{2\pi p}\Big(\frac{p}{e}\Big)^p \leqslant p! \leqslant p^p. \tag{182}$$

In addition $t^{1/t} \leqslant e^{1/e}$ for $t > 0$, so $(r/2)^{2/r} \leqslant e^{1/e}$. Combining the previous inequalities, we obtain

$$\begin{aligned}
\mathbb{E}[|X|^p] &\leqslant \sigma^2\big(K\log(K/\sigma)\big)^{p-2}e^{1+1/e}p!/(6\sqrt{\pi}) \\
&\leqslant \sigma^2\Big(K\log\Big(\frac{K}{\sigma}\Big)\Big)^{p-2}p!/2,
\end{aligned} \tag{183}$$

where we used that $e^{1+1/e}/(3\sqrt{\pi}) = 0.738\ldots \leqslant 1$. By the fourth point, this implies that X is $(\sigma^2, K\log(K/\sigma))$-sub-gamma. For the last statement, using the inequality $(\frac{p}{e})^p \leqslant p!$ for $p \geqslant 2$, we obtain

$$\mathbb{E}[|X|^p] \leqslant \Big(\frac{Kp}{2e}\Big)^p \leqslant \Big(\frac{K}{2}\Big)^p p! = \frac{1}{2}\frac{K^2}{2}\Big(\frac{K}{2}\Big)^{p-2}p!, \tag{184}$$

so by the fourth point X is $(K^2/2, K/2)$-sub-gamma. □

Finally, we will also use the following consequence of Bennett's inequality, which shows that bounded variables are sub-gamma.

Lemma 36. *Let X be a random variable such that $\mathbb{E}[X^2] \leqslant \sigma^2$ and $X \leqslant b$ almost surely, for some $\sigma^2 > 0$ and $b > 0$. Then*

1. *$X - \mathbb{E}[X]$ is $(\sigma^2, 3b)$-sub-gamma.*
2. *For all $\lambda \in [0, b^{-1}]$, $\log \mathbb{E}e^{\lambda X} \leqslant \lambda\mathbb{E}[X] + \sigma^2/b^2$.*

Proof. By homogeneity we assume that $b = 1$. Let $X' = X - \mathbb{E}[X]$. Using Bennett's inequality [BLM13, Theorem 2.9], one has, for all $\lambda > 0$,

$$\log \mathbb{E}e^{\lambda X'} \leqslant \sigma^2\phi(\lambda), \quad \phi(\lambda) = e^\lambda - \lambda - 1. \tag{185}$$

Moreover, for every $\lambda \in [0, 1/3)$

$$\phi(\lambda) = \sum_{k\geqslant 2}\frac{\lambda^k}{k!} = \frac{\lambda^2}{2}\sum_{k\geqslant 0}\frac{\lambda^k}{(k+2)!/2} \leqslant \frac{\lambda^2}{2}\sum_{k\geqslant 0}\frac{\lambda^k}{3^k} = \frac{\lambda^2}{2(1-3\lambda)},$$

where we used that $(k+2)!/2 = \prod_{j=3}^{k+2} j \geqslant 3^k$ for $k \geqslant 1$. The first point is proved. For the second point, we start from (185) and use that $\phi(\lambda) \leqslant \phi(1) = e - 2 \leqslant 1$ for all $\lambda \in [0, 1]$. □

B Polar coordinates and spherical caps

B.1 Polar coordinates

Depending on the situation, it may be more convenient to express the position of θ relative to θ^* (with direction $u^* = \theta^*/\|\theta^*\|$) in either of the following two equivalent ways: (1) in terms of the component $\langle u^*, \theta\rangle$ parallel to u^* and of the orthogonal component $\theta - \langle u^*, \theta\rangle u^*$; or (2) in terms of the norm $\|\theta\|$ and of the direction $u = \theta/\|\theta\|$. The following lemma gathers inequalities relating the two representations.

Lemma 37. *Let $\theta, \theta^* \in \mathbb{R}^d$, and set $u = \theta/\|\theta\|$, $u^* = \theta^*/\|\theta^*\| \in S^{d-1}$ and $\theta_\perp = \theta - \langle u^*, \theta\rangle u^*$.*

1. *If $\langle u^*, u\rangle \geqslant 0$, then*
$$\frac{\|u - u^*\|}{\sqrt{2}} \leqslant \frac{\|\theta_\perp\|}{\|\theta\|} = \sqrt{1 - \langle u, u^*\rangle^2} \leqslant \|u - u^*\| \,. \tag{186}$$

2. *If $\|u - u^*\| \leqslant 1$, then $\|\theta\|/2 \leqslant \langle u^*, \theta\rangle \leqslant \|\theta\|$.*

3. *One has*
$$|\langle u^*, \theta - \theta^*\rangle| \leqslant \big|\|\theta\| - \|\theta^*\|\big| + \|\theta^*\| \cdot \frac{\|u - u^*\|^2}{2} \,. \tag{187}$$

4. *One has*
$$\big|\|\theta\| - \|\theta^*\|\big| \leqslant |\langle u^*, \theta - \theta^*\rangle| + \frac{\|\theta_\perp\|^2}{\|\theta\| + \|\theta^*\|} \,. \tag{188}$$

Proof. We start with the first point. By orthogonality,
$$\begin{aligned}\|\theta_\perp\|^2 &= \|\theta\|^2 - \langle u^*, \theta\rangle^2 = \|\theta\|^2\big[1 - \langle u^*, u\rangle^2\big] \\ &= \|\theta\|^2\big[1 - \langle u^*, u\rangle\big]\big[1 + \langle u^*, u\rangle\big] = \frac{1}{2}\|\theta\|^2\|u - u^*\|^2\big[1 + \langle u^*, u\rangle\big] \,.\end{aligned} \tag{189}$$

Hence, if $\langle u^*, u\rangle \geqslant 0$, then
$$\frac{1}{2}\|\theta\|^2\|u - u^*\|^2 \leqslant \|\theta_\perp\|^2 \leqslant \|\theta\|^2\|u - u^*\|^2 \,,$$
which together with the identity (189) proves the first claim. The second point follows from the fact that
$$\frac{\langle u^*, \theta\rangle}{\|\theta\|} = \langle u, u^*\rangle = 1 - \frac{1}{2}\|u - u^*\|^2 \in \Big[\frac{1}{2}, 1\Big] \,.$$

We now turn to the third point. Since $\langle u^*, \theta^*\rangle = \|\theta^*\|$, we have
$$\begin{aligned}|\langle u^*, \theta - \theta^*\rangle| = \big|\|\theta\|\langle u^*, u\rangle - \|\theta^*\|\big| &\leqslant \big|(\|\theta\| - \|\theta^*\|)\langle u^*, u\rangle\big| + \big|\|\theta^*\|(\langle u^*, u\rangle - 1)\big| \\ &\leqslant \big|\|\theta\| - \|\theta^*\|\big| + \|\theta^*\| \cdot \frac{\|u - u^*\|^2}{2} \,,\end{aligned}$$
where we used that $\|u - u^*\|^2 = 2(1 - \langle u, u^*\rangle)$. For the fourth point, note that
$$\begin{aligned}\big(\|\theta\| + \|\theta^*\|\big)\big|\|\theta\| - \|\theta^*\|\big| &= \big|\|\theta\|^2 - \|\theta^*\|^2\big| \\ &= \big|\|\theta_\perp\|^2 + \langle u^*, \theta\rangle^2 - \langle u^*, \theta^*\rangle^2\big| \\ &\leqslant \|\theta_\perp\|^2 + |\langle u^*, \theta + \theta^*\rangle| \cdot |\langle u^*, \theta - \theta^*\rangle| \\ &\leqslant \|\theta_\perp\|^2 + (\|\theta\| + \|\theta^*\|) \cdot |\langle u^*, \theta - \theta^*\rangle| \,;\end{aligned}$$
dividing by $\|\theta\| + \|\theta^*\|$ gives the claimed inequality. □

B.2 Spherical caps

In Section 6.3, we defined spherical caps through their angles as

$$\mathsf{C}(u,\varepsilon) = \left\{v \in S^{d-1},\ \langle u,v\rangle \geqslant 0,\ |\sin(u,v)| \leqslant \varepsilon\right\}, \tag{190}$$

for any $u \in S^{d-1}$ and $\varepsilon \in [0,1]$, where (u,v) denotes the angle between two unit vectors, that is $(u,v) = \arccos(\langle u,v\rangle)$. Spherical caps can be equivalently defined using the Euclidean distance by

$$\widetilde{\mathsf{C}}(u,r) = \left\{v \in S^{d-1},\ \|u-v\| \leqslant r\right\}. \tag{191}$$

The following result provides a formal statement of this equivalence.

Fact 6. *For every $\varepsilon \in [0,1]$, $\mathsf{C}(u,\varepsilon) = \widetilde{\mathsf{C}}(u,r_\varepsilon)$, where $r_\varepsilon = \sqrt{2\big(1-\sqrt{1-\varepsilon^2}\big)}$. Moreover, it holds that $\varepsilon \leqslant r_\varepsilon \leqslant \sqrt{2}\varepsilon$ and*

$$\mathsf{C}(u,\varepsilon/\sqrt{2}) \subset \widetilde{\mathsf{C}}(u,\varepsilon) \subset \mathsf{C}(u,\varepsilon). \tag{192}$$

Proof. This simply follows from the fact that for any two vectors u,v on the unit sphere, denoting by ϕ the angle between them, one has

$$\|u-v\|^2 = 2\big(1-\langle u,v\rangle\big) = 2\big(1-\cos\phi\big) = 2\Big(1-\sqrt{1-\sin^2\phi}\Big).$$

In addition, by concavity, for all $t \in [0,1]$, $1-t \leqslant \sqrt{1-t} \leqslant 1-t/2$. Finally, (192) follows from the first point of Lemma 37. □

C Proof of Proposition 1

In this section we provide the proof of Proposition 1 from Section 2.2, regarding the necessity of the two-dimensional margin assumption.

We start with the following simple fact.

Fact 7. *Let Z and Z' be real sub-exponential variables with $\|Z\|_{\psi_1} \leqslant K$, $\|Z'\|_{\psi_1} \leqslant K'$, for some $K, K' \geqslant e$. Then for all $\lambda \geqslant K'$, $\mathbb{E}[Z^2\mathbf{1}(Z' \geqslant \lambda)] \leqslant K^2 e^{-\lambda/K'}$.*

Proof. By the Cauchy-Schwarz inequality, the first point of Lemma 35 and Definition 5,

$$\mathbb{E}\big[Z^2\mathbf{1}(Z' \geqslant \lambda)\big] \leqslant \sqrt{\mathbb{E}Z^4}\sqrt{\mathbb{P}(Z' \geqslant \lambda)} \leqslant \sqrt{\left(\frac{4K}{2e}\right)^4}\sqrt{e^{-2\lambda/K'}} \leqslant (2/e)^2 K^2 e^{-\lambda/K'}. \qquad \square$$

Proof of Proposition 1. For every $v \in S^{d-1}$ and $c_0 \geqslant 1$,

$$\begin{aligned}
\langle H_X(\theta^*)v,v\rangle &= \mathbb{E}\big[\sigma'(B\langle u^*,X\rangle)\langle v,X\rangle^2\big] \\
&= \mathbb{E}\Big[\sigma'(B\langle u^*,X\rangle)\mathbf{1}\Big(|\langle u^*,X\rangle| > \frac{c_0\log B}{B}\Big)\langle v,X\rangle^2\Big] \\
&\quad + \mathbb{E}\Big[\sigma'(B\langle u^*,X\rangle)\mathbf{1}\Big(|\langle u^*,X\rangle| \leqslant \frac{c_0\log B}{B}\Big)\langle v,X\rangle^2\Big] \\
&\leqslant \frac{1}{B^{c_0}} + \mathbb{E}\Big[\mathbf{1}\Big(|\langle u^*,X\rangle| \leqslant \frac{c_0\log B}{B}\Big)\langle v,X\rangle^2\Big].
\end{aligned}$$

In view of Remark 1, we furthermore let $m = \max\left\{B^{-1}, \|u^* - v\|\right\}$. Conditioning on the value of $|\langle v, X\rangle|$, for every $C \geqslant 1$, the expectation in the second term above rewrites

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B}\Big)\langle v, X\rangle^2\Big] \\
&\quad = \mathbb{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| < \frac{m}{C}\Big)\langle v, X\rangle^2\Big] \\
&\quad + \mathbb{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| \geqslant \frac{m}{C}\Big)\langle v, X\rangle^2\Big]\,.
\end{aligned} \tag{193}$$

Regarding the first term, we find using Assumption 2 that

$$\begin{aligned}
\mathbb{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| < \frac{m}{C}\Big)\langle v, X\rangle^2\Big] &\leqslant \frac{m^2}{C^2}\mathbb{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B}\Big) \\
&\leqslant \frac{m^2}{B} \cdot \frac{c\, c_0 \log B}{C^2}\,.
\end{aligned}$$

Finally, we further decompose the second term in (193). For all $\lambda \geqslant m/C$,

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{1}\Big\{|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| \geqslant \frac{m}{C}\Big\}\langle v, X\rangle^2\Big] \\
&\quad = \mathbb{E}\Big[\mathbf{1}\Big\{|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| \in \Big[\frac{m}{C}, \lambda\Big)\Big\}\langle v, X\rangle^2\Big] \\
&\quad + \mathbb{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| \geqslant \lambda\Big)\langle v, X\rangle^2\Big] \\
&\quad \leqslant \lambda^2\, \mathbb{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| \geqslant \frac{m}{C}\Big) \\
&\quad + \mathbb{E}\big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B}\Big)\mathbf{1}\big(|\langle v, X\rangle| \geqslant \lambda\big)\langle v, X\rangle^2\big]\,.
\end{aligned} \tag{194}$$

We now bound the last term using Fact 7. From now on, we let $\eta = B^{-1}$ and $c = c_0 \log B$. Note that by the triangle inequality, on the event $\{|\langle u^*, X\rangle| \leqslant c\eta\}$, letting $w = (u^* - v)/\|u^* - v\| \in S^{d-1}$,

$$|\langle v, X\rangle| \leqslant |\langle u^*, X\rangle| + \|u^* - v\| \cdot |\langle w, X\rangle| \leqslant c\eta + m|\langle w, X\rangle|\,.$$

Hence,

$$\mathbf{1}(|\langle u^*, X\rangle| \leqslant c\eta)\mathbf{1}(|\langle v, X\rangle| \geqslant \lambda) \leqslant \mathbf{1}(m|\langle w, X\rangle| \geqslant \lambda - c\eta)\,.$$

Therefore, by Fact 7, for all $\lambda \geqslant mK + c\eta$,

$$\begin{aligned}
\mathbb{E}\big[\mathbf{1}(|\langle u^*, X\rangle| \leqslant c\eta)\mathbf{1}(|\langle v, X\rangle| \geqslant \lambda)\langle v, X\rangle^2\big] &\leqslant \mathbb{E}\big[\mathbf{1}(m|\langle w, X\rangle| \geqslant \lambda - c\eta)\langle v, X\rangle^2\big] \\
&\leqslant \frac{4}{e^2}K^2 \exp\Big(-\frac{\lambda - c\eta}{mK}\Big)\,.
\end{aligned}$$

In particular, letting $\lambda = 2mK(3\log(KB) + \log C_0)$, it holds that $\lambda - c\eta \geqslant \lambda/2 \geqslant mK$ and

$$\exp\Big(-\frac{\lambda - c\eta}{mK}\Big) \leqslant \exp\Big(-\frac{\lambda}{2mK}\Big) = \frac{1}{C_0 K^3 B^3}\,,$$

from which we deduce that the second term in (194) can be bounded as

$$\mathbb{E}\big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B}\Big)\mathbf{1}\big(|\langle v, X\rangle| \geqslant \lambda\big)\langle v, X\rangle^2\big] \leqslant \frac{4}{e^2} \cdot \frac{1}{C_0 K B^3} \leqslant \frac{4}{e^3} \cdot \frac{m^2}{C_0 B}\,, \tag{195}$$

since $B^{-3} \leqslant m^2/B$.

Putting everything together, we find that

$$\frac{m^2}{C_0 B} \leqslant \mathbb{E}\big[\sigma'(B\langle u^*, X\rangle)\langle v, X\rangle^2\big]$$
$$\leqslant \frac{1}{B^{c_0}} + \frac{m^2}{B}\cdot\frac{c\,c_0\log B}{C^2} + \lambda^2\,\mathbb{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0\log B}{B};\, |\langle v, X\rangle| \geqslant \frac{m}{C}\Big) + \frac{4}{e^3}\cdot\frac{m^2}{B}\,. \tag{196}$$

We now choose the values of the parameters c_0 and C in such a way that in the inequality above, the three terms which do not involve the probability describing the margin condition add up to at most $3m^2/(4C_0B)$. First, we set $c_0 = 3+\log(4C_0)$ so that $B^{-c_0} \leqslant B^{-3}/(4C_0) \leqslant m^2/(4C_0B)$. Next we let $C = 2\sqrt{c_0 C_0 c\log B}$, and finally in (195) we further bound $4/e^3 \leqslant 1/4$. Rearranging the terms in (196) yields

$$\mathbb{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0\log B}{B};\, |\langle v, X\rangle| \geqslant \frac{m}{C}\Big) \geqslant \frac{m^2}{4C_0\lambda^2 B}\,.$$

the result follows by further bounding λ^2. □

Acknowledgements. This research is supported by a grant of the French National Research Agency (ANR), “Investissements d’Avenir” (LabEx Ecodec/ANR-11-LABX-0047).

References

[AA84] Adelin Albert and John A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[AC11] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 2011.

[ACL19] Pierre Alquier, Vincent Cottet, and Guillaume Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *Ann. Statist.*, 47(4):2117–2144, 2019.

[ALMT14] Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference*, 3(3):224–294, 2014.

[AT07] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007.

[Bac10] Francis Bach. Self-concordant analysis for logistic regression. *Electron. J. Stat.*, 4:384–414, 2010.

[Bac14] Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.*, 15(1):595–627, 2014.

[Bac24] Francis Bach. *Learning Theory from First Principles.* MIT Press, 2024.

[BBL05] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.

[BC24] Yannick Baraud and Juntong Chen. Robust estimation of a regression function in exponential families. *J. Statist. Plann. Inference*, 233:106167, 2024.

[Ber44] Joseph Berkson. Application of the logistic function to bio-assay. *J. Amer. Statist. Assoc.*, 39(227):357–365, 1944.

[BJM06] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101(473):138–156, 2006.

[BL19] Sergey G. Bobkov and Michel Ledoux. One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Mem. Amer. Math. Soc.*, 2019.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, 2013.

[BM06] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.

[BM13] Francis Bach and Éric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems 26*, pages 773–781, 2013.

[Bou02] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[Bro86] Lawrence D. Brown. *Fundamentals of statistical exponential families with applications in statistical decision theory*, volume 9 of *IMS Lecture Notes - Monograph Series.* Institute of Mathematical Statistics, 1986.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine learning*, 8(3-4):231–357, 2015.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[Cat07] Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning.* Institute of Mathematical Statistics, 2007.

[Cat16] Olivier Catoni. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *Preprint arXiv:1603.05229*, 2016.

[CBL06] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006.

[CLL20] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with Lipschitz and convex loss functions. *Probab. Theory Related Fields*, 176:897–940, 2020.

[Cov65] Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965.

[CS20] Emmanuel J. Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.*, 48(1):27–42, 2020.

[DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer-Verlag, 1996.

[DVW21] Joseph De Vilmarest and Olivier Wintenberger. Stochastic online optimization using Kalman recursion. *J. Mach. Learn. Res.*, 22(223):1–55, 2021.

[EK18a] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Related Fields*, 170(1):95–175, 2018.

[EK18b] Noureddine El Karoui. Random matrices and high-dimensional statistics: beyond covariance matrices. In *Proc. International Congress of Mathematicians*, volume 4, pages 2875–2894, Rio de Janeiro, 2018.

[Fel68] William Feller. On the Berry-Esseen theorem. *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, 10(3):261–268, 1968.

[FKL+18] Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: the importance of being improper. In *Proc. 31st Conference on Learning Theory*, pages 167–208, 2018.

[GNT14] Olivier Guédon, Piotr Nayar, and Tomasz Tkocz. Concentration inequalities and geometry of convex bodies. *Analytical and probabilistic methods in the geometry of convex bodies*, 2:9–86, 2014.

[HKL14] Elad Hazan, Tomer Koren, and Kfir Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Proc. 27th Conference on Learning Theory*, pages 197–209, 2014.

[HM23] Daniel Hsu and Arya Mazumdar. On the sample complexity of estimation in logistic regression. *arXiv preprint arXiv:2307.04191*, 2023.

[JGR20] Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. In *Proc. 33rd Conference on Learning Theory*, pages 2085–2108, 2020.

[KN05] Sham M. Kakade and Andrew Y. Ng. Online bounds for Bayesian algorithms. In *Advances in Neural Information Processing Systems 17*, pages 641–648, 2005.

[Kol11] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *École d'Été de Probabilités de Saint-Flour*. Springer-Verlag Berlin Heidelberg, 2011.

[KS12] Bo'az Klartag and Sasha Sodin. Variations on the Berry–Esseen theorem. *Theory Probab. Appl.*, 56(3):403–419, 2012.

[Kuc24] Felix Kuchelmeister. On the probability of linear separability through intrinsic volumes. *arXiv preprint arXiv:2404.12889*, 2024.

[KvdG23] Felix Kuchelmeister and Sara van de Geer. Finite sample rates for logistic regression with small noise or few samples. *arXiv preprint arXiv:2305.15991*, 2023.

[Lat97] Rafał Latała. Estimation of moments of sums of independent real random variables. *Ann. Probab.*, 25(3):1502–1513, 1997.

[LCY00] Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts.* Springer Series in Statistics. Springer-Verlag New York, 2000.

[Mas07] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics.* Springer Berlin Heidelberg, 2007.

[Meh17] Nishant Mehta. Fast rates with high probability in exp-concave statistical learning. In *Proc. 20th International Conference on Artificial Intelligence and Statistics*, pages 1085–1093, 2017.

[MG22] Jaouad Mourtada and Stéphane Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *J. Mach. Learn. Res.*, 23(31):1–49, 2022.

[MN89] Peter McCullagh and John A. Nelder. *Generalized Linear Models.* Chapman and Hall/CRC, 2 edition, 1989.

[Mon18] Andrea Montanari. Mean field asymptotics in high-dimensional statistics: From exact results to efficient algorithms. In *Proc. International Congress of Mathematicians*, pages 2973–2994, 2018.

[Mou22] Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *Ann. Statist.*, 50(4):2157–2178, 2022.

[MT99] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

[OB21] Dmitrii Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. *Electron. J. Stat.*, 15(1):326–391, 2021.

[OLBC10] Frank W. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark. *NIST Handbook of Mathematical Functions.* Cambridge University Press, 2010.

[Oli16] Roberto I. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probab. Theory Related Fields*, 166(3):1175–1194, 2016.

[PZ23] Nikita Puchkin and Nikita Zhivotovskiy. Exploring local norms in exp-concave statistical learning. In *Proc. 36th Annual Conference on Learning Theory*, pages 1993–2013, 2023.

[QRZ24] Jian Qian, Alexander Rakhlin, and Nikita Zhivotovskiy. Refined risk bounds for unbounded losses via transductive priors. *arXiv preprint arXiv:2410.21621*, 2024.

[SAH19] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, pages 11982–11992, 2019.

[SC19] Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA*, 116(29):14516–14525, 2019.

[She10] Irina G. Shevtsova. An improvement of convergence rate estimates in the Lyapunov theorem. In *Doklady Mathematics*, volume 82, pages 862–864. Springer, 2010.

[SW14] Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8:45–114, 2014.

[Tal21] Michel Talagrand. *Upper and lower bounds for stochastic processes: Decomposition Theorems*. Springer, 2nd edition, 2021.

[TB24] Kai Tan and Pierre C. Bellec. Multinomial logistic regression: Asymptotic normality on null covariates in high-dimensions. In *Advances in Neural Information Processing Systems*, pages 70892–70925, 2024.

[TPT20] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Sharp asymptotics and optimal performance for inference in binary models. In *Proc. 23rd Conference on Artificial Intelligence and Statistics*, pages 3739–3749, 2020.

[Tsy04] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[Tyu12] I. S. Tyurin. A refinement of the remainder in the Lyapunov theorem. *Theory Probab. Appl.*, 56(4):693–696, 2012.

[vdHZCB23] Dirk van der Hoeven, Nikita Zhivotovskiy, and Nicolò Cesa-Bianchi. High-probability risk bounds via sequential predictors. *Preprint arXiv:2308.07588*, 2023.

[vdV98] Aad van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[vdVW09] Aad van der Vaart and Jon A. Wellner. A note on bounds for VC dimensions. *Institute of Mathematical Statistics collections*, 5:103, 2009.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[vH14] Ramon van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

[Vij21] Suhas Vijaykumar. Localization, convexity, and star aggregation. In *Advances in Neural Information Processing Systems 34*, pages 4570–4581, 2021.

[VM24] Kabir A. Verchand and Andrea Montanari. High-dimensional logistic regression with missing data: Imputation, regularization, and universality. *arXiv preprint arXiv:2410.01093*, 2024.

[Vov98] Vladimir Vovk. A game of prediction with expert advice. *J. Comput. System Sci.*, 56(2):153–173, 1998.

[Zha04] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.

[Zhi24] Nikita Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electron. J. Probab.*, 29:1–28, 2024.

[ZSC22] Qian Zhao, Pragya Sur, and Emmanuel J. Candes. The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. *Bernoulli*, 28(3):1835–1861, 2022.