INSTITUT POLYTECHNIQUE DE PARIS

ENSAE

IP PARIS

NNT : 2025IPPAG003

Thèse de doctorat

Finite-sample theory for maximum-likelihood estimation in logistic regression

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique

école doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 26 juin 2025, par

HUGO CHARDON

Composition du Jury :

M. Alexandre Tsybakov Professeur, ENSAE-CREST	Président
M. Francis Bach Directeur de recherche, INRIA	Rapporteur
Mme Sara van de Geer Professeur, ETH Zurich	Rapporteur
M. Alain Célisse Professeur, Université Paris 1 Panthéon-Sorbonne	Examinateur
M. Matthieu Lerasle Professeur, ENSAE-CREST	Directeur de thèse
M. Jaouad Mourtada Professeur, ENSAE-CREST	Co-directeur de thèse

Remerciements

Mes premiers remerciements vont naturellement à mes directeurs de thèse Matthieu et Jaouad. Merci de m'avoir fait confiance en fin de M2. Matthieu, en réalité, tu as su avant moi ce qui me plairait réellement, merci de m'avoir "rattrapé" pour me prendre en thèse avec toi et d'avoir proposé ce co-encadrement avec Jaouad. Jaouad, j'ai énormément appris sur le plan technique avec toi, des astuces à une façon générale de faire. J'ai grandement profité de ta vision large, et pour cela, merci.

It is a great honor to have Sara van de Geer and Francis Bach as reviewers for this thesis. Their work is a true source of inspiration for me and many aspiring researchers. Thank you for having taken the time to read this manuscript and for your warm and encouraging comments. Je remercie également très vivement Alexandre Tsybakov qui me fait l'honneur de présider le jury. Je veux remercier particulièrement Alain Célisse qui, en plus d'accepter de faire partie du jury, m'a donné tôt l'occasion de présenter mes travaux de thèse dans des cadres très agréables.

Je remercie chaleureusement les doctorants que j'ai croisés plus ou moins longtemps au CREST. En premier lieu mes chers co-bureaux Clara et Étienne qui, avec Amir, m'ont accueilli en 3010. Clara et Étienne, mes deux premières années auraient été bien différentes si je n'avais pas été motivé à venir jusque sur le plateau par l'idée de vous y retrouver ! Ce bureau a été un lieu convivial et chaleureux, et je vous remercie pour cela. On y a été rejoints par Nina puis Clémentine et Sirine. A vous toutes et tous, merci d'avoir contribué à maintenir un si bon environnement. Enfin, merci Nayel de m'avoir remis au vélo avec les trajets de Paris à Palaiseau. Nos discussions d'abord de maths puis sur tout le reste ont été très précieuses. Merci aux anciens doctorants, Julien, Flore, Gabriel, Badr. En particulier Nicolas S. pour tes conseils et ta présence dans les conférences où on se croise.

Merci également à tous les membres permanents de l'équipe de statistique du CREST pour la bienveillance dont vous faites preuve envers les doctorants. Merci donc Arnak, Nicolas C., Anna, Azadeh, Cristina. Merci en particulier à Victor-Emmanuel pour le temps considérable que tu passes à faire en sorte que les thèses se passent bien par ton implication dans l'école doctorale. Dans mon cas, je te remercie tout particulièrement pour ton soutien face aux soucis administratifs que j'ai rencontrés et, de façon plus générale, pour ton attitude encourageante. Merci aux nouveaux permanents Vincent et Austin pour votre proximité avec nous, doctorants.

I also want to thank Nikita Zhivotovskiy who will be hosting me in Berkeley next year. I could not hope for a better outcome of my Ph.D. than going to work with you. I also want to thank chair Haiyan Huang and co-chair Aditya Guntuboyina for giving me the opportunity to join their department.

Je souhaite également remercier plusieurs des professeurs qui ont marqué ma scolarité et mes études. En premier lieu, Guy Alarcon, mon professeur de mathématiques en classe de 1ère. Merci de m'avoir donné le goût des maths plus que scolaires. Merci également à Nicolas Tosel pour sa façon d'enseigner qui arrive à concilier le format des classes

préparatoires et des ouvertures à des sujets variés. Cette thèse m'a également donné à voir à quel point l'un de tes dictons est pertinent, à savoir, "Les maths, c'est pas un sport de spectateur".

Un autre volet mais pas des moindres, merci à ma deuxième famille, celle du judo. Merci à tout le cercle étendu de ce merveilleux club qu'est Ippon 5, et donc en premier lieu, merci à Fabrice de m'y avoir emmené il y a dix ans. Merci à Jean-Baptiste Gros pour nos interminables entraînements, je ne me lasserai jamais de faire vingt minutes de randori de suite avec toi ! Merci de m'avoir poussé à aller au Japon, en Suisse et en Auvergne. Merci à Thomas Perret d'avoir toléré mes retards et ma motivation fluctuante ainsi que de me donner autant de fil à retordre. Merci à Emmanuel Charlot pour le judo (maintenant j'ai retenu qu'il ne faut pas tomber) et surtout le reste, les interminables discussions à toute heure du jour et (surtout) de la nuit. Merci à Jane Bridge pour sa présence. J'ai beaucoup d'admiration pour ta capacité à apporter une telle expertise technique avec autant de simplicité. Merci à tous les membres du club pour l'ambiance générale qui y règne, sur le tapis et en dehors, au Japon, aux Geneveys-sur-Coffrane et à Mur-de-Barrez. Merci donc David, Maxime, Alexandre, Wassim, Cyrus, Paul, Olivier et Thomas, Félix et Ferdinand, Aron et Benjamin, Maxence, Lucie S., Hélène, Brieuc, Léa, Lucie D. ; et tous ceux que j'ai honteusement oubliés. Merci également à tous ceux que j'ai plaisir à retrouver en stage, Suisses et Cannois, en particulier Amadeus. Enfin, tout ce paragraphe n'existerait pas si je n'avais pas commencé le judo quand j'étais haut comme trois pommes (si si). Merci Maman de m'y avoir inscrit. J'ai bien sûr une pensée toute particulière pour mon premier professeur Michael Afonso. Merci de m'avoir initié si bien à cette discipline. Tu as fait naitre en moi le goût du beau judo. Ton exigence, tes conseils et ta façon de voir le judo m'imprègnent encore aujourd'hui.

Merci également à mes amis les plus proches de m'avoir aidé à garder la tête hors de l'eau quand c'était nécessaire. Merci Loïc pour ta présence rassurante et ton bon sens à toute épreuve, merci aussi de m'avoir poussé à venir au Brésil, ça nous fera des souvenirs pour toujours. Merci Pablo pour toutes ces discussions, nos aventures en France, en Belgique et aux États-Unis ! Merci également André pour les bons moments en marge du judo et pas que, mais surtout merci de m'y avoir ramené quand j'en ai eu besoin. Merci également à Maude, on s'est suivis de près du lycée au M2 de maths en passant par la prépa et l'ENSAE, merci d'avoir été une compagne de galère dans toutes celles que ces années nous ont offertes ! A ce propos, merci également Théo pour nos discussions variées, ton avis est toujours précieux. Merci Hugo T. pour nos aventures à vélo, j'espère qu'on en aura d'autres à l'avenir, en espérant que je sois capable de suivre ! Merci Balthazar pour ta sagesse et ta connaissance sans faille de Kaamelott. J'ai la chance que, contrairement au roi Loth, mon destin ne soit pas *"de m'entourer de nuls, de vrais nuls"*. Bien au contraire, de personnes extraordinaires.

Avant de passer à ma famille, je voudrais avoir une pensée toute particulière pour une personne très spéciale. Tatie, je te dois énormément, merci pour tout ce que tu m'as apporté dans mon enfance et après. J'ai le cœur bien gros en écrivant ces lignes que tu ne liras pas. J'aurais tant aimé que tu sois toujours parmi nous. Merci pour ces journées de bonheur passées chez toi. Toi et Tonton Gaby étiez un modèle de stabilité et de bienveillance.

Naturellement, je remercie ici ma famille pour leur soutien. Merci Papa et Maman de m'avoir donné un goût pour les sciences, par votre exemple et en éveillant ma curiosité petit. Merci aussi, et surtout, pour le reste. Entre autres choses, d'avoir cherché à m'intéresser à des choses plus variées et de faire en sorte que je n'aie pas deux mains gauche. Merci d'avoir toujours été présents matériellement et moralement dans des moments difficiles. Merci également à mes grand-parents pour leur présence continuellement bienveillante.

À la mémoire de Denise et Gabriel Rouxel, Tatie et Tonton Gaby.

Contents

Outline and summary of contributions

In Chapter 1, we give a detailed introduction to the topic of maximum-likelihood estimation in logistic regression and formulate precisely the questions that subsequent chapters provide answers to. We also give an overview of prior works on the topic from different perspectives. This introductory chapter also gives a full overview of the results of this thesis, as well as detailed sketches of proofs.

Chapter 2 explains the general approach via convex localization used in Chapters 3 and 6 to provide guarantees on the existence and performance of the maximum-likelihood estimator. We also sketch the main difficulties that will arise in subsequent chapters.

Chapters 3 and 4 both deal with the case of a well-specified logit model with a Gaussian design, but they address questions of different natures. In the former we provide sharp excess risk bounds for the MLE, and in the latter we show that if the sufficient sample size condition from Chapter 3 is not satisfied, then not only is the probability of existence of the MLE bounded away from 1, it actually vanishes.

Chapter 3 deals with the case of a well-specified model with a Gaussian design. It contains the core arguments of our analysis of logistic regression. It is also the setting where we obtain the sharpest results. It also contains on of the most technically involved arguments, regarding the uniform deviations of a collection of random matrices, in the proof of Theorem 3.2.

The purpose of Chapter 4 is to show that the sufficient condition on the sample size in the Gaussian well-specified model is actually also necessary. It builds on the geometric approach from the seminal work of Candès and Sur [CS20] to prove a fully non-asymptotic phase transition regarding the existence of the MLE.

Chapter 5 describes a class of probability distributions, that we call *regular*. This notion ensures a "near-Gaussian" for the MLE when the design is no longer assumed to be Gaussian. We introduce in particular a new "two-dimensional margin condition" that we prove to be necessary in some sense. We then give examples of distribution satisfying these assumptions, investigating in particular the case of i.i.d. coordinates.

Chapter 6 provides guarantees for the existence and behavior of the MLE when the design is no longer assumed to be Gaussian, but only regular, in the sense defined in Chapter 5. The results are exposed in increasing order of generality, starting with the case where the model is still assumed to be well-specified (Theorem 6.1) and then moving on to the most general case where the design is regular and no assumption is made on the conditional distribution of the outcome on the input (Theorem 6.2).

Finally, in Chapter 7, we are interested in the classification performance of the MLE as a plug-in classifier. We relate our results to Zhang's lemma, regarding the statistical consistency of predictors obtained as minimizers of convex surrogate risks, as well as the fast rates obtained under Tsybakov's *margin condition*. We show that in the ideal case, the MLE achieves, as a binary classifier, the fast parametric rate $1/n$.

Notation

$\mathbf{P}$	probability measure on a measurable space $(\Omega, \mathcal{A})$ that will generally not be specified
$\mathcal{P}(\mathcal{Z})$	set of all probability measures on the measurable space $\mathcal{Z}$ (with an implicit σ-algebra)
P_Z	distribution of the random variable $Z : (\Omega, \mathcal{A}, \mathbf{P}) \to \mathcal{Z}$, hence the push-forward measure of $\mathbf{P}$ by Z
$Z \sim P$	the random variable Z follows the distribution P. Equivalently $P_Z = P$
$\mathbf{E}Z$	expectation of the random variable $Z : \Omega \to \mathcal{Z}$, that is $\int_\Omega Z(\omega)\mathbf{P}(\mathrm{d}\omega) = \int_{\mathcal{Z}} zP_Z(\mathrm{d}z)$
$\mathsf{N}(m, \Sigma)$	Gaussian distribution with mean m and covariance matrix Σ on the Euclidean space that will be clear from context
$\chi^2(d)$	χ^2 distribution with d degrees of freedom
$P \ll Q$	the measure P is absolutely continuous with respect to the measure Q
$\frac{\mathrm{d}P}{\mathrm{d}Q}$	density of P with respect to Q
$D(P\|Q)$	Kullback-Leibler divergence from P to Q, defined as $\int \log(\frac{\mathrm{d}P}{\mathrm{d}Q})\mathrm{d}P$
i.i.d.	independent and identically distributed
$\mathbb{R}$	set of real numbers
$L^p(\mu)$	given $p \geqslant 1$ and a measure μ on a space $\mathcal{X}$, the space of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\int \lvert f\rvert^p \mathrm{d}\mu < +\infty$
$\lVert\cdot\rVert_p$	L^p-norm on the space $L^p(\mu)$ for some measure μ on a space $\mathcal{X}$ that will both be clear from context, $\lVert f\rVert_p = (\int \lvert f\rvert^p \mathrm{d}\mu)^{1/p}$. If ambiguous, we use $\lVert\cdot\rVert_{L^p(\mu)}$. Also denotes ℓ_p-norm on a finite-dimensional space
$\mathbf{1}(\cdot \in A)$	indicator function of the set A (with slight abuses such as $\mathbf{1}(x \geqslant 0)$ instead of $\mathbf{1}(x \in [0, +\infty))$. Sometimes denoted $\mathbf{1}\{\cdot \in A\}$
$\langle\cdot,\cdot\rangle$	canonical inner product on a Euclidean space that will be clear from context
$\lVert\cdot\rVert$	Euclidean norm
$\lVert x\rVert_M$	elliptic norm $\langle Mx, x\rangle = \lVert M^{1/2}x\rVert$, for every $x \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d\times d}$ positive symmetric
$\lVert A\rVert_{\mathrm{op}}$	operator norm of the matrix A, $\lVert A\rVert_{\mathrm{op}} = \sup_{\lVert v\rVert=1} \langle Av, v\rangle$
$\xrightarrow{(\mathrm{d})}$	convergence in distribution
$\xrightarrow{(\mathbf{P})}$	convergence in probability
$\xrightarrow{(\mathrm{a.s.})}$	almost sure convergence
log	natural logarithm

Chapter 1

Introduction

This thesis studies various aspects of random-design logistic regression from a non-asymptotic perspective. In this introduction, we set the stage for the discussions developed in the subsequent chapters. We also state all of the main results and give sketches of their proofs.

Contents

Historical perspective and practical importance. Logistic regression is a classical model describing the dependence of binary outcomes on multivariate features. Its simple description makes it very popular in a large variety of applications in experimental and social sciences. A prototypical example is in biomedical research: how can one predict whether a patient will develop a certain disease (e.g., lung cancer) based on genetic markers and other clinical characteristics of this person? Logistic regression (or the logistic model) is a way to tackle such a problem.

The logit model was first introduced by Berkson [Ber44] in the context of drug dosage clinical studies. From a biomedical perspective, the use of the *probit* model (where the link function is the cumulative of the Gaussian distribution) was more popular, but Berkson argued that the logistic model had the advantage of being more interpretable and computationally more tractable. He proposed to model the dependence of an outcome $Y = \pm 1$ on a real covariate X (with modern notation) through its conditional probability distribution as

$$\mathbf{P}(Y = 1|X = x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}} = \sigma(\theta_0 + \theta_1 x)\,, \tag{1.1}$$

where θ_0 and θ_1 are the unknown parameters to estimate, and

$$\sigma(t) = \frac{e^t}{1 + e^t} = \frac{1}{1 + e^{-t}}\,, \quad t \in \mathbb{R}\,, \tag{1.2}$$

is called the *sigmoid* function.

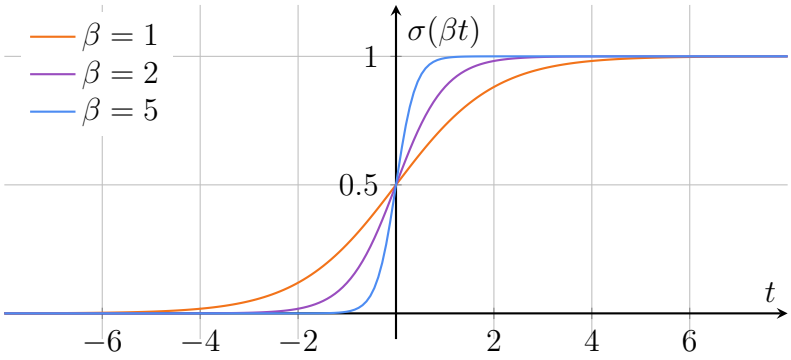

Figure 1.1: The sigmoid function $\sigma(\beta t) = 1/(1 + e^{-\beta t})$ for different values of the signal strength β.

Berkson puts forward the relevance of the logit function, which is the inverse of the sigmoid (1.2),

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right), \tag{1.3}$$

which in the logistic model (1.1) transforms the logarithm of the odds ratio $p/(1-p)$ into an affine function of the covariate x

$$\text{logit}(p) = \theta_0 + \theta_1 x\,, \quad p = \mathbf{P}(Y = 1|X = x)\,. \tag{1.4}$$

He therefore suggests the following procedure: given observations $(x_1, y_1), \ldots, (x_n, y_n)$, compute the corresponding log-odds ratios with the logit function (1.3), then fit the linear model (1.4) using the classical least-squares method.

Although Berkson properly introduced the (univariate) logistic model (1.1), he did not advocate for the use of the generic maximum-likelihood estimation method which had been known for many years in the context of parametric estimation in statistical models.

A few years later, Cox formalized logistic regression in terms closer to the modern setting in [Cox58]. Let us point out the main differences with the earlier work of Berkson. First, Cox considers the logit model for general purposes, as opposed to the description of Berkson which focuses on bioassays. Second, Cox highlights the more natural character of maximum-likelihood estimation in the logistic model. Finally, Cox also considers multivariate features[1] $x \in \mathbb{R}^d$, for some dimension $d \geqslant 1$.

Modern description. In its general, modern form, the logistic model can be described in the following way. 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:

$$\mathcal{P}_{\text{logit}} = \left\{ p_\theta : \theta \in \mathbb{R}^d \right\}, \quad \text{where} \quad p_\theta(y|x) = \sigma(y\langle \theta, x \rangle), \tag{1.5}$$

for all $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$, where σ denotes the sigmoid function (1.2), 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 $\mathcal{P}_{\text{logit}}$.

Historically, the logit model was studied under the assumption that the data were generated according to the model, as was usually the case in parametric estimation. More precisely, the statistical setting was the following: the statistician is given observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ in $\mathbb{R}^d \times \{-1, 1\}$ which are assumed to be independent and have common unknown distribution P, and this distribution is such that the model is well-specified, meaning that there exists $\theta^* \in \mathbb{R}^d$ such that for all i,

$$\mathbf{P}(Y_i = 1 \,|X_i) = \sigma(\langle \theta^*, X_i \rangle). \tag{1.6}$$

Then the statistician would fit the model using maximum-likelihood estimation, that is computing

$$\widehat{\theta}_n = \arg\max_{\theta \in \mathbb{R}^d} \prod_{i=1}^n p_\theta(Y_i|X_i) = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \log\left(1 + e^{-Y_i \langle \theta, X_i \rangle}\right), \tag{1.7}$$

for the purpose of inference on the the parameter θ^*, mainly to test whether particular coefficients are null. This point of view is motivated by the aim of testing whether a particular variable $x^{(j)}$ has a statistically significant effect on the binary outcome y—for instance, whether a given gene influences the probability of developing a certain disease.

However, this approach has an important limitation: in order to make sense, it requires assuming that the model is *well-specified*, meaning that the true data-generating distribution belongs to the model, or equivalently that (1.6) holds.

Beyond inference, logistic regression is also appealing as a prediction method, where the focus shifts from testing the significance of particular variables to accurately estimating the conditional probability $\mathbf{P}(Y = 1|X = x)$ for all x. This predictive perspective is the main focus of this thesis. In particular, we also consider the general framework of statistical learning, where the model is typically not assumed to be well-specified. In this setting, logistic regression serves as a way to approximate the true conditional distribution of Y given X by a distribution belonging to the logit model.

[1] Technically speaking, Cox considers the cases $d = 1$ and $d = 2$ but the shift to two-dimensional model can easily be generalized and lays the foundations for the general model.

1.1 Main questions

In this thesis, we investigate the predictive performance of the maximum likelihood estimator (MLE), in the spirit of statistical learning. We will thus 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 (1.7) 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 (1.7) 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 (1.7) 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 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 (1.5). 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:

$$\ell(\theta, (x, y)) = -\log p_\theta(y|x) = \log(1 + e^{-y\langle \theta, x \rangle}) \quad \text{and} \quad L(\theta) = \mathbf{E}[\ell(\theta, (X, Y))]\,. \tag{1.8}$$

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

$$\widehat{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\theta, (X_i, Y_i)) = \frac{1}{n} \sum_{i=1}^n \log(1 + e^{-Y_i \langle \theta, X_i \rangle})\,. \tag{1.9}$$

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). We provide non-asymptotic, high probability upper bounds on this quantity in Chapters 3 and 6.

The logistic loss also naturally arises in statistical learning theory [BBL05, Kol11, Bac24], as a convex surrogate of the classification error [Zha04, BJM06]. As such, we will also consider (in Chapter 7) the classification performance of the MLE as a plug-in classifier to correctly predict the label Y of a point X. By plug-in classifier we mean predictors of the form $f_\theta(x) = \mathrm{sign}(\langle \theta, x \rangle)$.

In this setting, performance is assessed by the classification risk, defined for every classifier $f : \mathbb{R}^d \to \mathbb{R}$ as $R(f) = \mathbf{P}(Yf(X) < 0)$ We will therefore measure the binary classification performance of $\widehat{\theta}_n$ by the excess risk of classification of $f_{\widehat{\theta}_n}$ with respect to the overall best classifier, namely

$$R(\widehat{\theta}_n) - \inf_f R(f) = \mathbf{P}(Y\langle \widehat{\theta}_n, X \rangle < 0) - \inf_f \mathbf{P}(Yf(X) < 0)\,,$$

where the infimum is taken over all measurable functions from $\mathbb{R}^d$ to $\mathbb{R}$.

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, called design distribution; 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 $\mathbf{P}(Y = 1|X) = \sigma(\langle \theta^*, X \rangle)$.

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. We will finally investigate the classification performance of the plug-in estimator associated with the MLE which, as mentioned above, is closely related to the estimation of the direction of θ^*.

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

$$2n\big\{L(\widehat{\theta}_n) - L(\theta^*)\big\} \xrightarrow{\text{(d)}} \chi^2(d)\,, \tag{1.10}$$

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

$$\liminf_{n\to\infty} \mathbf{P}\Big(L(\widehat{\theta}_n) - L(\theta^*) \leqslant \frac{d + 2\log(1/\delta)}{n}\Big) \geqslant 1 - \delta\,, \tag{1.11}$$

with the convention that $L(\widehat{\theta}_n) - L(\theta^*) = +\infty$ if the dataset is linearly separated. Note that the convergence (1.10) holds only in the well-specified case, and that in the misspec-

ified 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 (1.11) is sharp, in light of the convergence in distribution (1.10) 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 (1.11) to occur—in particular, it provides no information on the sample size required for the existence of the MLE.

1.2.2 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[2]: 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 $\mathbf{P}(Y = 1|X) = \sigma(\langle \theta^*, X \rangle)$, 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

$$\lim_{n\to\infty} \mathbf{P}(\text{MLE exists}) = \begin{cases} 0 & \text{if } \gamma > h(\beta) \\ 1 & \text{if } \gamma < h(\beta)\,. \end{cases} \tag{1.12}$$

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 (1.12) 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 (1.12) to provide an accurate approximation. As an

[2] 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.

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 (1.12) 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 (1.12) 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 behaves in a similar way as for Gaussian design. Likewise, the characterization (1.12) 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 (1.12) 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.

1.2.3 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

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant \frac{e^{cB} d}{n} \,. \tag{1.13}$$

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 (1.10) and (1.11). 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 (1.13) 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

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant \frac{B^3 d \log(1/\delta)}{n} \,. \tag{1.14}$$

Like the bound (1.13), 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 (1.13) 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 (1.14) is larger than the asymptotic risk (1.11) by a factor of B^3, which suggests possible improvements.

Recently and while the article [CLM24] 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 noise 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

$$\left\| \frac{\widehat{\theta}_n}{\|\widehat{\theta}_n\|} - \frac{\theta^*}{\|\theta^*\|} \right\| \leqslant c\sqrt{\frac{d\log n + \log(1/\delta)}{Bn}}\,, \quad \left| \|\widehat{\theta}_n\| - \|\theta^*\| \right| \leqslant cB^{3/2}\sqrt{\frac{d\log n + \log(1/\delta)}{n}}\,. \tag{1.15}$$

While the bound (1.15) 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

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c'\,\frac{d\log n + \log(1/\delta)}{n} \tag{1.16}$$

for some constant $c' > 0$. This guarantee matches the asymptotic risk (1.11) 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 (1.15), 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.

In all the results discussed above, the dependence with respect to both the dimension and sample size were essentially optimal (up to logarithmic factors) and sub-optimality came from the dependence on the signal strength B, in the high signal regime. This is why in this thesis we focus on the dependence on B, especially when B is large. This is the regime where the probabilistic predictions of the MLE become unreliable, as the signal strength itself is hard to estimate.

1.2.4 Binary classification

So far we discussed previously established results on the performance of the MLE in terms of logistic loss, which is a natural measure of quality when estimating probabilities. As mentioned above, the logistic loss is also a natural convex surrogate for the binary loss in supervised classification. It satisfies the definition of *calibration* introduced by [Zha04] and [BJM06] and therefore falls into the scope of the celebrated Zhang's lemma. This result allows one to bound the classification error of a predictor in terms of its excess risk under the convex surrogate loss. When specified for logistic regression, in the Gaussian, well-specified setting, it leads to the following bound

$$R(\theta) - R(\theta^*) \leqslant C\sqrt{L(\theta) - L(\theta^*)}\,, \tag{1.17}$$

where $R(\theta)$ denotes the classification risk of the linear predictor associated with the vector θ, that is $R(\theta) = \mathbf{P}(Y\langle\theta, X\rangle < 0)$, and C is a universal constant. Using non-asymptotic results for the logistic excess risk of the MLE, e.g., (1.14) from [OB21], this yields a classification risk scaling as $\sqrt{d/n}$, which is the typical slow rate provided by Vapnik-Chervonenkis theory. In short, (1.17) ensures that the MLE $\widehat{\theta}_n$ is consistent when used as a plug-in for binary classification, but in terms of quantitative bounds, it significantly degrades the rates obtained for the excess logistic risk mentioned in the previous section.

We emphasize that compared to the previous section where we discussed the sharpness of the results with respect to the signal strength B, the issue here has to do with the very rate of convergence with respect to the sample size n, whereas in the previous section, the dependence with respect to the dimension d and the sample size n were sharp (up to logarithmic terms), and only the dependence with respect to the signal strength left space for improvement.

1.3 Convex localization, empirical gradient and empirical Hessians (Chapter 2)

1.3.1 Recent approach: the local Bernstein condition

We now discuss the global approach that we will follow to obtain finite-sample guarantees for both existence and performance of the MLE in logistic regression. Recall that we defined the loss $\ell(\theta,(x,y)) = \log(1+e^{-y\langle\theta,x\rangle})$ for all $\theta \in \mathbb{R}^d$ and all $(x,y) \in \mathbb{R}^d \times \{-1,1\}$, as well as the risk and empirical risk, respectively as

$$L(\theta) = \mathbf{E}[\ell(\theta,(X,Y))] \quad \text{and} \quad \widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta,(X_i,Y_i))\,,$$

where $(X,Y),(X_1,Y_1),\dots,(X_n,Y_n)$ are i.i.d. with common unknown distribution P.

Our strategy relies on the following simple observation: since $\widehat{\theta}_n$ is the unique minimizer of $\widehat{L}_n$ (see the discussion at the beginning of Section 10.1), one has $\widehat{L}_n(\widehat{\theta}_n) \leqslant \widehat{L}_n(\theta^*)$. Therefore, any compact convex set Θ containing θ^* and such that for all θ on its boundary, $\widehat{L}_n(\theta) > \widehat{L}_n(\theta^*)$; necessarily contains $\widehat{\theta}_n$.

As highlighted by [CLL20], the convexity of the loss can be leveraged to localize the empirical risk minimizer based on this observation. Their result relies on a local version of Bernstein's condition, which bears on the curvature of the loss function. The global Bernstein condition was first introduced in [BM06] to obtain fast rates in empirical risk minimization. The weaker, local version of [CLL20] applies to other loss functions than the logistic one, but for the sake of clarity, we express this condition within our setting, in particular assuming (without loss of generality) that the design is isotropic: $\mathbf{E}XX^\top = I_d$. The local Bernstein condition [CLL20, Assumption 4] states that the risk function is locally strongly convex around its minimum. Specifically, there exist positive numbers A and $r(A)$ such that for all $\theta \in \mathbb{R}^d$ satisfying $\|\theta-\theta^*\| = r(A)$, it holds that

$$\|\theta-\theta^*\|^2 \leqslant A\big[L(\theta)-L(\theta^*)\big]\,. \tag{1.18}$$

There are two important things to note here. First, the larger A, the looser the downstream localization of $\widehat{\theta}_n$ (in the sense that the region which contains it with high probability will be large). Second, A is related to the Hessian matrices of the risk L in the neighborhood of θ^*. Indeed, if for all θ' on the segment going from θ^* to θ, $\nabla^2 L(\theta') \succcurlyeq A^{-1}\cdot I_d$, then (1.18) holds. This supports and quantifies the reasonable claim that "the more convex the loss is, the easier it is to localize the ERM". In [CLL20], the authors note in particular that for the logistic loss, the curvature condition (1.18) holds for A of order e^B, which explains the bound (1.13). To see this, observe that

$$\nabla^2 L(\theta) = \mathbf{E}\big[\sigma'(\langle\theta,X\rangle)XX^\top\big]\,.$$

One can check that the derivative σ' of the sigmoid is even and that for all $t\in\mathbb{R}$, $\sigma'(t)$ is of order $e^{-|t|}$. In the interest of clarity, let us discuss the case where X is Gaussian, namely $X \sim \mathsf{N}(0,I_d)$. In this case, $|\langle\theta,X\rangle|$ is typically of order $\|\theta\|$, so that if θ is in a neighborhood of θ^*, the typical value of $\sigma'(\langle\theta,X\rangle)$ is indeed close to e^{-B}. Then, as $\mathbf{E}XX^\top = I_d$, we find that $\nabla^2 L(\theta) \gtrsim e^{-B} I_d$.

1.3.2 Strengthened local curvature: empirical gradient and Hessians

To avoid the aforementioned exponential dependence with respect to B, our strategy is the following. In order to localize $\widehat{\theta}_n$, we still want to exhibit a compact convex set Θ_n (as small as possible) on the boundary of which

$$\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) > 0\,, \tag{1.19}$$

which would show that $\widehat{\theta}_n \in \Theta_n$. To do this, instead of controlling the uniform deviation of $\widehat{L}_n$ from L over a neighborhood of θ^*, we directly use a second-order Taylor expansion *of the empirical risk* around θ^*, which gives, for all θ (with a slight abuse of notation),

$$\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) = \langle \nabla \widehat{L}_n(\theta^*), \theta - \theta^* \rangle + \int_{\theta^*}^{\theta} \big\langle \nabla^2 \widehat{L}_n(\theta')(\theta - \theta^*), \theta - \theta^* \big\rangle \mathrm{d}\theta'\,. \tag{1.20}$$

Here, by integrating from θ^* to θ, we mean letting $\theta_t = (1-t)\theta^* + t\theta$ for every $t \in [0,1]$ and integrating with respect to t. In this expansion, the only quantity that depends on the path from θ^* to θ is in the argument of the integrated empirical Hessian. In fact, we can rewrite (1.20) as

$$\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) = \langle \nabla \widehat{L}_n(\theta^*), \theta - \theta^* \rangle + \big\langle \widetilde{H}_n(\theta)(\theta - \theta^*), \theta - \theta^* \big\rangle\,, \tag{1.21}$$

with

$$\widetilde{H}_n(\theta) = \int_0^1 (1-t) \nabla^2 \widehat{L}_n\big((1-t)\theta^* + t\theta\big) \mathrm{d}t\,.$$

It becomes clear now that the relevant sets Θ_n for (1.19) to hold should be ellipsoids for the Hessian matrix $\nabla^2 L(\theta^*)$. Indeed, if (i) the first order term is not to large, and (ii) the second term behaves like the quadratic form associated with $\nabla^2 L(\theta^*)$, then (1.19) will hold for θ on the boundary of a $\nabla^2 L(\theta^*)$-ellipsoid. The idea outlined here will be made formal in Chapter 2, more precisely in Lemma 2.1 therein.

Let us discuss the technical challenges that will arise from these observations. First, for any $\theta \in \mathbb{R}^d$, if n is large enough, the empirical Hessian $\widehat{H}_n(\theta) = \nabla^2 \widehat{L}_n(\theta)$ should approximate well the population Hessian $H(\theta) = \nabla^2 L(\theta)$, in the sense that with high probability, $\widehat{H}_n(\theta) \succcurlyeq \frac{1}{2} H(\theta)$, say. Second, for θ close to θ^*, by regularity of the map $\theta \mapsto H(\theta)$, one can reasonably expect that $H(\theta) \succcurlyeq \frac{1}{2} H(\theta^*)$. In other words, we want a uniform lower bound on the empirical Hessians over a neighborhood Θ of θ^*, which is itself such that for all $\theta \in \Theta$, $H(\theta)$ is almost constant equal to $H(\theta^*)$. By suitably choosing a good approximation H of $H(\theta^*)$, we may instead work with H. If we can prove that for n large enough (in a sense that should be precisely quantified in terms of B, d, and level of confidence δ), it holds with probability at least $1-\delta$ that *simultaneously* for all $\theta \in \Theta$, $\nabla^2 \widehat{L}_n(\theta) \succcurlyeq cH$, then by (1.21), we have by the Cauchy-Schwarz inequality,

$$\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) \geqslant -\big\|\nabla \widehat{L}_n(\theta^*)\big\|_{H^{-1}} \|\theta - \theta^*\|_H + \frac{c}{2} \|\theta - \theta^*\|_H^2\,.$$

Now if the rescaled empirical gradient is upper-bounded as

$$\big\|\nabla \widehat{L}_n(\theta^*)\big\|_{H^{-1}} \leqslant \psi_n\,,$$

then the previous inequality becomes

$$\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) \geqslant \|\theta - \theta^*\|_H \Big(\frac{c}{2} \|\theta - \theta^*\|_H - \psi_n \Big) . \tag{1.22}$$

Of course, the rate ψ_n also depends on B, d and δ, although our aim is to weaken the dependence on B as much as possible—ideally to obtain a bound that does not depend on B.

It is clear from (1.22) that for any $r > 2\psi_n/c$, any θ on the boundary of an H-ellipsoid of radius r satisfies $\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) > 0$. This shows that (i) the MLE $\widehat{\theta}_n$ exists and (ii) it is localized within an H-ellipsoid of radius $O(\psi_n)$, namely that

$$\|\widehat{\theta}_n - \theta^*\|_H \leqslant \frac{2\psi_n}{c} . \tag{1.23}$$

If in addition, on the same region where we were able to uniformly bound from below the empirical Hessians, we can uniformly upper bound the *population* Hessians by a factor of H, then by a Taylor expansion of L this time (thus with no first-order term, as $\nabla L(\theta^*) = 0$), we deduce that the excess risk of $\widehat{\theta}_n$ is itself bounded as $O(\psi_n^2)$.

In fact, the excess risk of any approximate minimizer of $\widehat{L}_n$ will enjoy the same risk bound, which is important in practical applications, as it supports the idea that there is no need to optimize numerically below the statistical error.

The take-home message of this discussion is the following: in order to prove that the MLE exists and to localize it, we will prove that the empirical risk $\widehat{L}_n$ is with high probability sufficiently strongly convex in a large enough ellipsoid around θ^*. To summarize,

- we need to bound from above the norm of the rescaled empirical gradient as

$$\big\|\nabla \widehat{L}_n(\theta^*)\big\|_{H^{-1}} = \big\|H^{-1/2} \nabla \widehat{L}_n(\theta^*)\big\| \leqslant \psi_n \, , \tag{1.24}$$

  where $H \asymp \nabla^2 L(\theta^*)$. The high-probability upper bound ψ_n on the gradient will give the downstream bound $O(\psi_n^2)$ on the excess risk of the MLE. Given the previous results mentioned in Section 1.2.3, we thus aim for ψ_n of order $\sqrt{d/n}$ with no dependence on B, at least in the ideal case of a well-specified model with a Gaussian design.

- We also need to uniformly bound from below the empirical Hessians in the region Θ where the population Hessian is almost constant and within factors of H. More precisely, we need to identify a critical sample size $N = N(B, d, \delta)$ such that for all $n \geqslant N$,

$$\mathbf{P}\Big(\inf_{\theta \in \Theta} \nabla^2 \widehat{L}_n(\theta) \succcurlyeq cH \Big) \geqslant 1 - \delta \, , \tag{1.25}$$

  where Θ is an H-ellipsoid, namely $\Theta = \{\theta : \|\theta - \theta^*\|_H \leqslant r_0\}$ for some r_0 as large as possible.

The localization of $\widehat{\theta}_n$ is thus done in two steps, each yielding a sample size condition.

First, n has to be large enough for (1.25) to hold. This is the first sample size condition. Second, in order to combine (1.25) with (1.22), we need that $\psi_n \leqslant cr_0/2$. This is the second sample size condition. In short, these conditions together quantify how large n should be for $\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*)$ to exhibit the expected quadratic behavior, where the second-order term dominates the first-order term.

As discussed in Section 1.2.3, we will consider three settings of increasing generality, whose respective assumptions will have an influence on the nature and difficulty of both tasks above. To set the stage, we discuss in the next section what we can hope for in the ideal case of a well-specified model with a Gaussian design.

1.3.3 What to expect in the ideal case?

It is now clear from the discussion above that the Hessian of the population risk at θ^* plays a central role in the analysis of the MLE. In fact, we only need an approximation H of $\nabla^2 L(\theta^*)$, for which the two important bounds (1.24) and (1.25) hold. Also, the region Θ on which (1.25) holds is bound to be such that for all $\theta \in \Theta$, $H(\theta) \asymp H(\theta^*) \asymp H$, which is itself an H-ellipsoid, as will be made clear below.

It is worth mentioning now the following observations. Given a pair $z = (x, y) \in \mathbb{R}^d \times \{-1, 1\}$ and a parameter $\theta \in \mathbb{R}^d$, the gradient and Hessian of the logistic loss evaluated at θ for the observation z are given by

$$\nabla \ell(\theta, z) = -y\sigma(-y\langle \theta, x\rangle)x\,, \quad \nabla^2 \ell(\theta, z) = \sigma'(\langle \theta, x\rangle)xx^\top. \tag{1.26}$$

In particular, the derivative of the sigmoid σ' being even, the Hessian of the logistic loss does not depend on y. As such, the distribution of the empirical Hessian will not depend on the distribution of Y, only on that of X, that is, the design. The distribution of the empirical gradient on the other hand will heavily depend on the distribution of Y, and more precisely on the conditional distribution of Y given X, and thus, on whether or not the model is well-specified.

Let us consider now the ideal case where the design is Gaussian (specifically, $X \sim \mathsf{N}(0, I_d)$) and the model is well-specified with parameter $\theta^* \in \mathbb{R}^d \setminus \{0\}$. In this case, we find the following. Let $u^* = \theta^*/\|\theta^*\|$ denote the direction of θ^*, and as before $B = \|\theta^*\|$ (assuming a strong signal, see the previous section). By rotational invariance of the distribution of X, it is clear that $H(\theta^*)$ has two eigenvalues, with eigenspaces $\mathbb{R}u^*$ and its orthogonal hyperplane $(\mathbb{R}u^*)^\perp$, namely

$$H(\theta^*) = g_1(B)u^*u^{*\top} + g_2(B)(I_d - u^*u^{*\top})\,, \tag{1.27}$$

where

$$g_1(B) = \mathbf{E}[\sigma'(\langle \theta^*, X\rangle)\langle u^*, X\rangle^2]\,, \quad \text{and} \quad g_2(B) = \mathbf{E}[\sigma'(\langle \theta^*, X\rangle)]\,.$$

This is precisely where we make a first key observation: instead of using the fact that $|\langle \theta^*, X\rangle| = B|\langle u^*, X\rangle|$ is of order B with high probability (hence $\sigma'(\langle \theta^*, X\rangle)$ is of order e^{-B}) we rather rely on the fact that $|\langle \theta^*, X\rangle|$ is of order 1 with probability $1/B$. Therefore, recalling that $\sigma'(t) \asymp e^{-|t|}$,

$$\sigma'(\langle \theta^*, X\rangle)\langle u^*, X\rangle^2 \asymp \frac{1}{B^2}\,, \quad \text{with probability } \frac{1}{B}\,, \tag{1.28}$$

and similarly for g_2, so in the end

$$g_1(B) \asymp \frac{1}{B^3} \quad \text{and} \quad g_2(B) \asymp \frac{1}{B}\,. \tag{1.29}$$

We thus define once and for all the matrix H as

$$H = \frac{1}{B^3}u^*u^{*\top} + \frac{1}{B}(I_d - u^*u^{*\top})\,. \tag{1.30}$$

The same structure is shared by $H(\theta)$ for other values of θ, namely,

$$H(\theta) = g_1(\|\theta\|)uu^\top + g_2(\|\theta\|)(I_d - uu^\top)\,,$$

where $u = \theta/\|\theta\|$. In addition, it can be deduced from this characterization of the structure of the Hessians that the region Θ where $H(\theta) \asymp H(\theta^*)$—hence $H(\theta) \asymp H$—is indeed an H-ellipsoid of the form

$$\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \lesssim \frac{1}{\sqrt{B}}\right\}. \tag{1.31}$$

Equivalently, in polar coordinates, this corresponds to vectors $\theta = \|\theta\|u \in \mathbb{R}^d$ such that

$$\|u - u^*\| \lesssim \frac{1}{B} \quad \text{and} \quad \|\theta\| \asymp \|\theta^*\|. \tag{1.32}$$

This follows from Lemma 8.3. With these estimates at hand, we are in a position to identify the best possible guarantees we can hope for; again, at least in the well-specified, Gaussian case.

In this ideal case, we seek sharp bounds with respect to all parameters involved. As the model is well-specified, the Hessian $H(\theta^*)$ is the covariance matrix of the gradient $\nabla\ell(\theta^*, Z)$, which implies that

$$H^{-1/2}\nabla\widehat{L}_n(\theta^*) = \frac{1}{n}\sum_{i=1}^{n} H^{-1/2}\nabla\ell(\theta^*, Z_i)$$

is a sum of i.i.d., centered, and nearly isotropic random vectors, as $H(\theta^*) \asymp H$. In particular, we seek to bound it as

$$\left\|\nabla\widehat{L}_n(\theta^*)\right\|_{H^{-1}} \leqslant C\sqrt{\frac{d + \log(1/\delta)}{n}}, \tag{1.33}$$

with probability larger than $1 - \delta$, where C is a universal constant. Therefore, given the shape (1.31) of the set Θ, this will naturally give rise to the sample size condition

$$C\sqrt{\frac{d + \log(1/\delta)}{n}} \leqslant \frac{C'}{\sqrt{B}}, \tag{1.34}$$

or equivalently, $n \gtrsim B(d + \log(1/\delta))$.

To summarize, it is reasonable to expect the following. If $X \sim \mathsf{N}(0, I_d)$ and the model is well-specified with true parameter $\theta^* = Bu^*$, then, for some *universal* constant C, for all $\delta \in (0,1)$, if $n \geqslant C\,B(d + \log(1/\delta))$, then with probability at least $1 - \delta$: the MLE $\widehat{\theta}_n$ exists and it satisfies

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant C\,\frac{d + \log(1/\delta)}{n}. \tag{1.35}$$

This will hold if we are able to prove that if $n \geqslant C\,B(d + \log(1/\delta))$, (1.33) holds and, letting $\Theta = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant c'/\sqrt{B}\}$ for some absolute constant c', to prove that the uniform lower bound

$$\mathbf{P}\Big(\inf_{\theta\in\Theta, v\in S^{d-1}} \langle \widehat{H}_n(\theta)v, v\rangle \geqslant c\langle Hv, v\rangle\Big) \geqslant 1 - \delta, \tag{1.36}$$

holds for some universal c. We precisely establish this in Chapter 3, Theorem 3.1. In this result, the risk bound (1.35) is optimal in light of the convergence in distribution (1.10) and the tail bound on the chi-squared distribution. We will show that the sample size condition is also optimal, both with respect to δ (Theorem 1.1) and with respect to B and

d jointly. We further discuss this last point in Section 1.5 and in Chapter 4 which presents a non-asymptotic version of the phase transition result of Candès and Sur [CS20].

The uniform lower bound (1.36) is proved in Theorem 3.2 and we emphasize that, despite this being the seemingly simplest case (Gaussian design, well-specified model), Theorem 3.2 is one of the most difficult results of this thesis as it features no unnecessary dependence on B, even logarithmic. Section 1.4.1 provides a detailed sketch of the proof of this result, which can be of independent interest.

1.3.4 Empirical processes, and influence of the assumptions

The sharp result of Theorem 3.1, stated informally in (1.35), is obtained by combining the tight bounds (1.33) on the empirical gradient and the uniform bound (1.36) on empirical Hessians. As briefly discussed, we want to investigate how these bounds are degraded when relaxing the assumptions of Theorem 3.1; first the Gaussian design assumption and second, allowing the model to be misspecified. One of our contributions is precisely to identify sufficient and (almost) necessary conditions on the design X for the MLE to perform almost as well as it would under a Gaussian design assumption, when the model is still well-specified. We gather these assumptions in the definition of *regular* distributions in Chapter 5 (Definition 5.1), also discussed in Section 1.6 of this introduction. Here, we discuss how the various assumptions impact the difficulty of bounding the empirical gradient and Hessians. Indeed, proving the desired bounds (1.24) and (1.25) amounts to bounding the supremum or infimum of some empirical processes whose distributions are defined by the assumptions on X and Y.

As noted before in the short discussion following (1.26), the distribution of empirical Hessians only depends on the design (distribution of X), not on the distribution of Y. In contrast, the distribution of the empirical gradient is highly sensitive to the model specification.

We now highlight the key differences between the well-specified and misspecified cases in terms of bounding the empirical gradient.

Model specification and the empirical gradient. In order to prove (1.24), we have to bound the supremum of a linear process indexed by the unit sphere, as

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \sigma(-Y_i \langle \theta^*, X_i \rangle) H^{-1/2} X_i \right\| = \sup_{v \in S^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \sigma(-Y_i \langle \theta^*, X_i \rangle) \langle v, H^{-1/2} X_i \rangle \,. \tag{1.37}$$

Uniformity can be obtained using standard techniques but we will need to do this very carefully in order to obtain sharp dependence with respect to B in the final bounds (see the next two sections). In particular, the subtle part here is not to obtain uniformity over the sphere, it is to obtain the right pointwise bounds.

Let us first discuss how a well-specified model allows to obtain a sharp bound like (1.33).

Well-specified case. The important observation is the following. Using that

$$\sigma_i = \sigma(-Y_i \langle \theta^*, X_i \rangle) \leqslant 1$$

and that the X_i's are light-tailed actually leads to a suboptimal bound—even in the Gaussian case. We provide more detail on this point in Chapter 2, Section 2.2. There, we explain in particular why a standard sub-Gaussian bound fails to capture the correct

magnitude of the fluctuations. In fact, taking advantage of the fact that σ_i can be very small turns out to be more important than using the light tails of the X_i's. More precisely, the key observation is the following: if $Y_i\langle\theta^*, X_i\rangle \geqslant 0$, then the sigmoid σ_i is bounded as

$$\sigma_i = \sigma(-Y_i\langle\theta^*, X_i\rangle) = \sigma(-|\langle\theta^*, X_i\rangle|) \leqslant \exp(-|\langle\theta^*, X_i\rangle|)\,, \tag{1.38}$$

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

$$\mathbf{P}(Y_i\langle\theta^*, X_i\rangle < 0|X_i) = \sigma(-|\langle\theta^*, X_i\rangle|) \leqslant \exp(-|\langle\theta^*, X_i\rangle|)\,. \tag{1.39}$$

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. As we will see in greater detail in Chapter 2, Section 2.2, the right property to establish is that $H^{-1/2}\nabla\widehat{L}_n(\theta^*)$ is *sub-gamma* with the right parameters in all directions (by which we mean showing that for all $v \in S^{d-1}$, the random variable $\langle v, H^{-1/2}\nabla\widehat{L}_n(\theta^*)\rangle$ is sub-gamma).

We briefly recall this notion here, and refer to [BLM13, §2.4] for an in-depth overview. Let $\sigma^2, b > 0$. A real random variable ξ is (σ^2, b)-sub-gamma if for all $\lambda \in (0, 1/b)$,

$$\mathbf{E}\exp(\lambda\xi) \leqslant \exp\left(\frac{\sigma^2\lambda^2}{2(1-b\lambda)}\right). \tag{1.40}$$

This property is the natural way to ensure that a random variable satisfies Bernstein's inequality. Indeed, if ξ satisfies (1.40), then for every $t \geqslant 0$, one has

$$\mathbf{P}(\xi \geqslant \sigma\sqrt{2t} + bt) \leqslant e^{-t}\,. \tag{1.41}$$

A useful characterization of the sub-gamma property is through the growth of the moments: if for all integer $p \geqslant 2$,

$$\mathbf{E}|\xi|^p \leqslant \frac{\sigma^2}{2} b^{p-2} p!\,, \tag{1.42}$$

then ξ is (σ^2, b)-sub-gamma.

The core argument to obtain a sharp bound on the supremum (1.37) is then twofold.

First, observe that the matrix $H^{-1/2}$ has a very particular structure. Indeed,

$$H^{-1/2} = B^{3/2}u^*u^{*\top} + B^{1/2}(I_d - u^*u^{*\top})\,,$$

so its operator norm is $B^{3/2}$ but the corresponding eigenspace associated to this eigenvalue is only one-dimensional: it is the line $\mathbb{R}u^*$. In any orthogonal subspace, the norm of the restriction of $H^{-1/2}$ is $B^{1/2}$, which is much smaller. This suggests the following decomposition:

$$\begin{aligned}\|H^{-1/2}\nabla\widehat{L}_n(\theta^*)\| &= \sup_{v\in S^{d-1}} \langle v, H^{-1/2}\nabla\widehat{L}_n(\theta^*)\rangle \\ &\leqslant B^{3/2}|\langle u^*, \nabla\widehat{L}_n(\theta^*)\rangle| + B^{1/2}\sup_{\substack{w\in S^{d-1}\\ \langle u^*, w\rangle = 0}} \langle w, \nabla\widehat{L}_n(\theta^*)\rangle\,.\end{aligned} \tag{1.43}$$

We will thus carefully bound $|\langle u^*, \nabla\widehat{L}_n(\theta^*)\rangle|$ and, for any w orthogonal to u^*, $\langle w, \nabla\widehat{L}_n(\theta^*)\rangle$. Then, we will obtain uniformity for the second term in (1.43) using a standard single-scale

approximation (ε-net) argument. We do so by showing that these variables satisfy the sub-gamma property, using the moment characterization (1.42).

The second step, when the model is well-specified, is to properly take advantage of this assumption. We now explain how the observations above, (1.38) and (1.39), allow to do so. For any $v \in S^{d-1}$ and any integer $p \geqslant 2$, one has

$$|\langle v, \nabla \ell(\theta^*, Z)\rangle|^p \leqslant \sigma(-Y\langle \theta^*, X\rangle)^p |\langle v, X\rangle|^p \leqslant \sigma(-Y\langle \theta^*, X\rangle)|\langle v, X\rangle|^p \,. \tag{1.44}$$

We then write (see (1.38))

$$\sigma(-Y\langle \theta^*, X\rangle) \leqslant \sigma(-|\langle \theta^*, X\rangle|) + \mathbf{1}(Y\langle \theta^*, X\rangle < 0) \,. \tag{1.45}$$

Now, using the key remark (1.39) and conditioning on X, we deduce that

$$\mathbf{E}[\sigma(-Y\langle \theta^*, X\rangle)\,|X] \leqslant 2\exp(-|\langle \theta^*, X\rangle|) \,.$$

Finally, taking expectation in (1.44) yields

$$\mathbf{E}\big|\big\langle v, \nabla \ell(\theta^*, Z)\big\rangle\big|^p \leqslant 2\mathbf{E}\big[\exp(-|\langle \theta^*, X\rangle|)|\langle v, X\rangle|^p\big] \,. \tag{1.46}$$

This reduces the problem of bounding the empirical gradient to an inequality only involving X and not Y. Moreover, (1.46) sheds light on a crucial observation: it is the behavior of $\langle u^*, X\rangle$ near 0—and more precisely at a scale of $1/B$—that will lead to optimal bounds. Indeed, similarly to (1.28), we use that the typical value of $\exp(-|\langle \theta^*, X\rangle|)$ is of order e^{-B}. It is only of order 1 when $|\langle \theta^*, X\rangle| \lesssim 1$, which happens with probability $1/B$. Using this observation, we bound

$$\mathbf{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle u^*, X\rangle|^p\big] \quad \text{and} \quad \mathbf{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle w, X\rangle|^p\big] \,, \tag{1.47}$$

for $\langle u^*, w\rangle = 0$, in a way that ensures the right sub-gamma behavior.

To summarize, the key distinction between the well-specified and misspecified cases boils down to whether (1.46) holds or not. When this bound holds, the degradation from Gaussian to *regular* designs (Definition 5.1) is only logarithmic in B, due to the fact that the projections $\langle u^*, X\rangle$ and $\langle w, X\rangle$ with $\langle u^*, w\rangle = 0$ are no longer independent. In the end, in the case of a well-specified model with a Gaussian design, we manage to show that the ideal bound (1.33) holds, which is

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

The detailed derivation can be found in Chapter 3, Section 3.4 for the Gaussian case. We also give a more complete sketch of the proof of this result in Section 1.4. For the regular, well-specified case, see Section 1.7.1 in this introduction and Chapter 6, Section 6.5.1 (Lemma 6.2). The latter uses the small-ball (or margin) condition of Assumption 5.2 that precisely ensures the right behavior of $\langle u^*, X\rangle$ near 0.

Misspecified model. Let us briefly explain how the misspecified case is handled (the detailed treatment is in Chapter 6, Section 6.5.2). When the model is no longer assumed to be well-specified, the bound on the conditional probability of misclasiffication (1.39) does not hold, and neither does (1.46) which reduced the problem to bounding

a function of X only. The decomposition (1.45) though still holds. The first tirm is handled as in the regular, well-specified case where we bounded the moments (1.47). We thus have to circumvent the issue related to the second term and bound

$$\mathbf{E}\big[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\cdot|\langle u^*, X\rangle|^p\big] \quad \text{and} \quad \mathbf{E}\big[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\cdot|\langle w, X\rangle|^p\big]\,, \tag{1.48}$$

still with $\langle u^*, w\rangle = 0$. The main idea is to use that

$$\theta^* \in \arg\min_{\theta\in\mathbb{R}^d} L(\theta) \quad \text{so} \quad \nabla L(\theta^*) = 0\,. \tag{1.49}$$

In this case, instead of bounding the *conditional* probability of misclassification by θ^* (as in (1.39)), we bound the total probability $\mathbf{P}(Y\langle\theta^*, X\rangle < 0)$. Essentially, the idea is to use that $\nabla L(\theta^*) = 0$ to prove an identity involving exponential moments of type $\mathbf{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle v, X\rangle|^p\big]$ that we know how to handle from the previous case. Doing so, we bound the moments (1.48) of order 0, 1, and 2. Note that order 0 is simply $\mathbf{P}(Y\langle\theta^*, X\rangle < 0)$. Then, we rely on an improved control of the moments of a sub-exponential variable (Point 6 of Lemma 8.1) which shows that such a variable is actually sub-gamma, *with the right variance parameter*, which is why we only bounded the moments (1.48) of order up to 2.

Recall that a centered random variable ξ is sub-exponential if the sequence of its L^p norms grows at most linearly with p, namely if there exists $K \geqslant 1$ such that for all $p \geqslant 1$, $(\mathbf{E}|\xi|^p)^{1/p} \leqslant Kp$ (see Definition 8.1 in Chapter 8, or [Ver18, Definition 2.7.5] for equivalent definitions). Such a variable satisfies that $|\xi| \lesssim Kt$ with probability at least $1 - e^{-t}$ for any $t > 0$. It is easy to show that such a variable is $(K^2/2, K/2)$ sub-gamma, and that if $\xi_1, \dots, \xi_n$ are independent copies of ξ, then with probability at least $1 - e^{-t}$,

$$\left|\frac{1}{n}\sum_{i=1}^n \xi_i\right| \leqslant K\sqrt{\frac{t}{n}} + \frac{K}{2}\cdot\frac{t}{n}\,.$$

For large n (*i.e.*, $n \gg t$) the right-hand side of the inequality above is of order $K\sqrt{t/n}$, which is structurally not the right order, because of the central limit theorem. Point 6 of Lemma 8.1 establishes that instead, one has

$$\left|\frac{1}{n}\sum_{i=1}^n \xi_i\right| \leqslant \sigma\sqrt{\frac{2t}{n}} + K\log\Big(\frac{K}{\sigma}\Big)\frac{t}{n}\,,$$

where σ^2 is the actual variance of ξ. When applied to the empirical gradient of the logistic loss, this allows to improve the downstream bound by some power of B.

The final bound on the empirical gradient is given in Proposition 6.2 (Chapter 6). We discuss these ideas in greater detail in Section 1.7 of this introduction. The complete proof of Proposition 6.2 can be found in Chapter 6, Section 6.5.2.

Design assumptions and empirical Hessians. The desired bound (1.25) on empirical Hessians is very different in its structure from previously established results in the non-asymptotic analysis of random matrices [Ver12, Tro12, Tro15]. Indeed, (1.25) is a high-probability lower bound on

$$\inf_{\theta\in\Theta, v\in S^{d-1}} \frac{1}{n}\sum_{i=1}^n \sigma'(\langle\theta, X_i\rangle)\langle v, X_i\rangle^2\,, \tag{1.50}$$

with an unusual non-linearity due to σ'. If there was no infimum with respect to θ in (1.50), the task at hand would be to control the smallest eigenvalue of an empirical covariance matrix ($\widehat{H}_n(\theta)$ for a fixed θ). This can be done using Oliveira's bound [Oli16, Theorem 3.1]. In fact, it is not very difficult to see that for each θ, the random matrix

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

satisfies the L^4/L^2 norm equivalence assumed in [Oli16, Theorem 3.1]. A careful computation of the parameter in this equivalence enables us to deduce from this result the following pointwise bound. Fix $\theta \in \mathbb{R}^d$ and let $\delta \in (0,1)$. Then, if $n \geqslant C\|\theta\|(d+\log(1/\delta))$ (for some constant $C > 0$), one has

$$\mathbf{P}\left(\widehat{H}_n(\theta) \succcurlyeq \left[1 - C\sqrt{\frac{\|\theta\| \vee 1(d+\log(1/\delta))}{n}}\right] H(\theta)\right) \geqslant 1-\delta\,. \tag{1.51}$$

There is therefore an additional difficulty due to the desired uniformity with respect to θ, which furthermore appears through a non-polynomial dependence because of σ'. In comparison, (1.51) would be enough when working with the square loss rather than the logistic loss since the Hessian of the square risk is constant. In fact, [Oli16] is motivated by the analysis of random-design linear regression. This result was refined in [Mou22] to obtain non-vacuous bounds in the regime $\delta \to 0$, while [Zhi24] obtained a dimension-free version of [Oli16, Theorem 3.1].

Our task can thus be seen as providing a uniform version of (1.51) with respect to θ. This pointwise bound follows easily from Oliveira's inequality after a careful computation of a certain parameter; the major challenge here is to obtain an non-standard uniform bound on a collection of random matrices.

We emphasize again that the Hessian of the logistic loss for any (x, y) does not depend on y, thus for every θ, the distribution of the empirical Hessian $\widehat{H}_n(\theta)$ only depends on the distribution of $X_1, \dots, X_n$. In particular, the task of bounding the Hessians does not depend on whether the model is well-specified or not, as opposed to the empirical gradient. We will therefore only distinguish between the cases of Gaussian and regular designs.

This distinction calls for some comments. The Gaussian design is a special case of regular design, and as such, any result that holds for regular designs would hold in particular for the Gaussian design. The reason why we consider it separately is because in this case we obtain sharper guarantees, involving universal constants rather than polylogarithmic factors in B, both in the condition on the sample size and in the definition of the neighborhood Θ. Let us further explain the distinction.

In both cases, we prove the following: for n large enough (depending on B, d, t),

$$\mathbf{P}\left(\forall \theta \in \Theta,\ \widehat{H}_n(\theta) \succcurlyeq c_0 H\right) \geqslant 1 - e^{-t}\,, \tag{1.52}$$

where

$$\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant \frac{c_1}{\sqrt{B}}\right\}, \tag{1.53}$$

for some constants c_0, c_1 that should not depend (or weakly depend) on n, B, d, where the matrix H is the proxy for $\nabla^2 L(\theta^*)$ defined in (1.30).

In the Gaussian case, we prove that (1.52) holds as soon as $n \geqslant CB(d + \log(1/\delta))$ on the neighborhood (1.68) with c_1 being indeed an absolute constant. This proves the ideal bound (1.36) discussed earlier, thereby establishing that (1.51) does hold uniformly with respect to θ, over a region Θ that is as large as one could hope for. This is done in Theorem 3.2, restated as Theorem 1.2 in Section 1.4.1. Our proof relies on the so-called PAC-Bayesian inequality, which allows to control suprema of "smoothed" processes. Section 1.4.1 provides a very detailed sketch of the proof of Theorem 3.2, which is one of the main contributions of this thesis. In short, we actually apply the PAC-Bayes inequality to an auxiliary process whose (negative) smoothed version *is* the process of interest, that is

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

and with non-usual smoothing distributions that provide an optimal bound with no unnecessary dependence (even logarithmic) on B.

In the regular case (Theorem 6.3), the guarantees are slightly degraded, in the sense that we require $n \geqslant CB(d\log B + t)$ for (1.52) to hold on a region Θ that is slightly smaller than what it was in the Gaussian case, more precisely, it can still be written as (1.68), only this time c_1 is not an absolute constant, it is instead of order $1/\log B$. The proof of Theorem 6.3 is simpler than that of Theorem 3.2, after observing that the design must satisfy a certain property regarding the marginal distribution of its two-dimensional projections. This property is satisfied by the Gaussian distribution, and it is what motivates Assumption 5.3 in the definition of regular designs in Chapter 5 (Definition 5.1), see also Section 1.6.1 in this introduction.

Using this property, we relate the quadratic forms associated with the empirical hessians to empirical frequencies of events belonging to a certain Vapnik-Chervonenkis (VC) class. Then, we control the infimum of these empirical frequencies using tools from empirical processes theory. A key step in doing so is that the VC class of interest enjoys some sort of product structure with respect to the class of halfspaces. Its VC dimension thus remains bounded as $O(d)$.

Let us sketch the idea very briefly. This time, we directly bound

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

by a constant factor of $\langle Hv, v\rangle$. To do so, we observe that for every $v \in S^{d-1}$,

$$\langle Hv, v\rangle = \frac{\langle u^*, v\rangle^2}{B^3} + \frac{1-\langle u^*, v\rangle^2}{B} \lesssim \frac{1}{B}\cdot\max\left\{\frac{1}{B}, \|u^* - v\|\right\}^2.$$

Then, we use that for every $\theta \in \Theta$, $\sigma'(\langle\theta, X_i\rangle) \gtrsim \mathbf{1}(|\langle\theta, X_i\rangle| \leqslant 1)$, hence (letting $u = \theta/\|\theta\|$),

$$\begin{aligned}
&\sigma'(\langle\theta, X_i\rangle)\langle v, X_i\rangle^2\\
&\quad\gtrsim \mathbf{1}(|\langle\theta, X_i\rangle| \leqslant 1)\langle v, X_i\rangle^2\\
&\quad\gtrsim \mathbf{1}\Big(|\langle u, X_i\rangle| \leqslant \frac{c'}{B}; |\langle v, X_i\rangle| \geqslant \frac{\max\{B^{-1}, \|u^* - v\|\}}{c'}\Big)\max\left\{\frac{1}{B}, \|u^* - v\|\right\}^2\\
&\quad\gtrsim \mathbf{1}\Big(|\langle u, X_i\rangle| \leqslant \frac{c'}{B}; |\langle v, X_i\rangle| \geqslant \frac{\max\{B^{-1}, \|u^* - v\|\}}{c'}\Big)B\langle Hv, v\rangle\,.
\end{aligned}$$

This roughly explains the motivation of Assumption 5.3. Indeed, if the event above has probability at least c''/B for every $v \in S^{d-1}$, and that the empirical frequencies of these events are uniformly close to their expectation, this will show the desired bound

$$\inf_{\theta,v} \frac{1}{n} \sum_{i=1}^{n} \sigma'(\langle \theta, X_i \rangle) \langle v, X_i \rangle^2 \gtrsim \langle Hv, v \rangle \,.$$

In a nutshell, the differences between the cases of Gaussian and regular designs are the following: in the Gaussian case, the proof uses very technical, non-standard arguments to bound the process of interest. In the regular case, the difficulty is to relate the problem to more classical inequalities in empirical processes theory. As such, the identification of the *two-dimensional margin condition* (Assumption 5.3) as an important structural property constitutes one of our contributions.

1.4 Sharp bounds in the Gaussian, well-specified setting (Chapter 3)

Here we consider what we earlier referred to as the ideal case, that is when the design is Gaussian and the logit model is well-specified. This section provides an overview of Chapter 3, whose main result is Theorem 3.1, reproduced below. This result provides optimal guarantees (up to absolute constants) for the existence and accuracy of the MLE.

Theorem 1.1 (Theorem 3.1, Chapter 3)**.** *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 $\delta \in (0,1)$: if*

$$n \geqslant c_1 B(d + \log(1/\delta)) \,, \tag{1.54}$$

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

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c_2 \frac{d + \log(1/\delta)}{n} \,. \tag{1.55}$$

Moreover, for any $d \geqslant 70$ and $\delta \in (0, 1/2]$, if $n \leqslant c_3 B(d + \log(1/\delta))$ then

$$\mathbf{P}(\text{MLE exists}) \leqslant 1 - \delta \,. \tag{1.56}$$

This result removes a $\log n$ factor from the bound (1.16) deduced from the work of [KvdG23] in the case of a probit model. We note also that the proof of Theorem 1.1 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 $1 - e^{-t}$,

$$\left\| \frac{\widehat{\theta}_n}{\|\widehat{\theta}_n\|} - \frac{\theta^*}{\|\theta^*\|} \right\| \leqslant c_4 \sqrt{\frac{d+t}{Bn}} \,, \qquad \left| \|\widehat{\theta}_n\| - \|\theta^*\| \right| \leqslant c_4 \sqrt{\frac{B^3(d+t)}{n}} \,. \tag{1.57}$$

As such, it answers in the affirmative a question from [HM24] 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 (1.11) 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 (1.12) from Candès and Sur [CS20]. On the other hand, in the regime where $n \ll Bd$, Theorem 3.1 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 (1.12). We address this question in Chapter 4 where we show that in this regime $n \ll Bd$, not only is the probability of existence of the MLE bounded away from 1, it is actually close to 0. We formally establish this in Theorem 4.2 which can be seen as a quantitative finite-sample version of the phase transition (1.12) from [CS20]. We give more detail on this point in Section 1.5 of this introduction.

1.4.1 Sketch of the proof of Theorem 1.1

The proof of Theorem 1.1 follows the structure outlined in Section 1.3.3, and thus relies on a bound on the empirical gradient and a uniform lower bound on empirical Hessians. We assume throughout that $\|\theta^*\| \geqslant e$, so that $B = \|\theta^*\|$. The low-signal regime is covered by Theorem 1.7. Recall the proxy for $\nabla^2 L(\theta^*)$,

$$H = \frac{1}{B^3} u^* u^{*\top} + \frac{1}{B}(I_d - u^* u^{*\top}), \quad u^* = \frac{\theta^*}{\|\theta^*\|}. \tag{1.58}$$

The first step to prove Theorem 1.1 is the following result, which is an optimal deviation bound for the empirical gradient.

Proposition 1.1 (Proposition 3.1, Chapter 3)**.** *Assume that X is a standard Gaussian vector and the model is well-specified. Let H be the matrix defined in (1.58). There exists an absolute constant C such that for any $t > 0$, if $n \geqslant C\,B(d+t)$ then with probability at least $1 - e^{-t}$,*

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

Sketch of proof. The proof follows the ideas exposed in Section 1.3.4. We start from the decomposition

$$\|H^{-1/2}\nabla \widehat{L}_n(\theta^*)\| \leqslant B^{3/2}|\langle u^*, \nabla \widehat{L}_n(\theta^*)\rangle| + B^{1/2} \sup_{\substack{w \in S^{d-1} \\ \langle u^*, w\rangle = 0}} \langle w, \nabla \widehat{L}_n(\theta^*)\rangle. \tag{1.59}$$

We now bound the moments of $|\langle u^*, \nabla \widehat{L}_n(\theta^*)\rangle|$ and $|\langle w, \nabla \widehat{L}_n(\theta^*)\rangle|$ for any w orthogonal to u^* to show that these variables satisfy the sub-gamma property.

For any $v \in S^{d-1}$ and any integer $p \geqslant 2$, one has

$$|\langle v, \nabla \widehat{L}_n(\theta^*)\rangle|^p \leqslant \sigma(-Y\langle \theta^*, X\rangle)^p |\langle v, X\rangle|^p \leqslant \sigma(-Y\langle \theta^*, X\rangle)|\langle v, X\rangle|^p. \tag{1.60}$$

We then write (see (1.38))

$$\sigma(-Y\langle \theta^*, X\rangle) \leqslant \sigma(-|\langle \theta^*, X\rangle|) + \mathbf{1}(Y\langle \theta^*, X\rangle < 0). \tag{1.61}$$

Now, using the key remark that $\mathbf{P}(Y\langle\theta^*, X\rangle < 0\,|X) \leqslant \exp(-|\langle\theta^*, X\rangle|)$ and conditioning on X, we deduce that

$$\mathbf{E}[\sigma(-Y\langle\theta^*, X\rangle)\,|X] \leqslant 2\exp(-|\langle\theta^*, X\rangle|)\,.$$

Finally, taking expectation in (1.44) yields

$$\mathbf{E}\big|\big\langle v, \nabla\ell(\theta^*, Z)\big\rangle\big|^p \leqslant 2\mathbf{E}\big[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p\big]\,. \tag{1.62}$$

This reduces the problem of bounding the empirical gradient to an inequality only involving X and not Y. Using this observation, we bound

$$\mathbf{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle u^*, X\rangle|^p\big] \quad \text{and} \quad \mathbf{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle w, X\rangle|^p\big]\,, \tag{1.63}$$

for $\langle u^*, w\rangle = 0$, in a way that ensures the right sub-gamma behavior. The typical value of $\exp(-|\langle\theta^*, X\rangle|)$ is of order e^{-B} and it is of order 1 only when $|\langle\theta^*, X\rangle| \lesssim 1$, which happens with probability $1/B$. Thus, the first expectation in (1.63) will scale as $p!/B^{p+1}$; and by independence, the second will be of order $p!/B$. More precisely, if $\xi \sim \mathsf{N}(0,1)$ and $\beta \geqslant 1$,

$$\mathbf{E}\big[\exp(-\beta|\xi|)|\xi|^p\big] \leqslant \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} |x|^p \exp(-\beta|x|)\mathrm{d}x = \sqrt{\frac{2}{\pi}}\frac{p!}{\beta^{p+1}}\,. \tag{1.64}$$

From there, if $\langle u^*, w\rangle = 0$, we use the independence of $\langle u^*, X\rangle$ and $\langle w, X\rangle$ and classical estimates for the moments of the standard Gaussian distribution,

$$\mathbf{E}\big[\exp(-|\langle\theta^*, X\rangle|)|\langle w, X\rangle|^p\big] = \mathbf{E}[\exp(-|\langle\theta^*, X\rangle|)]\mathbf{E}|\langle w, X\rangle|^p \leqslant \frac{p!}{B}\,. \tag{1.65}$$

In the direction of u^*, using (1.64),

$$\mathbf{E}\big[\exp(-|\langle\theta^*, X\rangle|)|\langle u^*, X\rangle|^p\big] \leqslant \frac{p!}{B^{p+1}}\,.$$

From now on and until the end of this section on the Gaussian design setting, $c, c', C\ldots$ will denote absolute, positive constants whose value may change from one instance to the other. We deduce from the previous inequality and (1.62) that for all $p \geqslant 2$,

$$\mathbf{E}|\langle u^*, \nabla\ell(\theta^*, Z)\rangle|^p \leqslant \frac{c}{B^3}\cdot\frac{p!}{B^{p-2}}\,,$$

thus by Bernstein's inequality (1.41), with probability at least $1 - e^{-t}$,

$$|\langle u^*, \nabla\widehat{L}_n(\theta^*)\rangle| \leqslant c\sqrt{\frac{t}{B^3 n}} + c\frac{t}{Bn} \leqslant c'\sqrt{\frac{t}{B^3 n}}\,, \tag{1.66}$$

where the last inequality follows from the condition $n \geqslant CBt$. We now deal with the uniform bound on the subspace $(\mathbb{R}u^*)^\perp$. The bound (1.65) (combined with (1.62)) shows that for every $w \in S^{d-1}$ orthogonal to u^*, $\langle w, \nabla\ell(\theta^*, Z)\rangle$ is $(c/B, c)$-sub-gamma, thus with probability at least $1 - e^{-t}$,

$$|\langle w, \nabla\widehat{L}_n(\theta^*)\rangle| \leqslant c\sqrt{\frac{t}{Bn}} + c\frac{t}{n}\,.$$

Then, taking a union bound over a 1/2-net of S^{d-2} whose cardinality is at most 5^d (see [Ver18, Lemma 4.2.13]), we find that with probability at least $1 - e^{-t}$,

$$\sup_{\substack{w \in S^{d-1} \\ \langle u^*, w \rangle = 0}} \langle w, \nabla \widehat{L}_n(\theta^*) \rangle \leqslant c \left[\sqrt{\frac{d+t}{Bn}} + \frac{d+t}{n} \right] \leqslant C \frac{d+t}{n} , \tag{1.67}$$

since $n \geqslant CB(d+t)$. Plugging (1.66) and (1.67) in (1.59) proves the claim. ■

The second step in the proof of Theorem 1.1 is a uniform lower bound on the empirical Hessians.

Theorem 1.2 (Theorem 3.2, Chapter 3)**.** *Assume that* $X \sim \mathsf{N}(0, I_d)$. *There exist* universal *constants* $c_0, c, C > 0$ *such that the following holds. Let*

$$\Theta = \left\{ \theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant \frac{c_0}{\sqrt{B}} \right\} . \tag{1.68}$$

For any $t > 0$, *if* $n \geqslant CB(d+t)$, *then with probability at least* $1 - e^{-t}$, *simultaneously for all* $\theta \in \Theta$,

$$\widehat{H}_n(\theta) \succcurlyeq cH .$$

This result is one of the most important contributions of this thesis, and can be of independent interest. Its proof relies on the so-called PAC-Bayesian inequality and combines some of the most technically involved arguments of this manuscript. The PAC-Bayesian, or variational, inequality allows to control the smoothed version of a stochastic process. It can be viewed as a more flexible version of the standard single-scale approximation (ε-net) argument. In some cases, it allows to recover results otherwise only known using generic chaining [Tal96, Tal21], such as the Gaussian complexity of ellipsoids [Zhi24, Example 2]. This method was pioneered by Catoni, Giulini, Audibert [AC10, AC11, Cat16, CG17] and McAllester [McA99], in the context of Bayesian learning and robust statistics. It was later used to control the lower tail of empirical covariance matrices [Oli16, Mou22] and to obtain two-sided bounds for random matrices [Zhi24].

PAC-Bayes inequality. The PAC-Bayes inequality relies on the variational representation of the logarithmic Laplace transform (also known as duality formula for relative entropy). Given two probability measures ν and μ on a same measurable space, such that ν is dominated by μ, we denote by $D(\nu \| \mu)$ the Kullback-Leibler divergence from ν to μ, that is

$$D(\nu \| \mu) = \int \log \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \right) \mathrm{d}\nu . \tag{1.69}$$

The variational representation is the following identity. Given a reference measure μ and a measurable function f, one has

$$\log \int e^f \mathrm{d}\mu = \sup_{\nu \ll \mu} \left\{ \int f \mathrm{d}\nu - D(\nu \| \mu) \right\} , \tag{1.70}$$

where the supremum spans all measures ν dominated by μ.

We now state the PAC-Bayesian inequality and give a simple example of its application, before moving on to the sketch of proof of Theorem 1.2.

Lemma 1.1 (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 $Z_1, \dots, Z_n$ be independent copies of the process Z. Let also $\lambda > 0$ be such that $\mathbf{E}\exp(\lambda Z(\omega)) < \infty$ 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,*

$$\frac{1}{n}\sum_{i=1}^{n}\int_E Z_i(\omega)\rho(\mathrm{d}\omega) \leqslant \frac{1}{\lambda}\int_E \log\Big(\mathbf{E}\big[e^{\lambda Z(\omega)}\big]\Big)\rho(\mathrm{d}\omega) + \frac{D(\rho\|\pi) + t}{\lambda n}\,. \tag{1.71}$$

In this lemma, π is referred to as the "prior" distribution, and the distributions ρ as "posteriors", by analogy with Bayesian statistics. These are actually *probability* distributions; however, to avoid any confusion when using this lemma, we will write expectations with respect to these measures as explicit integrals instead of the expectation symbol $\mathbf{E}$, that we reserve for integration with respect to the distribution of the variables $Z(\omega)$.

We give below a short sketch of the proof of this result. By independence, the tensorization is straightforward so we sketch the proof for a single process $(Z(\omega))_{\omega\in E}$.

Sketch of proof of PAC-Bayes inequality. Let $\lambda > 0$ be such that $\mathbf{E}\exp(\lambda Z(\omega)) < \infty$ for all ω and let $\psi(\lambda, \omega) = \log(\mathbf{E}[\exp(\lambda Z(\omega))])$. Let also

$$Z'(\omega, \lambda) = \lambda Z(\omega) - \psi(\lambda, \omega)\,.$$

Then by definition, for all ω, $\mathbf{E}[e^{Z'(\omega,\lambda)}] = 1$, so by Fubini's theorem,

$$\mathbf{E}\Big[\int e^{Z'(\omega,\lambda)}\pi(\mathrm{d}\omega)\Big] = 1\,. \tag{1.72}$$

By the variational representation (1.70),

$$\int e^{Z'(\omega,\lambda)}\pi(\mathrm{d}\omega) = \exp\left(\sup_{\rho\ll\pi}\Big\{\int Z'(\omega,\lambda)\rho(\mathrm{d}\omega) - D(\rho\|\pi)\Big\}\right).$$

Therfore, combining (1.72) and Markov's inequality, we deduce that for all $t > 0$,

$$\begin{aligned}\mathbf{P}\left(\sup_{\rho\ll\pi}\Big\{\int Z'(\omega,\lambda)\rho(\mathrm{d}\omega) - D(\rho\|\pi)\Big\} > t\right)\\ \leqslant \mathbf{E}\left[\exp\left(\sup_{\rho\ll\pi}\Big\{\int Z'(\omega,\lambda)\rho(\mathrm{d}\omega) - D(\rho\|\pi)\Big\}\right)\right]\cdot e^{-t} = e^{-t}\,.\end{aligned}$$

Finally, using that

$$\int Z'(\omega,\lambda)\rho(\mathrm{d}\omega) = \lambda\int Z(\omega)\rho(\mathrm{d}\omega) - \int\psi(\lambda,\omega)\rho(\mathrm{d}\omega)\,,$$

the last inequality can be equivalently stated as: with probability at least $1 - e^{-t}$, *simultaneously* for all ρ dominated by π,

$$\int Z(\omega)\rho(\mathrm{d}\omega) \leqslant \frac{1}{\lambda}\left[\int\psi(\lambda,\omega)\rho(\mathrm{d}\omega) + D(\rho\|\pi) + t\right]. \qquad \blacksquare$$

This result provides a particularly flexible tool to control certain empirical processes, and in particular those indexed by ellipsoids.

A simple example using PAC-Bayes inequality. Let us illustrate the use of the PAC-Bayesian inequality with a simple example, where we bound the norm of the sum of anisotropic sub-Gaussian vectors. This is a well-known result, we only include it to illustrate the use of the PAC-Bayesian inequality.

Proposition 1.2. *Let Γ be a $d \times d$ positive, symmetric matrix and let ξ be a random vector in $\mathbb{R}^d$ such that for all $v \in S^{d-1}$ and all $\lambda > 0$,*

$$\mathbf{E} \exp(\lambda \langle v, \xi \rangle) \leqslant \exp(\lambda^2 \langle \Gamma v, v \rangle) .$$

Let $\xi_1, \dots, \xi_n$ be independent copies of ξ. Then for all $t > 0$, with probability greater than $1 - e^{-t}$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\| \leqslant 2 \sqrt{\frac{\mathrm{Tr}(\Gamma) + 2\|\Gamma\|_{\mathrm{op}} t}{n}} .$$

Proof. Let $Z(\theta) = \langle \theta, \xi \rangle$ for every $\theta \in \mathbb{R}^d$ and let $t > 0$. For all $\lambda > 0$,

$$\log(\mathbf{E}[\exp(\lambda Z(\theta))]) \leqslant \langle \Gamma\theta, \theta \rangle .$$

Writing $S_n = \sum_{i=1}^n \xi_i$, we want to apply the PAC-Bayes inequality to $Z_1(\theta) + \cdots + Z_n(\theta) = \langle \theta, S_n \rangle$ in order to bound $\|S_n\| = \sup_{v \in S^{d-1}} \langle v, S_n \rangle$. As the process of interest is linear, a natural choice of smoothing distributions is a collection $(\rho_v)_{v \in S^{d-1}}$ such that for all $v \in S^{d-1}$, ρ_v is centered at v. Let therefore $\rho_v = \mathsf{N}(v, \beta^{-1} I_d)$ for some $\beta > 0$ to be chosen later, and let $\pi = \mathsf{N}(0, \beta^{-1} I_d)$ be the prior distribution. It is clear that for all $\lambda > 0$ and all $v \in S^{d-1}$,

$$\int_{\mathbb{R}^d} \langle \theta, S_n \rangle \rho_v(\mathrm{d}\theta) = \langle v, S_n \rangle ,$$

hence, by Lemma 3.2, it holds with probability greater than $1 - e^{-t}$ that

$$\lambda \langle v, S_n \rangle \leqslant n\lambda^2 \int_{\mathbb{R}^d} \langle \Gamma\theta, \theta \rangle \rho_v(\mathrm{d}\theta) + D(\rho_v \| \pi) + t ,$$

simultaneously for all $v \in S^{d-1}$. By standard computations, for all v, $D(\rho_v \| \pi) = \beta/2$ and

$$\int_{\mathbb{R}^d} \langle \Gamma\theta, \theta \rangle \rho_v(\mathrm{d}\theta) = \langle \Gamma v, v \rangle + \frac{\mathrm{Tr}(\Gamma)}{\beta} .$$

By Cauchy-Schwarz inequality $\langle \Gamma v, v \rangle \leqslant \|\Gamma\|_{\mathrm{op}} \|v\|^2 = \|\Gamma\|_{\mathrm{op}}$ so

$$\lambda \|S_n\| \leqslant n\lambda^2 \Big(\|\Gamma\|_{\mathrm{op}} + \frac{\mathrm{Tr}(\Gamma)}{\beta} \Big) + \frac{\beta}{2} + t ,$$

and, rearranging,

$$\|S_n\| \leqslant n\lambda \|\Gamma\|_{\mathrm{op}} + \frac{t}{\lambda} + \frac{n\lambda \mathrm{Tr}(\Gamma)}{\beta} + \frac{\beta}{2\lambda} . \tag{1.73}$$

Now, optimizing the right-hand-side we get

$$\inf_{\lambda > 0} n\lambda \|\Gamma\|_{\mathrm{op}} + \frac{2t}{\lambda} = 2\sqrt{n \|\Gamma\|_{\mathrm{op}} t} , \tag{1.74}$$

and

$$\inf_{\lambda, \beta > 0} \Big\{ \frac{n\lambda \mathrm{Tr}(\Gamma)}{\beta} + \frac{\beta}{2\lambda} \Big\} = \sqrt{2n \mathrm{Tr}(\Gamma)} .$$

Plugging this in (1.73), we find that

$$\|S_n\| \leqslant \sqrt{2n\mathrm{Tr}(\Gamma)} + 2\sqrt{n\|\Gamma\|_{\mathrm{op}}t} \leqslant 2\sqrt{n\big(\mathrm{Tr}(\Gamma) + 2\|\Gamma\|_{\mathrm{op}}t\big)}\,,$$

where the last inequality uses the concavity of the square root. ■

In the previous simple example, we were able to find smoothing distributions such that the smoothed process was exactly the process of interest. This is not the case in general, and if we want to bound a variable of the form

$$\sup_{t\in T} Z(t)\,,$$

we want to choose smoothing distributions ρ_t, $t \in T$, such that

$$Z(t) \lesssim \int_T Z(\theta)\rho_t(\mathrm{d}\theta)\,,$$

and that this holds uniformly, that is

$$\sup_{t\in T} Z(t) \leqslant C \sup_{t\in T} \int_T Z(\theta)\rho_t(\mathrm{d}\theta)\,, \tag{1.75}$$

ideally for some absolute constant C. A simple way to ensure this is to choose very concentrated smoothing distributions; but in doing so, the "complexity" term $D(\rho_t\|\pi)$ will likely increase. Thus, when using the PAC-Bayes inequality, a trade-off must be made between complexity and precision of the approximation.

Major challenges of Theorem 1.2, and a sketch of proof. Let us highlight the major difficulties that will arise in the proof of Theorem 1.2. We want to bound from below

$$\big\langle H^{-1/2}H_n(\theta)H^{-1/2}v, v\big\rangle = \frac{1}{n}\sum_{i=1}^n \sigma'(\langle\theta, X_i\rangle)\big\langle H^{-1/2}v, X_i\big\rangle^2 \geqslant c \tag{1.76}$$

for some absolute constant c, with high probability, uniformly over $\theta \in \Theta$ and $v \in S^{d-1}$. We do so by resorting to the PAC-Bayes inequality with a suitable choice of prior distribution π and posterior distributions $\rho_{\theta,v}$ for $(\theta, v) \in \Theta \times S^{d-1}$.

The main difficulty is to obtain a bound that does not feature any unnecessary dependence with respect to B. Our uniform bound on empirical Hessians in the regular case is suboptimal only because of an extra $\log B$ factor, making the desired improvement very subtle. We thus need to be extremely careful with the choice of prior and posterior distributions to avoid this. Let us explain what could lead to a potential superfluous dependence scaling as $\log B$. The smoothing distributions usually have a typical "radius" $\varepsilon > 0$, analogue of the approximation scale in the classic ε-net (union bound) technique. To illustrate this idea, think of the uniform distribution on a Euclidean ball of radius ε, or the Gaussian distribution with covariance $\varepsilon^2 I_d$.

In the PAC-Bayes inequality, the delicate issue mainly arises from the "complexity" term $D(\rho_{\theta,v}\|\pi)$ which needs to be uniformly bounded. In particular, the support of the prior π has to be large enough to contain those of all posteriors distributions (to ensure domination). But, for a typical smoothing radius ε, the ratio of posterior and prior densities has a tendency to scale as $(c/\varepsilon)^d$, which leads to a divergence of order

$O(d\log(1/\varepsilon))$. In our case, one of the major difficulties is to ensure that this term remains bounded by $O(d)$, which means that ε has to be an absolute constant, and cannot scale as a power of $1/B$. Given the highly anisotropic shape of the ellipsoid Θ (1.68), we will have to choose anisotropic smoothing and prior distributions in order to capture the geometry of Θ without leading to a too large divergence term.

We now give a detailed sketch of the proof of Theorem 1.2.

Sketch of the proof of Theorem 1.2. The first important idea will be to apply the PAC-Bayes inequality (with the index set $\Theta \times S^{d-1}$) to an auxiliary process $Z(\theta, v)$ which is such that for all $\theta \in \Theta$ and $v \in S^{d-1}$,

$$\sigma'(\langle \theta, X\rangle)\langle H^{-1/2}v, X\rangle^2 \geqslant c \iint_{\mathbb{R}^d \times S^{d-1}} Z(\theta', v')\rho_{\theta,v}(\mathrm{d}\theta', \mathrm{d}v')\,, \tag{1.77}$$

for a suitable choice of posterior distributions $\rho_{\theta,v}$ for $(\theta, v) \in \Theta \times S^{d-1}$. This way, we avoid having to control an error term between the process of interest and its smoothed version, since the process of interest is itself the smoothed one. We therefore have to control the Laplace transform of the auxiliary process in order to apply PAC-Bayes inequality.

We observed earlier that $\sigma'(t) \geqslant c\mathbf{1}(|t| \leqslant 1)$ for all real t. We will also need to restrict ourselves to the high-probability event where $\|X_i\| \leqslant C\sqrt{d}$, for some absolute constant $C \geqslant 2$. This suggests defining for all $\theta \in \mathbb{R}^d$ and $v \in S^{d-1}$,

$$Z(\theta, v) = \mathbf{1}\big(|\langle \theta, X\rangle| \leqslant 1, \|X\| \leqslant C\sqrt{d}\big)\langle H^{-1/2}v, X\rangle^2\,. \tag{1.78}$$

We will later show that this process satisfies (1.77) for the collection of posterior distributions that we will introduce in due time. For now, we simply mention that as $Z(\theta, v)$ is a product of a function of θ and a function of v, it is natural to choose $\rho_{\theta,v}$ as a product measure $\rho_\theta \otimes \rho_v$.

Applying PAC-Bayes inequality to $-Z(\theta, v)$, we find that for all $\lambda > 0$, it holds with probability at least $1 - e^{-t}$, simultaneously for all $(\theta, v) \in \mathbb{R}^d \times S^{d-1}$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \iint_{\Theta\times S^{d-1}} &-\lambda Z(\theta', v')\rho_\theta(\mathrm{d}\theta')\rho_v(\mathrm{d}v') \\ &\leqslant \iint_{\Theta\times S^{d-1}} \log(\mathbf{E}[e^{-\lambda Z(\theta', v')}])\rho_\theta(\mathrm{d}\theta')\rho_v(\mathrm{d}v') + \frac{D(\rho_{\theta,v}\|\pi) + t}{n}\,.\end{aligned} \tag{1.79}$$

Now, by convexity, for all θ, v,

$$\mathbf{E}[e^{-\lambda Z(\theta', v')}] \leqslant 1 - \lambda \mathbf{E}Z(\theta, v) + \frac{\lambda^2}{2}\mathbf{E}Z(\theta, v)^2\,.$$

Then, we show that $\mathbf{E}Z(\theta, v)$ is bounded from below by an absolute constant, and that a condition of L^4/L^2 moments equivalence is satisfied, as discussed in Section 1.3.4, namely that $\mathbf{E}Z(\theta, v)^2 \leqslant CB$. This will show that

$$\mathbf{E}[e^{-\lambda Z(\theta', v')}] \leqslant 1 - c\lambda + CB\lambda^2\,. \tag{1.80}$$

Then, taking the logarithm, integrating, plugging in (1.79) and dividing by $\lambda > 0$, this will show that

$$\frac{1}{n}\sum_{i=1}^n \iint_{\Theta\times S^{d-1}} -Z(\theta', v')\rho_\theta(\mathrm{d}\theta')\rho_v(\mathrm{d}v') \leqslant -c + C\lambda B + \frac{D(\rho_{\theta,v}\|\pi) + t}{\lambda n}\,.$$

Therefore, if for all θ, v, it holds that $D(\rho_{\theta,v}\|\pi) \leqslant C'd$, optimizing over λ in the right hand side of the last inequality will lead to the right term, as

$$\inf_{\lambda>0}\Big\{\lambda B + \frac{d+t}{\lambda n}\Big\} = 2\sqrt{\frac{B(d+t)}{n}}\,.$$

Finally, combining with the smoothing inequality (1.77) will show that

$$\frac{1}{n}\sum_{i=1}^{n}\sigma'(\langle\theta, X_i\rangle)\langle H^{-1/2}v, X_i\rangle^2 \geqslant c - C\sqrt{\frac{B(d+t)}{n}}\,. \tag{1.81}$$

The desired result thus follows as $n \geqslant C'B(d+t)$, with C' an absolute constant.

Let us briefly sketch how to obtain the bound on the Laplace transform (1.80). For the first term, we have to bound from below

$$\mathbf{E}\big[\mathbf{1}\big(|\langle\theta, X\rangle| \leqslant 1, \|X\| \leqslant C\sqrt{d}\big)\langle H^{-1/2}v, X\rangle^2\big]\,.$$

This is equivalent to showing that the matrices $\widetilde{H}(\theta) = \mathbf{E}\big[\mathbf{1}\big(|\langle\theta, X\rangle| \leqslant 1, \|X\| \leqslant C\sqrt{d}\big)XX^\top\big]$ uniformly satisfy $\widetilde{H}(\theta) \succcurlyeq cH$. This follows by a rotation invariance argument and one-dimensional Gaussian estimates. For the second term, it uses two arguments: first, that the population Hessians $H(\theta)$ are uniformly close to H on Θ, and second, the exponential moment inequalities

$$\mathbf{E}[e^{-B|\xi|}|\xi|^k] \leqslant \frac{k!}{B^{k+1}}\,, \quad \xi \sim \mathsf{N}(0,1)\,.$$

We thus obtain that $\mathbf{E}Z(\theta, v)^2 = O(B)$.

The structure of the proof is now clear, and the technically difficult parts are now to show that for a proper choice of posterior distributions, the smoothing inequality (1.77) and the bound on the relative entropy $D(\rho_{\theta,v}\|\pi) \lesssim d$ hold for all parameter θ in the region Θ and all v on the unit sphere. We now define the prior and posterior distributions.

Prior and posterior distributions. The prior and all posteriors are defined as product measures in the following way.

Prior distribution: truncated anisotropic Gaussian. We write the prior as $\pi = \pi_\Theta \otimes \pi_S$, where π_S is the uniform distribution on the unit sphere and π_Θ is a truncated Gaussian. Such distributions are defined as follows: let γ denote a Gaussian measure on $\mathbb{R}^d$, and let S be a subset of $\mathbb{R}^d$. The corresponding truncated Gaussian distribution $\overline{\gamma}$ is defined through its density as

$$\mathrm{d}\overline{\gamma}(\theta) = \frac{\mathbf{1}(\theta \in S)}{\gamma(S)}\mathrm{d}\gamma(\theta)\,, \quad \text{for all } \theta \in \mathbb{R}^d\,. \tag{1.82}$$

The prior π_Θ should have the same shape as Θ itself, and be large enough to contain Θ itself as well as all the supports of the posteriors ρ_θ. Recall that

$$\Theta = \Big\{\theta :\, \|\theta - \theta^*\|_H \leqslant \frac{c_0}{\sqrt{B}}\Big\}$$

which is an ellipsoid centered at θ^*, whose main axis is $\mathbb{R}\theta^*$, with length $2c_0B$, and all other axes in orthogonal directions having constant length $2c_0$. To that end, let γ denote the Gaussian measure $\mathsf{N}(0,\Gamma)$ with covariance

$$\Gamma = B^2 u^* u^{*\top} + \frac{1}{d}(I_d - u^* u^{*\top})\,. \tag{1.83}$$

This covariance structure ensures that the orthogonal (to θ^*) component of γ is of constant order of magnitude, while the component in the direction of θ^* is of order B. We define π_Θ as the truncated Gaussian obtained from γ with support

$$S_\pi = \Big\{\theta : \|\theta\|_H \leqslant \frac{C_\pi}{\sqrt{B}}\Big\}\,, \tag{1.84}$$

where $C_\pi > 0$ is an absolute constant ensuring that this set contains the supports of all the posteriors ρ_θ for $\theta \in \Theta$.

Posterior distributions. For every $(\theta, v) \in \Theta \times S^{d-1}$, write $\rho_{\theta,v} = \rho_\theta \otimes \rho_v$. We let the "sphere part" ρ_v be the uniform measure on the spherical cap of radius ε around v, that is $\mathsf{C}(v,\varepsilon) = \{v' \in S^{d-1},\ \|v' - v\| \leqslant \varepsilon\}$.

The other component ρ_θ has to satisfy the following requirements: its support should be large enough for the divergence $D(\rho_\theta\|\pi_\Theta)$ to remain of order $O(d)$, without any dependence on B. At the same time, the way it spreads around θ—in particular in the direction of θ itself—has to ensure that the smoothing inequality (1.77) holds. Note that the smoothing can be thought of as a convolution, in our case with an indicator; hence, if we want the result of the convolution to be concentrated, the distribution ρ_θ itself cannot be too concentrated. Indeed, recall that we want ρ_θ to satisfy, for all $x \in \mathbb{R}^d$,

$$\int_{\mathbb{R}^d} \mathbf{1}(|\langle\theta', x\rangle| \leqslant 1)\rho_\theta(\mathrm{d}\theta') \lesssim \exp(-|\langle\theta, x\rangle|)\,. \tag{1.85}$$

Let us explain with more details. We let ρ_θ be the distribution of $U\theta + Z$, where U is uniform on $[1-c, 1+c]$ for a small absolute constant c; while Z is independent of U and follows an isotropic truncated Gaussian distribution in the hyperplane orthogonal to θ, with covariance $\frac{1}{d}I_{d-1}$, supported on a $(d-1)$-dimensional Euclidean ball of radius r. In terms of support, this is not very different from a truncated Gaussian with mean θ and covariance

$$cB^2 uu^\top + \frac{r}{d}(I_d - uu^\top)\,, \quad u = \frac{\theta}{\|\theta\|}\,, \tag{1.86}$$

and in fact, only the component in the direction of θ would differ. The reason why we choose a strongly non-Gaussian smoothing in the direction of θ is that with a Gaussian smoothing, (1.85) would not hold at all because the Gaussian distribution has strong concentration properties. Roughly speaking, if we let θ' follow a truncated Gaussian distribution $\overline{\gamma}_\theta$ centered at θ and with covariance given by (1.86), then we would typically obtain

$$\int_{\mathbb{R}^d} \mathbf{1}(|\langle\theta', x\rangle| \leqslant 1)\overline{\gamma}_\theta(\mathrm{d}\theta') = \mathbf{P}_{\theta'\sim\overline{\gamma}_\theta}(|\langle\theta', x\rangle| \leqslant 1) \asymp \frac{1}{\max\{1, \|\theta\|\}}\,,$$

which is much larger than $\exp(-|\langle\theta, x\rangle|)$ that we are aiming for in (1.85). Conversely, (1.85) means that for a random vector $\theta' \sim \rho_\theta$, for all x, the probability that $|\langle\theta', x\rangle| \leqslant 1$ is exponentially small, which precisely means that the variable $\langle\theta', x\rangle$ cannot concentrate too much. On the other hand, a (truncated) Gaussian would make the computation of the divergence $D(\rho_\theta\|\pi_\Theta)$ easier as π_Θ is of this nature.

Smoothing inequality. To prove (1.77), we first integrate with respect to ρ_θ, and then ρ_v. We thus prove (1.85) first, and to that end we write

$$\begin{aligned}\int_\Theta \mathbf{1}\{|\langle\theta', x\rangle| \leqslant 1\}\, \rho_\theta(\mathrm{d}\theta') &= \mathbf{P}(|U\langle\theta, x\rangle + \langle Z, x\rangle| \leqslant 1) \\ &= \mathbf{E}\big[\mathbf{P}\left(|U\langle\theta, x\rangle + \langle Z, x\rangle| \leqslant 1|U\right)\big]\,. \end{aligned} \tag{1.87}$$

The key idea here is that with a proper truncation in the definition of the distribution of Z, its density is bounded by a constant factor of a Gaussian density. Therefore, using that $\|x\| \leqslant C\sqrt{d}$, we deduce that the density of $\langle x, Z\rangle$ is bounded by the density of a (one-dimensional) Gaussian with constant variance. We then deduce that

$$\mathbf{P}\left(|U\langle\theta, x\rangle + \langle Z, x\rangle| \leqslant 1|U\right) \leqslant C_1 \exp(-cU^2\langle\theta, x\rangle^2) \leqslant C_2\sigma'(\langle\theta, x\rangle)\,.$$

This in turn yields

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \iint_{\Theta\times S^{d-1}} Z_i(\theta', v')\rho_\theta(\mathrm{d}\theta')\rho_v(\mathrm{d}v') \\ &\qquad\leqslant C\int_{\mathsf{C}(v,\varepsilon)} \big\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v', v'\big\rangle \rho_v(\mathrm{d}v') \\ &\qquad\leqslant C\big\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\big\rangle + C\frac{\varepsilon^2}{d-1}\mathrm{Tr}\left(H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}\right)\,.\end{aligned}$$

An additional difficulty arises from the fact that the second term above has to be uniformly bounded over Θ. Straightforward computations show that

$$\mathrm{Tr}\big(H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}\big) \leqslant C'Bd\cdot\frac{1}{n}\sum_{i=1}^n \exp\big(-c|\langle\theta, X_i\rangle|\big)\mathbf{1}(\|X_i\| \leqslant C\sqrt{d})\,,$$

We thus have to show that if $n \geqslant C''B(d+t)$, then with probability at least $1-e^{-t}$,

$$\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n \exp\big(-c|\langle\theta, X_i\rangle|\big)\mathbf{1}\big(\|X_i\| \leqslant C\sqrt{d}\big) \leqslant \frac{c'}{B}\,.$$

Recall that Θ can be described in polar coordinates as the set of vectors θ such that

$$\|u - u^*\| \lesssim \frac{1}{B}\,, \quad \|\theta\| \asymp B\,,$$

so the previous inequality can be stated in terms of spherical cap as

$$\sup_{u\in\mathsf{C}(u^*,c/B)}\frac{1}{n}\sum_{i=1}^n \exp\big(-cB|\langle u, X_i\rangle|\big)\mathbf{1}\big(\|X_i\| \leqslant C\sqrt{d}\big) \leqslant \frac{c'}{B}\,. \tag{1.88}$$

It turns out that bounding this supremum also has to be done using PAC-Bayes inequality (by smoothing using uniform distributions on small spherical caps). Indeed, the entropic approach using Dudley's integral (see e.g., [Ver18, Theorem 8.1.3]) leads to a bound of order $1/B + \sqrt{d/n}$. Since we want a bound of order $1/B$, this would entail $n \gtrsim B^2 d$. In Lemma 3.10, we prove that (1.88) holds with probability at least $1-e^{-t}$ if $n \geqslant CB(d+t)$. We thus obtain that

$$\frac{1}{n}\sum_{i=1}^n \iint_{\Theta\times S^{d-1}} Z_i(\theta', v')\rho_\theta(\mathrm{d}\theta')\rho_v(\mathrm{d}v') \leqslant C\big\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\big\rangle + C'\varepsilon^2 \tag{1.89}$$

Relative entropy. The last step to prove Theorem 1.2 is to show that for all $\theta \in \Theta$ and $v \in S^{d-1}$, $D(\rho_{\theta,v}\|\pi) \leqslant Cd$. First, as these are product measures,

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

The second term writes

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

and can be handled using a covering argument [Mou22, §4.4], which gives us

$$D(\rho_v\|\pi_S) \leqslant d\cdot\log\left(1+\frac{2}{\varepsilon}\right)\,. \tag{1.90}$$

Bounding the first term is more delicate given the unusual nature of the posteriors ρ_θ. We do this using two successive approximations. First, we show that the density of ρ_θ is bounded by the density of a truncated d-dimensional Gaussian distribution with covariance given by (1.86), and then use a conditioning property that allows to bound from above the Kullback-Leibler divergence from a truncated Gaussian to an other one by a constant factor of the divergence between their non-truncated versions. Recall that we could not directly set ρ_θ as a truncated Gaussian because we needed sufficient variability in the direction of θ.

The two approximation steps can be described as follows. Define the Gaussian distribution γ_θ with covariance given by (1.86)

$$\gamma_\theta = \mathsf{N}(0,\Gamma_\theta)\,,\quad \Gamma_\theta = cB^2uu^\top + \frac{r}{d}(I_d - uu^\top)\,,$$

then, define $\overline{\gamma}_\theta$ as the truncation of γ_θ with support

$$S_\theta = \left\{\theta' :\, \|\theta'\|_{H_\theta} \leqslant \frac{c_1}{\sqrt{B}}\right\},\quad H_\theta = \frac{1}{B^3}uu^\top + \frac{1}{B}(I_d - uu^\top)\,.$$

Note that this is very similar to the prior π_Θ, as S_θ is a rotation (and scaling by a constant factor to ensure inclusion) of the support S_π of π_Θ. Indeed, the matrix H_θ has the same anisotropic structure as H, as it is the product of a rotation and H. The two steps are then

1. bounding the density of ρ_θ by that of $\overline{\gamma}_\theta$, using that their supports are similar;
2. bounding the divergence $D(\overline{\gamma}_\theta\|\pi_\Theta)$ by a constant factor of $D(\gamma_\theta\|\gamma)$, using that π_Θ is a truncation of γ, eand a conditioning property. Then, $D(\gamma_\theta\|\gamma)$ admits a closed form as it involves Gaussian distribution and we show that it is at most $O(d)$.

For the second point, the conditioning property is the following. Given two probability measures P,Q and an event A such that $\mathsf{P}(A) > 0$, one has

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

Finally we explicitly compute the divergence $D(\gamma_\theta\|\gamma)$ which, as $\det(\Gamma) = \det(\Gamma_\theta)$, 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 $\Gamma^{-1} \preccurlyeq C\, Bd \cdot H$,

$$\mathrm{Tr}(\Gamma^{-1/2}\Gamma_\theta\Gamma^{-1/2}) \leqslant C' Bd\, \mathrm{Tr}(H^{1/2}\Gamma_\theta H^{1/2}) \leqslant C'' d\,,$$

and, by definition of Θ,

$$\|\theta - \theta^*\|^2_{\Gamma^{-1}} \leqslant C_1' Bd \|\theta - \theta^*\|^2_H \leqslant C_1'' d\,.$$

To conclude the proof, observe that in the result of the smoothing step (1.89), the smoothing radius ε can be chosen as a (small) absolute constant. Since in addition the divergence satisfies $D(\rho_{\theta,v}\|\pi) \leqslant Cd$, the desired bound (1.81) holds, which proves the result. ■

1.5 Linear separation and conic geometry (Chapter 4)

1.5.1 Interplay between dimension and signal strength

In the discussion following Theorem 1.1 (see Chapter 3, Theorem 3.1 for a full exposition), we observed that this result establishes that the MLE exists with high probability in the regime $n \gg Bd$ (here, the notation $a \gg b$ stands for $a > Cb$ for a large, but absolute constant C). In contrast, in the opposite regime $n \ll Bd$, Theorem 1.1 only establishes that the probability of existence of the MLE is bounded away from 1, without guaranteeing that it is close to 0.

In light of the phase transition phenomenon established by Candès and Sur in [CS20], this suggests that there remains a gap in fully understanding the behavior of the MLE, particularly regarding how the interplay between dimension and signal strength affects its existence.

We start by recalling the result of Candès and Sur [CS20]. Their setting is that of so-called "proportional asymptotics" or "high-dimensional" asymptotics, where one considers a sequence of parameters $(d_n, \theta_n^*)_{n\geqslant 1}$ with $d_n/n \to \gamma \in (0,1)$ and $\beta_n = \|\theta_n^*\| \to \beta \in \mathbb{R}^+$. For every (d_n, θ_n^*), we are given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of size n of i.i.d. data from a well-specified logit model with parameter θ_n^* and isotropic Gaussian design $X_i \sim \mathsf{N}(0, I_{d_n})$. In this setting, Candès and Sur established the following result.

Theorem 1.3 ([CS20], Theorem 2.2)**.** *In the setting described above, one has*

$$\mathbf{P}(\text{MLE exists}) \underset{n\to\infty}{\longrightarrow} \begin{cases} 1 \text{ if } \gamma < h(\beta) \\ 0 \text{ if } \gamma > h(\beta)\,, \end{cases} \tag{1.91}$$

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 $\mathbf{P}(Y'_\beta = 1|X') = \sigma(\beta X')$, and let $V_\beta = Y'_\beta X'$. In addition, let $Z \sim \mathsf{N}(0,1)$ be independent of V_β. Then,

$$h(\beta) = \min_{t\in\mathbb{R}} \mathbf{E}\big[(tV_\beta - Z)_+^2\big]\,. \tag{1.92}$$

On the positive side, this result shows a sharp transition which depends on only two parameters: the limiting aspect ratio γ and the signal strength β. This results in a neat splitting of the phase space (γ, β), illustrated in Figure 1.2.

On the other side, a limitation of the proportional asymptotic framework is that it does not allow the signal strength to grow with n or d which is thus assumed fixed, that

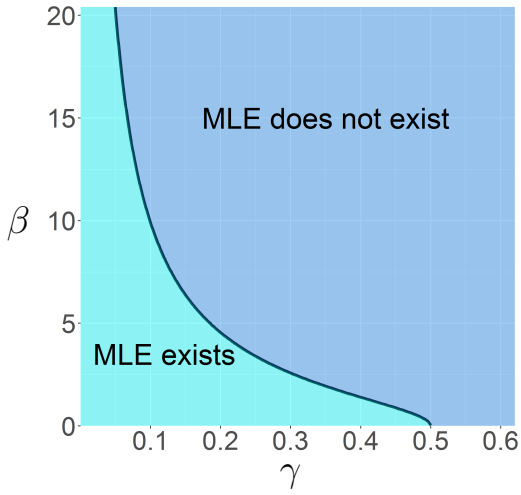

Figure 1.2: Phase transition diagram (taken from Figure 1(a) in [CS20]).

is $B = O(1)$ with our notation. However, an estimate of the boundary function h defined in (1.92) gives $h(\beta) \asymp 1/\beta$ for $\beta \gtrsim 1$. Together with (1.91), this suggests that in the non-asymptotic case that we want to address, the transition does happen around the critical sample size Bd.

In Chapter 4, outlined here, we address the limitation of Theorem 1.1 regarding the regime $n \ll Bd$ and complement it by a stronger result on non-existence of the MLE, Theorem 4.2, one of whose notable features is that is does not assume that $B = O(1)$.

As mentioned earlier in Section 10.1, the existence of the MLE is a purely geometric question, equivalent to the dataset not being linearly separated. More precisely, the MLE exists if and only if there exists no $\theta \in \mathbb{R}^d \setminus \{0\}$ such that $Y_i\langle\theta, X_i\rangle \geqslant 0$ for all $i \in \{1, \dots, n\}$.

Intuitively, the distinct effects of dimension and signal strength are easy to understand: if $n < d$, the family $(X_1, \dots, X_n)$ in $\mathbb{R}^d$ has rank at most $n < d$ and can thus be separated by a hyperplane. Under mild assumptions on the design, the rank of $(X_1, \dots, X_n)$ is indeed $\min\{n, d\}$ almost surely, so when $n > d$, these vectors typically span $\mathbb{R}^d$, and there is no "algebraic" reason for linear separability. This is where the signal strength comes into play: if the model is well-specified with strong signal—meaning that $\mathbf{P}(Y = 1|X) = \sigma(\langle\theta^*, X\rangle)$ with $\|\theta^*\| = B \gg 1$—then θ^* itself is likely to separate the dataset. Indeed, observe that

$$\mathbf{P}(Y\langle\theta^*, X\rangle < 0) = \mathbf{E}\big[\mathbf{P}(Y\langle\theta^*, X\rangle < 0\,|X)\big] \leqslant \mathbf{E}\big[\exp(-|\langle\theta^*, X\rangle|)\big] \leqslant \frac{1}{B}\,.$$

Hence, the probability that the dataset is linearly separated satisfies

$$\mathbf{P}(\forall i,\, Y_i\langle\theta^*, X_i\rangle \geqslant 0) = \big(1 - \mathbf{P}(Y\langle\theta^*, X\rangle < 0)\big)^n \geqslant \Big(1 - \frac{1}{B}\Big)^n \geqslant \exp\Big(-c\frac{n}{B}\Big)\,.$$

In short, it is not surprising that if $n \ll \max\{B, d\}$, there is a high probability that the MLE does not exist. The more interesting regime, discussed below, is when $\max\{B, d\} \ll n \ll Bd$, where it is the *interaction* between B and d that accounts for the non existence of the MLE.

We now state our result, before giving some proof ideas.

Theorem 1.4 (Chapter 4, Theorem 4.2)**.** *Let $d \geqslant 70$, and assume that $X \sim \mathsf{N}(0, I_d)$ and that the logistic model is well-specified. For every $\kappa \geqslant 1$, if $n \leqslant Bd/(200\kappa)$ then*

$$\mathbf{P}(\text{MLE exists}) \leqslant \exp\Big(- \max\Big\{\kappa\sqrt{d}, \frac{\kappa^2 d}{B^2}\Big\}\Big) + 6e^{-d/24}\,. \tag{1.93}$$

This can be seen as a quantitative version of the convergence to 0 in the phase transition (1.91) from [CS20]. Some quantitative observations can be made about this result. First, the parameter κ quantifies how small n is compared to the critical threshold Bd. In view of the previous discussion, a particularly interesting regime is when $n \asymp d$ and $B \gg 1$ (so $\kappa \asymp B$). In this regime, Theorem 1.4 shows that the probability of existence of the MLE is smaller than $\exp(-cd)$ for some constant c. Also, this regime still allows the configuration where $n \gg B$, in which case θ^* is unlikely to linearly separate the dataset. It is therefore another *random* vector θ (close to θ^*) that will separate the data. Such a θ can be found due to the abundance of room in high-dimensional spaces.

Before giving some proof ideas, let us summarize the main message of Theorem 1.4. In the previous section, Theorem 1.1 showed that if $n \gg Bd$, then the MLE exists with probability at least $1 - e^{-n/B}$ (which is itself at least $1 - e^{-d}$). Theorem 1.4 fills the gap in the opposite regime, showing that for every $\kappa \geqslant 1$, if $n \ll Bd/\kappa$, the probability that the MLE exists is *at most* $c\exp(-\max\{\kappa\sqrt{d}, \kappa^2 d/B^2\})$. In short,

- if $n \gg \kappa Bd$, then $\mathbf{P}(\text{MLE exists}) \geqslant 1 - e^{-\kappa d}$ (Theorem 1.1);
- if $n \ll Bd/\kappa$, then $\mathbf{P}(\text{MLE exists}) \leqslant c\exp(-\max\{\kappa\sqrt{d}, \kappa^2 d/B^2\})$ (Theorem 1.4).

Also, in the interesting regime where $1 \ll \max\{B, d\} \ll n \ll Bd$, it is a *random* vector θ close to θ^* that separates the data, not θ^* itself.

1.5.2 Some proof ideas

In this section, we elaborate on some of the main ideas underlying the proof of Theorem 1.4. The detailed proof can be found in Chapter 4, Section 4.3. We emphasize in particular the key structural differences with the proof of Candès and Sur.

Relation with conic geometry. The proof of Theorem 1.4 is of a very different nature from that of Theorem 1.1. To establish that the MLE exists with high probability, we essentially showed that the (random) function it minimizes, $\widehat{L}_n$, is locally strongly convex with high probability when n is large enough. In contrast, to show that if n is small, then with high probability, the MLE *does not exist* (Theorem 1.4), we rely on arguments concerning the stochastic geometry of cones. The connection between linear separation and conic geometry was put forward in the seminal work of Candès and Sur [CS20]. It uses the rotational invariance of the standard Gaussian distribution and the fact that the logit model is a single-index model. More precisely, if $X \sim \mathsf{N}(0, I_d)$ and $Y|X$ follows the logit model with parameter θ^* such that $B = \|\theta^*\|$, we can assume without loss of generality that

$$\mathbf{P}(Y = 1\,|X) = \sigma(BX^1)\,. \tag{1.94}$$

Here and in the rest of this section, superscripts denote coordinates indices for vectors in $\mathbb{R}^d$, *i.e.*, X_i^j will denote the j-th coordinate of X_i for every $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, d\}$. Similarly, $\theta^1, \dots, \theta^d$ will denote the coordinates of a vector $\theta \in \mathbb{R}^d$.

Our starting point to prove Theorem 1.4 is the same as that of [CS20]. Using (1.94), one can 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. Recall that a cone is a subset that is stable under multiplication by non-negative scalars. A notable fact at this point is that although we study a random set of n points in $\mathbb{R}^d$, the relevant property to study involves cones in $\mathbb{R}^n$, not $\mathbb{R}^d$.

With the formulation (1.94), for every i, Y_i depends on X_i only through its first coordinate X_i^1, and is independent of the $d-1$ others. We want to rewrite the property that there exists a non-zero vector $\theta \in \mathbb{R}^d$ such that for all i, $\langle \theta, Y_i X_i \rangle \geqslant 0$. Equivalently, for all i,

$$\theta^1 Y_i X_i^1 + \sum_{j=2}^{d} \theta^j Y_i X_i^j \geqslant 0\,.$$

From there, the key observation is the following. 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 $(Y_i X_i^1, Y_i X_i^2, \dots, Y_i X_i^d)$ and $(Y_i X_i^1, X_i^2, \dots, X_i^d)$ have the same distribution. Therefore, in a more compact form,

$$\mathbf{P}\big(\exists \theta \neq 0 : \forall i, Y_i \langle \theta, X_i \rangle \geqslant 0\big) = \mathbf{P}\Big(\exists \theta \neq 0 : \theta^1 \boldsymbol{V} + \sum_{j=2}^{d} \theta^j \boldsymbol{X}^j \in \mathbb{R}^n_+\Big)\,, \tag{1.95}$$

where $\mathbb{R}^n_+ = [0, +\infty)^n$ is the non-negative orthant,

$$\boldsymbol{V} = (Y_1 X_1^1, \dots, Y_n X_n^1) \in \mathbb{R}^n\,, \tag{1.96}$$

and, for every $j \in \{2, \dots, d\}$,

$$\boldsymbol{X}^j = (X_1^j, \dots, X_n^j) \in \mathbb{R}^n\,. \tag{1.97}$$

This forms a collection of $d-1$ independent standard Gaussian vectors in $\mathbb{R}^n$, hence $\mathcal{L} = \operatorname{span}\{\boldsymbol{X}^2, \dots, \boldsymbol{X}^d\}$ follows the uniform distribution on the set of all $(d-1)$-dimensional subspaces of $\mathbb{R}^n$. In the end, we are naturally led to study the event where the random subspace $\mathcal{L}$ non-trivially intersects the random cone

$$\Lambda = \mathbb{R}\boldsymbol{V} + \mathbb{R}^n_+\,, \tag{1.98}$$

since

$$\mathbf{P}(\text{MLE does not exist}) \geqslant \mathbf{P}(\Lambda \cap \mathcal{L} \neq \{0\})\,. \tag{1.99}$$

Phase transition and the kinematic formula. The last paragraph reduced the problem of existence of the MLE to that of the intersection of a random cone Λ in $\mathbb{R}^n$ with a uniform random subspace $\mathcal{L}$.

Now, it follows from the work [ALMT14] that the probability of such an event depends on the dimension of the random subspace $\mathcal{L}$, and on a certain geometric parameter of the cone Λ called "statistical dimension".

The phase transition result that we referred is the *approximate kinematic formula*, based on the more general kinematic formula from conic geometry. This result involves some geometric quantities called (conic) intrinsic volumes. We refer to [ALMT14, § 5.1–5.2] for a more detailed exposition and thorough references on this rich topic.

Let us briefly explain this notion.

Definition 1.1 (Conic intrinsic volumes)**.** Let $\mathbf{C}$ be a polyhedral cone[3] in $\mathbb{R}^n$ and $\boldsymbol{Z} \sim \mathsf{N}(0, I_n)$. For every $k \in \{0, \dots, n\}$, the k-th (conic) intrinsic volume of $\mathbf{C}$ is the quantity $v_k(\mathbf{C})$ defined as

$$v_k(\mathbf{C}) = \mathbf{P}\big(\Pi_{\mathbf{C}}\boldsymbol{Z} \text{ lies in the relative interior of a } k\text{-dimensional face of } \mathbf{C}\big)\,,$$

where $\Pi_{\mathbf{C}}$ denotes the Euclidean metric projection on $\mathbf{C}$.

This notion generalizes that of algebraic dimension, in the sense that if E denotes a linear subspace of dimension p, then $v_k(E) = \mathbf{1}(k = p)$ for every $k \in \{0, \dots, n\}$. It is clear that, for any polyhedral cone $\mathbf{C} \subset \mathbb{R}^n$,

$$\sum_{k=0}^{n} v_k(\mathbf{C}) = 1\,,$$

and therefore, we can naturally associate to every such $\mathbf{C}$ a probability distribution on $\{0, \dots, n\}$. We can now define the parameter alluded to earlier, known as statistical dimension.

Definition 1.2 (Statistical dimension)**.** Let $\mathbf{C}$ be a polyhedral cone in $\mathbb{R}^n$ with intrinsic volumes $v_0(\mathbf{C}), \dots, v_n(\mathbf{C})$. Let V be a random variable taking values in $\{0, \dots, n\}$ whose distribution is given by $\mathbf{P}(V = k) = v_k(\mathbf{C})$ for every $k \in \{0, \dots, n\}$. The *statistical dimension* of $\mathbf{C}$ is defined as the expectation of V, namely

$$\delta(\mathbf{C}) = \mathbf{E}V = \sum_{k=1}^{n} k v_k(\mathbf{C})\,.$$

As before, this is a natural generalization of the usual algebraic dimension, in the sense that $\delta(E) = \dim(E)$ for every linear subspace E. In addition, one can show that the statistical dimension can be equivalently defined in terms of Gaussian projections as

$$\delta(\mathbf{C}) = \mathbf{E}\|\Pi_{\mathbf{C}}\boldsymbol{Z}\|^2\,. \tag{1.100}$$

The kinematic formula. We can now state the kinematic formula. This result gives an explicit formula for the probability that a cone $\mathbf{C}$ non-trivially intersects a random rotation of an other cone $\mathbf{K}$ in terms of their intrinsic volumes.

Fact 1.1 (The kinematic formula for cones, [ALMT14], Fact 2.1)**.** *Let $\mathbf{C}$ and $\mathbf{K}$ be closed convex cones in $\mathbb{R}^n$, one of which is not a subspace. Draw a random orthogonal basis $Q \in \mathbb{R}^{n\times n}$. Then*

$$\mathbf{P}\big(\mathbf{C} \cap Q\mathbf{K} \neq \{0\}\big) = \sum_{i=0}^{n} \big(1 + (-1)^{i+1}\big) \sum_{j=i}^{n} v_i(\mathbf{C}) \cdot v_{n+i-j}(\mathbf{K}).$$

[3] A cone is polyhedral if it can be obtained as a finite intersection of halfspaces. The definition of intrinsic volumes extends to general convex cones by an approximation argument with respect to the conic Hausdorff metric, for which the set of polyhedral cones is dense in the set of all closed convex cones. See [ALMT14, § 5.1] and references therein.

We are interested in the case where $\mathbf{K}$ is a linear subspace. This simplifies the expression given by the kinematic formula but still requires computing all the intrinsic volumes of $\mathbf{C}$. It turns out that in many situations, the distribution of intrinsic volumes (the random variable V in Definition 1.2) concentrates around its expectation $\delta(\mathbf{C})$. This property gives rise to the *approximate* kinematic formulas. In short, these formulas state that if the sum of the statistical dimensions of the two cones exceeds the ambient dimension, then they non-trivial intersect with high probability (over random rotations of one of them). Otherwise, if $\delta(\mathbf{C}) + \delta(\mathbf{K})$ is too small compared to the ambient dimension, then with high probability, their intersection reduces to $\{0\}$. This is the prototypical example of *phase transition*: the probability of intersection is either close to 0 or to 1 depending on the relative values of the dimensions. As one can expect, the tighter the concentration of the intrinsic volumes, the sharper the resulting phase transition.

Application to linear separation. In light of the property (1.99) and the discussion above, it is clear that in order to control the probability of existence of the MLE, we must combine two steps:

1. conditionally on the cone Λ, apply a phase transition result (that is, an approximate kinematic formula) showing that the probability that a random subspace does not intersect Λ is small;

2. 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], namely Theorem I therein. Applied to a *deterministic* cone and a random subspace, this result can be restated as follows.

Theorem 1.5 (adapted from [ALMT14], Theorem I)**.** *Let $\mathbf{C}$ be a convex cone in $\mathbb{R}^n$. Let $p \in \{0, \dots, n\}$ and let $\mathcal{L}$ be a random subspace drawn uniformly from all subspaces of dimension p. Then for every $t > 0$,*

$$p + \delta(\mathbf{C}) > n + \sqrt{nt} \quad \Longrightarrow \quad \mathbf{P}(\mathcal{L} \cap \mathbf{C} \neq \{0\}) \geqslant 1 - 4e^{-t/8}\,, \tag{1.101}$$

$$p + \delta(\mathbf{C}) < n - \sqrt{nt} \quad \Longrightarrow \quad \mathbf{P}(\mathcal{L} \cap \mathbf{C} \neq \{0\}) \leqslant 4e^{-t/8}\,. \tag{1.102}$$

This result is one of the possible approximate kinematic formulas. It derives from a Hoeffding-type concentration inequality for the distribution of the intrinsic volumes of the cone $\mathbf{C}$. This accounts for the fluctuations of order $\sqrt{nt}$ and $\sqrt{nt}$ between (1.101) and (1.102).

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 (1.91) in the asymptotic regime, several refinements are required in order to obtain the quantitative bound of Theorem 1.4.

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, as the suboptimal fluctuations between the two regimes of Theorem 1.5 can no longer be handled by the convergence $n, d_n \to +\infty$. We use the following version of the approximate kinematic formula, which is the "Bernstein" version of Theorem 1.5.

Theorem 1.6 ("Bernstein" approximate kinematic formula ([ALMT14], Theorem 7.1))**.** *Let $\mathcal{L}$ be a random subspace of $\mathbb{R}^n$ drawn uniformly from all subspaces of dimension p and let $\mathbf{C} \subset \mathbb{R}^n$ be a cone. For all $t > 0$, if*

$$p + \delta(\mathbf{C}) > n + t\,, \tag{1.103}$$

then

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

Compared to the "Hoeffding" version of the approximate kinematic formula in Theorem 1.5, there is no additional $\sqrt{n}$ margin in (1.103) compared to (1.101). This means that the "transition width" is of order $\sqrt{\min\{\delta(\mathbf{C}), n - \delta(\mathbf{C})\}}$. For the cone of interest Λ (1.98), we will show that this quantity is (with high probability, see below) of order $\sqrt{n/B}$, instead of $\sqrt{n}$, which will enable us to obtain the desired sharp finite-sample phase transition.

This brings us to the second structural difference compared to [CS20]: we must establish a refined high-probability control on the statistical dimension of the random cone Λ. This requires a high-probability bound on the sum of i.i.d. random variables that (as shown in [CS20]) controls this dimension. Let us sketch the argument.

The first step is to express the *random* statistical dimension of Λ (1.98) as a function of $\boldsymbol{V}$ (1.96). To do so, we use the metric characterization (1.100) of the statistical dimension. This yields (by conditioning on $\boldsymbol{V}$)

$$\delta(\Lambda) = n - \mathbf{E}_{\boldsymbol{Z}}\left[\operatorname{dist}(\boldsymbol{Z}, \Lambda)^2 \,|\boldsymbol{V}\right], \tag{1.104}$$

where $\mathbf{E}_{\boldsymbol{Z}}$ denotes expectation with respect to $\boldsymbol{Z} = (Z_1, \dots, Z_n) \sim \mathsf{N}(0, I_n)$ (and is independent of Λ). In what follows, we let

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

In view of (1.103), we thus want to show that with probability as large as possible, $F(\boldsymbol{V})$ is at most a constant fraction of d. 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

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

where h is the phase transition function (1.92) from Theorem 1.3. In addition, since $h(B) \lesssim 1/B$, it follows that $\mathbf{E}[F(\boldsymbol{V})] \lesssim n/B \ll d/\kappa$. We now reduce the problem at hand to that of bounding the sum of certain i.i.d. variables. To do so, we bound

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

where we define $\psi(v) = \mathbf{E}\left[(v - Z)_+^2\right]$ with $Z \sim \mathsf{N}(0,1)$, for every $v \in \mathbb{R}$. Recall the model specification (1.94). We write for short $U_i = X_i^1$ for every $i \in \{1, \dots, n\}$. These variables are i.i.d. with distribution $\mathsf{N}(0,1)$. We then show that for all $\lambda \geqslant 0$,

$$\psi(-\lambda V_i) \leqslant e^{-\lambda^2 U_i^2/2} + \mathbf{1}(Y_i U_i \leqslant 0) + \lambda^2 U_i^2 \mathbf{1}(Y_i U_i \leqslant 0)\,. \tag{1.105}$$

For all $i \in \{1, \dots, n\}$ and $\lambda \in \mathbb{R}$, let

$$\zeta_{i,\lambda} = e^{-\lambda^2 U_i^2/2}, \quad \varepsilon_i = \mathbf{1}(Y_i U_i \leqslant 0), \quad \psi_i = U_i^2 \mathbf{1}(Y_i U_i \leqslant 0). \tag{1.106}$$

The first two variables are light-tailed, namely sub-gamma, which allows to bound $\sum_i \zeta_{i,\lambda}$ and $\sum_i \varepsilon_i$ using Bernstein's inequality. In contrast, bounding $\sum_i \psi_i$ is more delicate, and in particular, will not follow from Bernstein's inequality. Indeed, recall that a real random variable ξ is sub-gamma if there exists $\sigma^2, b > 0$ such that for all $\lambda \in [0, 1/b)$,

$$\mathbf{E} \exp(\lambda \xi) \leqslant \exp\left(\frac{\sigma^2 \lambda^2}{2(1 - b\lambda)}\right).$$

In particular, this bound means that for small values of λ (say, $\lambda \leqslant 1/(2b)$) the Laplace transform of ξ is similar to that of a sub-Gaussian variable. This yields two regimes for the deviations of ξ: sub-Gaussian for small deviations and sub-exponential for large deviations. The variables ψ_i's (1.106) do not satisfy this condition at all, as their moments $\|\psi_i\|_s$ grow at least linearly with s, even for small values of s. In short, we cannot bound their Laplace transform and subsequently that of their sum via independence. Essentially, the heavier tails of the ψ_i's prevent us from using the usual tensorization argument.

To circumvent this issue, we first obtain a tight control on the moments of the individual summands, and then apply a sharp estimate of Latała [Lat97] on moments of sums of independent random variables. The result we use is the following.

Lemma 1.2 ([Lat97], Corollary 1)**.** *Let $\xi, \xi_1, \dots, \xi_n$ be i.i.d. nonnegative random variables. Then for any $p \geqslant 1$,*

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

In order to use this result, we show that for all $s \geqslant 1$,

$$\|\psi_i\|_s \lesssim \min\left\{ \frac{s^2}{B^2}, s \right\}. \tag{1.107}$$

Plugging this in Latała's estimate and optimizing, we deduce that for every $p \geqslant 1$,

$$\|\psi_1 + \cdots + \psi_n\|_p \lesssim \frac{1}{B^2} \max\left\{ \frac{n}{B}, \min\left\{ p^2 \left(\frac{n}{B}\right)^{1/p}, B^2 p \left(\frac{n}{p}\right)^{1/B^2} \right\} \right\}. \tag{1.108}$$

Finally, for $p = \max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}$, we deduce that

$$\|S_n\|_p \lesssim \frac{1}{B^2} \max\left\{ \frac{d}{C_0 \kappa}, \kappa^2 d \right\} \lesssim \frac{\kappa^2 d}{B^2}. \tag{1.109}$$

In the end, this yields the desired high-probability bound on the statistical dimension $\delta(\Lambda)$, and in turn, allows to conditionally apply the approximate kinematic formula (Theorem 1.6) on the event where the statistical dimension behaves suitably, which has itself a high probability.

The bound (1.109) is the core reason why the final bound in Theorem 1.4 features a term scaling as $\exp(-c \max\{\kappa\sqrt{d}, \kappa^2 d/B^2\})$, and not simply $\exp(-cd)$.

1.6 Beyond Gaussian designs: sufficient and necessary regularity conditions (Chapter 5)

The previous results (Theorem 1.1 and 1.4) are specific to the case of a Gaussian design, which can be seen as the most favorable case. Indeed, in this setting, the behavior of the MLE is quite simple to describe: it either exists and is optimal with high probability (Theorem 1.1), this is when $n \gg Bd$; or, with high probability, it does not exist (in the opposite regime $n \ll Bd$, Theorem 1.4).

This raises the following natural question: which properties of the Gaussian distribution are responsible for the previously described behavior of the MLE? In particular its optimal performance in Theorem 1.1. Equivalently, for which distributions of the design does the MLE behave (at least in the well-specified case) similarly as if the design were Gaussian?

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 Chapter 5, we identify suitable assumptions on the design distribution leading to a near-Gaussian behavior. We state them below and comment on their roles in the relevant situations.

1.6.1 Regularity assumptions

Aside from light tails (Assumption 1.1), the assumptions include a condition on the behavior of one-dimensional linear projections of the design near 0 (Assumption 1.2), which is related to standard margin conditions in the classification literature [MT99, Tsy04]. However, as shown in Proposition 5.1, another assumption is necessary to obtain a near-Gaussian behavior (in a suitable sense); this non-standard condition (Assumption 1.3) 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".

We now give the precise statements of the regularity assumptions. The first assumption on the design is standard and states that the design X is light-tailed; we refer to Definition 8.1 in Appendix 8.1 for the definition of the ψ_1-norm.

Assumption 1.1. 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 $v \in S^{d-1}$.

This ensures that for all $v \in S^{d-1}$, $\mathbf{P}(|\langle v, X\rangle| \geqslant Kt) \leqslant e^{-t}$. This assumption is used, among other things, to extend the other two assumptions to local neighborhoods of the parameter u^* used in their definitions.

Note that in the following assumptions, we introduce some parameters, which are a direction $u^* \in S^{d-1}$, and a *scale* $\eta \in [0, 1/e]$. The discussion here bears on geometric properties of some probability measures on $\mathbb{R}^d$, that can be of independent interest beyond the topic of logistic regression. Nonetheless, in order to grasp the meaning of these assumptions, one can think of the direction u^* as that of the optimal parameter θ^* in the logistic model, and the scale η as the inverse of the signal strength, that is $\eta = B^{-1}$.

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]. We discuss this topic in greater details in Chapter 7. The discussion regarding high probability bounds on the empirical gradient in Section 1.3.4 highlighted the importance of the behavior of X near 0 in the direction of θ^*, and more precisely at a scale of $1/B$, where B is the parameter norm (see also Section 2.2 in Chapter 2). This observation motivates the definition of Assumption 1.2.

Assumption 1.2. Let $u^* \in S^{d-1}$ and $\eta \in (0, 1]$. For some $c \geqslant 1$, one has for every $t \geqslant \eta$ that

$$\mathbf{P}\big(|\langle u^*, X\rangle| \leqslant t\big) \leqslant ct\,. \tag{1.110}$$

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 1.3. 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

$$\mathbf{P}\big(|\langle u^*, X\rangle| \leqslant c\eta,\ |\langle v, X\rangle| \geqslant c^{-1}\max\{\eta, \|u^* - v\|\}\big) \geqslant \eta/c\,. \tag{1.111}$$

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 5.3 is that for every $v \in S^{d-1}$, one has

$$\mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant c\eta,\ |\langle v, X\rangle| \geqslant c^{-1}\max\Big\{\eta, \sqrt{1 - \langle u^*, v\rangle^2}\Big\}\Big) \geqslant \eta/c\,. \tag{1.112}$$

This only changes the value of the parameter c from (1.111) 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 almost as well as it would in the Gaussian case. We argue here that Assumption 1.3 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 (1.27) and (1.29), within constant factors of the matrix

$$H = \frac{1}{B^3} u^* u^{*\top} + \frac{1}{B}(I_d - u^* u^{*\top})\,, \tag{1.113}$$

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 $n \to \infty$:

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

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

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

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 1.1, 1.2 and 1.3 are sufficient to prove that this inequality holds, and even that the Hessian $H_X(\theta)$ is locally equivalent to H, see Lemma 6.8. Actually, the main results in Sections 6.2.1 and 6.2.2 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 5.3 is necessary to bound the Hessian $H_X(\theta^*)$ in the sense that if Assumptions 5.1 and 5.2 hold and $H_X(\theta^*) \succcurlyeq CH$, then Assumption 5.3 must also hold.

Proposition 1.3 (Chapter 5, Proposition 5.1)**.** *Let X be a random vector satisfying Assumption 1.1 with parameter K and Assumption 1.2 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 1.3 holds with parameters $u^* = \theta^*/\|\theta^*\|$, $\eta = 1/B$, and*

$$c' = O\big(C_0 K^2 \log^2(C_0 K B)\big)$$

We refer to Proposition 5.1 in Chapter 5 for an exact statement, whose proof can be found in Section 5.4.

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

Definition 1.3. 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, $\mathbf{E}[XX^\top] = I_d$) and satisfies Assumptions 5.2 and 5.3 with parameters u^*, η, c.

1.6.2 Examples of regular distributions

We now give some examples of probability measures on $\mathbb{R}^d$ satisfying the assumptions presented in the previous section.

Constant scales. The simplest example is the following. At a lower-bounded scale (bounded signal strength for the application to logistic regression), all sub-exponential distributions are regular. More precisely, if X is an isotropic, sub-exponential vector, then for any $u^* \in S^{d-1}$ and any $\eta \in (0, e^{-1}]$, X is $(u^*, \eta, c_{K,\eta})$-regular, where $c_{K,\eta} = \max\{2K \log(2K)/\eta, 2K^4\}$. The main idea here 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 6.1 and 6.2 depend exponentially on c, leading to an exponential dependence on the signal strength $B = \eta^{-1}$. For this reason, Proposition 5.2 is mainly relevant in the case of constant signal strength.

Log-concave distributions. Recall that the motivation behind the notion of regular design distributions is to identify the essential properties of the standard Gaussian measure that explain the good behavior of the MLE. 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 1.4. *Assume that X has a centered isotropic (that is, $\mathbf{E}X = 0$ and $\mathbf{E}XX^\top = I_d$) 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 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 [LV07]. 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.

Product measures. 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. This topic is discussed in much greater detail in Chapter 5, Section 5.2.3.

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

Assumption 1.4. The random vector $X = (X_1, \dots, X_d)$ is such that: $X_1, \dots, X_d$ are i.i.d., with $\mathbf{E}[X_j] = 0$, $\mathbf{E}[X_j^2] = 1$ and $\|X_j\|_{\psi_1} \leqslant K$ (for some $K \geqslant e$) for $j = 1, \dots, d$.

It is easy to see that such a vector is sub-exponential. The main issue which we focus on is to investigate whether Assumptions 1.2 and 1.3 are satisfied. Specifically, the question can be asked in the following way: given a direction $u^* \in S^{d-1}$, what is the smallest scale down to which Assumptions 1.2 and 1.3 hold?

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 $\mathbf{P}(X_j = 1) = \mathbf{P}(X_j = -1) = 1/2$ for $1 \leqslant j \leqslant d$. This design is even sub-Gaussian (as the coordinates of X are independent and bounded), but the question of its regularity is heavily dependent on the direction.

Despite these facts, the behavior of the MLE under a Bernoulli design can be drastically different from the case of a Gaussian design. Indeed, 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 1.1). To see this, consider a design X distributed as described above, and a well-specified logit model, where the parameter θ^*

is in a coordinate direction. Then, it is easy to see that for a sample of size n from this model, if $n \leqslant 0.1 \exp(B)$ then $\mathbf{P}(\text{MLE exists}) \leqslant 0.1$ (see Fact 5.1 for a formal statement).

This exponential dependence on the norm B comes from the fact that X is not regular at small scales in the direction $u^* = e_1$, the first vector from the canonical basis of $\mathbb{R}^d$. 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 1.3 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.

It should be noted though that this exponential dependence comes from the fact that in this example, u^* is among the worst possible directions. Indeed, for a more "diffuse" direction, such as

$$u^* = \Big(\frac{1}{\sqrt{d}}, \dots, \frac{1}{\sqrt{d}}\Big), \tag{1.114}$$

regularity can hold at much smaller scales.

We thus mostly focus on this typical diffuse direction in Section 5.2.3 (and in the rest of this discussion). For this direction, Assumption 1.2 is a condition on the one-dimensional marginal $\langle u^*, X\rangle = \frac{1}{\sqrt{d}} \sum_{j=1}^d X_j$. Under Assumption 1.4, 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 $1/\sqrt{d}$. In particular, We show in Chapter 5 (Lemma 5.1) that if X satisfies Assumption 1.4 with sub-exponential norm at most K, and if $d \geqslant K^6$, then Assumption 1.2 holds with $\eta = K^3/\sqrt{d}$ and $c = 1$ for the direction (1.114).

However, this is not sufficient to ensure a near-Gaussian behavior. Indeed, Proposition 1.3 shows that the *two-dimensional* margin condition, Assumption 1.3, is necessary for the model to exhibit the right Fisher information at θ^*.

As it turns out, regularity in the diffuse direction (1.114) holds at scales down to $1/\sqrt{d}$. The remaining difficulty towards this is to show that the two-dimensional margin condition holds at such scales. This is done in Lemma 5.2, reproduced below.

Lemma 1.3 (Chapter 5, Lemma 5.2)**.** *Let X have i.i.d. coordinates, with $\mathbf{E}[X_1] = 0$, $\mathbf{E}[X_1^2] = 1$ and $\mathbf{E}[X_1^8] \leqslant \kappa^8$ 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*

$$\mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant \eta,\ |\langle v, X\rangle| \geqslant 0.2 \max\{\eta, \|u^* - v\|_2\}\Big) \geqslant \frac{\eta}{70\,000\,\kappa^4}\,. \tag{1.115}$$

In particular, if X satisfies Assumption 5.4, then Assumption 5.3 holds for any $\eta \in [18K^3/\sqrt{d}, e^{-1}]$ with $c = 21\,000$.

Lemma 1.3 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 5.3.3.

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 5.4, 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.

1.7 Risk bounds for non-Gaussian designs and misspecified models (Chapter 6)

This section introduces the results of Chapter 6 providing guarantees on the existence and performance of the MLE when the design is no longer Gaussian but only regular, in the sense defined in the previous section. We first consider the case where the model is still assumed to be well-specified before moving on to the most general case where we make no assumption on the conditional distribution of Y given X.

1.7.1 Regular design, well-specified model

Under the regularity assumptions on the design gathered in Definition 1.3, but still in the well-specified case, Theorem 6.1 from Chapter 6 (reproduced below as Theorem 1.7) shows that the MLE behaves similarly as in the Gaussian case, up to poly-logarithmic factors in the norm B.

Theorem 1.7 (Chapter 6, Theorem 6.1)**.** *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 5.1, 5.2 and 5.3 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*

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

The guarantees of Theorem 1.7 almost match (up to poly-logarithmic factors in B) those of Theorem 1.1 in the Gaussian case, which are optimal as discussed in Chapters 3 and 4 (see also Section 1.4 and 1.5 of this introduction). In fact, one can almost recover (again up to $\log^4(B)$ factors) the guarantees of Theorem 1.1 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.

1.7.2 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 were those from [OB21], namely the excess risk bound (1.14) of order $B^3 d \log(1/\delta)/n$, when the design is Gaussian but the model may be misspecified. Our main result in this setting is Theorem 6.2 (Chapter 6, Section 6.2.2, reproduced below as Theorem 1.8) which improves these guarantees by removing polynomial factors in B.

Theorem 1.8 (Chapter 6, Theorem 6.2)**.** *Suppose that* X *satisfies Assumptions 5.1, 5.2 and 5.3 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*

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

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*

$$\mathbf{P}(\text{MLE exists}) \leqslant 1 - e^{-t} . \tag{1.116}$$

In addition, for the same distribution,

$$\liminf_{n \to \infty} \mathbf{P}\left(L(\widehat{\theta}_n) - L(\theta^*) \geqslant c_3 \frac{d + Bt}{n} \right) \geqslant e^{-t} . \tag{1.117}$$

here and as in Theorem 1.7, the constants c_1 and c_2 depend only 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 (1.14) 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" 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 provided by Theorem 1.8 is best possible (up to polylog(B) factors) in the general misspecified setting, both in the condition for existence of the MLE and for its excess risk bound. We prove this by exhibiting a conditional distribution of Y given X for which this factor is unavoidable. It is known from asymptotic theory (see e.g., [vdV98, Example 5.25 p. 55]) that in the misspecified case,

$$\sqrt{n} H(\theta^*)^{1/2}(\widehat{\theta}_n - \theta^*) \xrightarrow[n\to\infty]{\text{(d)}} \mathsf{N}(0, \Gamma)\,, \quad \Gamma = H(\theta^*)^{-1/2} G H(\theta^*)^{-1/2}\,, \tag{1.118}$$

where $H(\theta^*) = \nabla^2 L(\theta^*)$ is the population Hessian and $G = \mathbf{E}[\nabla\ell(\theta^*)\nabla\ell(\theta^*)^\top]$ is the the covariance of the gradient at θ^*. Then, consider the following setting: the design is Gaussian, $X \sim \mathsf{N}(0, I_d)$ and the distribution of $Y\,|X$ is defined as follows. Let $u^* \in S^{d-1}$ and $p \in (0, 0.07)$ such that

$$\mathbf{P}(Y\langle u^*, X\rangle < 0\,|X) = p\,. \tag{1.119}$$

Then

1. the signal strength $B = \max\{e, \|\theta^*\|\}$ is related to the probability of misclassification by
$$\frac{1}{2B^2} \leqslant \mathbf{P}(Y\langle u^*, X\rangle < 0) \leqslant \frac{1}{B^2}\,.$$
2. The covariance of the gradient $G = \mathbf{E}[\nabla\ell(\theta^*, Z)\nabla\ell(\theta^*, Z)^\top]$ satisfies
$$\|H^{-1/2} G H^{-1/2}\|_{\text{op}} \geqslant \frac{B}{8}\,.$$

Thus, the risk bound from Theorem 1.8 is asymptotically optimal (up to poly-logarithmic factors), in light of the convergence in distribution (1.118).

1.7.3 Some proof ideas

We now give some ideas of the proofs of Theorems 1.7 and 1.8. These proofs follow the structure outlined in Section 1.3.3 and therefore involve bounding from above the empirical gradient $\|H^{-1/2}\nabla\widehat{L}_n(\theta^*)\|$ and uniformly bounding from below the empirical Hessians $\widehat{H}_n(\theta)$ over a neighborhood of θ^*.

We start with the bound on the Hessians, as it is common to the proofs of Theorems 1.7 and 1.8. In contrast, bounding the gradient relies heavily on whether the model is well-specified or not, as we emphasized in Section 1.3.4.

Uniform lower bound on empirical Hessians. The proofs of Theorems 1.7 and 1.8 rely on the same uniform lower bound on empirical Hessians, given in Theorem 6.3 and reproduced below.

Theorem 1.9 (Chapter 6, Theorem 6.3)**.** *Let X be a random vector satisfying Assumptions 1.1 and 1.3 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 the regularity parameters c and K for which, letting

$$\Theta = \left\{ \theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant \frac{c_0}{\log(B)\sqrt{B}} \right\}, \tag{1.120}$$

the following holds: for any $t > 0$, if

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

then with probability at least $1 - e^{-t}$, simultaneously for all $\theta \in \Theta$,

$$\widehat{H}_n(\theta) \succcurlyeq c_2 H\,.$$

Note that compared to Theorem 1.2, the neighborhood Θ on which the uniform lower bound holds is shrinked by a factor $\log B$, and that the sample size condition is slightly worse, also by a $\log B$ factor.

Sketch of proof. The proof of Theorem 1.9 relies on (and in fact motivates) the two-dimensional margin condition of Assumption 1.3. We want to bound from below

$$\inf_{\theta\in\Theta, v\in S^{d-1}} \frac{1}{n}\sum_{i=1}^n \sigma'(\langle\theta, X_i\rangle)\langle v, X_i\rangle^2$$

by a constant factor of $\langle Hv, v\rangle$. To do so, we observe that for every $v \in S^{d-1}$,

$$\langle Hv, v\rangle = \frac{\langle u^*, v\rangle^2}{B^3} + \frac{1 - \langle u^*, v\rangle^2}{B} \lesssim \frac{1}{B}\cdot\max\left\{\frac{1}{B}, \|u^* - v\|\right\}^2.$$

Then, we use that for every $\theta \in \Theta$, $\sigma'(\langle\theta, X_i\rangle) \gtrsim \mathbf{1}(|\langle\theta, X_i\rangle| \leqslant 1)$, hence (letting $u = \theta/\|\theta\|$),

$$\begin{aligned}
&\sigma'(\langle\theta, X_i\rangle)\langle v, X_i\rangle^2 \\
&\quad\gtrsim \mathbf{1}(|\langle\theta, X_i\rangle| \leqslant 1)\langle v, X_i\rangle^2 \\
&\quad\gtrsim \mathbf{1}\Big(|\langle u, X_i\rangle| \leqslant \frac{c'}{B}; |\langle v, X_i\rangle| \geqslant \frac{\max\{B^{-1}, \|u^* - v\|\}}{c'}\Big)\max\left\{\frac{1}{B}, \|u^* - v\|\right\}^2 \\
&\quad\gtrsim \mathbf{1}\Big(|\langle u, X_i\rangle| \leqslant \frac{c'}{B}; |\langle v, X_i\rangle| \geqslant \frac{\max\{B^{-1}, \|u^* - v\|\}}{c'}\Big)B\langle Hv, v\rangle\,.
\end{aligned}$$

The result will thus follow if we are able to prove that

$$\inf_{\substack{u,v\in S^{d-1}\\ \|u-u^*\|\leqslant\frac{c_0}{B\log B}}} \frac{1}{n}\sum_{i=1}^n \mathbf{1}\Big(|\langle u, X_i\rangle| \leqslant \frac{c'}{B}; |\langle v, X_i\rangle| \geqslant \frac{\max\{B^{-1}, \|u^* - v\|\}}{c'}\Big) \geqslant \frac{1}{c'B}\,, \tag{1.121}$$

with probability at least $1 - \delta$, provided that $n \geqslant CB(d\log B + \log(1/\delta))$. There, we implicitly used the "polar" property of Θ, which is that for all $\theta = \|\theta\|u \in \Theta$,

$$\|u - u^*\| \leqslant \frac{c_0'}{B\log B}\,, \quad \|\theta\| \asymp \|\theta^*\|\,. \tag{1.122}$$

To prove (1.121), we proceed with two steps. First, we extend Assumption 1.3 to all directions $u \in S^{d-1}$ close to u^*, *i.e.*, satisfying (1.122). This is done by bounding the deviations of $\langle u - u^*, X_i\rangle$ which are $K/(B\log B)$-sub-exponential thanks to Assumption 1.1. This way, each set

$$A_{u,v} = \left\{ x : |\langle u, x\rangle| \leqslant \frac{c'}{B}; |\langle v, x\rangle| \geqslant \frac{\max\{B^{-1}, \|u^* - v\|\}}{c'} \right\}, \tag{1.123}$$

for $v \in S^{d-1}$ and $u \in S^{d-1}$ satisfying (1.122), satisfies that $P_X(A_{u,v}) \geqslant 1/(c''B)$.

Second, we show that 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 $m, M > 0$ is a Vapnik-Chervonenkis (VC) class with VC dimension at most $O(d)$. For this, remark that each of these sets is the union of two intersections of 3 halfspaces. The class of all halfspaces in $\mathbb{R}^d$ has VC dimension d, which combined with [vdVW09, Theorem 1.1] shows the desired upper bound on the VC dimension of the class of interest.

We then conclude using standard tools from empirical processes theory. ■

Note that one of the most important ideas is actually not in this sketch of proof and resides in the identification of Assumption 1.3 as the "right" assumption to allow such a control.

Upper bounds on the empirical gradient. To prove Theorems 1.7 and 1.8, we combine Theorem 1.9 with a bound on the empirical gradient, depending on whether the model is well-specified or misspecified.

Well-specified case. For the well-specified case, we rely on the following bound.

Proposition 1.5. *Assume that X satisfies Assumptions 5.1 and 5.2 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.

Sketch of proof. The first steps are the same as in the proof of Proposition 1.1 in the Gaussian case. We start from the decomposition

$$\|H^{-1/2}\nabla \widehat{L}_n(\theta^*)\| \leqslant B^{3/2}|\langle u^*, \nabla \widehat{L}_n(\theta^*)\rangle| + B^{1/2} \sup_{\substack{w \in S^{d-1} \\ \langle u^*, w\rangle = 0}} \langle w, \nabla \widehat{L}_n(\theta^*)\rangle\,.$$

and then bound the moments of these variables to show that they satisfy the sub-gamma property. The key step now is to use that $\mathbf{P}(Y\langle \theta^*, X\rangle < 0\,|X) \leqslant \exp(-|\langle \theta^*, X\rangle|)$ which shows that for all $v \in S^{d-1}$,

$$\mathbf{E}\big|\big\langle v, \nabla \ell(\theta^*, Z)\big\rangle\big|^p \leqslant 2\mathbf{E}\big[\exp(-|\langle \theta^*, X\rangle|)|\langle v, X\rangle|^p\big]\,. \tag{1.124}$$

This reduces the problem of bounding the empirical gradient to an inequality only involving X and not Y. Using this observation, we bound

$$\mathbf{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle u^*, X\rangle|^p\big] \quad \text{and} \quad \mathbf{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle w, X\rangle|^p\big]\,, \tag{1.125}$$

for all $p \geqslant 2$. This is where the proof becomes different from the Gaussian case. First, $\langle u^*, X\rangle$ and $\langle w, X\rangle$ are no longer independent. Second, we cannot rely on a closed form

to bound these moments. The lack of independence is handled using Hölder's inequality and the sub-exponential tails of the marginals of X. For any $s \in (0,1)$,

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

Letting $s = 1/\log(B)$, we find

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

To handle the first moment in (1.125), we use Assumption 1.2 that controls the behavior of $\langle u^*, X\rangle$ near 0. As we do not assume that this variable has a density, we instead sum over a geometric grid, which yields

$$\begin{aligned}\mathbf{E}[\exp(-B|\langle u^*, X\rangle|)] &\leqslant \mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{1}{B}\Big) + \sum_{k\geqslant 0} e^{-2^k}\mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{2^{k+1}}{B}\Big) \\ &\leqslant \frac{c}{B}\left(1 + \sum_{k\geqslant 0} 2^{k+1}e^{-2^k}\right) \leqslant \frac{5c}{B}\,.\end{aligned}$$

For higher-order moments, we use that

$$\mathbf{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle u^*, X\rangle|^p\big] \leqslant \sup_{t>0}\big\{t^p e^{-Bt/2}\big\}\big[\exp(-B|\langle u^*, X\rangle|/2)\big]\,,$$

and repeat the argument. This shows the sub-gamma behavior of the variables of interest, with only a logarithmic (in B) degradation compared to the Gaussian case. We then conclude using the same ε-net argument as in (1.67). ■

Misspecified case. The most important difference is that we can no longer reduce the problem to bounding a function of X by using the bound $\mathbf{E}[\sigma(-Y\langle\theta^*, X\rangle)\,|X] \leqslant \exp(-|\langle\theta^*, X\rangle|)$. We can still bound

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

but we now need to bound directly $\mathbf{E}[|\langle v, X\rangle|^p \mathbf{1}(Y\langle\theta^*, X\rangle < 0)]$, for either $v = u^*$ or v orthogonal to u^*. We only bound the first orders (up to $p = 2$) and then use the refined control of the moments of a sub-exponential variable alluded to at the end of Section 1.3.4, which is the 6th point of Lemma 8.1.

The core argument here is to use that $\nabla L(\theta^*) = 0$, hence $\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=1} L(t\theta^*) = 0$. Using the symmetry of σ, this shows that which writes

$$\mathbf{E}\big[|\langle\theta^*, X\rangle|\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\big] = \mathbf{E}\big[|\langle\theta^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big]\,. \tag{1.126}$$

Using the computations from the previous case (with $p = 1$), we obtain

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

Moreover, as

$$1 \leqslant |\langle\theta^*, X\rangle| + \mathbf{1}(|\langle\theta^*, X\rangle| \leqslant 1)\,,$$

we have

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

Applying the previous inequality and using Assumption 1.2, this proves that

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

We then bound the second moments (with some slight technical subtleties) and end in a position to apply Point 6 from Lemma 8.1 for each $v \in S^{d-1}$, and conclude again with an ε-net argument.

1.8 Fast rates for binary classification (Chapter 7)

So far, we were interested in the performance of the maximum-likelihood estimator as a predictor for the conditional probabilities of the outcome. The logistic loss also naturally arises in statistical learning theory [BBL05, Kol11, Bac24], as a convex surrogate for the classification error [Zha04, BJM06].

In short, it adresses the following problem. Binary classification aims at founding a predictor $f : \mathbb{R}^d \to \{-1, 1\}$ with small prediction risk

$$R(f) = \mathbf{P}(Y \neq f(X)) = \mathbf{P}(Yf(X) < 0)\,.$$

Equivalently, by taking the sign of f as a predictor, one can allow f to take arbitrary real values, in which case $R(f)$ is defined as $\mathbf{P}(Yf(X) < 0)$. A natural approach [Vap00, DGL96] is then to minimize the empirical counterpart of R over a subset $\mathcal{F}$ of all measurable functions from $\mathbb{R}^d$ to $\mathbb{R}$. A classical choice, which is relevant here, is the linear class $\mathcal{F}_{\text{lin}} = \{x \mapsto \langle\theta, x\rangle,\ \theta \in \mathbb{R}^d\}$. This leads to empirical risk minimization (ERM) over the linear class

$$\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Y_i\langle\theta, X_i\rangle < 0)\,. \tag{1.127}$$

A critical issue is that this is a non-convex optimization problem, because $\theta \mapsto \mathbf{1}(Y\langle\theta, X\rangle < 0)$ (with fixed X and Y) is not a convex function. This has both practical and theoretical implications: this problem is—except for special cases—extremely hard to solve numerically and harder to analyze than a convex problem. To circumvent this issue, one can replace the non-convex function $t \mapsto \mathbf{1}(t < 0)$ by a convex surrogate ϕ, and define

$$\widehat{\theta}_n^{\phi} \in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i\langle\theta, X_i\rangle)\,, \tag{1.128}$$

which is thus an ERM for the risk associated to ϕ, $R^{\phi}(\theta) = \mathbf{E}[\phi(Y\langle\theta, X\rangle)]$, called ϕ-risk for short. A critical issue is to ensure that the solution to the surrogate problem (1.128) is a good approximate solution to the original problem, namely to bound

$$R(\widehat{\theta}_n^{\phi}) - R^* = \mathbf{P}(Y\langle\widehat{\theta}_n^{\phi}, X\rangle < 0) - \inf_{f:\mathbb{R}^d\to\mathbb{R}} R(f)\,, \quad \text{where } R^* = \inf_{f:\mathbb{R}^d\to\mathbb{R}} R(f)\,,$$

is the (classification) Bayes risk. One of the natural choices for ϕ is the logistic loss $\phi(t) = \log(1 + e^{-t})$, in which case $\widehat{\theta}_n^{\phi}$ is the MLE from logistic regression $\widehat{\theta}_n$ defined in (1.7).

A seminal result due to Zhang [Zha04] (of which Bartlett, Jordan and McAuliffe provide a more general version in [BJM06]) shows that under suitable conditions on the surrogate function ϕ, the excess classification risk is controlled by the excess ϕ-risk. More precisely, Zhang's lemma applied to the logistic loss can be stated as follows. Suppose, for clarity, that the model is well-specified with parameter θ^*. Then there exists a numerical constant C such that for all θ,

$$R(\theta) - R(\theta^*) \leqslant C\sqrt{L(\theta) - L(\theta^*)}\,, \tag{1.129}$$

where as before, $L(\theta) = \mathbf{E}[\log(1 + \exp(-Y\langle\theta, X\rangle))]$ is the logistic risk. In particular, using previous non-asymptotic results for the logistic excess risk of the MLE, e.g., (1.14) from [OB21], this yields a classification risk scaling as $\sqrt{d/n}$, which is the typical slow rate provided by Vapnik-Chervonenkis theory for the empirical risk minimizer with respect to the binary loss.

In short, (1.129) ensures that the MLE $\widehat{\theta}_n$ is consistent when used as a plug-in for binary classification, but in terms of quantitative bounds, it significantly degrades the rate of convergence to 0.

In binary classification, the difficulty of the problem depends on how close the regression function $\zeta(X) = \mathbf{P}(Y = 1\,|X)$ is to the decision boundary $\{x : \zeta(x) = 1/2\}$. When $\zeta(X)$ tends to concentrate near $1/2$, prediction becomes intrinsically harder. The margin condition controls this behavior by bounding the probability that $\zeta(X)$ lies close to $1/2$. This condition, known as the Tsybakov noise condition [MT99, Tsy04], ensures that the mass near the boundary decays as a power of the distance, and thereby quantifies how informative the covariates are near the threshold. More precisely, the pair (X, Y) satisfies the margin condition with constant A and exponent α if there exists $\varepsilon_0 \in (0, 1/2]$ such that for all $\varepsilon \in (0, \varepsilon_0]$,

$$\mathbf{P}(|\zeta(X) - 1/2| \leqslant \varepsilon) \leqslant A\varepsilon^{\alpha}\,. \tag{1.130}$$

This margin condition implies a Bernstein condition (a lower bound on the curvature of the excess risk around its minimum) which in turn allows for faster rates of convergence of the excess classification risk. Specifically, if the margin condition holds with exponent α, then one can achieve excess risk bounds of order $(d/n)^{(\alpha+1)/(\alpha+2)}$. As it turns out, in the well-specified logistic model with Gaussian design and signal strength B, the margin condition holds with exponent 1 and constant proportional to $1/B$. As a result, one can obtain a classifier $\widetilde{\theta}_n$ satisfying

$$R(\widetilde{\theta}_n) - R(\theta^*) \leqslant C\Big(\frac{d}{n}\Big)^{2/3}\,,$$

either in expectation or with high probability.

Our main contribution is an improvement on Zhang's lemma for the linear class, by relating the excess risk in classification to the error on the estimation of the direction of the true parameter, which we do using the margin condition (1.130) (Propositions 1.6 and 1.7). In turn, combined with our results on the localization of the MLE, these yield fast classification rates for the (plug-in classifier of) the MLE (Theorems 1.10 and 1.11).

Although assuming that the model is well-specified is not required, for the sake of concreteness we state our results in this setting. Throughout this section, we assume that

the pair (X, Y) follows the logit model with parameter θ^*, which has a norm bounded away from 0. We let as before $B = \|\theta^*\| \geqslant e$. We first investigate the case where the design is Gaussian, and then generalize this to the case of regular designs, with a slight strengthening of the one-dimensional margin condition (Assumption 1.2), corresponding to a standard margin condition.

Gaussian design. Here, we assume the setting of Theorem 3.1. The following result gives a sharp bound on the excess risk in terms of estimation error of the direction of the true parameter.

Proposition 1.6. *Let $X \sim \mathsf{N}(0, I_d)$ and let (X, Y) follow the logit model with parameter $\theta^* = \|\theta^*\| u^*$. Then for any $u \in S^{d-1}$,*

$$R(u) - R(u^*) \leqslant 4 \|\theta^*\| \cdot \|u - u^*\|^2 . \tag{1.131}$$

Now, as a byproduct of Theorem 1.1, we obtained a bound on the estimation error of the direction u^* of θ^*. Combined with Proposition 1.6, this leads to the following bound on the binary excess risk of the MLE.

Theorem 1.10. *Grant the assumptions of Theorem 1.1. Then with probability at least $1 - \delta$ the MLE $\widehat{\theta}_n$ exists and satisfies the classification risk bound*

$$R(\widehat{\theta}_n) - R(\theta^*) \leqslant C \frac{d + \log(1/\delta)}{n} . \tag{1.132}$$

This result shows that when it exists, the MLE is also an optimal linear classifier (up to absolute constant) for the binary loss.

Regular designs. The setting here is related to that of Theorem 1.7: the design X is regular (Definition 1.3) and the pair (X, Y) follows the logit model (1.5) with parameter θ^*. Specifically, we assume that the design satisfies a slightly stronger version of Assumption 1.2, which is closely related to the standard margin assumption from the literature on supervised classification. This assumption is the following.

Assumption 1.5. There exists $c > 0$ such that for all $t > 0$, $\mathbf{P}(|\langle u^*, X \rangle| \leqslant t) \leqslant ct$.

This assumption extends Assumption 1.2 to arbitrary scales and not just down to $1/B$, which was sufficient to prove Theorems 1.7 and 1.8. In this setting, we obtain the following generalization of Proposition 1.6.

Proposition 1.7. *Let X be isotropic, K-sub-exponential and satisfying Assumption 7.1 with constant c. Then, for all $u \in S^{d-1}$ such that $\|u - u^*\| \leqslant 1/e$,*

$$R(u) - R(u^*) \leqslant 10 c K^2 \|\theta^*\| \cdot \|u - u^*\|^2 \log^2 \left(\frac{1}{\|u - u^*\|} \right) . \tag{1.133}$$

This result features a logarithmic degradation compared to Proposition 1.6, in the same way that Theorem 1.7 suffered a logarithmic degradation compared to Theorem 1.1. Combining Proposition 1.7 with Theorem 1.7, we obtain the following result regarding the binary excess risk of the MLE.

Theorem 1.11. *Let X be a K-sub-exponential, isotropic random vector, satisfying Assumption 1.5 with constant c and Assumption 1.3 with constant c and scale $1/B$. There are some constants $c_1, c_2 > 0$, depending only on c and K such that for every $\delta \in (0,1)$, if $n \geqslant c_1 B \log^4(B)(d + \log(1/\delta))$, then with probability at least $1 - \delta$, the MLE $\widehat{\theta}_n$ exists and satisfies*

$$R(\widehat{\theta}_n) - R(\theta^*) \leqslant c_2 K^2 \log^2(B) \frac{d + \log(1/\delta)}{n} \log^2\left(\frac{n}{d}\right).$$

Chapter 2

Convex localization through the control of gradient and Hessians

Contents

In this chapter, we describe the general scheme of proof that we use to establish Theorems 3.1, 6.1 and 6.2.

2.1 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 2.1, 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. It is not specific to logistic regression and can therefore be adapted to other M-estimation (empirical risk minimization) problems.

Lemma 2.1. *Assume that there exists a positive-definite matrix $H \in \mathbb{R}^{d\times d}$ and real numbers $r_0, c_0, c_1, \psi_n > 0$ such that the following conditions hold:*

(i) $\|\nabla \widehat{L}_n(\theta^*)\|_{H^{-1}} \leqslant \psi_n$;

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

(iii) *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 $\psi_n < c_0 r_0/2$, then the empirical risk $\widehat{L}_n$ admits a unique global minimizer $\widehat{\theta}_n$, which satisfies

$$\left\|\widehat{\theta}_n - \theta^*\right\|_H \leqslant \frac{2\psi_n}{c_0} \qquad \text{and} \qquad L(\widehat{\theta}_n) - L(\theta^*) \leqslant \frac{2c_1\psi_n^2}{c_0^2}\,. \tag{2.1}$$

If in addition $\psi_n < c_0 r_0/4$, 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

$$L(\widetilde{\theta}_n) - L(\theta^*) \leqslant \frac{c_1}{2}\left\|\widetilde{\theta}_n - \theta^*\right\|_H^2 \leqslant \max\left\{\frac{8c_1\psi_n^2}{c_0^2}, \frac{2c_1}{c_0}\left[\widehat{L}_n(\widetilde{\theta}_n) - \widehat{L}_n(\widehat{\theta}_n)\right]\right\}. \tag{2.2}$$

Let us comment on the implications of this result. Lemma 2.1 (proved in Section 2.4) reduces the proof of existence and risk bounds for the MLE 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 $\theta \in \Theta = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant r_0\}$.

Once these steps are completed, they naturally yield the condition of existence of the MLE in the following way. First, a condition on the sample size n will naturally appear to obtain a uniform lower bound on the empirical Hessians (condition (ii) in Lemma 2.1). Second, a prerequisite for the existence of the MLE is that the upper bound on the empirical gradient, ψ_n, has itself to be less than a constant. This imposes a second condition on n. If both sample size conditions are satisfied—hence the MLE exists on a high probability event—the risk bound is given by (2.1). In particular, the upper bound ψ_n on the rescaled gradient not only yields a condition on the sample size for the MLE to exist, it also gives the downstream rate for its excess risk.

In short, we want ψ_n to be as small as possible, and r_0 as large as possible, to derive a weak sufficient condition for the existence of $\widehat{\theta}_n$ and a sharp bound on its excess risk.

The nature of these tasks will depend on the combination of two factors: whether the design is Gaussian or not, and whether the model is well-specified or not. The task of bounding the Hessian will only be affected by the design assumption, as for any $\theta \in \mathbb{R}^d$ and $z = (x, y) \in \mathbb{R}^d \times \{-1, 1\}$, the Hessian of the logistic loss is given by

$$\nabla^2 \ell(\theta, z) = \sigma'(\langle \theta, x \rangle) x x^\top ,$$

which thanks to the fact that σ' is even, does not depend on y. As a consequence, in particular, for any $\theta \in \mathbb{R}^d$, the distribution of the empirical Hessian $\widehat{H}_n(\theta)$ does not depend on the distribution of Y. Conversely, the gradient of the logistic loss is

$$\nabla \ell(\theta, z) = -y\sigma(-y\langle \theta, x \rangle) x ,$$

and therefore, the distribution of the empirical gradient $\nabla \widehat{L}_n(\theta^*)$ crucially depends on the conditional distribution of Y given X.

Another observation that can be deduced from Lemma 2.1 is the fact that any approximate minimizer $\widetilde{\theta}_n$ of the empirical risk $\widehat{L}_n$ enjoys the same risk bound as the MLE $\widehat{\theta}_n$ thanks to (2.2). A practical implication is that when computing the MLE on large data sets using convex optimization algorithms (e.g., stochastic gradient descent, as in [Bac10, BM13]), there is no need to optimize below statistical error, a key insight from [BB08] in more general large-scale machine learning frameworks.

Although the matrix H (and the corresponding parameters c_0, c_1, r_0, ψ_n) from Lemma 2.1 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 (but its intrinsic structure only depends on the design, as mentioned earlier). Indeed, in order to obtain sharp bounds we would like c_0, c_1 from conditions (ii) and (iii) to be of constant order, and indeed in the Gaussian case these will be universal constants.

In fact, such a choice for H is somewhat canonical in the following sense: 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 $\theta \in \Theta$. Furthermore, by the law of numbers $\nabla^2 \widehat{L}_n(\theta)$ should be close to its expectation $\mathbf{E}[\nabla^2 \widehat{L}_n(\theta)] = \nabla^2 L(\theta)$ for large n, 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

$$\Theta' = \left\{ \theta \in \mathbb{R}^d : c_2^{-1} \nabla^2 L(\theta^*) \preccurlyeq \nabla^2 L(\theta) \preccurlyeq c_2 \nabla^2 L(\theta^*) \right\} \tag{2.3}$$

with $c_2 = c_1/c_0$.

It follows from these considerations that, in order to apply Lemma 2.1 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 (2.3) where the conditions of Lemma 2.1 could be expected to hold.

To get a reasonable idea of the behavior of the population Hessian, we focus below on the Gaussian case. 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

$$\nabla^2 L(\theta) = c_0(\|\theta\|) u u^\top + c_1(\|\theta\|)(I_d - u u^\top) , \tag{2.4}$$

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

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

In addition, one may verify (see Lemma 2.2 in Section 2.4) that for some numerical constants c_0', c_0'', c_1', c_1'':

$$\frac{c_0'}{(\beta+1)^3} \leqslant c_0(\beta) \leqslant \frac{c_0''}{(\beta+1)^3}\,, \qquad \frac{c_1'}{\beta+1} \leqslant c_1(\beta) \leqslant \frac{c_1''}{\beta+1}\,. \tag{2.5}$$

We will therefore set H to be the matrix

$$H = \frac{1}{B^3}u^*u^{*\top} + \frac{1}{B}(I_d - u^*u^{*\top})\,, \quad u^* = \frac{\theta^*}{\|\theta^*\|} \in S^{d-1}\,, \quad B = \max\{e, \|\theta^*\|\}\,, \tag{2.6}$$

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 (2.3) (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 $r_0 \asymp 1/\sqrt{B}$. Equivalently, in polar coordinates, this corresponds to vectors $\theta = u\|\theta\| \in \mathbb{R}^d$ such that

$$\|u - u^*\| \lesssim \frac{1}{B} \quad \text{and} \quad \|\theta\| \asymp \|\theta^*\|\,. \tag{2.7}$$

Now that we have a good approximation of the Hessian of the logistic risk, as well as a reasonable idea of the neighborhood of θ^* where the risk should behave quadratically in a way dictated by $\nabla^2 L(\theta^*)$ (hence by H), we sketch the analysis of the gradient rescaled by $H^{-1/2}$ and of the empirical hessians. More precisely, we show in the next section why a natural approach fails to capture the right behavior of the empirical gradient even in the case of a Gaussian design and a well-specified model. Then in Section 2.3, we take a first look at the empirical Hessians.

2.2 Structure of the empirical gradient

In this section we focus on the case where the design is Gaussian and the model is well-specified, in order to highlight the difficulties that we will face. The quantity that we want to bound is

$$\|\nabla \widehat{L}_n(\theta^*)\|_{H^{-1}} = \left\|\frac{1}{n}\sum_{i=1}^n \nabla \ell(\theta^*, (X_i, Y_i))\right\|_{H^{-1}} = \left\|\frac{1}{n}\sum_{i=1}^n \sigma(-Y_i\langle\theta^*, X_i\rangle)H^{-1/2}X_i\right\|. \tag{2.8}$$

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 (2.8) are H^{-1}-sub-Gaussian. By a standard deviation bound for sub-Gaussian vectors (recalled in Chapter 1 as Proposition 1.2), this implies that for all $t > 0$, with probability at least $1 - e^{-t}$,

$$\|\nabla \widehat{L}_n(\theta^*)\|_{H^{-1}} \lesssim \sqrt{\frac{\mathrm{Tr}(H^{-1}) + \|H^{-1}\|_{\mathrm{op}}t}{n}} \lesssim \sqrt{\frac{Bd + B^3 t}{n}}\,. \tag{2.9}$$

Unfortunately, this bound features a suboptimal dependence on the norm B. Indeed, in the well-specified case, the covariance of the gradient at θ^* is also the Hessian $\nabla^2 L(\theta^*)$, namely

$$\nabla^2 L(\theta^*) = \mathbf{E}[\nabla^2 \ell(\theta^*, Z)] = \mathbf{E}[\nabla \ell(\theta^*, Z) \nabla \ell(\theta^*, Z)^\top] \,.$$

Therefore, the vectors $H^{-1/2} \nabla \ell(\theta^*, Z_i)$ should be approximately isotropic and the correct bound should feature a first-order term scaling as $O(\sqrt{d/n})$ without any dependence in B; and a second-order term with a weak dependence in B.

To investigate this flaw, we have to check which step was loose in the obtention of (2.9). In light of (2.8), the temptation was to use that the X_i's are light-tailed, which is the case in particular if they are Gaussian vectors. It turns out that this is actually a misleading property, as it relies on bounding the sigmoid terms by 1, a step which is highly sub-optimal. 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

$$\mathbf{P}(Y_i \langle \theta^*, X_i \rangle < 0 | X_i) = \sigma(-|\langle \theta^*, X_i \rangle|) \leqslant \exp(-|\langle \theta^*, X_i \rangle|) \,. \tag{2.10}$$

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. This suggests that the correct upper bound on

$$\left\| \frac{1}{n} \sum_{i=1}^n \sigma(-Y_i \langle \theta^*, X_i \rangle) H^{-1/2} X_i \right\| = \sup_{v \in S^{d-1}} \frac{1}{n} \sum_{i=1}^n \sigma(-Y_i \langle \theta^*, X_i \rangle) \langle v, H^{-1/2} X_i \rangle \,, \tag{2.11}$$

should feature two terms, the first one being of exact order $O(\sqrt{d/n})$. To obtain this, a Bernstein-type bound would be ideal since it would naturally feature the true variance in the first-order term. Now that we know what is the ideal that we are chasing, it becomes clear that a first improvement with respect to (2.9) would be to use a Talagrand-type inequality for unbounded processes, namely Adamczak's inequality [Ada08]. Letting $G_i = \sigma(-Y_i \langle \theta^*, X_i \rangle) H^{-1/2} X_i$, using Adamczak's inequality would result in the following: with probability $1 - \delta$,

$$\left\| \frac{1}{n} \sum_{i=1}^n G_i \right\| \lesssim \sqrt{\frac{d + \log(1/\delta)}{n}} + \log n \, \frac{B^{1/2} d + B^{3/2} \log(1/\delta)}{n} \,. \tag{2.12}$$

A notable fact compared to the sub-Gaussian bound (2.9) is that the first-order term $\sqrt{\frac{d+\log(1/\delta)}{n}}$ is optimal, but the second order term still exhibits a poor dependence in B ($B^{3/2}$ in front of the level-of-confidence term), and an extra logarithmic factor in n.

The key observation to get a tight bound is that the sub-exponential (ψ_1) norm of the G_i's is much smaller than their sub-Gaussian (ψ_2) norm, more precisely,

$$\sup_{v \in S^{d-1}} \frac{\|\langle v, G \rangle\|_{\psi_2}}{\|\langle v, G \rangle\|_{L^2}} \asymp B^{3/2} \quad \text{and} \quad \sup_{v \in S^{d-1}} \frac{\|\langle v, G \rangle\|_{\psi_1}}{\|\langle v, G \rangle\|_{L^2}} \asymp B^{1/2} . \tag{2.13}$$

In addition, we observe the following consequence of the particular structure of the Hessian's proxy H defined in (2.6). Let $v \in S^{d-1}$ and decompose it as $v = \langle u^*, v\rangle u^* + \sqrt{1-\langle u^*, v\rangle^2} w$ where $w \in S^{d-1}$ is such that $\langle u^*, w\rangle = 0$. Then, we find that

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

This suggests that a better-suited approach is to control the moments

$$\mathbf{E}|\langle u^*, \nabla \ell(\theta^*, Z)\rangle|^p \quad \text{and} \quad \mathbf{E}|\langle w, \nabla \ell(\theta^*, Z)\rangle|^p\,, \tag{2.14}$$

for all $p \geqslant 1$. Specifically, the random variables $\langle v, \nabla \ell(\theta^*, Z)\rangle$ (with v ranging in S^{d-1}) can be shown to satisfy the *sub-gamma* property [BLM13, §2.4], which we recall in Definition 8.2. This final refinement compared to the sub-exponential property allows to apply Bernstein's inequality to the sum of the random variables $\langle v, \nabla \ell(\theta^*, Z_i)\rangle$. Then, going from direction-dependent bounds to a bound on the supremum over the sphere (2.11) can be done easily with a standard single-scale approximation argument, namely an ε-net. This approach remains tight despite its simple looks because although the matrix H is not isotropic, there is only one single problematic direction, which is that of θ^*. In a sense, it is "lightly" anisotropic, which enables the ε-net method to give the right second-order term. Also, this approach allows to get rid of the extra $\log n$ factor.

In a nutshell, once the structure of the Hessian at θ^* has been correctly identified, the issue in controlling the norm of the rescaled gradient is not handling the supremum of a process, it is to obtain the right point-wise bounds.

The discussion above sheds light on two important remarks. First, the analysis outlined above highlights the fact that even in the most simple setting (Gaussian design and well-specified model), a careful approach is needed to avoid any unnecessary dependence on the parameter norm B. Second, the crucial step in this analysis is to avoid bounding the sigmoid by 1, but instead taking advantage of the fact that it is dramatically smaller in most configurations, in the sense of (2.10). This property relies on the behavior of X near 0 in the direction of θ^*, meaning that it is related to the behavior of the quantity $\mathbf{P}(|\langle u^*, X\rangle| \leqslant t)$ for small values of t. The Gaussian distribution naturally exhibits the right behavior, and in this case, we obtain the exact rate $\psi_n = O(\sqrt{(d+\log(1/\delta))/n})$ (up to an absolute constant). This result is stated in Chapter 3, Proposition 3.1.

To conclude this section, we briefly discuss the other two cases: that of a non-Gaussian design but still assuming a well-specified model, and the most general case of a misspecified model. The specific difficulties arising in these settings will be discussed in greater detail in Chapter 6. In short, when the model is still assumed to be well-specified, the high-level argument sketched above remains valid; and in particular, the key remark (2.10) combined with the one-dimensional margin assumption (Assumption 5.2, in the definition of *regular* designs, see Chapter 6, Section 5.1) allow to bound the moments (2.14) almost similarly as in the Gaussian case. When the model is allowed to be misspecified, the key bound (2.10) no longer holds, which calls for another approach whose nature is different from that of the well-specified case. We provide a more detailed and technical discussion in Chapter 6.

2.3 Empirical Hessians

We now turn to the second component of the proof scheme that emerges from Lemma 2.1, namely a high-probability lower bound on the Hessians of the empirical risk:

$$\widehat{H}_n(\theta) = \nabla^2 \widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \sigma'(\langle \theta, X_i\rangle) X_i X_i^\top \,, \tag{2.15}$$

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 2.1 that an "ideal" guarantee would be of the form: for n large enough (depending on B, d, t),

$$\mathbf{P}\Big(\forall \theta \in \Theta,\ \widehat{H}_n(\theta) \succcurlyeq c_0 H\Big) \geqslant 1 - e^{-t} \,, \tag{2.16}$$

where

$$\Theta = \Big\{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_H \leqslant \frac{c_1}{\sqrt{B}}\Big\} \,,$$

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 (2.6).

As is clear from the expression (2.15), 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.

The desired inequality (2.16) can be restated as follows: for all $t > 0$ and all n large enough, it holds with probability at least $1 - e^{-t}$ that

$$\inf_{\theta \in \Theta, v \in S^{d-1}} \frac{1}{n}\sum_{i=1}^n \sigma'(\langle \theta, X_i\rangle)\langle v, X_i\rangle^2 \geqslant c_0 .$$

In other words, it is a uniform control on an empirical process index by the product set $\Theta \times S^{d-1}$. This is a highly subtle task because of the double indexation of the process and in particular its non-linear nature induced by σ'. If there was no such non-linearity, or if it had only mild consequences, the empirical Hessian matrices $\widehat{H}_n(\theta)$ would essentially all be equivalent to the empirical covariance of $X_1, \dots, X_n$. The behavior of these matrices—especially that of their lower tails—has attracted significant attention recently [Oli16, Mou22, Zhi24], but handling uniformity over a collection of random matrices requires a specific approach, that we now briefly outline.

2.4 Proof of Lemma 2.1 and additional results

We start with the proof of the localization lemma (Lemma 2.1).

Proof of Lemma 2.1. Let r be arbitrary such that $2\psi_n/c_0 < r < r_0$, which exists since $r_0 > 2\psi_n/c_0$ by assumption. For any $\theta \in \mathbb{R}^d$ such that $\|\theta - \theta^*\|_H = r$, a Taylor expansion

of order 2 shows that

$$\begin{aligned}
&\widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) \\
&\quad = \langle \nabla \widehat{L}_n(\theta^*), \theta - \theta^* \rangle + \int_0^1 (1-t) \big\langle \nabla^2 \widehat{L}_n\big((1-t)\theta^* + t\theta\big)(\theta - \theta^*), \theta - \theta^* \big\rangle \mathrm{d}t \\
&\quad \geqslant -\|\nabla \widehat{L}_n(\theta^*)\|_{H^{-1}} \|\theta - \theta^*\|_H + \frac{c_0}{2} \|\theta - \theta^*\|_H^2 \qquad (2.17) \\
&\quad \geqslant -\psi_n r + c_0 \frac{r^2}{2} > 0 \,, \qquad (2.18)
\end{aligned}$$

where inequality (2.17) comes from the fact that $\nabla^2 \widehat{L}_n((1-t)\theta^* + t\theta) \succcurlyeq c_0 H$ by assumption (ii), and (2.18) from the condition $r > 2\psi_n/c_0$. 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 $r \in (2\psi_n/c_0, r_0)$, we deduce that $\|\widehat{\theta}_n - \theta^*\|_H \leqslant 2\psi_n/c_0$.

The excess risk bound (2.1) 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

$$\max\{4\psi_n/c_0, 2\sqrt{\varepsilon/c_0}\} < r < r_0 \,.$$

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

$$\widehat{L}_n(\theta) - \widehat{L}_n(\widehat{\theta}_n) \geqslant \widehat{L}_n(\theta) - \widehat{L}_n(\theta^*) \geqslant -\psi_n r + c_0 r^2/2 \geqslant c_0 r^2/4 > \varepsilon \,,$$

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 $r \to \max\{4\psi_n/c_0, 2\sqrt{\varepsilon/c_0}\}$ 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 (2.4). We provide explicit constants for the estimates (2.5) on the components $c_0(\cdot), c_1(\cdot)$ of the Hessian. Lemma 2.2 below shows that

$$\frac{2\sqrt{2}}{3e^4\sqrt{\pi}} \min\Big(1, \frac{1}{\beta^3}\Big) \leqslant c_0(\beta) \leqslant 2\sqrt{\frac{2}{\pi}} \min\Big(1, \frac{1}{\beta^3}\Big) \,; \qquad (2.19)$$

$$\frac{1}{2e^4}\sqrt{\frac{2}{\pi}} \min\Big(1, \frac{1}{\beta}\Big) \leqslant c_1(\beta) \leqslant \sqrt{2} \min\Big(1, \frac{1}{\beta}\Big) \,. \qquad (2.20)$$

Lemma 2.2. *Let $G \sim \mathsf{N}(0,1)$. For any $\beta > 0$ and integer $k \geqslant 0$,*

$$\sqrt{\frac{2}{\pi}} \frac{2^{k+1}}{k+1} \min\left(\frac{1}{4e^4\beta^{k+1}}, \frac{\sigma'(2)}{e^2}\right) \leqslant \mathbf{E}[\sigma'(\beta G)|G|^k] \leqslant \sqrt{\frac{2}{\pi}} \min\left(\Gamma\Big(\frac{k+1}{2}\Big), \frac{k!}{\beta^{k+1}}\right) .$$

Proof. We have

$$\mathbf{E}[\sigma'(\beta G)|G|^k] = \sqrt{\frac{2}{\pi}} \int_0^{+\infty} x^k \sigma'(\beta x) \exp\Big(- \frac{x^2}{2}\Big) \mathrm{d}x \,.$$

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

$$\mathbf{E}[\sigma'(\beta G)|G|^k] \leqslant \sqrt{\frac{2}{\pi}} \min\left(\int_0^{+\infty} x^k \exp\Big(- \frac{x^2}{2}\Big) \mathrm{d}x, \int_0^{+\infty} x^k \exp\big(- \beta x\big) \mathrm{d}x \right).$$

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

$$\begin{aligned}
\mathbf{E}[\sigma'(\beta G)|G|^k] &\geqslant \sqrt{\frac{2}{\pi}} \int_0^2 x^k \sigma'(\beta x) e^{-x^2/2} \mathrm{d}x \\
&\geqslant \sqrt{\frac{2}{\pi}} \max\left\{ \frac{1}{4e^2} \int_0^2 x^k e^{-\beta x} \mathrm{d}x,\, \sigma'(2\beta) \int_0^2 x^k e^{-x^2/2} \mathrm{d}x \right\} \\
&= \sqrt{\frac{2}{\pi}} \max\left\{ \frac{1}{4e^2 \beta^{k+1}} \int_0^2 x^k e^{-x} \mathrm{d}x, \sigma'(2\beta) \int_0^2 x^k e^{-x^2/2} \mathrm{d}x \right\} \\
&\geqslant \sqrt{\frac{2}{\pi}} \frac{2^{k+1}}{k+1} \max\left(\frac{1}{4e^4 \beta^{k+1}}, \frac{\sigma'(2\beta)}{e^2} \right).
\end{aligned}$$

To get the lower bound, we use the first bound when $\beta > 1$ and the second one when $\beta \leqslant 1$. ∎

2.4.1 Concentration of sub-gamma random vectors

In all of the settings we consider, the bounds on empirical gradients are based on the following classical deviation inequality for sub-gamma random vectors. As mentioned before, the difficult part is to show that the gradients enjoy the right sub-gamma property. Once this is established, uniformity—hence bounds on the norm—follows from Lemma 2.3 below.

Lemma 2.3. *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 8.2). Then, for any $t > 0$, it holds with probability at least $1 - \exp(-t)$ that*

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

Proof. By Point 3 in Lemma 8.1, 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 8.1, for any $v \in V \cap S^{d-1}$ and any $t > 0$, it holds

$$\mathbf{P}\left(\frac{1}{n} \sum_{i=1}^n \langle v, Z_i \rangle > \nu \sqrt{\frac{2t}{n}} + K \frac{t}{n} \right) \leqslant \exp(-t) \,. \tag{2.21}$$

To make this bound uniform and conclude the proof, we use an ε-net argument. Let $\mathcal{N}$ 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 $\mathcal{N}$, so, for every $v \in B_V$, there exists $v' \in \mathcal{N}$ such that $\|v - v'\| \leqslant 1/2$. Therefore, for any $v \in B_V$,

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

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

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

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

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

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

Chapter 3

Risk bounds in the Gaussian, well-specified model

Abstract

This chapter focuses on the obtention risk bounds for the maximum-likelihood estimator (MLE) in logistic regression in the stylized case of a well-specified model with a Gaussian design. In this setting, we provide upper and lower bounds on the excess risk of the MLE that match up to universal constants. In particular, contrarily to recent results, there are no residual logarithmic factors in our main result, Theorem 3.1, neither for the sample size condition, nor for the corresponding bound on the excess risk. This proof follows the scheme outlined in the previous chapter by combining bounds on empirical gradient and Hessians. Despite the small number of parameters describing the problem, this setting already requires a very careful analysis to obtain sharp bounds.

Contents

3.1 Introduction

In this chapter, we consider the ideal case of a well-specified logit model with a Gaussian design. Specifically, we are given i.i.d. copies of a pair (X, Y) such that $X \sim \mathsf{N}(0, I_d)$ and $\mathbf{P}(Y = 1 \,|X) = \sigma(\langle \theta^*, X \rangle) = 1/(1 + e^{-\langle \theta^*, X \rangle})$ for some unknown parameter $\theta^* \in \mathbb{R}^d$. We will further assume that $\|\theta^*\| \geqslant e$, so that $B = \|\theta^*\|$ as in Section 1.4. The low signal regime will be addressed later, as in this case, there is almost no improvement between non-Gaussian and Gaussian design cases.

In this setting, the problem at hand—existence and performance of the maximum-likelihood estimator—is described in terms of a small number of parameters: the ambient dimension d, the sample size n, the signal strength B and the desired level of confidence $1 - \delta$. As it is a basic problem, we aim for results that exhibit optimal dependence with respect to all of these parameters.

Outline. This chapter is organized as follows. In Section 3.2 we state the main result of the chapter, which is Theorem 3.1. Then, in Section 3.3 we provide the main results needed to apply the localization result from the previous chapter, Lemma 2.1. These are: a bound on the empirical gradient at θ^* (Section 3.3.1, Proposition 3.1) and a uniform lower bound on the empirical Hessians (Section 3.3.2, Theorem 3.2). Then, we prove Proposition 3.1 in Section 3.4 and Theorem 3.2 in Section 3.3.2. Finally, Appendix 3.A provides the proof of Theorem 3.1.

We finally mention that Theorem 3.2, which provides a uniform lower bound on empirical Hessians, is one of the important contributions of this manuscript, and its proof is one of the most technically involved.

3.2 Main result

The main result of this chapter is Theorem 3.1 below. It provides a sharp condition (up to universal constant factors) on the sample size 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 Appendix 3.A.

Theorem 3.1. *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*

$$n \geqslant c_1 B(d + t)\,, \tag{3.1}$$

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

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c_2 \frac{d + t}{n}\,. \tag{3.2}$$

Moreover, for any $d \geqslant 53$ and $t > 0$, if $n \leqslant c_3 B(d+t)$ then the MLE exists with probability at most $1 - e^{-t}$.

It follows from Theorem 3.1 that, up to numerical constants, the condition (3.1) 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 (1.11) in

the regime where B, d are fixed while $n \to \infty$, which is optimal in light of the convergence in distribution (1.10). 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 2.1, the proof of Theorem 3.1 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 $1 - e^{-t}$,

$$\left\| \frac{\widehat{\theta}_n}{\|\widehat{\theta}_n\|} - \frac{\theta^*}{\|\theta^*\|} \right\| \leqslant c_4 \sqrt{\frac{d+t}{Bn}}, \qquad \big| \|\widehat{\theta}_n\| - \|\theta^*\| \big| \leqslant c_4 \sqrt{\frac{B^3(d+t)}{n}}. \tag{3.3}$$

The proof of Theorem 3.1 can be found in Section 3.A (combining results from Sections 3.4 and 3.5), while the scheme of proof is described in Chapter 2. In particular, a key structural result in the analysis is Theorem 3.2, 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 3.1 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 3.1. This can be seen as a quantitative version of the convergence to 1 in the phase transition (1.12) for the existence of the MLE established by Candès and Sur [CS20]. On the other hand, if $n \ll Bd$, then Theorem 3.1 (with $t \to 0$) 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 (1.12).

In Chapter 4 we establish that this is actually the case, which can similarly be interpreted as a quantitative version of convergence to 0 in the phase transition (1.12). The proof of this result is very different in nature from the proof of Theorem 3.1 as it is a purely geometric argument. For this reason, a whole chapter is devoted to this aspect of the Gaussian setting.

3.3 Main ingredients: bounds on gradient and empirical hessians

As explained in Chapter 2, there are two main steps to establish existence and risk bounds of the MLE with high probability. These are an upper bound on the H^{-1}-norm of the empirical gradient at θ^* and a uniform lower bound on the empirical Hessians in a neighborhood of θ^*. We start with a bound on the gradient in the next section.

3.3.1 Upper bound on the empirical gradient

Proposition 3.1 below provides an optimal bound on the deviation of the empirical gradient at θ^*, which derives from a deviation bound for sub-gamma random vectors (Lemma 2.3). We refer to the more detailed discussion in the previous chapter, Section 2.2 but in short, the difficult part is not to prove Lemma 2.3, it is to show that the one-dimensional projections of $\nabla \widehat{L}_n(\theta^*)$ enjoy the right property.

Proposition 3.1. *Assume that X is Gaussian and the model is well-specified. Let H be the matrix defined in (2.6). For any $t > 0$, if $n \geqslant 4B(d \log 5 + t)$ then with probability at*

least $1 - 2e^{-t}$,

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

This result is proved in Section 3.4. The proof follows the sketch given in Chapter 2, Section 2.2 by showing that the projections of the empirical gradient exhibit the right sub-gamma behavior (characterized by their moments), thanks to the low conditional probability of misclassification by θ^* in the strong signal regime. The proof of Proposition 3.1 highlights the somewhat counterintuitive fact that it is the behavior of the projections of X near 0 (more precisely at a scale of $1/B$) that yields the desired sub-gamma behavior, rather than their light tails.

3.3.2 Uniform lower bound on empirical Hessians

In this section, we provide the second key component to establish Theorem 3.1, a uniform lower bound on empirical Hessians. We emphasize the fact that Theorem 3.2 below is sharp up to universal constants. Its proof can be of independent interest from a random matrices perspective, and is one of the most technically involved contributions of this thesis.

Theorem 3.2. *Assume that* $X \sim \mathsf{N}(0, I_d)$. *For any* $t > 0$, *if* $n \geqslant 320000B(d+t)$ *then with probability at least* $1 - 2e^{-t}$,

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

The proof of Theorem 3.2 is provided in Section 3.5. 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 6.3. The reason for this is that the techniques used to establish Theorem 6.3 (specifically, the use of Vapnik-Chervonenkis arguments) inherently lead to additional logarithmic factors, hence the proof of Theorem 3.2 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 (2.15) 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 3.5 for more details on this point.

3.4 Proof of Proposition 3.1 (gradient)

Reduction to exponential moments. We start by reducing the proof to the obtention of bounds on the moments of one-dimensional projections. As alluded to in

the discussion of Section 2.2 in Chapter 2, the bound on the empirical gradient derives from the fact that the vector $H^{-1/2}\nabla\widehat{L}_n(\theta^*)$ is sub-gamma, using a classical deviation inequality (Lemma 2.3) recalled in Section 2.4.1. We are therefore left with the task of showing that all the projections $\langle v, H^{-1/2}\nabla\widehat{L}_n(\theta^*)\rangle$ for $v \in S^{d-1}$ indeed exhibit the right sub-gamma behavior. To prove this, we first use the definition of H to decompose this variable as follows. Recall the definition of H from Chapter 2,

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

where $u^* = \theta^*/\|\theta^*\|$, that we will use throughout. Let $v \in S^{d-1}$ and let $w \in S^{d-1}$ be such that $\langle u^*, w\rangle = 0$. Then, we can write $v = \langle u^*, v\rangle u^* + \sqrt{1-\langle u^*, v\rangle^2}\, w$, so that

$$\langle v, H^{-1/2}\nabla\widehat{L}_n(\theta^*)\rangle \leqslant B^{3/2}|\langle u^*, \nabla\widehat{L}_n(\theta^*)\rangle| + B^{1/2}|\langle w, \nabla\widehat{L}_n(\theta^*)\rangle|\,. \tag{3.4}$$

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 2.3, it is sufficient to prove that the variables $\langle v, \nabla\ell(\theta^*, Z_i)\rangle$ are sub-gamma and, in view of the decomposition (3.4), it is furthermore enough to prove this for $v = u^*$ and for v in any orthogonal direction, *i.e.*, when $\langle u^*, v\rangle = 0$. These random variables are centered, thus, by Point 4 in Lemma 8.1, this property can be obtained by proving a proper upper bound on the moments of these random variables. In addition, we point out that the independence of orthogonal projections of a standard Gaussian vector will be instrumental in obtaining sharp bounds.

The following result shows that the task at hand can be further reduced to that of bounding exponential moments, owing to the fact that the model is well-specified. This reduction will also be used in the non-Gaussian setting, when the model is still assumed to be well-specified (Chapter 6, Section 6.5.1).

Lemma 3.1. *Let $Z = (X, Y)$ denote a random variable taking value in $\mathbb{R}^d \times \{-1, 1\}$. Assume that this pair follows the logit model (1.5) with parameter θ^*. Let also $u^* = \theta^*/\|\theta^*\|$. It holds for any $p \geqslant 2$ and any $v \in S^{d-1}$ that*

$$\mathbf{E}|\langle v, \nabla\ell(\theta^*, Z)\rangle|^p \leqslant 2\mathbf{E}\big[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p\big]\,. \tag{3.5}$$

Proof. Recall that $\nabla\ell(\theta^*, Z) = -Y\sigma(-Y\langle\theta^*, X\rangle)X$. Since $\sigma(\cdot) \leqslant 1$, $\sigma(\cdot)^p \leqslant \sigma(\cdot)$, hence, for any $v \in S^{d-1}$,

$$\big|\big\langle v, \nabla\ell(\theta^*, Z)\big\rangle\big|^p \leqslant |\langle v, X\rangle|^p\sigma(-Y\langle\theta^*, X\rangle)\,. \tag{3.6}$$

Moreover,

$$\begin{aligned}\sigma(-Y\langle\theta^*, X\rangle) &= \sigma(-|\langle\theta^*, X\rangle|)\mathbf{1}\{Y\langle\theta^*, X\rangle > 0\} + \sigma(\langle\theta^*, X\rangle)\mathbf{1}\{Y\langle\theta^*, X\rangle < 0\}\\ &\leqslant \sigma\big(-|\langle\theta^*, X\rangle|\big) + \mathbf{1}\{Y\langle\theta^*, X\rangle < 0\}\,.\end{aligned} \tag{3.7}$$

Now, conditioning on X one has

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

where the last inequality holds since $\sigma(-t) \leqslant \exp(-t)$ for all $t \geqslant 0$. Using this and taking the conditional expectation on X in (3.7), we deduce that

$$\begin{aligned}\mathbf{E}[\sigma(-Y\langle\theta^*, X\rangle)\,|X] &\leqslant \exp(-|\langle\theta^*, X\rangle|) + \mathbf{P}(Y\langle\theta^*, X\rangle < 0\,|X) \\ &\leqslant 2\exp(-|\langle\theta^*, X\rangle|)\,. \end{aligned} \tag{3.8}$$

Finally, taking expectation in (3.6) yields

$$\begin{aligned}\mathbf{E}\big|\big\langle v, \nabla\ell(\theta^*, Z)\big\rangle\big|^p &\leqslant \mathbf{E}\big[|\langle v, X\rangle|^p \mathbf{E}[\sigma(-Y\langle\theta^*, X\rangle)\,|X]\big] \\ &\leqslant 2\mathbf{E}\big[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p\big]\,. \end{aligned}$$

■

We have therefore reduced the problem of bounding the empirical gradient to an inequality only involving X and not Y. We now move on to the proof of Proposition 3.1

Proof of Proposition 3.1. By Lemma 3.1, we have to bound the random variables

$$\mathbf{E}[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p]\,,$$

where X is a standard Gaussian vector. 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

$$\mathbf{E}[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p] \leqslant \mathbf{E}|\langle v, X\rangle|^p = \frac{\sqrt{2}^p}{\sqrt{\pi}}\Gamma\left(\frac{p+1}{2}\right) \leqslant \frac{p!}{\sqrt{\pi}}\,. \tag{3.9}$$

By Point 4 in Lemma 8.1, the variable $\langle v, \nabla\ell(\theta^*, Z_i)\rangle$ is thus $(2/\sqrt{\pi}, 1)$ sub-gamma. Hence, by Lemma 2.3, for any $t > 0$, with probability larger than $1 - \exp(-t)$, simultaneously for all $v \in S^{d-1}$,

$$\langle v, \nabla\widehat{L}_n(\theta^*)\rangle \leqslant 2\left(\frac{2}{\pi^{1/4}}\sqrt{\frac{d\log 5 + t}{n}} + \frac{d\log 5 + t}{n}\right) \leqslant 6\sqrt{\frac{d+t}{n}}\,, \tag{3.10}$$

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 $6e^{3/2} < 27$ 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

$$\begin{aligned}\mathbf{E}\big[\exp(-|\langle\theta^*, X\rangle|)|\langle u^*, X\rangle|^k\big] &= \mathbf{E}\big[\exp(-B|\langle u^*, X\rangle|)|\langle u^*, X\rangle|^k\big] \\ &\leqslant \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} |x|^k \exp(-B|x|)\mathrm{d}x = \sqrt{\frac{2}{\pi}}\frac{k!}{B^{k+1}}\,. \end{aligned} \tag{3.11}$$

Thus, (3.11) and (3.5) show that

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

This shows, by Bernstein's inequality recalled in (2.21) that, for any $t > 0$, with probability larger than $1 - \exp(-t)$

$$|\langle u^*, \nabla\widehat{L}_n(\theta^*)\rangle| \leqslant \frac{1}{B^{3/2}}\sqrt{\frac{t}{n}}\left[2\left(\frac{8}{\pi}\right)^{1/4} + \sqrt{\frac{Bt}{n}}\right] \leqslant \frac{3}{B^{3/2}}\sqrt{\frac{t}{n}}\,, \tag{3.12}$$

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

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

We bound the first term in the right-hand side using (3.11) with $k = 0$ and the second one with (3.9) to get

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

By Point 4 in Lemma 8.1, the vectors $\langle v, \nabla \ell(\theta^*, Z_i) \rangle$ are $(8/(\pi B), 1)$ sub-gamma. Hence, by Lemma 2.3, for any $t > 0$, with probability larger than $1 - \exp(-t)$, simultaneously for any $w \in S^{d-1}$ such that $\langle w, u^* \rangle = 0$,

$$\langle w, \nabla \widehat{L}_n(\theta^*) \rangle \leqslant 8\sqrt{\frac{d \log 5 + t}{\pi B n}} + \frac{2(d \log 5 + t)}{n} \leqslant 6\sqrt{\frac{d+t}{Bn}} \,. \tag{3.13}$$

Plugging (3.12) and (3.13) into (3.4) concludes the proof of the second part of the proposition. ■

The next section is devoted to the proof of Theorem 3.2 on empirical Hessian matrices.

3.5 Proof of Theorem 3.2 (Hessian matrices)

This section is devoted to the proof of the uniform lower bound on empirical Hessian matrices, namely Theorem 3.2. A detailed, high-level sketch of this proof is provided in the second part of Section 1.4.1, where we highlight the key steps.

In this section, we define the ellipsoid

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

We would like to show that with high probability, one has $\widehat{H}_n(\theta) \succcurlyeq cH$ for every $\theta \in \Theta$, where c is an absolute constant. This property amounts to

$$\lambda_{\min}(H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}) \geqslant c$$

for every $\theta \in \Theta$, and therefore to

$$\inf_{\theta \in \Theta,\, v \in S^{d-1}} \langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v \rangle = \inf_{\theta \in \Theta,\, v \in S^{d-1}} \left\{ \frac{1}{n}\sum_{i=1}^n \sigma'(\langle \theta, X_i \rangle)\langle v, H^{-1/2}X_i \rangle^2 \right\} \geqslant c \,. \tag{3.14}$$

We are therefore led to control the infimum of an empirical process indexed by $(\theta, v) \in \Theta \times S^{d-1}$.

The main challenge towards such a control is the presence of the nonlinear terms $\sigma'(\langle \theta, X_i \rangle)$, and the fact that we aim for a fully sharp estimate. This rules out the use of VC arguments as in the regular case (Chapter 6, Theorem 6.3), since such methods would necessarily produce additional logarithmic factors in the low-noise regime we consider.

We will achieve this by using the PAC-Bayes inequality; see e.g. [CG17, Proposition 2.1]. The use of this inequality to control empirical processes was pioneered by Audibert and Catoni [AC11] in the context of robust linear regression, and has since

found applications to matrix concentration [Oli16, Cat16, Mou22, Zhi24] and robust covariance matrix estimation [CG17, Giu18, AZ24, OR24, MZ25], among others. In our logistic regression setting, it underpins the only approach we know of for establishing the optimal estimate of Theorem 3.2 on the empirical Hessian.

In order to state this inequality, recall that if μ and ν denote two probability measures on the same measurable space $(E, \mathcal{E})$ such that ν is absolutely continuous with respect to μ, the Kullback-Leibler divergence (relative entropy) between ν and μ is defined by

$$D(\nu\|\mu) = \int_\Omega \log\Big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\Big)\mathrm{d}\nu\,. \tag{3.15}$$

Lemma 3.2 (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 $Z_1, \dots, Z_n$ be independent copies of the process Z. Let also $\lambda > 0$ be such that $\mathbf{E}[\exp(\lambda Z(\omega))] < \infty$ 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,*

$$\frac{1}{n}\sum_{i=1}^n \int_E Z_i(\omega)\rho(\mathrm{d}\omega) \leqslant \frac{1}{\lambda}\int_E \log\Big(\mathbf{E}\big[e^{\lambda Z(\omega)}\big]\Big)\rho(\mathrm{d}\omega) + \frac{D(\rho\|\pi) + t}{\lambda n}\,. \tag{3.16}$$

Since we aim to obtain a lower bound on the empirical process (3.14), it will be more convenient to apply Lemma 3.2 to the process $-Z$, which yields the following inequality:

$$\frac{1}{n}\sum_{i=1}^n \int_E Z_i(\omega)\rho(\mathrm{d}\omega) \geqslant -\frac{1}{\lambda}\int_E \log\mathbf{E}\big[\exp(-\lambda Z(\omega))\big]\rho(\mathrm{d}\omega) - \frac{D(\rho\|\pi) + t}{\lambda n} \tag{3.17}$$

for every $\lambda > 0$ such that $\mathbf{E}[\exp(-\lambda Z(\omega))] < \infty$, which always holds if $Z \geqslant 0$.

Inequality (3.17) depends on the choice of a fixed distribution π on the index space E, which we call the "prior". In addition, it yields a lower bound on a *smoothed* version of the process Z, obtained by averaging with respect to distributions ρ on E referred to as "posteriors". Although the posterior distribution ρ can in principle be arbitrary, the bound (3.17) depends on the relative entropy $D(\rho\|\pi)$ between ρ and the prior π. This prevents one from choosing posterior distributions that are (arbitrarily close to) Dirac masses at parameters, as such choices would render the bound vacuous. Here, the relative entropy $D(\rho\|\pi)$ quantifies the "complexity" of the posterior ρ, and constitutes the price of requiring that the bound hold simultaneously for all posteriors.

In practice, we must select a prior distribution π on a parameter set $E \subset \mathbb{R}^d \times S^{d-1}$; a process Z to which the PAC-Bayes inequality (3.17) is applied; and for each $(\theta, v) \in \Theta \times S^{d-1}$, a posterior/smoothing distribution $\rho_{\theta,v}$ on E. We must then obtain the following:

1. a lower bound on the *log-Laplace transform*: $-\log\mathbf{E}[\exp(-\lambda Z(\omega))]$ for every $\omega \in E$;
2. an upper bound on the *relative entropy* $D(\rho_{\theta,v}\|\pi)$ for every $(\theta, v) \in \Theta \times S^{d-1}$;
3. a control on the *smoothing approximation error*, namely a lower bound on the quantity of interest $\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle$ in terms of the smoothed process

$$\frac{1}{n}\sum_{i=1}^n \int_E Z_i(\omega)\rho_{\theta,v}(\mathrm{d}\omega) \qquad \text{for all } (\theta, v) \in \Theta \times S^{d-1}\,.$$

In particular, the above outline suggests a trade-off in the choice of the posterior $\rho_{\theta,v}$: on the one hand, it must be concentrated enough near its corresponding parameter (θ, v) that the smoothing error is small; on the other hand, it must be diffuse enough that the relative entropy $D(\rho_{\theta,v}\|\pi)$ is not too large, uniformly over $(\theta, v) \in \Theta \times S^{d-1}$. The latter condition also requires a suitable choice of prior π.

Now, from (3.14), the term we aim to control corresponds to an empirical process:

$$\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle = \frac{1}{n}\sum_{i=1}^{n} Z_i^{\mathrm{int}}(\theta, v)\,, \quad \text{where} \quad Z_i^{\mathrm{int}}(\theta, v) = \sigma'(\langle\theta, X_i\rangle)\langle v, H^{-1/2}X_i\rangle^2\,.$$

Hence, a natural approach is to apply the PAC-Bayes inequality to $Z = Z^{\mathrm{int}}$, the process of interest, and then to control the difference $Z(\theta, v) - \int_E Z(\omega)\rho_{\theta,v}(\mathrm{d}\omega)$ between this process and its smoothed version.

We will follow a different approach: we will apply the PAC-Bayes inequality to *an auxiliary process* Z^{aux} *whose smoothed version yields* (a lower bound on) *the process of interest* Z^{int}. Specifically, the sigmoid $\sigma'(\cdot)$ in the process of interest $Z^{\mathrm{int}}(\theta, v) = \sigma'(\langle\theta, X\rangle)\langle v, H^{-1/2}X\rangle^2$ will be replaced by a suitable indicator. In addition, the posterior distribution ρ_θ on θ will critically involve a non-Gaussian component, in order to ensure the previous property.

Concretely, we will show that for a suitable choice of posteriors $\rho_{\theta,v} = \rho_\theta \otimes \rho_v$ (see Definition 3.1 for the definition of ρ_θ), we have for every $(\theta, v) \in \Theta \times S^{d-1}$,

$$\begin{aligned}
&\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\sigma'(\langle\theta, X_i\rangle)\langle v, H^{-1/2}X_i\rangle^2 \\
&\quad \geqslant \frac{c}{n}\sum_{i=1}^{n}\int_{\Theta\times S^{d-1}} \mathbf{1}\big\{|\langle\theta', X_i\rangle| \leqslant 1,\ \|X_i\| \leqslant 2\sqrt{d}\big\}\langle v', H^{-1/2}X_i\rangle^2\rho_\theta(\mathrm{d}\theta')\rho_v(\mathrm{d}v') - (\mathsf{Rem})
\end{aligned}$$

for some constant $c > 0$. A key step to achieve this is the smoothing result of Lemma 3.4 below. In addition, the remainder term (Rem) above comes from the effect of smoothing over v, and depends on $X_1, \ldots, X_n$. It is itself bounded through a separate application of PAC-Bayes, this time over spherical caps in S^{d-1} localized around $u^* = \theta^*/\|\theta^*\|$ (Lemma 3.10).

We next proceed to the remainder of the proof, following the outline described above. First, we define the auxiliary process Z to which the PAC-Bayes inequality is applied, and control its log-Laplace transform. Second, we define posterior distributions ρ_θ over parameters $\theta' \in \mathbb{R}^d$, and establish an approximation guarantee for smoothing under such posterior distributions. Third, we define posterior distributions ρ_v over directions $v' \in S^{d-1}$, and control the smoothing approximation error under these posterior distributions. Fourth, we define the prior distribution π over pairs $(\theta', v') \in \mathbb{R}^d \times S^{d-1}$, and bound the relative entropy $\mathrm{KL}(\rho_\theta \otimes \rho_v, \pi)$ between the posterior and prior. We then put things together to conclude the proof. To keep the exposition streamlined, some technical lemmas used in this section are deferred to Section 3.5.1.

Auxiliary process and control of its log-Laplace transform

We first define the index set E and auxiliary process Z to which the PAC-Bayes inequality (3.17) will be applied. We let $E = \Theta' \times S^{d-1}$, where Θ' is the enlarged ellipsoid defined

by

$$\Theta' = \Big\{\theta' \in \mathbb{R}^d : \|\theta' - \theta^*\|_H \leqslant \frac{1}{10\sqrt{B}}\Big\} \supset \Theta\,. \tag{3.18}$$

In addition, for $\omega = (\theta', v') \in \Theta' \times S^{d-1}$ we let

$$Z(\omega) = \mathbf{1}\Big\{|\langle\theta', X\rangle| \leqslant 1; \|X\| \leqslant 2\sqrt{d}\Big\}\,\big\langle v', H^{-1/2}X\big\rangle^2\,. \tag{3.19}$$

Likewise, for $i = 1, \dots, n$, we define $Z_i(\omega)$ by replacing X with X_i in (3.19).

We now control the log-Laplace transform of the process Z, which will provide a lower bound on the first term of the right-hand side of (3.17).

Lemma 3.3. *For any $\omega = (\theta', v') \in E = \Theta' \times S^{d-1}$ and $\lambda \geqslant 0$, we have*

$$-\log \mathbf{E}\big[\exp(-\lambda Z(\omega))\big] \geqslant 0.03\lambda - 3\lambda^2 B\,.$$

Proof. Since $\exp(-s) \leqslant 1 - s + s^2/2$ for any $s \geqslant 0$, we have

$$\mathbf{E}\big[\exp(-\lambda Z(\omega))\big] \leqslant 1 - \lambda\mathbf{E}[Z(\omega)] + \frac{\lambda^2}{2}\mathbf{E}[Z(\omega)^2]\,.$$

Since in addition $-\log(1-s) \geqslant s$ for any $s \geqslant 0$, we deduce that

$$-\log \mathbf{E}\big[\exp(-\lambda Z(\omega))\big] \geqslant \lambda\mathbf{E}[Z(\omega)] - \frac{\lambda^2}{2}\mathbf{E}[Z(\omega)^2]\,. \tag{3.20}$$

In light of (3.20), in order to prove Lemma 3.3 it suffices to establish a lower bound on $\mathbf{E}[Z(\omega)]$ and an upper bound on $\mathbf{E}[Z(\omega)^2]$.

Let $u' = \theta'/\|\theta'\|$ ($u' \in S^{d-1}$ can be arbitrary if $\theta' = 0$), and define the matrices

$$\widetilde{H}(\theta') = \mathbf{E}\Big[\mathbf{1}\Big\{\langle\theta', X\rangle \leqslant 1; \|X\| \leqslant 2\sqrt{d}\Big\}XX^\top\Big]\,; \tag{3.21}$$

$$H_{\theta'} = \frac{1}{B^3}u'u'^\top + \frac{1}{B}\big(I_d - u'u'^\top\big)\,. \tag{3.22}$$

Lemmas 3.8 and 3.7 below imply respectively that, since $\theta' \in \Theta'$,

$$H_{\theta'} \succcurlyeq 0.75\,H \quad \text{and} \quad \widetilde{H}(\theta') \succcurlyeq 0.05\,H_{\theta'} \succcurlyeq 0.03\,H\,. \tag{3.23}$$

With these prerequisites in place, we bound $\mathbf{E}[Z(\omega)]$ from below. Using (3.23), we have

$$\begin{aligned}\mathbf{E}[Z(\omega)] &= \mathbf{E}\Big[\mathbf{1}\Big\{|\langle\theta', X\rangle| \leqslant 1; \|X\| \leqslant 2\sqrt{d}\Big\}\,\langle v', H^{-1/2}X\rangle^2\Big] \\ &= \langle H^{-1/2}\widetilde{H}(\theta')H^{-1/2}v', v'\rangle \geqslant 0.03\|v'\|^2 = 0.03\,.\end{aligned} \tag{3.24}$$

Next, we bound $\mathbf{E}[Z(\omega)^2]$ from above. Denoting $v = H_{\theta'}^{1/2}H^{-1/2}v'$, we have

$$\mathbf{E}[Z(\omega)^2] \leqslant \mathbf{E}\big[\mathbf{1}\{|\langle\theta', X\rangle| \leqslant 1\}\langle H^{-1/2}v', X\rangle^4\big] = \mathbf{E}\big[\mathbf{1}\{|\langle\theta', X\rangle| \leqslant 1\}\langle H_{\theta'}^{-1/2}v, X\rangle^4\big]\,.$$

In addition, since

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

using that $\langle u', X\rangle$ and $\langle v - \langle u', v\rangle u', X\rangle$ are independent centered Gaussian random variables, and bounding

$$|\langle u', v\rangle| \leqslant \|v\| \quad \text{and} \quad \|v - \langle u', v\rangle u'\| \leqslant \|v\|\,,$$

we obtain that

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

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

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

In addition, the first inequality in (3.23) (and the fact that $\|v'\| = 1$) implies that

$$\|v\|^4 = \langle H^{-1/2} H_{\theta'} H^{-1/2} v', v'\rangle^2 \leqslant \Big(\frac{4}{3}\|v'\|^2\Big)^2 = \frac{16}{9}\,.$$

Finally, it follows from Point 1 in Lemma 8.4 that $\|\theta'\| \geqslant 0.9B$ (as $B = \|\theta^*\| \geqslant e$). Combining the previous inequalities, we obtain

$$\mathbf{E}\big[Z(\omega)^2\big] \leqslant \frac{16}{9}\sqrt{\frac{2}{\pi}}\bigg[B^6 \times \frac{1}{5B^5}\Big(\frac{10}{9}\Big)^5 + 6B^4 \times \frac{1}{3B^3}\Big(\frac{10}{9}\Big)^3 + B^2 \times \frac{10}{9B}\bigg] \leqslant 6B\,. \quad (3.25)$$

Plugging the bounds (3.24) and (3.25) into (3.20) concludes the proof. ∎

Posterior distributions and smoothing over parameters θ

As mentioned previously, the posteriors we will use on $E = \Theta' \times S^{d-1}$ will be of the form $\rho_{\theta,v} = \rho_\theta \otimes \rho_v$ for $\theta \in \Theta$ and $v \in S^{d-1}$, where ρ_θ is a distribution on Θ' and ρ_v a distribution on S^{d-1} (with a slight abuse of notation, we use similar notation for both posterior distributions).

In this section, for every $\theta \in \Theta$, we define the posterior distribution ρ_θ over parameters $\theta' \in \Theta'$, and establish an approximation result for smoothing under such distributions.

Definition 3.1. 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 $Z' \sim \mathsf{N}(0, (I_d - uu^\top)/(2\cdot 10^4 \cdot d))$ given that $\|Z'\| \leqslant 1/100$, where $u = \theta/\|\theta\|$.

The motivation behind the choice of the posterior (or smoothing distribution) ρ_θ in Definition 3.1 is twofold. On the one hand, it is sufficiently spread out that, for every $\theta \in \Theta$, the relative entropy between ρ_θ and a suitably chosen prior is controlled: Lemma 3.6 below shows 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 3.4 below.

Lemma 3.4. *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*

$$\sigma'(\langle \theta, x\rangle) \geqslant \frac{1}{15} \int_{\mathbb{R}^d} \mathbf{1}\big\{|\langle \theta', x\rangle| \leqslant 1\big\} \rho_\theta(\mathrm{d}\theta') \,. \tag{3.26}$$

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 3.1 was Gaussian (say, if $U \sim \mathsf{N}(1, 0.01^2)$): in this case, the smoothed indicator would be of order $(1 + |\langle \theta, x\rangle|)^{-1}$, which is much larger than $\sigma'(|\langle \theta, x\rangle|)$.

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 3.1. By Lemma 3.8, since $\|\theta - \theta^*\|_H \leqslant 1/100\sqrt{B}$, 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 8.4, as $\theta \in \Theta$,

$$\|\theta' - \theta\|_H \leqslant \frac{0.015}{\sqrt{B}} \qquad \text{and} \qquad \|\theta' - \theta^*\|_H \leqslant \frac{0.025}{\sqrt{B}} \,. \tag{3.27}$$

We now prove inequality (3.26). Let $x \in \mathbb{R}^d$ be such that $\|x\| \leqslant 2\sqrt{d}$. We have

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

If $Z' \sim \mathsf{N}(0, (I_d - uu^\top)/(2 \cdot 10^4 \cdot d))$, we have, as $\mathbf{P}(\|Z'\| \leqslant 1/100) \geqslant 1 - 10^4 \mathbf{E}\|Z'\|^2 \geqslant 3/4$,

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

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

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

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

We apply this result with

$$\sigma^2 = \mathrm{Var}(\langle g', x\rangle) \leqslant \frac{\|x\|^2}{2 \cdot 10^4 \cdot d} \leqslant \frac{1}{5000}, \qquad b = \frac{1}{U} \leqslant \frac{1}{0.99}, \qquad a = U\langle \theta, x\rangle \,.$$

This shows that

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

This proves the lower bound (3.26). ∎

Posterior distributions and smoothing over directions v

We now define, for every $v \in S^{d-1}$, the posterior ρ_v over directions $v' \in S^{d-1}$. We then control the approximation error that arises from smoothing over ρ_v.

Definition 3.2. Let $\varepsilon \in (0,1)$. For any $v \in S^{d-1}$, let ρ_v denote the uniform distribution on the spherical cap $\mathsf{C}(v,\varepsilon)$ of radius ε around v, defined by

$$\mathsf{C}(v,\varepsilon) = \left\{ v' \in S^{d-1} : \langle v, v' \rangle \geqslant \sqrt{1-\varepsilon^2} \right\}. \tag{3.28}$$

We next show that the empirical Hessian can be lower-bounded in terms of the smoothed process and a remainder term R_n.

Lemma 3.5. *Let $\rho_{\theta,v} = \rho_\theta \otimes \rho_v$ denote the posterior distribution defined as the product of the posterior ρ_θ of Definition 3.1 and the posterior ρ_v of Definition 3.2. Then, for any $(\theta, v) \in \Theta \times S^{d-1}$,*

$$\langle H^{-1/2} \widehat{H}_n(\theta) H^{-1/2} v, v \rangle \geqslant \frac{1}{15n} \sum_{i=1}^n \int_{\Theta' \times S^{d-1}} Z_i(\omega) \rho_{\theta,v}(\mathrm{d}\omega) - 22\varepsilon^2 B R_n \,, \tag{3.29}$$

where

$$R_n = \sup_{u \in \mathsf{C}(u^*, 1/10B)} \left\{ \frac{1}{n} \sum_{i=1}^n \exp\big(-0.49\, B |\langle u, X_i \rangle| \big) \mathbf{1}\big(\|X_i\| \leqslant 2\sqrt{d} \big) \right\}.$$

Proof. By Lemma 3.4, for every $v' \in S^{d-1}$ one has

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

Integrating over $v' \sim \rho_v$, we obtain

$$\begin{aligned}
&\frac{1}{15n} \sum_{i=1}^n \int_{\Theta' \times S^{d-1}} Z_i(\omega) \rho_{\theta,v}(\mathrm{d}\omega) \\
&\quad \leqslant \frac{1}{n} \sum_{i=1}^n \int_{S^{d-1}} \sigma'(\langle \theta, X_i \rangle) \mathbf{1}\Big\{ \|X_i\| \leqslant 2\sqrt{d} \Big\} \langle X_i, H^{-1/2} v' \rangle^2 \rho_v(\mathrm{d}v') \\
&\quad = \int_{\mathsf{C}(v,\varepsilon)} \langle H^{-1/2} \overline{H}_n(\theta) H^{-1/2} v', v' \rangle \rho_v(\mathrm{d}v') \,,
\end{aligned} \tag{3.30}$$

where

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

Using the computations from [Mou22, Eqs. (42) and (43)] and Fact 8.1, we get

$$\begin{aligned}
&\int_{\mathsf{C}(v,\varepsilon)} \langle H^{-1/2} \overline{H}_n(\theta) H^{-1/2} v', v' \rangle \rho_v(\mathrm{d}v') \\
&\qquad \leqslant \big\langle H^{-1/2} \widehat{H}_n(\theta) H^{-1/2} v, v \big\rangle + \frac{2\varepsilon^2}{d-1} \mathrm{Tr}\left(H^{-1/2} \overline{H}_n(\theta) H^{-1/2} \right) .
\end{aligned} \tag{3.31}$$

We now control the trace term in (3.31). First, by Lemma 3.8, one has $H^{-1} \preccurlyeq 1.03 H_\theta^{-1}$ for any $\theta \in \Theta$, thus

$$\begin{aligned}
&\operatorname{Tr}\big(H^{-1/2}\overline{H}_n(\theta)H^{-1/2}\big) = \operatorname{Tr}\big(\overline{H}_n(\theta)^{1/2}H^{-1}\overline{H}_n(\theta)^{1/2}\big) \\
&\leqslant 1.03\operatorname{Tr}(\overline{H}_n(\theta)^{1/2}H_\theta^{-1}\overline{H}_n(\theta)^{1/2}) = 1.03\operatorname{Tr}\big(H_\theta^{-1/2}\overline{H}_n(\theta)H_\theta^{-1/2}\big)\,.
\end{aligned}$$

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

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

Now, we use the inequalities $\sigma'(t) \leqslant e^{-|t|}$ and $t^2e^{-t} = (t/2)^2e^{t/2}\cdot 2^2e^{-t/2} \leqslant 4e^{-2}2^2e^{-t/2}$ for all $t \geqslant 0$ and, by Lemma 8.4, $\|\theta\| \geqslant 0.99B$ (as $\theta \in \Theta$ and $B = \|\theta^*\| \geqslant e$) to get

$$B^2\langle u, X_i\rangle^2\sigma'\big(\langle\theta, X_i\rangle\big) \leqslant 2.2\exp\big(-0.49B|\langle u, X_i\rangle|\big)\,.$$

Plugging this into the previous inequality gives (using that $d \geqslant 2$)

$$\begin{aligned}
\operatorname{Tr}\big(H^{-1/2}\overline{H}_n(\theta)H^{-1/2}\big) &\leqslant 1.03B(2.2+4d)\cdot\frac{1}{n}\sum_{i=1}^n \exp\big(-0.49B|\langle u, X_i\rangle|\big)\mathbf{1}(\|X_i\| \leqslant 2\sqrt{d}) \\
&\leqslant 5.5Bd\cdot R_n\,, \qquad (3.32)
\end{aligned}$$

where the last inequality comes from the fact that $u \in \mathsf{C}(u^*, 1/10B)$ as (by Lemma 8.4 and using that $\theta \in \Theta$) $\|u - u^*\| \leqslant 0.1/B$.

Combining inequalities (3.30), (3.31) and (3.32) gives (bounding $d - 1 \geqslant d/2$):

$$\frac{1}{15n}\sum_{i=1}^n \int_{\Theta'\times S^{d-1}} Z_i(\omega)\rho_{\theta,v}(\mathrm{d}\omega) \leqslant \big\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\big\rangle + 22\varepsilon^2 BR_n\,,$$

which concludes the proof. ∎

Note that Lemma 3.5 features an empirical remainder term R_n, which must also be controlled. Since this term is bounded using a second application of the PAC-Bayes inequality over a different space than $\Theta' \times S^{d-1}$, we defer the proof of this bound (Lemma 3.10 below) to a paragraph following the conclusion of the proof.

Prior distribution and bound on the relative entropy

We now define the prior distribution π and bound from above the relative entropy term $D(\rho_{\theta,v}\|\pi)$, where $\rho_{\theta,v}$ was defined in Lemma 3.5.

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 $\mathsf{N}(\mu, \Sigma \,|\, S)$ denote the Gaussian distribution $\mathsf{N}(\mu, \Sigma)$ conditioned on S, that is the distribution ν with density

$$\mathrm{d}\nu = \frac{\mathbf{1}_S}{\gamma(S)}\,\mathrm{d}\gamma\,, \qquad (3.33)$$

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

Definition 3.3. The prior distribution π on $\Theta \times S^{d-1}$ is the product measure $\pi = \pi_\Theta \otimes \pi_S$, where π_S is the uniform distribution on S^{d-1} and $\pi_\Theta = \mathsf{N}(\theta^*, \Gamma \,|\, \Theta')$, where

$$\Gamma = \frac{1}{10^4}\left[B^2 u^* u^{*\top} + \frac{1}{2d}\big(I_d - u^* u^{*\top}\big)\right].$$

The relative entropy term is bounded in the following lemma:

Lemma 3.6. *Let $\theta \in \Theta$ and $v \in S^{d-1}$. Let $\rho_{\theta,v}$ denote the prior defined in Lemma 3.5 and π denote the prior distribution of Definition 3.3. Then,*

$$D(\rho_{\theta,v} \| \pi) \leqslant \left(6.5 + \log\left(1 + \frac{2}{\varepsilon}\right)\right) d\,. \tag{3.34}$$

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 8.1, this yields

$$D(\rho_v \| \pi_S) \leqslant d \log\left(1 + \frac{2}{\varepsilon}\right). \tag{3.35}$$

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 (3.38), 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 $\tilde{\rho}_\theta = \mathsf{N}(\theta, \Gamma_\theta \,|\, \mathcal{E}_\theta)$ 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 3.4,

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

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

$$\Gamma_\theta = \frac{1}{10^4}\left(B^2 u u^\top + \frac{1}{2d}(I_d - u u^\top)\right), \qquad u = \frac{\theta}{\|\theta\|}\,.$$

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

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

We start with point 1. We have on one hand that the density f_θ of ρ_θ satisfies, for every $\theta_0 = tu + z$,

$$f_\theta(\theta_0) \leqslant \frac{1}{p0.02\|\theta\|}\Big(\frac{10^4 d}{\pi}\Big)^{\frac{d-1}{2}} \exp\Big(-\frac{d\|z\|^2}{10^4}\Big)\mathbf{1}\Big(t/\|\theta\| \in [0.99, 1.01]; \|z\| \leqslant 1/100\Big)$$

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

$$\tilde{f}_\theta(\theta_0) = \frac{1}{\gamma_\theta(\mathcal{E}_\theta)} \cdot \frac{e^{-\frac{(t-\|\theta\|)^2}{2(B/100)^2}}}{B/100} \cdot \frac{(2\cdot 10^4 d)^{\frac{d-1}{2}}}{(2\pi)^{d/2}} \exp\big(-10^4 d\|z\|^2\big)\mathbf{1}(\theta_0 \in \mathcal{E}_\theta)\,.$$

For any t such that $|t/\|\theta\| - 1| \leqslant 1/100$ and $\theta \in \Theta$ so, by Lemma 8.4, $\|\theta\|/B \in [0.99, 1.01]$, we deduce that

$$\frac{f_\theta(\theta_0)}{\tilde{f}_\theta(\theta_0)} \leqslant \frac{1}{2p}\frac{B}{\|\theta\|}\sqrt{\frac{\pi}{2}}\gamma_\theta(\mathcal{E}_\theta)\exp\Big(\frac{(t-\|\theta\|)^2}{2(B/100)^2}\Big)\mathbf{1}\left(\frac{t}{\|\theta\|} \in [0.99, 1.01]; \theta_0 \in \mathcal{E}_\theta\right) \leqslant 1.5\cdot\mathbf{1}(\theta_0 \in \mathcal{E}_\theta)\,. \tag{3.36}$$

Let us move to point 2. Fix $\theta \in \Theta$ so by Lemma 8.4, $\|u - u^*\| \leqslant 1/50B$. We have, if $N_\theta \sim \mathsf{N}(\theta, \Gamma_\theta)$, by Chebychev's inequality,

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

Besides,

$$\begin{aligned}\mathbf{E}\big[\|N_\theta - \theta\|_H^2\big] &= \mathrm{Tr}\big(H^{1/2}\Gamma_\theta H^{1/2}\big) = \frac{1}{10^4}\left[\Big(B^2 - \frac{1}{2d}\Big)\|H^{1/2}u\|^2 + \frac{1}{2d}\mathrm{Tr}(H)\right] \\ &\leqslant \frac{1}{10^4}\left[\Big(B^2 - \frac{1}{2d}\Big)\Big(\frac{1}{B^3} + \frac{\|u-u^*\|^2}{2B}\Big) + \frac{1}{2d}\Big(\frac{1}{B^3} + \frac{d-1}{B}\Big)\right] \leqslant \frac{2}{10^4 B}\,.\end{aligned}$$

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,

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

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

$$\begin{aligned}\mathbf{E}\|N - \theta\|_H^2 &= \|\theta - \theta^*\|_H^2 + \mathrm{Tr}\big(H^{1/2}\Gamma H^{1/2}\big) \\ &\leqslant \frac{1}{10^4 B} + \frac{1}{10^4}\left[\Big(B^2 - \frac{1}{2d}\Big)\|H^{1/2}u^*\|^2 + \frac{1}{2d}\mathrm{Tr}(H)\right] \\ &= \frac{1}{10^4 B}\Big(2 + \frac{d-1}{2d}\Big) \leqslant \frac{2.5}{10^4 B}\,.\end{aligned} \tag{3.37}$$

This concludes the proof of Point 2.

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

$$\begin{aligned}D(\rho_\theta\|\pi_\Theta) &= \int_{\mathcal{E}_\theta} \log\left(\frac{\mathrm{d}\rho_\theta}{\mathrm{d}\pi_\Theta}\right)\mathrm{d}\rho_\theta \leqslant \int_{\mathcal{E}_\theta} \log\left(\frac{1.5\mathrm{d}\tilde{\rho}_\theta}{\mathrm{d}\pi_\Theta}\right)1.5\,\mathrm{d}\tilde{\rho}_\theta \\ &= 1.5(\log 1.5 + D(\tilde{\rho}_\theta\|\pi_\Theta))\,.\end{aligned} \tag{3.38}$$

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

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

Using Lemma 3.9 and Point 2 to bound the first term in the left-hand-side and Point 2 for the second, we get,

$$D(\rho_\theta \| \pi_\Theta) \leqslant 1.5 \log(5) + 3D(\gamma_\theta \| \gamma)\,. \tag{3.39}$$

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 $\Gamma^{-1} \preccurlyeq 2 \cdot 10^4 dBH$, we have on one side, by (3.37),

$$\mathrm{Tr}(\Gamma^{-1/2}\Gamma_\theta\Gamma^{-1/2}) \leqslant 2 \cdot 10^4 dB\mathrm{Tr}(H^{1/2}\Gamma_\theta H^{1/2}) \leqslant 3d\,,$$

and, on the other side,

$$\|\theta - \theta^*\|^2_{\Gamma^{-1}} \leqslant 2 \cdot 10^4 dB\|\theta - \theta^*\|^2_H \leqslant 2d\,.$$

Thus, $D(\gamma_\theta \| \gamma) \leqslant 2d$ and, by (3.39),

$$D(\rho_\theta \| \pi_\Theta) \leqslant 6.5d\,. \tag{3.40}$$

Combining this inequality with (3.35) concludes the proof. ■

Conclusion of the proof

We apply the PAC-Bayes inequality (3.17) to the process $Z = (Z(\omega))_{\omega \in E}$ defined by (3.19) on $E = \Theta' \times S^{d-1}$, with prior distribution π on E from Definition 3.3 and parameter $\lambda > 0$ (to be chosen later). This implies that with probability at least $1 - e^{-t}$, for every $\theta \in \Theta$ and $v \in S^{d-1}$,

$$\frac{1}{n}\sum_{i=1}^n \int_E Z_i(\omega)\rho_{\theta,v}(\mathrm{d}\omega) \geqslant -\frac{1}{\lambda}\int_E \log \mathbf{E}\big[\exp(-\lambda Z(\omega))\big]\rho_{\theta,v}(\mathrm{d}\omega) - \frac{D(\rho_{\theta,v}\|\pi) + t}{\lambda n}\,, \tag{3.41}$$

where $\rho_{\theta,v} = \rho_\theta \otimes \rho_v$ (see Definitions 3.1 and 3.2).

Next, we bound the left-hand side of (3.41) using Lemma 3.5, and control the two terms in the right-hand side of (3.41) using Lemmas 3.3 and 3.6 respectively. This yields the following: with probability at least $1 - e^{-t}$, for every $\theta \in \Theta$ and $v \in S^{d-1}$,

$$\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle \geqslant \frac{1}{15}\big(0.03 - 3\lambda B\big) - \frac{(6.5 + \log(1 + 2\varepsilon^{-1}))d + t}{15\lambda n} - 22\varepsilon^2 B R_n\,, \tag{3.42}$$

where R_n is defined in Lemma 3.5. Now by Lemma 3.10, as $n \geqslant 1.1(d+t)$, with probability at least $1 - e^{-t}$ one has $R_n \leqslant 4/B$.

Hence, with probability at least $1 - 2e^{-t}$, one has for every $\theta \in \Theta$ and $v \in S^{d-1}$,

$$\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle \geqslant 0.002 - \frac{\lambda B}{5} - \frac{(6.5 + \log(1 + 2\varepsilon^{-1}))d + t}{15\lambda n} - 88\varepsilon^2\,. \tag{3.43}$$

We choose $\lambda = 0.002/B$ and $\varepsilon = 0.001$, so that whenever $n \geqslant 500B(d+t)$ one has

$$\frac{\lambda B}{5} \leqslant 0.0004\,, \quad \frac{(6.5 + \log(1+2\varepsilon^{-1}))d + t}{15\lambda n} \leqslant 0.0004\,, \quad 88\varepsilon^2 \leqslant 0.0002\,,$$

and therefore the lower bound (3.43) becomes

$$\langle H^{-1/2}\widehat{H}_n(\theta)H^{-1/2}v, v\rangle \geqslant 0.001\,,$$

which concludes the proof.

3.5.1 Technical lemmas for the proof of Theorem 3.2

This section gathers technical tools used repeatedly in the proofs.

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

Lemma 3.7. *For any $\theta \in \mathbb{R}^d$, let*

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

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

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

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

$$\begin{aligned}\langle\widetilde{H}(\theta)v, v\rangle = \langle u, v\rangle^2 \mathbf{E}\big[\mathbf{1}\big\{|\langle\theta, X\rangle| \leqslant 1;\ \|X\| \leqslant 2\sqrt{d}\big\}\langle u, X\rangle^2\big] \\ + (1 - \langle u, v\rangle^2)\mathbf{E}[\mathbf{1}\Big\{|\langle\theta, X\rangle| \leqslant 1;\ \|X\| \leqslant 2\sqrt{d}\Big\}]\,. \quad (3.44)\end{aligned}$$

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 $|\langle\theta, X\rangle| \leqslant 1$ and $\|X\| \leqslant 2\sqrt{d}$, so

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

By Lemma 8.4, as $\|\theta - \theta^*\|_H \leqslant 1/10\sqrt{B}$,

$$\frac{1}{2} \leqslant 0.9 \cdot B \leqslant \|\theta\| \leqslant 1.1 \cdot B\,. \quad (3.45)$$

By Markov's inequality, we have thus

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

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

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

Thus

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

We use the same arguments to bound the first expectation in the right hand side of (3.44), and find that

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

We have

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

Thus

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

Plugging these bounds into (3.44) yields

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

Lemma 3.8. *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

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

We have

$$\begin{aligned}|\langle v, u\rangle| &\leqslant |\langle v, u^*\rangle| + \|u - \langle u, u^*\rangle u^*\|\|v - \langle v, u^*\rangle u^*\|\,, \\ v - \langle u, v\rangle u &= (v - \langle u^*, v\rangle u^*) - \langle v - \langle u^*, v\rangle u^*, u\rangle u + \langle u^*, v\rangle(u^* - \langle u, u^*\rangle u)\,.\end{aligned}$$

By Lemma 8.4, $\|u - \langle u, u^*\rangle u^*\| \leqslant \frac{r}{(1-r)B}$. Using Cauchy-Schwarz inequality and $(a+b)^2 \leqslant (1+r)a^2 + (1+r^{-1})b^2$, we deduce

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

as well as

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

Hence,

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

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

Lemma 3.9. *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

$$D(\mathsf{P}\|\mathsf{Q}) = \int_A p \log\Big(\frac{p}{q}\Big)\mathrm{d}\mu + \int_{A^c} p\log\Big(\frac{p}{q}\Big)\mathrm{d}\mu\,. \tag{3.46}$$

By symmetry we do the computations on the event A.

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

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. ■

Finally, the following lemma bounds the remainder term R_n from Lemma 3.5, which arises from smoothing over the direction $v \in S^{d-1}$.

Lemma 3.10. *If $n \geqslant 1.1B(d+t)$, then with probability at least $1-e^{-t}$, one has*

$$R_n = \sup_{u\in\mathsf{C}(u^*,1/10B)} \frac{1}{n}\sum_{i=1}^n \exp\big(-0.49\,B|\langle u, X_i\rangle|\big)\mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\} \leqslant \frac{4}{B}\,.$$

Proof. We define the process $W = (W(u))_{u\in S^{d-1}}$ by (denoting $b = 0.49B$)

$$W(u) = \exp\big(-b|\langle u, X\rangle|\big)\mathbf{1}\Big\{\|X\| \leqslant 2\sqrt{d}\Big\}\,,$$

and likewise define $W_i(u)$ for $i = 1,\dots,n$ by replacing X by X_i above. We need to bound

$$\sup_{u\in\mathsf{C}(u^*,1/10B)} \frac{1}{n}\sum_{i=1}^n W_i(u)\,,$$

which we achieve by applying the PAC-Bayes inequality of Lemma 3.2, with $\lambda = 1$.

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.

Bound on the relative entropy. We prove in this paragraph that

$$D(\rho_u\|\pi) \leqslant 1.1 \cdot d\,. \tag{3.47}$$

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 surface 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

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

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

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

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,

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

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 (3.35).

Bound on the Laplace transform. In this paragraph, we prove that

$$\log \mathbf{E}[\exp(W(u))] \leqslant \frac{3}{B}\,. \tag{3.48}$$

Indeed, since $0 \leqslant W(u) \leqslant \exp(-b|\langle u, X\rangle|) \leqslant 1$, one has (using that the density of $\langle u, X\rangle$ is bounded by $1/\sqrt{2\pi}$)

$$\begin{aligned}\log \mathbf{E}[\exp(W(u))] &\leqslant \log\big\{1 + (e-1)\,\mathbf{E}[W(u)]\big\} \leqslant (e-1)\,\mathbf{E}[W(u)]\\ &\leqslant (e-1)\mathbf{E}[\exp(-b|\langle u, X\rangle|)] \leqslant \frac{e-1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-b|s|}\mathrm{d}s = \frac{(e-1)\sqrt{2/\pi}}{b}\,,\end{aligned}$$

which proves (3.48) after substituting $b = 0.49B$ and bounding the numerical constant.

Control of the smoothed process. In this paragraph, we show that

$$\int_{\mathsf{C}(u,1/10B)} W_i(u')\rho_u(\mathrm{d}u') \geqslant 0.82 \cdot W_i(u)\,. \tag{3.49}$$

We have by Jensen's inequality,

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

On the other hand, using [Mou22, Eq (42) and (43)] and Fact 8.1 for integration over spherical caps, one has under the event $\{\|X_i\| \leqslant 2\sqrt{d}\}$ that

$$\int_{\mathsf{C}(u,1/10B)} \langle u', X_i\rangle^2 \rho_u(\mathrm{d}u') \leqslant \langle u, X_i\rangle^2 + \frac{2}{100(d-1)B^2}\|X_i\|^2 \leqslant \langle u, X_i\rangle^2 + \frac{0.16}{B^2}\,.$$

Combining the previous two inequalities gives

$$\int_{\mathsf{C}(u,1/10B)} W_i(u')\rho_u(\mathrm{d}u') \geqslant \mathbf{1}\Big\{\|X_i\| \leqslant 2\sqrt{d}\Big\} \exp\bigg(-0.49\,B\bigg[|\langle u, X_i\rangle| + \frac{0.4}{B}\bigg]\bigg) \geqslant 0.82\,W_i(u)\,.$$

Conclusion of the proof. By the PAC-Bayes inequality (Lemma 3.2), for any $t \geqslant 0$, we have with probability at least $1 - e^{-t}$, simultaneously for all $u \in \mathsf{C}(u^*, 1/10B)$,

$$\frac{0.82}{n}\sum_{i=1}^n W_i(u) \leqslant \frac{1}{n}\sum_{i=1}^n \int_{\mathsf{C}(u,1/10B)} W_i(u')\rho_u(\mathrm{d}u') \leqslant \frac{3}{B} + \frac{1.1d+t}{n} \leqslant \frac{4}{B}\,,$$

where the last inequality comes from the condition on n. ∎

Appendix 3.A: Proof of Theorem 3.1

The proof of Theorem 3.1 follows the clear path given by Lemma 2.1 (localization) and combines Proposition 3.1 and Theorem 3.2. We now proceed with the proof.

By Proposition 3.1, since $n \geqslant 4B(d\log 5 + t)$, with probability larger than $1 - 2e^{-t}$,

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

Moreover, let

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

Theorem 3.2 ensures that, as $n \geqslant 500B(d+t)$, it holds with probability at least $1 - 2e^{-t}$ that, simultaneously for all $\theta \in \Theta$, $\widehat{H}_n(\theta) \succcurlyeq \frac{1}{1000}H$. Therefore, by Lemma 2.1, 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 3.11 and 2.1, 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 4.2 (Chapter 4) with Fact 4.1 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, $\mathbf{P}(Y\langle\theta^*, X\rangle < 0) = \mathbf{E}\sigma(-|\langle\theta^*, X\rangle|)$. In addition, for all real t, $\sigma(-|t|) \leqslant \min\{1/2, e^{-|t|}\}$, hence

$$\mathbf{P}(Y\langle\theta^*, X\rangle < 0) \leqslant \frac{1}{\max\{2, \|\theta^*\|\}} \leqslant \frac{e}{2B}\,.$$

We use Fact 4.1 with $p = e/(2B)$ to conclude that if $n \leqslant Bt/e$, then

$$\mathbf{P}(\{\text{MLE does not exist}\}) \geqslant \exp(-t)\,. \tag{3.50}$$

With these results at hand, one has that whenever

$$n \leqslant \frac{B}{2}\Big(\frac{d}{200} + \frac{t}{e}\Big) \leqslant \max\Big\{\frac{Bd}{200}, \frac{Bt}{e}\Big\}\,, \tag{3.51}$$

either $n \leqslant Bt/e$ or $n \leqslant Bd/200$. The former is already dealt with by (3.50). The latter, by Theorem 4.2 (with parameter $\kappa = 1$), implies that

$$\mathbf{P}(\{\text{MLE does not exist}\}) \geqslant 1 - \exp\Big(-\max\Big\{\sqrt{d}, \frac{d}{B^2}\Big\}\Big) - 6e^{-d/24}\,. \tag{3.52}$$

As $d \geqslant 70$ and $t \geqslant 1$, one has $6e^{-d/24} \leqslant 1/2$ and $1/2 - e^{-\sqrt{d}} \geqslant e^{-t}$. Hence, taking the minimum of the two lower bounds (3.50) and (3.52) shows that

$$\mathbf{P}(\{\text{MLE does not exist}\}) \geqslant \exp(-t)$$

and concludes the proof of Theorem 3.1.

Appendix 3.B: Technical tools

In the previous proofs, we used the following lemma linking the Hessians $\nabla^2 L(\theta) = H(\theta) = \mathbf{E}[\sigma'(\langle\theta, X\rangle)XX^\top]$ to H to conclude the proof.

Lemma 3.11. *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. Recall the proxy H_θ for $H(\theta)$ defined in (3.22) as

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

We write that, for any $v \in S^{d-1}$

$$\langle H_\theta^{-1/2}H(\theta)H_\theta^{-1/2}v, v\rangle = B^3\langle u, v\rangle^2\mathbf{E}[\sigma'(\langle\theta, X\rangle)\langle u, X\rangle^2] + B(1 - \langle u, v\rangle^2)\mathbf{E}[\sigma'(\langle\theta, X\rangle)]\,.$$

We start with the upper bound. By Lemma 2.2,

$$\mathbf{E}[\sigma'(\langle\theta, X\rangle)|\langle u, X\rangle|^k] \leqslant \sqrt{\frac{2}{\pi}}\min\Big\{\Gamma\Big(\frac{k+1}{2}\Big), \frac{\Gamma(k+1)}{\|\theta\|^{k+1}}\Big\}\,.$$

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 8.4, $\|\theta\| \geqslant 0.9 \cdot B$, 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 3.8 we also have $H_\theta \preccurlyeq 1.3 \cdot H$.

We now turn to the lower bound. By Lemma 2.2,

$$\mathbf{E}[\sigma'(\langle \theta, X\rangle) |\langle u, X\rangle|^k] \geqslant \sqrt{\frac{2}{\pi}} \frac{2^{k+1}}{k+1} \min\left(\frac{1}{4e^4 \|\theta\|^{k+1}}, \frac{\sigma'(2)}{e^2} \right) .$$

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 8.4, $\|\theta\| \geqslant 0.9 \cdot B$, 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 3.8 we also have $H_\theta \succcurlyeq 0.76 \cdot H$. ■

Chapter 4

Phase transition for linear separation

Abstract

In this chapter, we give a stronger version of the second part of Theorem 3.1 regarding the non-existence of the maximum-likelihood estimator in logistic regression. More precisely, the main result of this chapter, Theorem 4.2, can be seen as a quantitative, finite-sample version of the convergence to 0 for the probability of existence of the MLE in the proportional asymptotic regime studied by Candès and Sur in [CS20].

Contents

4.1 Introduction

In this chapter, we complement the analysis of Chapter 3 regarding the well-specified logit model with Gaussian design. There, our main result was Theorem 3.1 which established the following. Given a dimension $d \geqslant 1$, an unknown parameter $\theta^* \in \mathbb{R}^d$ and i.i.d. data $(X_1, Y_1), \dots, (X_n, Y_n)$ from a well-specified logit model with parameter θ^* (such that $B = \|\theta^*\| \geqslant e$) and design $X_i \sim \mathsf{N}(0, I_d)$, as well as a level of confidence $1 - \delta \in (0, 1)$, there exist positive absolute constants C, C', c such that

1. if $n \geqslant CB(d + \log(1/\delta))$, then with probability at least $1 - \delta$, the MLE exists and satisfies the risk bound $L(\widehat{\theta}_n) - L(\theta^*) \leqslant C'(d + \log(1/\delta))/n$;
2. if $n \leqslant cB(d + \log(1/\delta))$, then the MLE exists with probability *at most* $1 - \delta$.

While the risk bound in the first point is optimal (up to the numerical constant) in light of the asymptotic behavior of the MLE (1.11), the second point is not completely consistent with the convergence to 0 in the phase transition (1.12) of Candès and Sur [CS20]. Indeed, when n is much smaller than Bd, one can expect that not only should the probability of existence of the MLE be bounded away from 1, it should actually be close to 0.

Asymptotic phase transition. We start by recalling the result of Candès and Sur [CS20]. Their setting is that of so-called "proportional" or "high-dimensional" asymptotics, where one considers a sequence of parameters $(d_n, \theta_n^*)_{n \geqslant 1}$ with $d_n/n \to \gamma \in (0, 1)$ and $\beta_n = \|\theta_n^*\| \to \beta \in \mathbb{R}^+$. For every (d_n, θ_n^*), we are given a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of size n of i.i.d. data from a well-specified logit model with parameter θ_n^* and isotropic Gaussian design $X_i \sim \mathsf{N}(0, I_{d_n})$. In this setting, Candès and Sur established the following result.

Theorem 4.1 ([CS20], Theorem 2.2)**.** *In the setting described above, one has*

$$\mathbf{P}(\text{MLE exists}) \underset{n \to \infty}{\longrightarrow} \begin{cases} 1 \text{ if } \gamma < h(\beta) \\ 0 \text{ if } \gamma > h(\beta)\,, \end{cases} \tag{4.1}$$

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 $\mathbf{P}(Y'_\beta = 1 | X') = \sigma(\beta X')$, and let $V_\beta = Y'_\beta X'$. In addition, let $Z \sim \mathsf{N}(0, 1)$ be independent of V_β. Then,

$$h(\beta) = \min_{t \in \mathbb{R}} \mathbf{E}\big[(tV_\beta - Z)_+^2\big]\,. \tag{4.2}$$

Now, it is worth mentioning that for some absolute constant $C > 1$, the transition function h satisfies that for all $\beta \geqslant 1$,

$$\frac{1}{C\beta} \leqslant h(\beta) \leqslant \frac{C}{\beta}\,.$$

This follows essentially from Fact 4.2 and bounding suitable expectations. Together with (4.1), this estimate suggests that the critical sample size is indeed Bd, which would complement the sufficient condition on the sample size in the first part of Theorem 3.1 and would fill the gap left by the second point (which showed that the probability of existence was bounded away from 1).

In this chapter, we precisely establish such a non-asymptotic phase transition, in the sense that if $n \ll Bd$, then the probability that the MLE exists is indeed close to 0. This statement is made precise in Theorem 4.2, which is the main result of this chapter.

Setting. 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 $\mathbf{P}(Y_i = 1|X_i) = \sigma(\langle \theta^*, X_i \rangle)$. 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.

4.2 The non-asymptotic phase transition for the existence of the MLE

Theorem 4.2 below shows that if $n \ll Bd$, then the probability of existence of the MLE indeed approaches 0, at an exponential rate with respect to the dimension. This can be seen as a quantitative version of the convergence to 0 in the phase transition (1.12) from [CS20]. Theorem 4.2 is proved in Section 4.3.

Theorem 4.2. *Let $d \geqslant 70$, and assume that $X \sim \mathsf{N}(0, I_d)$ and that the logistic model is well-specified. For every $\kappa \geqslant 1$, if $n \leqslant Bd/(200\kappa)$ then*

$$\mathbf{P}(\text{MLE exists}) \leqslant \exp\big(-\max\big\{\kappa\sqrt{d}, \kappa^2 d/B^2\big\}\big) + 6e^{-d/24}. \tag{4.3}$$

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 (4.3). Theorem 4.2 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 (4.3). In particular, in the regime where $n \asymp d$ and $B \gg 1$ (so that $\kappa \asymp B$), Theorem 4.2 shows that the probability of existence of the MLE is smaller than $\exp(-cd)$ for some constant c.

The proof of Theorem 4.2, in the next section, builds on the approach of Candès and Sur [CS20]. We also refer to Section 1.5 in the introduction (Chapter 1) for more

background and a detailed sketch of the proof. Specifically, 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 (1.12) in this asymptotic regime, several refinements are required in order to obtain the quantitative bound of Theorem 4.2. 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. Again, we refer to Chapter 1, Section 1.5 for a detailed discussion of this point.

This control 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.

4.3 Proof of Theorem 4.2

The proof of Theorem 4.2 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 $\delta(C) = \mathbf{E}\|\Pi_C \boldsymbol{Z}\|^2$ where Π_C is the Euclidean projection on $\mathbf{C}$ and $\boldsymbol{Z} \sim \mathsf{N}(0, I_n)$.

Lemma 4.1 (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*

$$n - k \leqslant \delta(\mathbf{C}) - t\,, \tag{4.4}$$

then

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

We can now proceed with the proof of Theorem 4.2. 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 $i \in \{1, \dots, n\}$, $\mathbf{P}(Y_i = 1 \,|X_i) = \sigma(BX_i^1)$ where X_i^j denotes the j-th coordinate of X_i for every $j \in \{1, \dots, d\}$. Below we let $U_i = X_i^1$ and $V_i = Y_i U_i$ for all $i \in \{1, \dots, n\}$. Let also $\boldsymbol{V} = (V_1, \dots, V_n) \in \mathbb{R}^n$ and $\Lambda = \mathbb{R}\boldsymbol{V} + \mathbb{R}^n_+$, which is a random cone in $\mathbb{R}^n$. The proof of Theorem 4.2 relies on the following observation.

Lemma 4.2. *Let $\mathcal{L}$ be a random subspace drawn uniformly from all subspaces of dimension $d-1$ in $\mathbb{R}^n$. Then*

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

The proof of this result is postponed to the end of the chapter (Appendix 4.A) 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 4.1 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 (4.4) regarding the statistical dimension of Λ is satisfied with high probability. Hereafter we denote by E the event $\{\Lambda \cap \mathcal{L} \neq \{0\}\}$, and for every $t \geqslant 0$, we define the event

$$A_t = \{n - d + 1 \leqslant \delta(\Lambda) - t\}\,. \tag{4.5}$$

Our main task in this proof is to show that

$$\mathbf{P}(A_{\alpha d}) \geqslant 1 - \exp\big(-\max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}\big) - 2e^{-\tau d} \tag{4.6}$$

for some $\tau \in (0, 1/2)$ and $\alpha \in (1/2, 1)$. We now establish (4.6) with explicit constants. Conditionning on $\boldsymbol{V}$, one has $\delta(\Lambda) = n - \mathbf{E}_{\boldsymbol{Z}}\big[\mathrm{dist}(\boldsymbol{Z}, \Lambda)^2 \,|\boldsymbol{V}\big]$, where $\mathbf{E}_{\boldsymbol{Z}}$ denotes the expectation with respect to $\boldsymbol{Z}$. Throughout the rest of this proof, we let

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

This way, we will prove (4.6) 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

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

where h is the phase transition function (4.2) 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 $\mathbf{E}[F(\boldsymbol{V})] \lesssim n/B \lesssim d$.

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

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

By Fact 4.2, for all $\lambda \geqslant 0$,

$$\psi(-\lambda V_i) \leqslant e^{-\lambda^2 U_i^2/2} + \mathbf{1}(Y_i U_i \leqslant 0) + \lambda^2 U_i^2 \mathbf{1}(Y_i U_i \leqslant 0)\,. \tag{4.7}$$

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

$$\zeta_{i,\lambda} = e^{-\lambda^2 U_i^2/2}\,, \quad \varepsilon_i = \mathbf{1}(Y_i U_i \leqslant 0)\,, \quad \psi_i = U_i^2 \mathbf{1}(Y_i U_i \leqslant 0)\,, \tag{4.8}$$

so that we can further bound

$$F(\boldsymbol{V}) \leqslant \min_{\lambda > 0} \Big\{ \sum_{i=1}^n \zeta_{i,\lambda} + \sum_{i=1}^n \varepsilon_i + \lambda^2 \sum_{i=1}^n \psi_i \Big\}\,. \tag{4.9}$$

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 $k \in \{1, 2\}$,

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

Since in addition $\zeta_{i,\lambda} \leqslant 1$ almost surely, by Lemma 8.2 and the second and third points of Lemma 8.1, for all $t \geqslant 0$, with probability larger than $1 - e^{-t}$,

$$\sum_{i=1}^n \zeta_{i,\lambda} \leqslant \frac{n}{\lambda} + \sqrt{\frac{2nt}{\lambda}} + \frac{t}{3}\,. \tag{4.10}$$

Regarding the second sum, inequality (3.11) shows that for every i, $\mathbf{E}[\varepsilon_i] = \mathbf{E}[\exp(-B|U_i|)] \leqslant B^{-1}$. 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

$$\sum_{i=1}^n \varepsilon_i \leqslant \frac{n}{B} + \sqrt{\frac{2nt}{B}} + \frac{t}{3}\,. \tag{4.11}$$

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

Lemma 4.3 ([Lat97], Corollary 1)**.** *Let $\xi, \xi_1, \dots, \xi_n$ be i.i.d. nonnegative random variables. Then for any $p \geqslant 1$,*

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

From now on, we let $S_n = \psi_1 + \cdots + \psi_n$ and $p \in [1, n]$. By Markov's inequality, $\mathbf{P}(S_n \leqslant e\|S_n\|_p) \geqslant 1 - e^{-p}$, 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 $s \geqslant 1$,

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

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

$$\|\psi\|_s = \mathbf{E}[U^{2s}]^{1/s} \leqslant (3/e)s\,. \tag{4.12}$$

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

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

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

$$\|\psi\|_s \leqslant \frac{9}{e^2}\min\Big\{\frac{s^2}{B^2B^{1/s}}, s\Big\}\,. \tag{4.13}$$

Upper bound on the supremum. Using the control on the moments of the ψ_i's (4.13), it follows from Latała's inequality (Lemma 4.3) that

$$\|S_n\|_p \leqslant \frac{18}{B^2}\sup\Big\{\min\Big\{ps\Big(\frac{n}{pB}\Big)^{1/s}, B^2p\Big(\frac{n}{p}\Big)^{1/s}\Big\}; 1\leqslant s\leqslant p\Big\}\,. \tag{4.14}$$

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,

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

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

$$\|S_n\|_p \leqslant \frac{18}{B^2} \max\left\{\frac{n}{B}, \min\left\{p^2\Big(\frac{n}{B}\Big)^{1/p},\, B^2 p\Big(\frac{n}{p}\Big)^{1/B^2}\right\}\right\}. \tag{4.15}$$

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

$$p = \max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}. \tag{4.16}$$

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

$$p^2\Big(\frac{n}{B}\Big)^{1/p} \leqslant \kappa^2 d \exp\Big(\frac{2}{e\sqrt{C_0}}\Big). \tag{4.17}$$

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

$$\frac{\log\big(d/(C_0\kappa)\big)}{\kappa\sqrt{d}} \leqslant \frac{2}{e\kappa^{3/2}\sqrt{C_0}}, \tag{4.18}$$

from which (4.17) follows, since $\kappa \geqslant 1$.

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

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

Using the same argument as before, we find that

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

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

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

Plugging this in (4.15) and using again the assumption $n/B \leqslant d/(C_0\kappa)$, we deduce that for $p = \max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}$,

$$\|S_n\|_p \leqslant \frac{18}{B^2}\max\Big\{\frac{d}{C_0\kappa}, L_0\kappa^2 d\Big\} = 18L_0\frac{\kappa^2 d}{B^2}\,. \tag{4.19}$$

High-probability bound on the statistical dimension. We now use Markov's inequality using the moment bound (4.19). This yields

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

Finally, we combine this with (4.10) and (4.11), 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

$$\sum_{i=1}^n \psi(-\lambda V_i) \leqslant \frac{n}{B} + \sqrt{\frac{2nt}{B}} + \frac{n}{\lambda} + \sqrt{\frac{2nt}{\lambda}} + \frac{2t}{3} + \frac{L\lambda^2\kappa^2}{B^2}d\,, \quad L = 18eL_0\,.$$

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

$$\sum_{i=1}^n \psi(-\lambda V_i) \leqslant \frac{d}{C_0\kappa} + \sqrt{\frac{2\tau}{C_0\kappa}}d + \frac{Bd}{C_0\kappa\lambda} + \sqrt{\frac{2B\tau}{C_0\kappa\lambda}}d + \frac{2}{3}\tau d + \frac{L\lambda^2\kappa^2}{B^2}d\,.$$

Then, by the arithmetic mean–geometric mean inequality,

$$\sum_{i=1}^n \psi(-\lambda V_i) \leqslant \left[\frac{3}{2C_0\kappa} + \frac{5}{3}\tau + \frac{2B}{C_0\kappa\lambda} + \frac{L\kappa^2\lambda^2}{B^2}\right]d\,. \tag{4.20}$$

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

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

with equality if and only if $\lambda = \lambda^*$, the value such that $\frac{3B}{C_0\kappa\lambda} = \frac{3L\kappa^2\lambda^2}{B^2}$, that is $\lambda^* = \frac{B}{\kappa(C_0L)^{1/3}}$. Simplifying the constants yields

$$\inf_{\lambda>0}\Big\{\frac{2B}{C_0\kappa\lambda} + \frac{L\kappa^2\lambda^2}{B^2}\Big\} = 3\cdot(18e)^{1/3}\frac{\exp\left(\frac{1}{2eC_0^{2/3}}\right)}{C_0^{2/3}} =: \frac{C_1}{C_0^{2/3}}\,.$$

We finally plug this in (4.20) and obtain that

$$\mathbf{P}\Big(\sum_{i=1}^n \psi(-\lambda^* V_i) \leqslant \left[\frac{5\tau}{3} + \frac{3}{2C_0} + \frac{C_1}{C_0^{2/3}}\right]d\Big) \geqslant 1 - e^{-\max\{\kappa\sqrt{d},\kappa^2 d/B^2\}} - 2e^{-\tau d}\,.$$

We then let $c_0(C_0) = 3/(2C_0) + C_1/C_0^{2/3}$ so that the last inequality rewrites

$$\mathbf{P}\big(F(\boldsymbol{V}) \leqslant \big[5\tau/3 + c_0(C_0)\big]d\big) \geqslant 1 - e^{-\max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}} - 2e^{-\tau d}\,.$$

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

$$\left(\frac{5\tau}{3} + c_0(C_0)\right)d \leqslant d - 1 - \alpha d\,, \tag{4.21}$$

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

$$\mathbf{P}(A_{\alpha d}) = \mathbf{P}\big(n - d + 1 \leqslant \delta(\Lambda) - \alpha d\big) \geqslant 1 - \exp\Big(-\max\Big\{\kappa\sqrt{d}, \frac{\kappa^2}{B^2}d\Big\}\Big) - 2e^{-\tau d}\,, \tag{4.22}$$

provided that α, τ and C_0 satisfy (4.21). We have proved (4.6).

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 4.1, on the event $A_{\alpha d}$,

$$\mathbf{P}(E\,|\boldsymbol{V}) \geqslant 1 - 4\exp\Big(-\frac{(\alpha d)^2}{8(\min\{\delta(\Lambda), n - \delta(\Lambda)\} + \alpha d)}\Big) \geqslant 1 - 4e^{-\alpha^2 d/8}\,. \tag{4.23}$$

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 (4.21),

$$\mathbf{P}(A_{\alpha d}) \geqslant \mathbf{P}\big(F(\boldsymbol{V}) \leqslant \big[5\tau/3 + c_0(C_0)\big]d\big) \geqslant 1 - e^{-\max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}} - 2e^{-\tau d}\,.$$

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 (4.23), one has

$$\begin{aligned}\mathbf{1}(A_{\alpha d}) \leqslant \mathbf{1}\big(\mathbf{P}(E\,|\boldsymbol{V}) \geqslant 1 - 4e^{-\alpha^2 d/8}\big) &= \mathbf{1}\big(\mathbf{P}(E\,|\boldsymbol{V}) + 4e^{-\alpha^2 d/8} \geqslant 1\big) \\ &\leqslant \mathbf{P}(E\,|\boldsymbol{V}) + 4e^{-\alpha^2 d/8}\,.\end{aligned}$$

Taking expectation with respect to $\boldsymbol{V}$, this implies that

$$\mathbf{P}(E) \geqslant \mathbf{P}(A_{\alpha d}) - 4e^{-\alpha^2 d/8} \geqslant 1 - e^{-\max\{\kappa\sqrt{d}, \kappa^2 d/B^2\}} - 2e^{-\tau d} - 4e^{-\alpha^2 d/8}\,. \tag{4.24}$$

We will thus choose $\tau \geqslant \alpha^2/8$ so that $e^{-\tau d} \leqslant e^{-\alpha^2 d/8}$, under the constraint (4.21). This constraint rewrites

$$\tau \leqslant \frac{3}{5}\left(1 - \alpha - \frac{1}{d} - c_0(C_0)\right).$$

We choose $\tau = \alpha^2/8$ and then saturate the constraint, that is, we choose $\alpha = \alpha(d, C_0)$, the largest solution of the equation $\alpha^2/8 = 3(1 - \alpha - 1/d - c_0(C_0))/5$. We further bound the numerical constants for the choice $C_0 = 200$ and under the assumption that $d \geqslant 70$, that is $\alpha = \alpha(d, C_0) > 0.5844...$ and in particular $\alpha^2/8 > 1/24$. The result then follows from (4.24).

Appendix 4.A: Remaining proofs and additional results

Proof of Lemma 4.2. By definition, there exists a separating hyperplane if there is some $\theta \in \mathbb{R}^d \setminus \{0\}$ such that for all $i \in \{1, \dots, n\}$,

$$Y_i \langle \theta, X_i \rangle \geqslant 0 \,. \tag{4.25}$$

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 $(Y_i X_i^1, Y_i X_i^2, \dots, Y_i X_i^d)$ and $(Y_i X_i^1, X_i^2, \dots, X_i^d)$ have the same distribution. Therefore,

$$\mathbf{P}\big(\exists \theta \in \mathbb{R}^d \setminus \{0\},\, \forall i, Y_i \langle \theta, X_i \rangle \geqslant 0\big) = \mathbf{P}\Big(\exists \theta \in \mathbb{R}^d \setminus \{0\},\, \theta^1 \boldsymbol{V} + \sum_{j=2}^d \theta^j \boldsymbol{X}^j \in \mathbb{R}_+^n\Big) \,. \tag{4.26}$$

Now, let $\mathcal{L} = \operatorname{span}\{\boldsymbol{X}^2, \dots, \boldsymbol{X}^d\}$. 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 (4.26) concludes the proof. ∎

Fact 4.1. *Let $p \in (0, 1/2)$ and $u^* \in S^{d-1}$ be such that $\mathbf{P}(Y\langle u^*, X\rangle < 0) \leqslant p$. 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

$$\mathbf{P}\big(\forall i \leqslant n,\; Y_i \langle u^*, X_i \rangle \geqslant 0\big) = \big(1 - \mathbf{P}(Y\langle u^*, X\rangle < 0)\big)^n \geqslant (1-p)^n = \exp\big(n \log(1-p)\big) \,.$$

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

$$\mathbf{P}\big(\forall i \leqslant n,\; Y_i \langle \theta^*, X_i \rangle > 0\big) \geqslant \exp\big(-2np\big) \geqslant \exp(-t) \,. \qquad ∎$$

Fact 4.2. *Let $\psi(s) = \mathbf{E}\big[(s - Z)_+^2\big]$ 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,

$$\psi(-s) = \mathbf{E}(-s - Z)_+^2 = \mathbf{E}(-(s+Z))_+^2 = \mathbf{E}(s+Z)_-^2 = \mathbf{E}(s-Z)_-^2 \,.$$

Also, observing that for all $x \in \mathbb{R}$, $x^2 = x_+^2 + x_-^2$, we have

$$\psi(-s) + \psi(s) = \mathbf{E}(s-Z)_-^2 + \mathbf{E}(s-Z)_+^2 = \mathbf{E}(s-Z)^2 = s^2 + 1 \,. \tag{4.27}$$

We start with the case where $s < 0$. In this case, denoting by g the density of $\mathsf{N}(0,1)$, one has

$$\begin{aligned}\psi(s) = \mathbf{E}(s-Z)_+^2 &= \mathbf{E}\left[(s-Z)^2\mathbf{1}\{s-Z>0\}\right] = \int_{-\infty}^{s}(s-z)^2 g(z)\mathrm{d}z \\ &= \int_0^{+\infty} z^2 g(s-z)\mathrm{d}z = \int_0^{+\infty} z^2 \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{s^2-2sz+z^2}{2}\Big)\mathrm{d}z \\ &= e^{-s^2/2}\int_0^{+\infty} z^2 e^{sz} g(z)\mathrm{d}z = e^{-s^2/2}\,\mathbf{E}\left[Z^2 e^{sZ}\mathbf{1}\{Z>0\}\right].\end{aligned}$$

Note that, since $s<0$, $e^{sZ}\mathbf{1}\{Z>0\} \leqslant \mathbf{1}\{Z>0\}$, which implies that

$$\mathbf{E}\left[Z^2 e^{-sZ}\mathbf{1}\{Z<0\}\right] \leqslant \mathbf{E}\left[Z^2\mathbf{1}\{Z<0\}\right] = \frac{1}{2},$$

which proves the first part of the result. Regarding the case where $s \geqslant 0$, we deduce from (4.27) that $\psi(s) = s^2+1-\psi(-s)$, and, from the previous point, that $0 \leqslant \psi(-s) \leqslant 1/2$. ■

Chapter 5

Regular designs: definition and examples

Abstract

In this chapter we formally introduce the notion of *regular designs*. This definition shall capture the essential properties of the Gaussian distribution that were used to obtain sharp guarantees in Chapter 3. After precisely defining the notion of regularity, we provide three classes of distributions satisfying the regularity conditions.

Contents

5.1 Regularity assumptions

In this section, we introduce formally the setup that we called "regular" in the introduction. It is characterized by three conditions, stated below as Assumption 5.1, 5.2, and 5.3. Examples of distributions satisfying these regularity assumptions are given and discussed in Section 5.2.

The first assumption on the design is standard and states that the design X is light-tailed; we refer to Definition 8.1 in Appendix 8.1 for the definition of the ψ_1-norm.

Assumption 5.1. 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]. We discuss this topic in greater details in Chapter 7.

Assumption 5.2. Let $u^* \in S^{d-1}$ and $\eta \in (0, 1]$. For some $c \geqslant 1$, one has for every $t \geqslant \eta$ that

$$\mathbf{P}\big(|\langle u^*, X\rangle| \leqslant t\big) \leqslant ct\,. \tag{5.1}$$

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 5.3. 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

$$\mathbf{P}\big(|\langle u^*, X\rangle| \leqslant c\eta,\ |\langle v, X\rangle| \geqslant c^{-1}\max\{\eta, \|u^* - v\|\}\big) \geqslant \eta/c\,. \tag{5.2}$$

Remark 2. 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 5.3 is that for every $v \in S^{d-1}$, one has

$$\mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant c\eta,\ |\langle v, X\rangle| \geqslant c^{-1}\max\Big\{\eta, \sqrt{1 - \langle u^*, v\rangle^2}\Big\}\Big) \geqslant \eta/c\,. \tag{5.3}$$

This only changes the value of the parameter c from (5.2) by a factor $\sqrt{2}$. This equivalent formulation turns out to be more convenient in some situations.

Let us now discuss the meaning and necessity of this assumption. We start with an intuitive discussion and proceed with a precise statement. First, the condition $|\langle u^*, X\rangle| \lesssim \eta$ amounts to $|\langle \theta^*, X\rangle| \lesssim 1$. The restriction to such values of X stems from the fact that values of X for which $|\langle \theta^*, X\rangle| \gg 1$ provide little information on the precise value of θ^*. Indeed, it follows from the form of the logistic model that for such X, one has $Y = \mathrm{sign}(\langle \theta^*, X\rangle)$ with high probability; but for other parameters θ close to θ^*, one also has $|\langle \theta, X\rangle| \gg 1$ and $Y = \mathrm{sign}(\langle \theta, X\rangle)$. Hence, the value of Y does not allow one to distinguish between the values θ and θ^*, both of which are highly consistent with this label.

We thus focus on values of X such that $|\langle \theta^*, X\rangle| \lesssim 1$, which is the first condition. The second condition is that for any other direction $v \in S^{d-1}$, we find sufficiently many

values of X for which additionally $|\langle v, X\rangle|$ is sufficiently large. This intuitively comes from the fact that, in order to distinguish θ^* from another parameter $\theta = \theta^* + \varepsilon v$ with $\varepsilon > 0$ from a data point (X, Y), we need the "predictions" of these two parameters at X, namely $\langle \theta^*, X\rangle$ and $\langle \theta, X\rangle = \langle \theta^*, X\rangle + \varepsilon\langle v, X\rangle$, to be sufficiently different. This precisely amounts to saying that $|\langle v, X\rangle|$ is not too small. The threshold for $|\langle v, X\rangle|$ depends on the alignment between v and u^*, and its value in (5.3) comes from the fact that, in the baseline case where $X \sim \mathsf{N}(0, I_d)$, one has

$$\mathbf{E}\Big[\langle v, X\rangle^2 \,\Big|\, |\langle u^*, X\rangle| = \eta\Big]^{1/2} = \sqrt{\langle u^*, v\rangle^2\eta^2 + (1 - \langle u^*, v\rangle^2)} \asymp \max\{\eta, \sqrt{1 - \langle u^*, v\rangle^2}\}\,.$$

We now argue more formally that Assumption 5.3 is necessary for the MLE to behave similarly as in the Gaussian case. To see why, let $H_X(\theta^*) = \nabla^2 L(\theta^*)$ denote the Hessian under the design X. When the logistic model is well-specified, $H_X(\theta^*)$ coincides with the Fisher information matrix at θ^*. Hence, the MLE is asymptotically normal as $n \to \infty$, with

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

Therefore, for the error $\widehat{\theta}_n - \theta^*$ of the MLE under design X to be as small as the error under a Gaussian design $G \sim \mathsf{N}(0, I_d)$, one must have $H_X(\theta^*) \succcurlyeq cH_G(\theta^*)$ for some constant $c > 0$. In addition, it follows from (2.4) and (2.5) that $H_G(\theta^*)$ is upper and lower bounded (up to absolute constant factors) by the matrix

$$H = \frac{1}{B^3} u^* u^{*\top} + \frac{1}{B}(I_d - u^* u^{*\top})\,, \tag{5.4}$$

where $B = \max\{\|\theta^*\|, e\}$. Hence, the previous condition is of the form $H_X(\theta^*) \succcurlyeq cH$.

The following result shows that, whenever Assumptions 5.1 and 5.2 hold, Assumption 5.3 is necessary (up to logarithmic factors) for the condition $H_X(\theta^*) \succcurlyeq cH$ to hold.

Proposition 5.1. *Let $\theta^* \in \mathbb{R}^d$, and set $u^* \in S^{d-1}$ such that $\theta^* = \|\theta^*\|u^*$ as well as $B = \max\{\|\theta^*\|, e\}$. Let X be an isotropic random vector satisfying Assumption 5.1 with parameter $K \geqslant e$ and Assumption 5.2 with parameters $u^*, \eta = B^{-1}, c$. If $H_X(\theta^*) \succcurlyeq c_0 H$ for some $c_0 \in (0, 1)$, then for some constants c_1, c_2, c_3 depending only on K, c, c_0, one has*

$$\mathbf{P}\left(|\langle u^*, X\rangle| \leqslant \frac{c_1 \log B}{B},\ |\langle v, X\rangle| \geqslant \frac{\max\{\|u^* - v\|, B^{-1}\}}{c_2\sqrt{\log B}}\right) \geqslant \frac{c_3}{B\log^2(B)}\,. \tag{5.5}$$

The proof of this result can be found in Section 5.4.

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

Definition 5.1. 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, $\mathbf{E}[XX^\top] = I_d$) and satisfies Assumptions 5.2 and 5.3 with parameters u^*, η, c.

5.2 Examples of regular design distributions

In the previous section, we introduced certain regularity assumptions (Assumptions 5.1, 5.2 and 5.3) 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 5.2.1), log-concave distributions (Section 5.2.2), and product measures (Section 5.2.3).

We recall that the regularity assumptions introduced in Section 5.1 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 6.2 to apply.

5.2.1 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 5.2. *Let X be an isotropic and K-sub-exponential random vector (Assumption 5.1). 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 5.2 (proved in Section 5.3.1) 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 6.1 and 6.2 depend exponentially on c, leading to an exponential dependence on the signal strength $B = \eta^{-1}$. For this reason, Proposition 5.2 is mainly relevant in the case of constant signal strength.

5.2.2 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 6.2. As we shall see in Section 5.2.3, 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 5.3. *Assume that X has a centered isotropic (that is, $\mathbf{E}[X] = 0$ and $\mathbf{E}[XX^\top] = I_d$) 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 5.3 is provided in Section 5.3.2. 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 [LV07]. 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.

5.2.3 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 5.4. The random vector $X = (X_1, \dots, X_d)$ is such that: $X_1, \dots, X_d$ are i.i.d., with $\mathbf{E}[X_j] = 0$, $\mathbf{E}[X_j^2] = 1$ and $\|X_j\|_{\psi_1} \leqslant K$ (for some $K \geqslant e$) for $j = 1, \dots, d$.

It is a simple fact (see Lemma 5.6 in Section 5.3.3 below) that such a random vector is $4K$-sub-exponential. Hence, the main question is whether Assumptions 5.2 and 5.3 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 $\mathbf{P}(X_j = 1) = \mathbf{P}(X_j = -1) = 1/2$ for $1 \leqslant j \leqslant d$ (that is, X is uniform on the discrete hypercube $\{-1, 1\}^d$). This design satisfies Assumption 5.4; 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 3.1). 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 5.1. *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 $\mathbf{P}(\text{MLE exists}) \leqslant 0.1$.*

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

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

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 $\mathbf{P}(Y\langle\theta^*, X\rangle \leqslant 0) \leqslant \exp(-B)$. 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 5.3 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 5.1) 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 5.2 on the one-dimensional marginal $\langle u^*, X\rangle = \frac{1}{\sqrt{d}}\sum_{j=1}^d X_j$. Under Assumption 5.4, 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 5.1. *Let X satisfy Assumption 5.4. 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*

$$\frac{t}{4} \leqslant \mathbf{P}\big(|\langle u, X\rangle| \leqslant t\big) \leqslant t\,. \tag{5.6}$$

In particular, if $d \geqslant K^6$ and $u^* = (1/\sqrt{d}, \dots, 1/\sqrt{d})$, then Assumption 5.2 holds with $\eta = K^3/\sqrt{d}$ and $c = 1$.

Lemma 5.1 (whose proof is provided in Section 5.3.3) 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 5.1 (see Proposition 5.1), Assumption 5.2 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 5.3 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 5.3 is of the form

$$\mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant c\eta,\ |\langle v, X\rangle| \geqslant \frac{1}{c}\Big) \geqslant \frac{\eta}{c} \tag{5.7}$$

for some constant c. In the case where $X \sim \mathsf{N}(0, I_d)$ is Gaussian, condition (5.7) 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 5.2, a natural attempt to establish condition (5.7) 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 5.1, one can show that condition (5.7) 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 (5.7) holds for small η whenever v is sufficiently diffuse that $\|v\|_3^3$ is small. Unfortunately, condition (5.7) 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 5.3.3. Since the estimate on η_d obtained with this approach is sub-optimal and is improved in Lemma 5.2 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 5.1 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 5.2 below shows that this is indeed the case:

Lemma 5.2. *Let $X = (X_1, \dots, X_d)$ have i.i.d. coordinates, with $\mathbf{E}[X_1] = 0$, $\mathbf{E}[X_1^2] = 1$ and $\mathbf{E}[X_1^8] \leqslant \kappa^8$ 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*

$$\mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant \eta,\ |\langle v, X\rangle| \geqslant 0.2\max\{\eta, \|u^* - v\|_2\}\Big) \geqslant \frac{\eta}{70\,000\,\kappa^4}\,. \tag{5.8}$$

In particular, if X satisfies Assumption 5.4, then Assumption 5.3 holds for any $\eta \in [18K^3/\sqrt{d}, e^{-1}]$ with $c = 21\,000$.

Lemma 5.2 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 5.3.3.

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 5.4, 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 5.3.3 for more details on this proof.

We can now gather the conclusions of Lemmas 5.1 and 5.2 into the following statement, which is the main result of the present section.

Proposition 5.4. *Let $X = (X_1, \dots, X_d)$ satisfy Assumption 5.4, 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 6.1 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 5.2. *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* $\mathbf{P}(\text{MLE exists}) \leqslant 0.1$.

Proof. The proof is the same as that of Fact 5.1, 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

$$|\langle \theta^*, X \rangle| = B \Big| \sum_{j=1}^{d} X_j \Big| / \sqrt{d} \geqslant B/\sqrt{d}\,,$$

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 5.2. It turns out that, for a "typical" direction $u^* \in S^{d-1}$, Assumption 5.2 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 5.3.4.

5.3 Proofs of results from Section 5.2

5.3.1 Proof of Proposition 5.2 (regularity at constant scales)

First, note that Assumption 5.2 holds with $c = \eta^{-1} < c(K, \eta)$, since $\mathbf{P}(|\langle u^*, X \rangle| \leqslant t) \leqslant 1 \leqslant \eta^{-1} \cdot t$ for any $t \geqslant \eta$. We now show that Assumption 5.3 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 $s, t > 0$,

$$\mathbf{P}\big(|\langle u^*, X \rangle| \leqslant s,\, |\langle v, X \rangle| \geqslant t\big) \geqslant \mathbf{P}(|\langle v, X \rangle| \geqslant t) - \mathbf{P}(|\langle u^*, X \rangle| > s)\,.$$

In order to lower bound the first term above, we apply the Paley-Zygmund inequality (5.22) to $Z = \langle v, X \rangle^2$ (with $\mathbf{E}[Z] = 1$), which gives

$$\mathbf{P}\Big(|\langle v, X \rangle| \geqslant \frac{1}{\sqrt{2}}\Big) = \mathbf{P}\Big(\langle v, X \rangle^2 \geqslant \frac{1}{2}\mathbf{E}[\langle v, X \rangle^2]\Big) \geqslant \frac{1}{4} \frac{\mathbf{E}[\langle v, X \rangle^2]^2}{\mathbf{E}[\langle v, X \rangle^4]} \geqslant \frac{1}{4K^4}\,,$$

where the last inequality follows from the fact that $\|\langle v, X\rangle\|_{\psi_1} \leqslant K$, which by Definition 8.1 implies that $\|\langle v, X\rangle\|_4 \leqslant 4K/(2e) \leqslant K$. In addition, since $\|\langle u^*, X\rangle\|_{\psi_1} \leqslant K$, Lemma 8.1 implies that

$$\mathbf{P}\big(|\langle u^*, X\rangle| > 2K\log(2K)\big) \leqslant e^{-2\times 2\log(2K)} = \frac{1}{16K^4}\,. \tag{5.9}$$

Combining the previous inequalities and using that $\|u^* - v\| \leqslant \sqrt{2}$ and $\eta \leqslant e^{-1}$, we obtain that

$$\begin{aligned}\mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{2K\log(2K)}{\eta}\cdot\eta,\, |\langle v, X\rangle| \geqslant \frac{\max\{\|u^* - v\|, \eta\}}{2}\Big) &\geqslant \frac{1}{4K^4} - \frac{1}{16K^4} \\ = \frac{3}{16K^4} \geqslant \frac{3e\eta}{16K^4} &\geqslant \frac{\eta}{2K^4}\,,\end{aligned}$$

which shows that Assumption 5.3 holds with $c = c_{K,\eta}$ given by (5.9).

5.3.2 Proof of Proposition 5.3 (regularity of log-concave distributions)

In this section, we show that centered isotropic log-concave distributions satisfy Assumptions 5.1, 5.2 and 5.3, 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 5.3. *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 5.4. *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 $t > 0$,*

$$\mathbf{P}\big(|\langle u, X\rangle| \leqslant t\big) \leqslant 2t\,. \tag{5.10}$$

In other words, X satisfies Assumption 5.2 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 $\mathbf{E}[\langle u, X\rangle] = 0$ and $\mathbf{E}[\langle u, X\rangle^2] = 1$. 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 (5.10). ∎

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 5.2. For small scales, the proof uses the fact that centered and isotropic low-dimensional log-concave densities are lower-bounded around the origin.

Fact 5.3 ([LV07], Theorem 5.14 (a))**.** *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 5.5. *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 $v \in S^{d-1}$,*

$$\mathbf{P}\Big(|\langle u,X\rangle| \leqslant \eta\,;\, |\langle v,X\rangle| \geqslant \frac{1}{C}\max\big\{\eta, \|u-v\|\big\}\Big) \geqslant \frac{\eta}{C}\,. \tag{5.11}$$

Proof. Let $u,v \in S^{d-1}$ be such that $\langle u,v\rangle \geqslant 0$. Using Remark 2, we work with the quantity $\sqrt{1-\langle u,v\rangle^2}$ rather than $\|u-v\|$. Based on Fact 5.3, 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 5.3, 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

$$f_0 \geqslant 4\varepsilon^2 c_2 g_0\,. \tag{5.12}$$

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 (5.11), our goal is to bound from below (the integral of) f as

$$\int_{|s|\leqslant\eta}\int_{|t|\geqslant\frac{\max\{\eta,\sin\phi\}}{C}} f(s,t)\mathrm{d}s\mathrm{d}t \geqslant \frac{\eta}{C}\,, \tag{5.13}$$

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 (5.13). 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

$$\mathbf{P}\left(|U| \leqslant \eta\,;\, |V| \geqslant \frac{\eta}{c}\right) \geqslant \mathbf{P}(|U| \leqslant \eta) - \mathbf{P}\left(|V| < \frac{\eta}{c}\right)$$

On one hand, as $\eta \leqslant \varepsilon$, $\mathbf{P}(|U| \leqslant \eta) = \eta/\varepsilon$. 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

$$\begin{aligned}\mathbf{P}\Big(|V| \leqslant \frac{\eta}{c}\Big|W\Big) &= \mathbf{P}\Big(\big|U\cos\phi - W\sin\phi\big| \leqslant \frac{\eta}{c}\Big|W\Big)\\ &= \mathbf{P}\Big(\Big|U - W\frac{\sin\phi}{\cos\phi}\Big| \leqslant \frac{\eta}{c\cos\phi}\Big|W\Big)\\ &\leqslant \mathbf{P}\Big(\Big|U - W\frac{\sin\phi}{\cos\phi}\Big| \leqslant 1.1\frac{\eta}{c}\Big|W\Big)\,, \end{aligned} \tag{5.14}$$

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

$$\mathbf{P}(|U-t| \leqslant r) \leqslant \frac{r}{\varepsilon}\,.$$

Thus

$$\mathbf{P}\Big(\Big|U - W\frac{\sin\phi}{\cos\phi}\Big| \leqslant 1.1\frac{\eta}{c}\Big|W\Big) \leqslant 1.1\frac{\eta}{c\varepsilon}\,.$$

It then follows that as soon as $c \geqslant 2.2$,

$$\mathbf{P}\left(|U| \leqslant \eta\,;\, |V| \geqslant \frac{\eta}{c}\right) \geqslant \frac{\eta}{\varepsilon} - 1.1\frac{\eta}{c\varepsilon} \geqslant \frac{\eta}{2\varepsilon}\,. \tag{5.15}$$

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

$$\mathbf{P}\Big(|W-t| \geqslant \frac{\varepsilon}{2}\Big) \geqslant \frac{1}{2}\,.$$

This implies (since W is independent of U and symmetric) that as soon as $c \geqslant 2/\varepsilon$,

$$\begin{aligned}\mathbf{P}\Big(|V| \geqslant \frac{\sin\phi}{c}\Big|U\Big) &= \mathbf{P}\Big(\big|W\sin\phi - U\cos\phi\big| \geqslant \frac{\sin\phi}{c}\Big|U\Big)\\ &= \mathbf{P}\Big(\Big|W - U\frac{\cos\phi}{\sin\phi}\Big| \geqslant \frac{1}{c}\Big|U\Big)\\ &\geqslant \mathbf{P}\Big(\Big|W - U\frac{\cos\phi}{\sin\phi}\Big| \geqslant \frac{\varepsilon}{2}\Big|U\Big) \geqslant \frac{1}{2}\,.\end{aligned}$$

It follows that

$$\begin{aligned}\mathbf{P}\left(|U| \leqslant \eta\,;\, |V| \geqslant \frac{\sin\phi}{c}\right) &= \mathbf{E}\Big[\mathbf{1}\big(|U| \leqslant \eta\big)\mathbf{P}\Big(|V| \geqslant \frac{\sin\phi}{c}\Big|U\Big)\Big]\\ &\geqslant \frac{1}{2}\mathbf{P}\big(|U| \leqslant \eta\big) = \frac{\eta}{2\varepsilon}\,.\end{aligned} \tag{5.16}$$

Combining (5.15) and (5.16) shows that

$$\mathbf{P}\left(|U| \leqslant \eta\,;\, |V| \geqslant \frac{\max\{\eta,\sin\phi\}}{c}\right) \geqslant \frac{\eta}{2\varepsilon}\,, \tag{5.17}$$

meaning that g satisfies (5.13). Using (5.12) and applying the same change of variables to g_0 and f_0, we conclude that

$$\mathbf{P}\left(|\langle u, X\rangle| \leqslant \eta\,;\, |\langle v, X\rangle| \geqslant \frac{\varepsilon}{2}\max\left\{\eta, \sqrt{1-\langle u,v\rangle^2}\right\}\right) \geqslant 2\varepsilon c_2\eta\,. \qquad \blacksquare$$

5.3.3 Proof of Proposition 5.4 (regularity for i.i.d. coordinates)

This section contains the proofs of the results from Section 5.2.3. Specifically, we show that random vectors X with i.i.d. sub-exponential coordinates (Assumption 5.4) satisfy Assumptions 5.1, 5.2 and 5.3 down to a scale $\eta \asymp 1/\sqrt{d}$ in the "diffuse" direction $u^* = (1/\sqrt{d}, \dots, 1/\sqrt{d})$.

Assumption 5.1

We first recall the standard fact that a random vector with independent sub-exponential coordinates is itself sub-exponential.

Lemma 5.6. *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 8.1, 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 8.1 implies that $\|\langle v, X\rangle\|_{\psi_1} \leqslant 2\sqrt[3]{2e}\max(K/\sqrt{2}, 2 \cdot K/2) = 2\sqrt[3]{2e}K \leqslant 4K$. ∎

Assumption 5.2: proof of Lemma 5.1

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 5.7 ([Tyu12], Theorem 1)**.** *Let $Z_1, \dots, Z_d$ be independent centered random variables with $\sum_{j=1}^d \mathbf{E}[Z_j^2] = 1$. Let $Z = \sum_{j=1}^d Z_j$ and $G \sim \mathsf{N}(0,1)$. Then, for every $t \in \mathbb{R}$, one has*

$$\big|\mathbf{P}(Z \leqslant t) - \mathbf{P}(G \leqslant t)\big| \leqslant 0.56 \cdot \sum_{j=1}^d \mathbf{E}\big[|Z_j|^3\big]\,. \tag{5.18}$$

We now proceed with the proof of Lemma 5.1, which states that Assumption 5.2 holds.

Proof of Lemma 5.1. First, applying Lemma 5.7 to $-Z_1, \dots, -Z_d$ and $-s$ gives a similar bound as (5.18) for $\mathbf{P}(Z < s)$. After taking differences we deduce that, under the assumptions of Lemma 5.7, for every $s, t \in \mathbb{R}$ with $s \leqslant t$, one has

$$\big|\mathbf{P}(s \leqslant Z \leqslant t) - \mathbf{P}(s \leqslant G \leqslant t)\big| \leqslant 1.12 \cdot \sum_{j=1}^d \mathbf{E}\big[|Z_j|^3\big]\,. \tag{5.19}$$

We apply this inequality to $t \in [K^3\|u\|_3^3, 1]$, $s = -t$ and $Z_j = u_j X_j$, so that $\mathbf{E}[Z_j] = 0$, $\sum_{j=1}^d \mathbf{E}[Z_j^2] = \sum_{j=1}^d u_j^2 = 1$, and $\mathbf{E}[|Z_j|^3] = |u_j|^3\|X_j\|_3^3 \leqslant |u_j|^3(3K/2e)^3$. As $Z = \langle u, X\rangle$,

$$\big|\mathbf{P}(|\langle u, X\rangle| \leqslant t) - \mathbf{P}(|G| \leqslant t)\big| \leqslant 1.12 \cdot \sum_{j=1}^d \Big(\frac{3K}{2e}\Big)^3 |u_j|^3 \leqslant \frac{K^3}{5}\|u\|_3^3\,, \tag{5.20}$$

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

$$\frac{2t}{\sqrt{2\pi e}} \leqslant \mathbf{P}(|G| \leqslant t) \leqslant \frac{2t}{\sqrt{2\pi}}\,.$$

Plugging these inequalities into (5.20) and using that $K^3\|u\|_3^3 \leqslant t$ gives

$$\Big(\frac{2}{\sqrt{2\pi e}} - \frac{1}{5}\Big)t \leqslant \mathbf{P}(|\langle u, X\rangle| \leqslant t) \leqslant \Big(\sqrt{\frac{2}{\pi}} + \frac{1}{5}\Big)t\,,$$

which implies (5.6) by further bounding the numerical constants. ∎

Assumption 5.3: proof of Lemma 5.2

We now establish the two-dimensional margin condition.

Proof of Lemma 5.2. As discussed in Section 5.2.3, 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 $s > 0$,

$$\mathbf{P}\big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X\rangle| \geqslant s\big) = \mathbf{P}\big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X^\sigma\rangle| \geqslant s\big)\,.$$

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 $\mathbf{P}_\sigma$ and $\mathbf{E}_\sigma$ 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

$$\begin{aligned}
\mathbf{P}\big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X\rangle| \geqslant s\big) &= \frac{1}{d!}\sum_{\sigma'\in\mathfrak{S}_d} \mathbf{P}\big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X^{\sigma'}\rangle| \geqslant s\big) \\
&= \mathbf{E}\big[\mathbf{P}\big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X^\sigma\rangle| \geqslant s \,|\, \sigma\big)\big] \\
&= \mathbf{E}\big[\mathbf{P}_\sigma\big(|\langle u^*, X\rangle| \leqslant \eta,\, |\langle v, X^\sigma\rangle| \geqslant s\big)\big] \\
&= \mathbf{E}\big[\mathbf{1}\{|\langle u^*, X\rangle| \leqslant \eta\}\mathbf{P}_\sigma\big(|\langle v, X^\sigma\rangle| \geqslant s\big)\big]\,. \qquad (5.21)
\end{aligned}$$

Hence, in order to lower bound the left-hand side of (5.21), it suffices to lower bound $\mathbf{P}_\sigma\big(|\langle v, X^\sigma\rangle| \geqslant s\big)$ 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 $0 < \mathbf{E}[Z^2] < +\infty$, one has

$$\mathbf{P}\Big(Z \geqslant \frac{1}{2}\mathbf{E}[Z]\Big) \geqslant \frac{1}{4}\frac{\mathbf{E}[Z]^2}{\mathbf{E}[Z^2]}\,. \tag{5.22}$$

Applying this inequality to the random variable $Z = \langle v, X^\sigma\rangle^2$ conditionally on X gives

$$\mathbf{P}_\sigma\Big(|\langle v, X^\sigma\rangle| \geqslant \frac{1}{\sqrt{2}}\mathbf{E}_\sigma[\langle v, X^\sigma\rangle^2]^{1/2}\Big) \geqslant \frac{1}{4}\frac{\mathbf{E}_\sigma[\langle v, X^\sigma\rangle^2]^2}{\mathbf{E}_\sigma[\langle v, X^\sigma\rangle^4]}\,. \tag{5.23}$$

We are therefore led to bound $\mathbf{E}_\sigma[\langle v, X^\sigma\rangle^2]^{1/2}$ from below and $\mathbf{E}_\sigma[\langle v, X^\sigma\rangle^4]^{1/4}$ from above, ideally to conclude that these two quantities are both of the order of the value from Lemma 5.2. 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 $v \in S^{d-1}$,

$$\mathbf{E}_\sigma\big[\langle v, X^\sigma\rangle^2\big] = \mathbf{E}_\sigma\bigg[\Big(\sum_{i=1}^d v_i X_{\sigma(i)}\Big)^2\bigg] = \sum_{1\leqslant i,j\leqslant d} v_i v_j \mathbf{E}_\sigma[X_{\sigma(i)}X_{\sigma(j)}]\,.$$

For $i = j$, since $\sigma(i)$ is uniformly distributed on $\{1, \dots, d\}$ one has

$$\mathbf{E}_\sigma[X_{\sigma(i)}X_{\sigma(j)}] = \mathbf{E}_\sigma[X_{\sigma(i)}^2] = \frac{1}{d}\sum_{k=1}^d X_k^2 = \mu_2\,.$$

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

$$\mathbf{E}_\sigma[X_{\sigma(i)}X_{\sigma(j)}] = \frac{1}{d(d-1)}\sum_{k\neq l} X_k X_l = \frac{1}{d(d-1)}\bigg\{\Big(\sum_{k=1}^d X_k\Big)^2 - \sum_{k=1}^d X_k^2\bigg\} = \frac{\phi^2 - \mu_2}{d-1}.$$

Combining the previous two equations, we get for any i, j that

$$\mathbf{E}_\sigma[X_{\sigma(i)}X_{\sigma(j)}] = \frac{\phi^2 - \mu_2}{d-1} + \Big(\mu_2 + \frac{\mu_2 - \phi^2}{d-1}\Big)\mathbf{1}(i = j)\,.$$

Hence,

$$\begin{aligned}\mathbf{E}_\sigma\big[\langle v, X^\sigma\rangle^2\big] &= \frac{\phi^2-\mu_2}{d-1}\Big(\sum_{i=1}^d v_i\Big)^2 + \Big(\mu_2 + \frac{\mu_2-\phi^2}{d-1}\Big)\sum_{i=1}^d v_i^2 \\ &= \big(\phi^2-\mu_2\big)\frac{d}{d-1}\langle u^*, v\rangle^2 + \mu_2 + \frac{\mu_2-\phi^2}{d-1} \\ &= \frac{d}{d-1}\Big[\mu_2(1-\langle u^*, v\rangle^2) + \phi^2\langle u^*, v\rangle^2 - \frac{\phi^2}{d}\Big] \\ &\geqslant \mu_2(1-\langle u^*, v\rangle^2) + \phi^2\langle u^*, v\rangle^2 - \frac{\phi^2}{d}\,.\end{aligned}$$

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

$$\mathbf{E}_\sigma\big[\langle v, X^\sigma\rangle^2\big] \geqslant 0.37\big[1-\langle u^*, v\rangle^2 + \langle u^*, v\rangle^2\phi^2\big]\,. \tag{5.24}$$

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

$$\begin{aligned}\mathbf{E}_\sigma\big[\langle v, X^\sigma\rangle^4\big] &= \mathbf{E}_\sigma\big[\big(\alpha\langle w, X^\sigma\rangle + \sqrt{1-\alpha^2}\langle u, X^\sigma\rangle\big)^4\big] \\ &\leqslant 8\,\mathbf{E}_\sigma\big[\alpha^4\langle w, X^\sigma\rangle^4 + (1-\alpha^2)^2\langle u, X^\sigma\rangle^4\big] \\ &= 8\big\{(1-\langle u^*, v\rangle^2)^2\,\mathbf{E}_\sigma\big[\langle w, X^\sigma\rangle^4\big] + \langle u^*, v\rangle^4\phi^4\big\}\,.\end{aligned} \tag{5.25}$$

In light of (5.25), it remains to show that $\mathbf{E}_\sigma[\langle w, X^\sigma\rangle^4] \lesssim_\kappa 1$.

We start by writing:

$$\mathbf{E}_\sigma\big[\langle w, X^\sigma\rangle^4\big] = \sum_{1\leqslant i,j,k,l\leqslant d} w_i w_j w_k w_l \mathbf{E}[X_{\sigma(i)}X_{\sigma(j)}X_{\sigma(k)}X_{\sigma(l)}]\,. \tag{5.26}$$

We abbreviate "pairwise distinct" (indices in $\{1,\dots,d\}$) by p.d., and denote for i,j,k,l p.d.,

$$\begin{aligned}\alpha_4 &= \mathbf{E}_\sigma[X_{\sigma(i)}^4] \\ \alpha_{31} &= \mathbf{E}_\sigma[X_{\sigma(i)}^3 X_{\sigma(j)}] \\ \alpha_{22} &= \mathbf{E}_\sigma[X_{\sigma(i)}^2 X_{\sigma(j)}^2] \\ \alpha_{211} &= \mathbf{E}_\sigma[X_{\sigma(i)}^2 X_{\sigma(j)} X_{\sigma(k)}] \\ \alpha_{1111} &= \mathbf{E}_\sigma[X_{\sigma(i)}X_{\sigma(j)}X_{\sigma(k)}X_{\sigma(l)}]\,;\end{aligned}$$

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 (5.26) depending on the distinct indices, we obtain

$$\mathbf{E}_\sigma\big[\langle w, X^\sigma\rangle^4\big] = \Big(\sum_i w_i^4\Big)\alpha_4 + 4\Big(\sum_{i,j\,\mathrm{p.d.}} w_i^3 w_j\Big)\alpha_{31} + 3\Big(\sum_{i,j\,\mathrm{p.d.}} w_i^2 w_j^2\Big)\alpha_{22} + \\ + 6\Big(\sum_{i,j,k\,\mathrm{p.d.}} w_i^2 w_j w_k\Big)\alpha_{211} + \Big(\sum_{i,j,k,l\,\mathrm{p.d.}} w_i w_j w_k w_l\Big)\alpha_{1111}\,. \tag{5.27}$$

We control the sum in (5.27) 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

$$|\alpha_{\iota_1,\ldots,\iota_r}| \leqslant \mu_4\,. \tag{5.28}$$

To show (5.28), first note that since $\sigma(1)$ is uniformly distributed in $\{1,\ldots,d\}$, we have

$$\alpha_4 = \mathbf{E}_\sigma[X_{\sigma(1)}^4] = \frac{1}{d}\sum_{i=1}^d X_i^4 = \mu_4\,.$$

Now for $\iota_1,\ldots,\iota_r$ as above, Hölder's inequality (with $\iota_1/4 + \cdots + \iota_r/4 = 1$) implies that

$$\begin{aligned}|\alpha_{\iota_1,\ldots,\iota_r}| &\leqslant \mathbf{E}_\sigma[|X_{\sigma(1)}|^{\iota_1}\cdots|X_{\sigma(r)}|^{\iota_r}] = \mathbf{E}_\sigma[(X_{\sigma(1)}^4)^{\iota_1/4}\cdots(X_{\sigma(r)}^4)^{\iota_r/4}] \\ &\leqslant \mathbf{E}_\sigma[X_{\sigma(1)}^4]^{\iota_1/4}\cdots\mathbf{E}_\sigma[X_{\sigma(r)}^4]^{\iota_r/4} = \mu_4\,.\end{aligned}$$

We now turn to the control of the coefficients in (5.27) 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 (5.27) 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

$$\sum_{i,j\,\mathrm{p.d.}} w_i^3 w_j = \Big(\sum_i w_i^3\Big)\Big(\sum_j w_j\Big) - \sum_i w_i^4 = -\sum_i w_i^4 \in [-1,0]\,.$$

For the third term,

$$\sum_{i,j\,\mathrm{p.d.}} w_i^2 w_j^2 = \Big(\sum_i w_i^2\Big)^2 - \sum_i w_i^4 = 1 - \sum_i w_i^4 \in [0,1]\,.$$

For the fourth term, by distinguishing the different possible configurations of $i, j, k \in \{1, \dots, d\}$,

$$\sum_{i,j,k \text{ p.d.}} w_i^2 w_j w_k = \Big(\sum_i w_i^2\Big)\Big(\sum_j w_j\Big)\Big(\sum_k w_k\Big) - \sum_i w_i^4 - \sum_{i,j \text{ p.d.}} w_i^2 w_j^2 - 2 \sum_{i,j \text{ p.d.}} w_i^3 w_j \,. \tag{5.29}$$

Plugging the previous identities in (5.29), we obtain

$$\begin{aligned}\sum_{i,j,k \text{ p.d.}} w_i^2 w_j w_k &= -\sum_i w_i^4 - \Big(1 - \sum_i w_i^4\Big) - 2\Big(-\sum_i w_i^4\Big) \\ &= 2\sum_i w_i^4 - 1 \in [-1, 1] \,.\end{aligned}$$

Finally, for the fifth term, we write (collecting the terms similarly to (5.27))

$$\begin{aligned}\sum_{i,j,k,l \text{ p.d.}} w_i w_j w_k w_l = \Big(\sum_i w_i\Big)\Big(\sum_j w_j\Big)\Big(\sum_k w_k\Big)\Big(\sum_l w_l\Big) - \sum_i w_i^4 \\ - 4 \sum_{i,j \text{ p.d.}} w_i^3 w_j - 3 \sum_{i,j \text{ p.d.}} w_i^2 w_j^2 - 6 \sum_{i,j,k \text{ p.d.}} w_i^2 w_j w_k \,.\end{aligned} \tag{5.30}$$

Using the identities for the previous four terms, equation (5.30) becomes

$$\begin{aligned}\sum_{i,j,k,l \text{ p.d.}} w_i w_j w_k w_l &= -\sum_i w_i^4 - 4\Big(-\sum_i w_i^4\Big) - 3\Big(1 - \sum_i w_i^4\Big) - 6\Big(2\sum_i w_i^4 - 1\Big) \\ &= -6\sum_i w_i^4 + 3 \in [-3, 3] \,.\end{aligned}$$

Finally, injecting the previous bounds into the decomposition (5.27), we obtain

$$\begin{aligned}\mathbf{E}_\sigma\big[\langle w, X^\sigma\rangle^4\big] &\leqslant \Big|\sum_i w_i^4\Big| \cdot |\alpha_4| + 4\Big|\sum_{i,j \text{ p.d.}} w_i^3 w_j\Big| \cdot |\alpha_{31}| + 3\Big|\sum_{i,j \text{ p.d.}} w_i^2 w_j^2\Big| \cdot |\alpha_{22}| \\ &\quad + 6\Big|\sum_{i,j,k \text{ p.d.}} w_i^2 w_j w_k\Big| \cdot |\alpha_{211}| + \Big|\sum_{i,j,k,l \text{ p.d.}} w_i w_j w_k w_l\Big| \cdot |\alpha_{1111}| \\ &\leqslant \big(1 + 4 \times 1 + 3 \times 1 + 6 \times 1 + 3\big)\mu_4 = 17\,\mu_4 \,,\end{aligned}$$

which combined with (5.25) gives

$$\mathbf{E}_\sigma\big[\langle v, X^\sigma\rangle^4\big] \leqslant 8\big\{17(1 - \langle u^*, v\rangle^2)^2\mu_4 + \langle u^*, v\rangle^4\phi^4\big\} \,. \tag{5.31}$$

Symmetric condition. So far, we have established a lower bound on the second moment $\mathbf{E}_\sigma[\langle v, X^\sigma\rangle^2]$ and an upper bound on the fourth moment $\mathbf{E}_\sigma[\langle v, X^\sigma\rangle^4]$, 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

$$\mathbf{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta,\ \mu_2(X) \geqslant 1/2,\ \mu_4(X) \leqslant 2\kappa^4\big) \,.$$

We start by writing

$$\begin{aligned}
&\mathbf{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta,\, \mu_2(X) \geqslant 1/2,\, \mu_4(X) \leqslant 2\kappa^4\big) \\
&= \mathbf{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta\big) - \mathbf{P}\big(\{|\langle u^*, X\rangle| \leqslant \eta\} \cap \{\mu_2(X) < 1/2 \text{ or } \mu_4(X) > 2\kappa^4\}\big) \\
&\geqslant \mathbf{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta\big) - \mathbf{P}\big(\mu_2(X) < 1/2\big) - \mathbf{P}\big(\mu_4(X) > 2\kappa^4\big)\,. \qquad (5.32)
\end{aligned}$$

Now, applying the Berry-Esseen inequality (Lemma 5.7) and proceeding as in the proof of Lemma 5.1, using that $\mathbf{E}[(|X_i|/\sqrt{d})^3] \leqslant \mathbf{E}[|X_i|^8]^{3/8}/d^{3/2} \leqslant \kappa^3/d^{3/2}$, we get

$$\mathbf{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta\big) \geqslant \frac{\eta}{\sqrt{2\pi e}} - \frac{2.24\kappa^3}{\sqrt{d}} \geqslant 0.24\,\eta - \frac{2.25\kappa^3}{\sqrt{d}}\,. \qquad (5.33)$$

We now upper bound $\mathbf{P}(\mu_4(X) > 2\kappa^4)$. Applying Chebyshev's inequality to $\sum_{i=1}^d X_i^4$ gives, for any $t > 0$,

$$\mathbf{P}\Big(\Big|\frac{1}{d}\sum_{i=1}^d X_i^4 - \mathbf{E}[X_1^4]\Big| > t\Big) \leqslant \frac{\mathbf{E}[X_1^8]}{d\cdot t^2} \leqslant \frac{\kappa^8}{d\cdot t^2}\,.$$

In particular, taking $t = \kappa^4$, applying the triangle inequality and using that $\mathbf{E}[X_1^4] \leqslant \mathbf{E}[X_1^8]^{1/2} \leqslant \kappa^4$ by assumption, we get

$$\mathbf{P}\big(\mu_4(X) > 2\kappa^4\big) \leqslant \frac{1}{d}\,. \qquad (5.34)$$

Likewise, Chebyshev's inequality implies that

$$\mathbf{P}\big(\mu_2(X) < 1/2\big) \leqslant \mathbf{P}\Big(\Big|\frac{1}{d}\sum_{i=1}^d X_i^2 - 1\Big| > \frac{1}{2}\Big) \leqslant \frac{4\,\mathbf{E}[X_1^4]}{d} \leqslant \frac{4\kappa^4}{d}\,. \qquad (5.35)$$

Plugging inequalities (5.33), (5.34) and (5.35) into (5.32) gives

$$\mathbf{P}\big(\eta/2 \leqslant |\langle u^*, X\rangle| \leqslant \eta,\, \mu_2(X) \geqslant 1/2,\, \mu_4(X) \leqslant 2\kappa^4\big) \geqslant 0.24\,\eta - \frac{2.25\kappa^3}{\sqrt{d}} - \frac{1}{d} - \frac{4\kappa^4}{d}\,,$$

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 $\mathbf{P}(E) \geqslant 0.12\eta$ by the above. In addition, it follows respectively from (5.24) and (5.31) that, under the event E,

$$\begin{aligned}
&\mathbf{E}_\sigma\big[\langle v, X^\sigma\rangle^2\big] \geqslant 0.37\big[1 - \langle u^*, v\rangle^2 + \langle u^*, v\rangle^2(\eta/2)^2\big]\,; \\
&\mathbf{E}_\sigma\big[\langle v, X^\sigma\rangle^4\big] \leqslant 8\big\{34\kappa^4(1 - \langle u^*, v\rangle^2)^2 + \langle u^*, v\rangle^4\eta^4\big\}\,.
\end{aligned}$$

Plugging these upper and lower bounds into (5.23) gives:

$$\begin{aligned}\mathbf{P}_\sigma\Big(|\langle v, X^\sigma\rangle| \geqslant \frac{0.6}{\sqrt{2}}\big[1-\langle u^*,v\rangle^2+\langle u^*,v\rangle^2\eta^2/4\big]^{1/2}\Big)\\ \geqslant \frac{1}{4}\frac{0.37^2\big[1-\langle u^*,v\rangle^2+\langle u^*,v\rangle^2\eta^2/4\big]^2}{8\big[34\kappa^4(1-\langle u^*,v\rangle^2)^2+\langle u^*,v\rangle^4\eta^4\big]}\\ \geqslant \frac{0.37^2}{32}\frac{(1-\langle u^*,v\rangle^2)^2+\langle u^*,v\rangle^4\phi^4/16}{34\kappa^4(1-\langle u^*,v\rangle^2)^2+\langle u^*,v\rangle^4\eta^4} \geqslant \frac{1}{8000\kappa^4}\,.\end{aligned}$$

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 (5.21) and the above, we obtain

$$\begin{aligned}\mathbf{P}\big(|\langle u^*,X\rangle|\leqslant\eta,\,|\langle v,X\rangle|\geqslant s\big) &= \mathbf{E}\big[\mathbf{1}\{|\langle u^*,X\rangle|\leqslant\eta\}\mathbf{P}_\sigma\big(|\langle v,X^\sigma\rangle|\geqslant s\big)\big]\\ &\geqslant \mathbf{E}\big[\mathbf{1}_E\cdot\mathbf{P}_\sigma\big(|\langle v,X^\sigma\rangle|\geqslant s\big)\big]\\ &\geqslant \frac{\mathbf{P}(E)}{8000\kappa^4}\geqslant\frac{0.12\eta}{8000\kappa^4}\geqslant\frac{\eta}{70\,000\kappa^4}\,.\end{aligned}$$

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 8.3, 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 5.2 is obtained by lower-bounding $0.6/(2\sqrt{2})>0.2$.

Finally, the last part of Lemma 5.2 follows from (5.8), since under Assumption 5.4 one has $\mathbf{E}[X_1^4]^{1/4}\leqslant\kappa=\frac{4}{2e}\|X_1\|_{\psi_1}\leqslant\frac{2}{e}K$, 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 5.2.3, 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 (5.7) 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 (5.7) holds with $\eta \asymp \sqrt{k/d}$.

Taking the best of the two guarantees above (depending on the value of $k = k(v)$), condition (5.7) holds down to $\eta \asymp \min(\sqrt{k/d}, 1/\sqrt{k}) \leqslant d^{-1/4}$ for any $v \in S^{d-1}$.

5.3.4 Improved regularity scales in generic directions?

We now discuss the phenomenon alluded to in Section 5.2.3, namely that Assumption 5.2 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 $\mathbf{E}[X_j] = 0$ and $\mathbf{E}[X_j^2] = 1$ for $j = 1, \dots, d$ and with finite fourth moment. Let*

$$\kappa = \left(\frac{1}{d} \sum_{j=1}^d \mathbf{E}[X_j^4] \right)^{1/4}.$$

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

$$\sup_{a,b \in \mathbb{R},\, a \leqslant b} \left| \mathbf{P}(a \leqslant \langle u, X\rangle \leqslant b) - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-s^2/2} \mathrm{d}s \right| \leqslant \frac{c \log^2(1/\varepsilon) \kappa^4}{d}. \tag{5.36}$$

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 $\mathbf{P}(|\langle u, X\rangle| \leqslant t)$ is of order t as long as $t \gtrsim \log^2(d)/d$ (hence, Assumption 5.2 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 5.2) 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 5.1) 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 5.3. Indeed, as previously discussed, a key difficulty is that the property (5.3) 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 5.2 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 5.4 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 5.3 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 6.1 and Proposition 5.1, 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.

5.4 Proof of Proposition 5.1

In this section we provide the proof of Proposition 5.1 from Section 5.1, regarding the necessity of the two-dimensional margin assumption.

Proof of Proposition 5.1. For every $v \in S^{d-1}$ and $c_0 \geqslant 1$,

$$\begin{aligned}
\langle H_X(\theta^*)v, v\rangle &= \mathbf{E}\big[\sigma'(B\langle u^*, X\rangle)\langle v, X\rangle^2\big] \\
&= \mathbf{E}\Big[\sigma'(B\langle u^*, X\rangle)\mathbf{1}\Big(|\langle u^*, X\rangle| > \frac{c_0 \log B}{B}\Big)\langle v, X\rangle^2\Big] \\
&\qquad + \mathbf{E}\Big[\sigma'(B\langle u^*, X\rangle)\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B}\Big)\langle v, X\rangle^2\Big] \\
&\leqslant \frac{1}{B^{c_0}} + \mathbf{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B}\Big)\langle v, X\rangle^2\Big].
\end{aligned}$$

In view of Remark 2, we furthermore let $m = \max\big\{B^{-1}, \|u^* - v\|\big\}$. Conditioning on the value of $|\langle v, X\rangle|$, for every $C \geqslant 1$, the expectation in the second term above rewrites

$$\begin{aligned}
&\mathbf{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B}\Big)\langle v, X\rangle^2\Big] \\
&\quad = \mathbf{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| < \frac{m}{C}\Big)\langle v, X\rangle^2\Big] \\
&\quad + \mathbf{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| \geqslant \frac{m}{C}\Big)\langle v, X\rangle^2\Big]. \qquad (5.37)
\end{aligned}$$

Regarding the first term, we find using Assumption 5.2 that

$$\mathbf{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| < \frac{m}{C}\Big)\langle v, X\rangle^2\Big] \leqslant \frac{m^2}{C^2}\mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B}\Big) \leqslant \frac{m^2}{B} \cdot \frac{c\, c_0 \log B}{C^2}\,.$$

Finally, we further decompose the second term in (5.37). For all $\lambda \geqslant m/C$,

$$\begin{aligned}
&\mathbf{E}\Big[\mathbf{1}\Big\{|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| \geqslant \frac{m}{C}\Big\}\langle v, X\rangle^2\Big] \\
&\quad = \mathbf{E}\Big[\mathbf{1}\Big\{|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| \in \Big[\frac{m}{C}, \lambda\Big)\Big\}\langle v, X\rangle^2\Big] \\
&\qquad + \mathbf{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| \geqslant \lambda\Big)\langle v, X\rangle^2\Big] \\
&\quad \leqslant \lambda^2\, \mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| \geqslant \frac{m}{C}\Big) \\
&\qquad + \mathbf{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B}\Big)\mathbf{1}\big(|\langle v, X\rangle| \geqslant \lambda\big)\langle v, X\rangle^2\Big]\,. \qquad (5.38)
\end{aligned}$$

We now bound the last term using Fact **??**. 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 **??**, for all $\lambda \geqslant mK + c\eta$,

$$\begin{aligned}
\mathbf{E}\big[\mathbf{1}(|\langle u^*, X\rangle| \leqslant c\eta)\mathbf{1}(|\langle v, X\rangle| \geqslant \lambda)\langle v, X\rangle^2\big] &\leqslant \mathbf{E}\big[\mathbf{1}(m|\langle w, X\rangle| \geqslant \lambda - c\eta)\langle v, X\rangle^2\big] \\
&\leqslant \frac{4}{e^2}K^2 \exp\Big(-\frac{\lambda - c\eta}{mK}\Big)\,.
\end{aligned}$$

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 (5.38) can be bounded as

$$\mathbf{E}\Big[\mathbf{1}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B}\Big)\mathbf{1}\big(|\langle v, X\rangle| \geqslant \lambda\big)\langle v, X\rangle^2\Big] \leqslant \frac{4}{e^2} \cdot \frac{1}{C_0 K B^3} \leqslant \frac{4}{e^3} \cdot \frac{m^2}{C_0 B}\,, \qquad (5.39)$$

since $B^{-3} \leqslant m^2/B$.

Putting everything together, we find that

$$\begin{aligned}
\frac{m^2}{C_0 B} &\leqslant \mathbf{E}\big[\sigma'(B\langle u^*, X\rangle)\langle v, X\rangle^2\big] \\
&\leqslant \frac{1}{B^{c_0}} + \frac{m^2}{B} \cdot \frac{c\, c_0 \log B}{C^2} + \lambda^2\, \mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B};\, |\langle v, X\rangle| \geqslant \frac{m}{C}\Big) + \frac{4}{e^3} \cdot \frac{m^2}{B}\,. \qquad (5.40)
\end{aligned}$$

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 $C = 2\sqrt{c_0C_0c\log B}$, and finally in (5.39) we further bound $4/e^3 \leqslant 1/4$. Rearranging the terms in (5.40) yields

$$\mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant \frac{c_0 \log B}{B}; |\langle v, X\rangle| \geqslant \frac{m}{C}\Big) \geqslant \frac{m^2}{4C_0\lambda^2 B}\,.$$

the result follows by further bounding λ^2. ■

Chapter 6

Risk bounds for logistic regression in general settings

Abstract

The analysis of the Gaussian design case in Chapter 3 revealed which essential properties of this somewhat ideal distribution were directly used, which led us to define a notion of *regular distributions* in the previous chapter (Definition 5.1). This class of distributions enables a near-Gaussian behavior for the maximum-likelihood estimator, at least in the well-specified case. In this chapter, we present our results under a regular design assumption, first in the case of a well-specified model (Theorem 6.1), where the degradation compared to the Gaussian design setting is only logarithmic in the signal strength; and then in the most general setting where the model may be misspecified (Theorem 6.2), and show that the slightly worse dependence on the signal strength in this case is in fact unavoidable.

Contents

6.1 Introduction and outline

In Chapter 3, we first investigated random-design logistic regression in the simplest setting, which is that of a well-specified logistic model with a Gaussian design. Despite its simple description, this "toy" problem required a careful analysis and highly involved arguments in order to obtain sharp bounds with respect to all the parameters of the problem, and in particular the signal strength.

In this chapter, we provide guarantees for the logistic MLE when the design is no longer assumed to be Gaussian, but instead *regular*, a notion introduced in the previous chapter and that shall guarantee a near-Gaussian behavior for the maximum-likelihood estimator. We first do so while still assuming a well-specified model in Theorem 6.1; and we finally address the case where the logistic model might be misspecified, in which case the guarantees are slightly degraded but still optimal (up to logarithmic factors) in Theorem 6.2.

Let us highlight the key technical differences with the well-specified Gaussian model studied in Chapter 3. As mentioned in Chapter 2, the strategy to characterize the existence and performance of the MLE remains the same: we want to use the convex localization argument of Lemma 2.1. We concluded from this result that there are two main steps to prove the desired guarantees: an upper bound on the empirical gradient at θ^* (rescaled by the inverse of its covariance); and a uniform lower bound on the empirical Hessians in a neighborhood of θ^*, which should be as large as possible to impose a weak sufficient condition on the sample size for the MLE to exist. What we mean by technical differences with respect to Chapter 3 is therefore the question of which specific properties from the assumptions therein were used to bound empirical gradients and Hessians.

First, we recall that the Hessian of the logistic loss does not depend on the label Y, therefore, the way we bound from below empirical Hessians depends only on whether the design is Gaussian or not. In fact, the Gaussian design is a particular case of regular design (Definition 5.1) so any bound that holds in the regular case would in particular hold for the Gaussian distribution. The motivation of Theorem 3.2 is that the problem of well-specified logistic regression with a Gaussian design can arguably be considered as a basic enough question to seek sharp bounds with respect to all parameters involved, including the signal strength B. To that end, Theorem 3.2 removes a logarithmic factor compared to the more general result of Theorem 6.3 stated and proved in this chapter, that applies to all regular distributions. We also emphasize that the proof techniques for Theorem 3.2 and Theorem 6.3 are very different in nature.

Second, the logit model on the conditional distribution of Y given X allowed us to bound the empirical gradient at θ^* by taking advantage of the small probability of misclassification by θ^*. The key step was Lemma 3.1 which in turn allowed us to prove that $\nabla \widehat{L}_n(\theta^*)$ satisfies the sub-gamma property. This was only due to the assumption that the model was well-specified, not that the design was Gaussian. Therefore, this will remain instrumental in proving Theorem 6.1 where only the Gaussian design assumption is removed compared to Chapter 3. We also lose the independence of projections of a standard Gaussian vector in orthogonal directions, as well as the boundedness of their density. In short, when the model is still well-specified but no longer Gaussian, we still use Lemma 3.1 which reduces the task of proving that the empirical gradient is sub-gamma to that of bounding the expectations $\mathbf{E}\big[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p\big]$ for all $p \geqslant 2$. In the regular, but non Gaussian case, this relies mainly on the one-dimensional margin assumption (Assumption 5.2) since, as already noted in the Gaussian case, it is the

behavior of $\langle u^*, X\rangle$ (where $u^* = \theta^*/\|\theta^*\|$) near 0 that yields the right sub-gamma property (Proposition 6.1, proved in Section 6.5.1).

The task of bounding the gradient becomes structurally different when the model is no longer assumed to be well-specified, as Lemma 3.1 does not apply anymore. It relied on the small *conditional* probability $\mathbf{P}(Y\langle\theta^*, X\rangle < 0\,|X)$, which, in the most general misspecified case, cannot be upper bounded using any information about the conditional distribution of Y given X. Instead we bound the *total* probability $\mathbf{P}(Y\langle\theta^*, X\rangle < 0)$. We discuss this point in more detail in Section 6.5.2.

Organization of the chapter. In Section 6.2.1 we extend the results of to the case of a regular design, while still assuming that the model is well-specified. Then, in Section 6.2.2 we consider the most general case, where the design is regular and, in addition, no assumption is made on the conditional distribution of Y given X. In Section 6.5, we prove the bounds on the empirical gradient stated in Section 6.3. Specifically, Section 6.5.1 highlights how the one-dimensional margin condition of Assumption 5.2 is used to prove that the empirical gradient is sub-gamma with the right parameters and in Section 6.5.2 we address the issues related to misspecification (while still using Assumption 5.2). Theorem 6.3 in Section 6.4 provides a uniform lower bound on empirical Hessians, whose proof in Section 6.6 uses the new two-dimensional margin condition of Assumption 5.3 and motivates its definition.

6.2 Risk bounds for the MLE under regular design assumption

In this section we provide our main results regarding the performance of the maximum-likelihood estimator when the design is not Gaussian, first in the case of a well-specified logistic model (Section 6.2.1) then in the most general case of a misspecified model (Section 6.2.2).

6.2.1 Well-specified model

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 6.A.1.

Theorem 6.1. *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 5.1, 5.2 and 5.3 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

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c_2 \log^4(B)\frac{d+t}{n}\,. \tag{6.1}$$

The guarantees of Theorem 6.1 almost match (up to poly-logarithmic factors in B) those of Theorem 3.1 in the Gaussian case, which are optimal as discussed in Chapter 3. In fact, one can almost recover (again up to $\log^4(B)$ factors) the guarantees of Theorem 3.1 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.

6.2.2 Misspecified model

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 (1.8)), namely

$$\theta^* \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} L(\theta)\,, \quad L(\theta) = \mathbf{E}\big[\log\big(1 + e^{-Y\langle \theta, X\rangle}\big)\big]\,. \tag{6.2}$$

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 $\mathbf{E}[XX^\top] = I_d$, which ensures strict convexity of L. Theorem 6.2 below is proved in Section 6.A.2.

Theorem 6.2. *Suppose that X satisfies Assumptions 5.1, 5.2 and 5.3 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

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c_2 \log^4(B) \frac{d + Bt}{n}\,. \tag{6.3}$$

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

$$\mathbf{P}(\text{MLE exists}) \leqslant 1 - e^{-t}\,. \tag{6.4}$$

In addition, for the same distribution,

$$\liminf_{n \to \infty} \mathbf{P}\Big(L(\widehat{\theta}_n) - L(\theta^*) \geqslant c_3 \frac{d + Bt}{n}\Big) \geqslant e^{-t}\,. \tag{6.5}$$

Theorem 6.2 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[1], feature a polynomial dependence on B but a stronger one than in Theorem 6.2: the condition for existence of the MLE writes (ignoring polylog(B) factors) $n \gtrsim B^8 dt$, and the risk is bounded by $B^3 dt/n$.

It should be noted that both the sample size needed for the MLE to exist and the bound on its excess risk in Theorem 6.2 exhibit a stronger dependence on B compared

[1]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.

to the well-specified case. As shown by (6.4) and (6.5), this stronger dependence on B is in fact necessary in the misspecified case. This shows that the non-asymptotic guarantees of Theorem 6.2 for the existence and the excess risk of the MLE are sharp, up to polylogarithmic factors in B. It should also 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. We explain below how the potential misspecification of the model accounts for this degradation. Specifically, we exhibit a conditional distribution of Y given X for which the term Bt is unavoidable. We discuss this point below.

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 6.1. *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*

$$\mathbf{P}(Y \langle u^*, X \rangle < 0 \,|X) = p \,. \tag{6.6}$$

Then

1. *the signal strength $B = \max\{e, \|\theta^*\|\}$ is related to the probability of misclassification by*
$$\frac{1}{2B^2} \leqslant \mathbf{P}(Y \langle u^*, X \rangle < 0) \leqslant \frac{1}{B^2} \,.$$
2. *The covariance of the gradient $G = \mathbf{E}[\nabla \ell(\theta^*, Z) \nabla \ell(\theta^*, Z)^\top]$ satisfies*
$$\|H^{-1/2} G H^{-1/2}\|_{\mathrm{op}} \geqslant \frac{B}{8} \,.$$

The proof of this result is postponed to Section 6.B. Essentially, the first point of this result shows that the sample size from Theorem 6.2 is optimal, while the second point shows the optimality of the risk bound established in Theorem 6.2.

6.3 Bounds on empirical gradients

6.3.1 Well-specified model

Proposition 6.1 below allows to bound the empirical gradient in the 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 in Chapter 2 remains valid. In particular, the conditional probability of misclassification by θ^* is still exponentially small, in the sense that

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

which enables the desired sub-gamma behavior. However, two important properties of the Gaussian distribution that we used in the proof of Proposition 3.1 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 $(2\pi)^{-1/2}$) 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 6.5.1.

Proposition 6.1. *Assume that X satisfies Assumptions 5.1 and 5.2 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 $1 - 2e^{-t}$,*

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

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

We now move on to the most general case where no assumption is made on the conditional distribution of Y given X, meaning that the model is allowed to be misspecified.

6.3.2 Misspecified model

Proposition 6.2. *Assume that X satisfies Assumptions 5.1 and 5.2 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 $1 - 2e^{-t}$,*

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

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

The fact that the model may be misspecified induces a significant change: the key bound (2.10) 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 $\mathbf{P}(Y_i\langle\theta^*, X_i\rangle < 0 \,|X_i)$ is available, it might be tempting to simply bound $|\sigma_i| \leqslant 1$, which as discussed in Chapter 2 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 $\mathbf{P}(Y\langle\theta^*, X\rangle < 0)$ of misclassification of θ^* must be small. The key result that expresses this intuition is Lemma 6.3, which shows that the probability of misclassification $\mathbf{P}(Y\langle\theta^*, X\rangle < 0)$ and the first moment $\mathbf{E}[|\langle u^*, X\rangle|\mathbf{1}(Y\langle\theta^*, X\rangle < 0)]$ 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 $\log(B)\sqrt{(d + Bt)/n}$.

The proof of Proposition 6.2 can be found in Section 6.5.2.

6.4 Uniform bound on Hessians

We now turn to the second key component needed to use Lemma 2.1. Theorem 6.3 below provides an almost optimal uniform lower bound on the empirical Hessians, of the

form (2.16), up to logarithmic factors in B. We note in passing that this control on the Hessians only requires Assumptions 5.1 and 5.3, while Assumption 5.2 was used in the control of the gradient discussed in Section 6.3.

Theorem 6.3. *Let X be a random vector satisfying Assumptions 5.1 and 5.3 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 6.3 is proved in Section 6.6, 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 (2.16). Specifically, the first step is to notice that if $X_1, \dots, X_n$ satisfy, for some constants c_0, c_1,

$$\inf_{\substack{u,v \in S^{d-1} \\ \|u-u^*\| \leqslant c_0/B \\ \langle u^*, v\rangle \geqslant 0}} \left[\sum_{i=1}^n \mathbf{1}\Big\{ |\langle u, X_i\rangle| \leqslant \frac{c_1}{B}\,;\, |\langle v, X_i\rangle| \geqslant \frac{\max\{B^{-1}, \|u^* - v\|\}}{c_1} \Big\} \right] \geqslant \frac{n}{2c_1 B}\,, \tag{6.7}$$

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 (6.7) holds with high probability over the random draw of $X_1, \dots, X_n$. To achieve this, observe first that condition (6.7) is essentially a variant of Assumption 5.3, 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 (6.7) (or rather, a slightly weaker version with additional $\log B$ factors) is thus established in two steps. First, we show that Assumption 5.3 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 (6.7) 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 6.3 applies to the general regular case, and in particular to the special case of a Gaussian design. (That the Gaussian design satisfies Assumption 5.3 may be verified using independence of orthogonal linear marginals, or alternatively follows from Proposition 5.3.) In fact, the identification of condition (6.7) as a structural property implying the "right" behavior of the empirical Hessian in the Gaussian case is what motivates the definition of Assumption 5.3.

At the same time, it should be noted that the guarantees of Theorem 6.3 feature additional $\log B$ factors, compared to "ideal" guarantees that led to Theorem 3.1 in the Gaussian case. In particular, the (sufficient) condition on the sample size n from Theorem 6.3 is stronger by a $\log B$ factor than the necessary condition presented at the end of Theorem 3.1.

Theorem 3.2 in Chapter 3 addressed this sub-optimality by using a very different proof scheme from the one outlined above. To the best of our knowledge, the PAC-Bayesian method was the only way to remove the unnecessary logarithmic factor and therefore obtain an optimal uniform lower bound.

6.5 Proofs of upper bounds on the empirical gradient

6.5.1 Proof of Proposition 6.1 (regular design, well-specified model)

We now prove Proposition 6.1, 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 5.1 and 5.2 with parameters u^*, $\eta = B^{-1}$ and $c \geqslant 1$. Note that the two-dimensional margin condition of Assumption 5.3 is needed only to bound from below the empirical Hessians. We now proceed with the proof.

Since the model is well-specified, Inequality (3.5) still holds, thereby reducing the problem to that of bounding $\mathbf{E}[\exp(-|\langle\theta^*, X\rangle|)|\langle v, X\rangle|^p]$ when the design X is regular, which is a key step in proving the result.

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

Lemma 6.2. *Let $\theta^* \in \mathbb{R}^d \setminus \{0\}$, $u^* = \theta^*/\|\theta^*\|$ and $B = \max(e, \|\theta^*\|)$. Suppose that X satisfies Assumptions 5.1 with parameter $K > 0$ and 5.2 with parameters $\eta = 1/B$ and $c \geqslant 1$. Then, for any $p \geqslant 0$, for any $v \in S^{d-1}$,*

$$\mathbf{E}\big[\exp\big(-|\langle\theta^*, X\rangle|\big)|\langle u^*, X\rangle|^p\big] \leqslant \frac{9c}{B}\left(\frac{K\log(B)}{2B}\right)^p p!\,. \tag{6.8}$$

$$\mathbf{E}\big[\exp\big(-|\langle\theta^*, X\rangle|\big)|\langle v, X\rangle|^p\big] \leqslant \frac{5ec}{B}\left(\frac{K\log(B)}{2}\right)^p p!\,. \tag{6.9}$$

Proof of Lemma 6.2. By Assumption 5.1, $\|\langle v, X\rangle\|_{\psi_1} \leqslant K$, so by Definition 8.1,

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

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 5.2, for any $b > 0$

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

This yields

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

This proves (6.8). For (6.9), Hölder's inequality implies that for any $\nu \in (0,1)$,

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

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

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

This concludes the proof of (6.9). ■

Conclusion of the proof. By (6.8) and Bernstein's inequality, we deduce that, for any $t > 0$, with probability larger than $1 - 2e^{-t}$,

$$|\langle u^*, \nabla \widehat{L}_n(\theta^*)\rangle| \leqslant \frac{K\log B}{B^{3/2}}\sqrt{\frac{t}{n}}\bigg(\sqrt{\frac{9c}{2}} + \sqrt{\frac{Bt}{4n}}\bigg) \leqslant 3.2\frac{K\log B}{B^{3/2}}\sqrt{\frac{ct}{n}}\,, \tag{6.10}$$

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

From (6.9) and Lemma 2.3, with probability larger than $1 - e^{-t}$, for any $w \in S^{d-1}$ such that $\langle u^*, w\rangle = 0$,

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

Plugging this upper bound and (6.10) into (3.4) concludes the proof of the proposition.

6.5.2 Proof of Proposition 6.2 (regular design, misspecified model)

We now turn to the proof of Proposition 6.2, which provides a deviation bound on the empirical gradient in the misspecified case. Specifically, X satisfies Assumptions 5.1 and 5.2 and the model might not be misspecified. The parameter θ^* is defined using the joint distribution of $Z = (X, Y)$ by

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

In this setting, we can no longer exploit the low conditional probability of misclasification by θ^* used in Chapter 3 and in the previous section, namely Inequality (3.5). However, the decomposition that led to this bound is still relevant, but has to be exploited using another, structurally different idea. We now explain this new approach in details.

Recall that for every pair $Z = (X, Y) \in \mathbb{R}^d \times \{-1, 1\}$, one has

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

Following the proof of Lemma 3.1, it still holds that for all $p \geqslant 1$ and all $v \in S^{d-1}$,

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

and that

$$\begin{aligned}\sigma(-Y\langle \theta^*, X\rangle) &= \sigma(-|\langle \theta^*, X\rangle|)\mathbf{1}\{Y\langle \theta^*, X\rangle > 0\} + \sigma(\langle \theta^*, X\rangle)\mathbf{1}\{Y\langle \theta^*, X\rangle < 0\} \\ &\leqslant \sigma\big(-|\langle \theta^*, X\rangle|\big) + \mathbf{1}\{Y\langle \theta^*, X\rangle < 0\} \\ &\leqslant \exp\big(-|\langle \theta^*, X\rangle|\big) + \mathbf{1}\{Y\langle \theta^*, X\rangle < 0\}\,. \end{aligned} \tag{6.11}$$

The key difference compared to the well-specified setting is that we can no longer bound directly $\mathbf{E}[\mathbf{1}\{Y\langle \theta^*, X\rangle < 0\}\,|X]$ because we make no assumption on the conditional distribution of Y given X. It follows that for all $v \in S^{d-1}$

$$\mathbf{E}\big|\big\langle v, \nabla \ell(\theta^*, Z)\big\rangle\big|^p \leqslant \mathbf{E}\big[|\langle v, X\rangle|^p \big(\exp\big(-|\langle \theta^*, X\rangle|\big) + \mathbf{1}\{Y\langle \theta^*, X\rangle < 0\}\big)\big]\,, \tag{6.12}$$

and our task is still to bound, for all $p \geqslant 2$, the expectations

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

The first expectation is bounded using Lemma 6.2. Therefore, we focus in this proof on the second expectation

$$\mathbf{E}\big[\mathbf{1}\{Y\langle \theta^*, X\rangle < 0\}\,|\langle v, X\rangle|^p\big]\,. \tag{6.13}$$

This is precisely where the approach become structurally different from the well-specified case.

When the model might be misspecified, the control of this last expectation is slightly worse than the one provided in Lemma 6.2 for the first expectation, yielding the extra $\sqrt{B}$ in front of $\sqrt{t}$ (which in general cannot be avoided). From a higher-level (but still technical) perspective, the way we prove Proposition 6.2 is quite different from the well-specified case. To prove Proposition 6.1, we bounded all the moments $\mathbf{E}|\langle v, \nabla \ell(\theta^*, Z)\rangle|^p$ for all $p \geqslant 2$ to prove the sub-gamma behavior of all the projections of the gradient. In the misspecified case, we rather rely on the sixth point in Lemma 8.1. This result shows that a sub-exponential variable satisfies a Bernstein-type inequality where the first-order term scales according to its L^2-norm rather than its ψ_1-norm (which appears with only an extra logarithmic factor in the second order term). In short, if $\xi_1, \dots, \xi_n$ denote independent sub-exponential variables with variance at most σ^2 and ψ_1-norm at most K, we refine the standard $1-\delta$-probability bound

$$\Big|\frac{1}{n}\sum_{i=1}^n \xi_i\Big| \lesssim K\Big(\sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n}\Big)$$

into

$$\Big|\frac{1}{n}\sum_{i=1}^n \xi_i\Big| \lesssim \sigma\sqrt{\frac{\log(1/\delta)}{n}} + K\log(K/\sigma)\frac{\log(1/\delta)}{n}\,. \tag{6.14}$$

In the proof of Lemma 8.1, we actually bound the moments of the ξ_i's of all orders larger than 2. Here, we do not directly bound the moments (6.13) of all orders, as we apply Lemma 8.1 where it is done to obtain the refined bound (6.14).

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 $p = 2$, which in turn will allow us to use Point 6 in Lemma 8.1.

Lemma 6.3. *Let $\theta^* \in \mathbb{R}^d$ such that $\|\theta^*\| \geqslant e$ and let $u^* = \theta^*/\|\theta^*\|$. Suppose that X satisfies Assumption 5.2 with parameters (u^*, B^{-1}, c). Then,*

$$\mathbf{E}\big[|\langle\theta^*, X\rangle|\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\big] \leqslant \frac{6c}{B^2}\,; \tag{6.15}$$

$$\mathbf{P}(Y\langle\theta^*, X\rangle < 0) \leqslant \frac{3.21c}{B}\,. \tag{6.16}$$

Remark 3. 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

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

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

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

which writes

$$\mathbf{E}\big[|\langle\theta^*, X\rangle|\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\big] = \mathbf{E}\big[|\langle\theta^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big]\,. \tag{6.17}$$

By (6.8) applied with $p = 1$, this shows (6.15). Moreover, as

$$1 \leqslant |\langle\theta^*, X\rangle| + \mathbf{1}(|\langle\theta^*, X\rangle| \leqslant 1)\,,$$

we have

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

Bounding the first term with (6.15) and the second using Assumption 5.2 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 6.4. *Let $\theta^* \in \mathbb{R}^d$ of direction $u^* \in S^{d-1}$. Suppose that X satisfies Assumptions 5.1 and 5.2 with parameters $K \geqslant e$, (u^*, B^{-1}, c). Then, for any $v \in S^{d-1}$,*

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

Proof. We start with the second inequality. As $\mathbf{E}[\langle v, X\rangle^2] = 1$, 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 (6.16), $\mathbf{P}(Y\langle\theta^*, X\rangle < 0) \leqslant \min\{1, 3.21c/B\}$. Hence, by the first part of Lemma 6.5,

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

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

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

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

so the inequality holds in this case. Hence, we may assume that $\|\theta^*\| \geqslant e$. In this case, by (6.15),

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

Thus, applying the second part of Lemma 6.5 below with

$$U = |\langle u^*, X\rangle|\mathbf{1}(Y\langle\theta^*, X\rangle < 0) \quad \text{and} \quad V = |\langle u^*, X\rangle|\,,$$

we get

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

Note that, together with Lemma 6.2 and (6.12), Lemma 6.4 shows that, for any $v \in S^{d-1}$,

$$\mathbf{E}[\langle u^*, \nabla\ell(\theta^*, Z)\rangle^2] \leqslant \frac{7.1cK\log(KB^2)}{B^2}\,. \tag{6.18}$$
$$\mathbf{E}[\langle v, \nabla\ell(\theta^*, Z)\rangle^2] \leqslant \frac{cK^2\log^2 B}{B}\,. \tag{6.19}$$

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 8.1. 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 (6.18) 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 5.1, $\|\langle u^*, \nabla\ell(\theta^*, Z)\rangle\|_{\psi_1} \leqslant K$. Hence, by Point 6 in Lemma 8.1, they are (ν^2, K') sub-gamma, with

$$\nu^2 = \frac{C^2K^2\log(B)}{B^2}\,, \quad K' = \max(e\nu, K)\log\left(\frac{B\max\{e\nu, K\}}{CK}\right) \leqslant c'K\log(B)\,, \tag{6.20}$$

where c' denote a function of c and K whose value may change from line to line. Therefore, by Lemma 8.1, for any $t > 0$, with probability larger than $1 - 2\exp(-t)$,

$$|\langle u^*, \nabla\widehat{L}_n(\theta^*)\rangle| \leqslant \frac{c'K}{B}\sqrt{\frac{t}{n}}\left(\sqrt{\log B} + B\log B\sqrt{\frac{t}{n}}\right) \leqslant c'\frac{K\log B}{B}\sqrt{\frac{t}{n}}\,. \tag{6.21}$$

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 (6.19). 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 5.1, $\|\langle v, \nabla\ell(\theta^*, Z)\rangle\|_{\psi_1} \leqslant K$. Hence, by Point 6 in Lemma 8.1, they are (ν^2, K') sub-gamma, with

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

Therefore, by Lemma 2.3, 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 (6.21) into (3.4) concludes the proof of the proposition.

We conclude this section with the following lemma that was used in the proof of Lemma 6.4.

Lemma 6.5. *Let U, V be nonnegative real random variables such that $\mathbf{E}[U] \leqslant \varepsilon$ and $\|V\|_{\psi_1} \leqslant K$ for some $\varepsilon, K > 0$.*

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

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 $u \in [0, 1]$)

$$\begin{aligned}\mathbf{E}\big[UV^2\big] &\leqslant \mathbf{E}\big[|V|^{2p}\big]^{1/p}\,\mathbf{E}\big[U^{p/(p-1)}\big]^{1-1/p} \leqslant \|V\|_{2p}^2\,\mathbf{E}[U]^{1-1/p} \\ &\leqslant \left(\frac{Kp}{2e}\right)^2\varepsilon^{1-1/p} = \frac{K^2\varepsilon}{4e^2}\varepsilon^{-1/p}p^2\,.\end{aligned}$$

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

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

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

$$\mathbf{E}[UV] \leqslant \mathbf{E}[V^p]^{1/p} \mathbf{E}[U^q]^{1/q} \leqslant \frac{Kp}{2e} \mathbf{E}[U^q]^{1/q} \,, \tag{6.22}$$

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

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

where we used that $\mathbf{E}[U] \leqslant \varepsilon$ 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

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

Plugging this inequality into (6.22) and letting $p \to 2\log(e \vee K/\varepsilon) \geqslant 2$, so that $\lim(K/\varepsilon)^{2/p} \leqslant e$, we get

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

which establishes the second point. ■

In the next section we prove the uniform lower bound on empirical Hessians, Theorem 6.3.

6.6 Proof of Theorem 6.3

In this section, we prove Theorem 6.3, namely the uniform lower bound on empirical Hessian matrices in the case of a regular design. Specifically, we assume that X satisfies Assumptions 5.1 and 5.3 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 $m, M > 0$,

$$\langle \widehat{H}_n(\theta) v, v \rangle \geqslant \frac{\sigma'(m(1+r)B) M^2}{n} \sum_{i=1}^n \mathbf{1}\{ |\langle u, X_i \rangle| \leqslant m, \ |\langle v, X_i \rangle| \geqslant M \} \,, \tag{6.23}$$

where we also used that, as $\|\theta-\theta^*\|_H \leqslant r/\sqrt{B}$, $\|\theta\| \leqslant (1+r)B$ by Lemma 8.4. It remains to bound from below the empirical process

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

uniformly over $\theta \in \Theta$ and $v \in S^{d-1}$, for a proper choice of m and M. We want to apply Lemma 6.6. For this, we have to estimate $\mathbf{P}(|\langle u, X_i\rangle| \leqslant m,\ |\langle v, X_i\rangle| \geqslant M)$ for each $\theta \in \Theta$ and $v \in S^{d-1}$. If $\|\theta^*\| \leqslant e$, $B = e$ so $\eta = 1/e$, so Proposition 5.2 shows that Assumption 5.3 is satisfied with constant $\max\{2eK\log(2K), 2K^4\} = 2K^4$. Therefore,

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

When $\|\theta^*\| \geqslant e$, the third point of Lemma 8.4 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 6.7, this implies that for all $\theta \in \Theta$ and $v \in S^{d-1}$, one has for all $t \geqslant 1/B$

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

This suggests to choose $m = \gamma/B$, $M = \max(1/B, \|u^*-v\|)/\gamma$ in (6.23), 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 $v \in S^{d-1}$,

$$\mathbf{P}\left(\big|\langle u, X\rangle\big| \leqslant \frac{\gamma}{B}\,;\, \big|\langle v, X\rangle\big| \geqslant \frac{\max\{1/B, \|u^*-v\|\}}{\gamma}\right) \geqslant \frac{1}{\gamma B}\,. \tag{6.24}$$

The next step to apply Lemma 6.6 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 has VC dimension 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 6.6 applies with $p = 1/(\gamma B)$ and VC dimension $165d$. It shows that, whenever

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

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 (6.23) shows that, on the same event,

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

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 $1+r+2/e \leqslant 2$. This concludes the proof of Theorem 6.3.

6.6.1 Technical lemmas for the proof of Theorem 6.3

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 6.6. *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*

$$n \geqslant \frac{1600\log(4400/p)\,d}{p}\,, \tag{6.25}$$

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

$$\inf_{A\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \in A) \geqslant \frac{p}{2}\,. \tag{6.26}$$

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 (6.26), 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

$$Z \leqslant 2\left(\mathbf{E}[Z] + \sqrt{\frac{pt}{n}} + \frac{t}{n}\right). \tag{6.27}$$

Hence, if $\mathbf{E}[Z] \leqslant p/8$, 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 $\mathbf{E}[Z] \leqslant p/8$. First, by symmetrization (e.g. [Kol11, Theorem 2.1]), one has

$$\mathbf{E}[Z] \leqslant \frac{2}{n}\,\mathbf{E}\Big[\sup_{A' \in \mathcal{A}'} \sum_{i=1}^n \varepsilon_i \mathbf{1}(X_i' \in A')\Big]\,, \tag{6.28}$$

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

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

where $S_n(\mathcal{A}') = \max_{x_1', \ldots, x_n'} |\{(\mathbf{1}(x_i' \in A'))_{1 \leqslant i \leqslant n} : A' \in \mathcal{A}'\}|$ 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:

$$\begin{aligned} R_n &= \mathbf{E}\Big[\sup_{A' \in \mathcal{A}'} \sum_{i=1}^n \varepsilon_i \mathbf{1}(X_i' \in A')\Big] = \mathbf{E}\Big[\mathbf{E}\Big[\sup_{A' \in \mathcal{A}'} \sum_{i=1}^n \varepsilon_i \mathbf{1}(X_i' \in A') \Big| X_1, \ldots, X_n\Big]\Big] \\ &\leqslant \sqrt{2\,\mathbf{E}\Big[\sup_{A' \in \mathcal{A}'} \sum_{i=1}^n \mathbf{1}(X_i' \in A')\Big] \log\Big[(n+1)\Big(\frac{en}{d}\Big)^d\Big]}\,. \end{aligned} \tag{6.29}$$

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

$$\begin{aligned} \mathbf{E}\Big[\sup_{A' \in \mathcal{A}'} \sum_{i=1}^n \mathbf{1}(X_i' \in A')\Big] &= np + \mathbf{E}\Big[\sup_{A' \in \mathcal{A}'} \sum_{i=1}^n \big\{\mathbf{1}(X_i' \in A') - \mathbf{P}(X_i' \in A')\big\}\Big] \\ &\leqslant np + 2\,\mathbf{E}\Big[\sup_{A' \in \mathcal{A}'} \sum_{i=1}^n \varepsilon_i \mathbf{1}(X_i' \in A')\Big] = np + 2R_n\,. \end{aligned} \tag{6.30}$$

Plugging (6.30) into (6.29) 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 (6.28) that $\mathbf{E}[Z] \leqslant 2R_n/n$, we get whenever $H_n \leqslant \varepsilon np$ with $\varepsilon = 1/1160$:

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

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 (6.25), 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 $v = 1160/p$). ∎

To apply Lemma 6.6 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 5.3 can be extended to all u in a neighborhood of u^* as shown in the following result.

Lemma 6.7. *Suppose that Assumptions 5.1 and 5.3 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*

$$\|u - u^*\| \leqslant \frac{2\eta}{K\log(c(c+1)/\eta)} \quad \text{and} \quad \langle u^*, v\rangle \geqslant 0\,, \tag{6.31}$$

one has

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

Proof. Let $u, v \in S^{d-1}$ satisfy (6.31). The triangle inequality

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

implies that

$$\begin{aligned}
&\mathbf{P}\Big(|\langle u, X\rangle| \leqslant (c+1)\eta\,;\, |\langle v, X\rangle| \geqslant \frac{\max\{\eta, \|u^* - v\|\}}{c}\Big) \\
&\geqslant \mathbf{P}\Big(|\langle u^*, X\rangle| \leqslant c\eta\,;\, |\langle v, X\rangle| \geqslant \frac{\max\{\eta, \|u^* - v\|\}}{c}\Big) - \mathbf{P}\big(|\langle u - u^*, X\rangle| > \eta\big)\,. \qquad (6.32)
\end{aligned}$$

Next, on the one hand, Assumption 5.3 asserts that

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

and on the other hand, Assumption 5.1 together with Point 1 in Lemma 8.1 implies that

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

Plugging the previous two inequalities into (6.32) concludes the proof, since

$$\frac{\eta}{c} - \frac{\eta}{c(c+1)} = \frac{\eta}{c+1}\,.$$

∎

Appendix 6.A: Proofs of the main results

In this section we prove Theorems 6.1 and 6.2. The proofs only gather the results of the previous sections regarding empirical gradients and Hessians, following the scheme given by Lemma 2.1. The technically difficult part was to establish bounds on gradients and Hessians.

We start with the proof of Theorem 6.1, which combines Proposition 6.1 and Theorem 6.3.

6.A.1 Proof of Theorem 6.1

By Proposition 6.1, we have, for any $n \geqslant B(d+t)$, for any $t > 0$, with probability at least $1 - 3e^{-t}$,

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

Moreover, by Theorem 6.3, if

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

then, with probability $1 - \exp(-t)$,

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

By Lemma 2.1, it follows that, if

$$n \geqslant 4\left(\frac{c'}{c_2 c_3}\right)^2 (\log B)^4 B(d+t) \,,$$

then with probability at least $1 - 4e^{-t}$,

$$\|\widehat{\theta}_n - \theta^*\|_H \leqslant \frac{2c' \log B}{c_0} \sqrt{\frac{d+t}{n}} \,. \tag{6.33}$$

By Lemmas 6.8 and 2.1, on the same event, there exists a constant c' that only depends on c and K such that

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c' (\log B)^4 \frac{d+t}{n} \,.$$

This concludes the proof of Theorem 6.1.

We move on to the proof of Theorem 6.2, which combines Proposition 6.2 with Theorem 6.3.

6.A.2 Proof of Theorem 6.2

By Proposition 6.2, if $n \geqslant B(d + Bt)$, then with probability larger than $1 - 3e^{-t}$,

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

Moreover, by Theorem 6.3, if

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

then, with probability $1 - \exp(-t)$,

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

By Lemma 2.1, it follows that, if

$$n \geqslant 4\left(\frac{c'}{c_2 c_3}\right)^2 (\log B)^4 B(d + Bt)\,, \tag{6.34}$$

with probability at least $1 - 4e^{-t}$,

$$\|\widehat{\theta}_n - \theta^*\|_H \leqslant \frac{2c'(\log B)^2}{c_0}\sqrt{\frac{d + Bt}{n}}\,. \tag{6.35}$$

By Lemmas 6.8 and 2.1, on the same event, there exists a function c' of c and K only such that

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

This establishes the first part of Theorem 6.2.

We now turn to the second part of the proof regarding the optimality of this result, namely the upper bound on the probability of existence of the MLE (6.4) and the asymptotic lower bound on its excess risk (6.5).

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 $\mathbf{P}(Y\langle u^*, X\rangle < 0|X)$ is constant (see (6.6)). The first point of Lemma 6.1 shows that for this distribution, $\mathbf{P}(Y\langle \theta^*, X\rangle < 0) \leqslant 1/B^2$. It then follows from Fact 4.1 that if $n \leqslant B^2 t/2$,

$$\mathbf{P}(\text{MLE exists}) \leqslant e^{-t}\,.$$

The conclusion follows from the same argument as in the proof of Theorem 6.1 in Chapter 3, see (3.51) and after.

We now turn to the optimality of our bound on the excess risk (6.5). It is known from asymptotic theory (see e.g. [vdV98, Example 5.25 p. 55]) that in the misspecified case,

$$\sqrt{n} H(\theta^*)^{1/2}(\widehat{\theta}_n - \theta^*) \xrightarrow[n\to\infty]{\text{(d)}} \mathsf{N}(0, \Gamma)\,, \quad \Gamma = H(\theta^*)^{-1/2} G H(\theta^*)^{-1/2}\,,$$

where $H(\theta^*) = \nabla^2 L(\theta^*)$ is the population Hessian and $G = \mathbf{E}[\nabla\ell(\theta^*)\nabla\ell(\theta^*)^\top]$ 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 (6.6)). Indeed, by Lemma 3.11, $\Gamma \succcurlyeq C_1^{-1} H^{-1/2} G H^{-1/2}$, with $C_1 = 2\sqrt{2/\pi}$, so by the second point of Lemma 6.1, 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 $c_1 d$[2]; and on the other hand, if $v \in S^{d-1}$ denotes an eigenvector of Γ associated to its largest eigenvalue,

$$\mathbf{P}(\|\xi\|^2 \geqslant c_1 \|\Gamma\|_{\mathrm{op}} t) \geqslant \mathbf{P}(\langle v, \xi\rangle^2 \geqslant c_1 \|\Gamma\|_{\mathrm{op}} t) \geqslant e^{-t}\,.$$

This concludes the proof of Theorem 6.2.

Appendix 6.B: Remaining proofs and additional results

6.B.1 Proof of Lemma 6.1

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

$$L(Q\theta) = L(\theta)\,. \tag{6.36}$$

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

$$\theta^* = \|\theta^*\| u^*\,. \tag{6.37}$$

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

$$\begin{aligned}
L(-\|\theta^*\|u^*) &= \mathbf{E}[\phi(Y\|\theta^*\|\langle u^*, X\rangle)] \\
&= (1-p)\mathbf{E}[\phi(\|\theta^*\||\langle u^*, X\rangle|)] + p\mathbf{E}[\phi(-\|\theta^*\||\langle u^*, X\rangle|)] \\
&> p\mathbf{E}[\phi(\|\theta^*\||\langle u^*, X\rangle|)] + (1-p)\mathbf{E}[\phi(-\|\theta^*\||\langle u^*, X\rangle|)] \\
&= \mathbf{E}[\phi(-Y\|\theta^*\|\langle u^*, X\rangle)] \\
&= L(\|\theta^*\|u^*)\,.
\end{aligned}$$

This proves (6.37).

It remains to show that $B = \|\theta^*\| \geqslant e$ and that $B \asymp p^{-1/2}$. In view of (6.6), $\mathbf{1}(Y\langle\theta^*, X\rangle < 0)$ does not depend on X. Hence (6.17) rewrites

$$p\mathbf{E}|\langle u^*, X\rangle| = \mathbf{E}\big[|\langle u^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big]\,.$$

In addition, since $\mathbf{E}|\langle u^*, X\rangle| = \sqrt{2/\pi}$, one has

$$p = \sqrt{\frac{\pi}{2}}\,\mathbf{E}\big[|\langle u^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big]\,. \tag{6.38}$$

[2] one way to see this is that, if $Z \sim \chi^2(d)$ and $\mathbf{M}Z$ denotes a median of Z, then $|\mathbf{E}Z - \mathbf{M}Z| \leqslant \sqrt{\mathrm{Var}(Z)}$ which rewrites $|\mathbf{M}Z - d| \leqslant \sqrt{2d}$, showing that $\mathbf{M}Z \asymp d$ when $d \gg 1$

On one hand, $\sigma(-t) \geqslant e^{-t}/2$ for all $t \geqslant 0$. Using that $\|\theta^*\| \leqslant B$, it follows from Lemma 6.9 that

$$\mathbf{E}\big[|\langle u^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big] \geqslant \frac{1}{2}\mathbf{E}\big[|\langle u^*, X\rangle| \exp(-B|\langle u^*, X\rangle|)\big] \geqslant \frac{1}{\sqrt{2\pi}B^2}\,. \tag{6.39}$$

Hence, using (6.38) 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 (6.39) and (6.38) and we have

$$B = \|\theta^*\| \geqslant e \quad \text{and} \quad p \geqslant \frac{1}{2B^2}. \tag{6.40}$$

We now prove the upper bound of the first point, which is a consequence of the exponential moment bound (3.11), since $B = \|\theta^*\| \geqslant e$. Using that $\sigma(-t) \leqslant e^{-t}$ for all $t \geqslant 0$, we deduce

$$\mathbf{E}\big[|\langle u^*, X\rangle|\sigma(-|\langle\theta^*, X\rangle|)\big] \leqslant \sqrt{\frac{2}{\pi}} \cdot \frac{1}{B^2}\,.$$

We plug this in (6.38) 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 $t \geqslant 0$,

$$\begin{aligned}\langle Gu^*, u^*\rangle &= \mathbf{E}\langle u^*, \nabla\ell(\theta^*, Z)\rangle^2 = \mathbf{E}\big[\sigma(-Y\langle\theta^*, X\rangle)^2\langle u^*, X\rangle^2\big] \\ &\geqslant \frac{1}{4}\mathbf{E}\big[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\langle u^*, X\rangle^2\big]\,.\end{aligned}$$

The distribution of $Y|X$ is designed so that $\mathbf{E}[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\,|X]$ is actually not a function of X, but constant and equal to p. More precisely,

$$\begin{aligned}\mathbf{E}\big[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\langle u^*, X\rangle^2\big] &= \mathbf{E}\big[\mathbf{E}\big[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\langle u^*, X\rangle^2\,|X\big]\big] \\ &= \mathbf{E}\big[\langle u^*, X\rangle^2\mathbf{E}\big[\mathbf{1}(Y\langle\theta^*, X\rangle < 0)\,|X\big]\big] \\ &= p\,\mathbf{E}\langle u^*, X\rangle^2 = p\,.\end{aligned}$$

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 $\langle H^{-1/2}GH^{-1/2}u^*, u^*\rangle \geqslant B/8$. This concludes the proof of Lemma 6.1.

6.B.2 Additional results

The following result shows that the population Hessians $H(\theta)$ are within a constant factor of H when θ ranges in an ellipsoid centered at θ^*.

Lemma 6.8. *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 5.1 with parameter $K > 0$ and 5.2 with parameters $\eta = 1/B$ and $c \geqslant 1$. Then, there exists c' depending on c and K such that*

$$\frac{1}{c'}H \preccurlyeq H(\theta) \preccurlyeq c'\log(B)^2 H\,.$$

Proof. In this proof we use the family of proxies H_θ for the true Hessians $H(\theta)$. These proxies were defined in Chapter 3, Equation (3.22).

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

$$\begin{aligned}\langle H_\theta^{-1/2} H(\theta) H_\theta^{-1/2} v, v\rangle &= B^3 \langle u, v\rangle^2 \mathbf{E}[\exp(-|\langle \theta, X\rangle|)\langle u, X\rangle^2] \\ &\quad + B(1 - \langle u, v\rangle^2)\mathbf{E}[\exp(-|\langle \theta, X\rangle|)\langle w, X\rangle^2]\,.\end{aligned}$$

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 8.4, $\|\theta\| \geqslant (1 - r)B$. Thus, by Lemma 6.2, 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

$$H(\theta) \preccurlyeq \frac{3c}{1 - r}(K \log((1 + r)B))^2 H_\theta\,, \tag{6.41}$$

from which the result follows in the case $B > e$, since by Lemma 3.8 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

$$\langle H(\theta)v, v\rangle = \mathbf{E}[\sigma'(\langle \theta, X\rangle)\langle v, X\rangle^2]\,.$$

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

$$\langle H(\theta)v, v\rangle \geqslant \sigma'(m(1 + r)B)M^2 \mathbf{P}\big(|\langle u, X\rangle| \leqslant m,\ |\langle v, X\rangle| \geqslant M\big)\,, \tag{6.42}$$

where we also used that, as $\|\theta - \theta^*\|_H \leqslant r/\sqrt{B}$, $\|\theta\| \leqslant (1 + r)B$ by Lemma 8.4.

If $\|\theta^*\| \leqslant e$, $B = e$, so Proposition 5.2 shows that Assumptions 5.2 holds with $c = e$ and Assumption 5.3 is satisfied with constant $\max\{2eK \log(2K), 2K^4\} = 2K^4$. Therefore,

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

Hence, choosing $m = 2K^4/B$ and $M = \max\{1/B, \|u^* - v\|\}/2K^4$ in (6.42), 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 8.4 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 6.7, this implies that for all $\theta \in \Theta$ and $v \in S^{d-1}$, one has for all $t \geqslant 1/B$

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

Hence, choosing $m = (c+1)/B$, $M = \max(1/B, \|u^* - v\|)/(c+1)$ in (6.23), we get that

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

■

Lemma 6.9. *Let $N \sim \mathsf{N}(0,1)$ and $B \geqslant e$. Then*

$$\mathbf{E}\big[|N| \exp(-B|N|)\big] \geqslant \frac{1}{\sqrt{2\pi} B^2}\,.$$

Proof. First, by symmetry,

$$\mathbf{E}\big[|N| \exp(-B|N|)\big] = \sqrt{\frac{2}{\pi}} \int_0^{+\infty} t e^{-Bt} e^{-t^2/2} \mathrm{d}t\,. \tag{6.43}$$

Then we proceed with an expansion of the integral.

$$\begin{aligned}\int_0^{+\infty} t e^{-Bt} e^{-t^2/2} \mathrm{d}t &= e^{B^2/2} \int_0^{+\infty} t e^{-\frac{(t+B)^2}{2}} \mathrm{d}t = e^{B^2/2} \int_B^{+\infty} (x-B) e^{-x^2/2} \mathrm{d}x \\ &= e^{B^2/2} \Big(\int_B^{+\infty} x e^{-x^2/2} \mathrm{d}x - B \int_B^{+\infty} e^{-x^2/2} \mathrm{d}x \Big) \\ &= 1 - B e^{B^2/2} \int_B^{+\infty} e^{-t^2/2} \mathrm{d}t\,. \end{aligned}$$

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

$$\begin{aligned} e^{B^2/2} \int_B^{+\infty} e^{-t^2/2} \mathrm{d}t &= \frac{1}{B} - \frac{1}{B^3} + \frac{3}{B^5} - 15 \int_B^{+\infty} \frac{1}{t^6} e^{-t^2/2} \mathrm{d}t \\ &\leqslant \frac{1}{B} - \frac{1}{B^3} + \frac{3}{B^5}\,. \end{aligned}$$

Then

$$\int_0^{+\infty} t e^{-Bt} e^{-t^2/2} \mathrm{d}t = 1 - B e^{B^2/2} \int_B^{+\infty} e^{-t^2/2} \mathrm{d}t \geqslant \frac{1}{B^2} - \frac{3}{B^4} = \frac{1}{B^2}\Big(1 - \frac{3}{B^2}\Big)\,.$$

Finally, as $B \geqslant e$, one has $3/B^2 \leqslant 3/e^2 \leqslant 1/2$; and combining with (6.43) proves the claim. ■

Chapter 7

Fast rates in classification

Abstract

The logistic loss is a classical convex surrogate for the binary loss in supervised classification. It is of practical interest as it leads to a smooth and convex optimization problem. A classical result relating the convex surrogate risk to the binary one is Zhang's lemma, that applies in particular to the logistic loss. In doing so, it degrades the rate of convergence by a power of 1/2, making it impossible to obtain a fast classification rate. Fast rates, that is, a bound on the excess risk that converges to 0 at a rate faster than $1/\sqrt{n}$ are possible under the *margin* condition, which is satisfied in the Gaussian case, and closely related to our definition of regular designs. In this chapter, we prove that the classification risk of the (plug-in of) the MLE enjoys the fast rate d/n. We extend this to the non-Gaussian case with only a logarithmic degradation, achieving the rate $\frac{d}{n}\log(\frac{n}{d})$.

Contents

7.1 Introduction

7.1.1 Logistic loss as a convex surrogate for classification

Logistic regression is a natural way to estimate the conditional distribution of a binary outcome $Y = \pm 1$ on covariates $X \in \mathbb{R}^d$. Indeed, the logarithmic loss enforces calibrated predictions, as it penalizes both overconfident and under-confident predictions. In the previous chapters (Chapters 3 and 6), we provided sharp guarantees on the performance of the MLE in terms of logarithmic loss, that is, for the problem of probability assignment.

The logistic loss also naturally arises in statistical learning as a convex surrogate for the binary loss in supervised classification [BJM06, Zha04]. Indeed, suppose we want to find a good binary linear classifier, that is, a vector $\theta \in \mathbb{R}^d$ with small classification risk

$$R(\theta) = \mathbf{P}(Y\langle \theta, X\rangle < 0)\,. \tag{7.1}$$

A natural way to do this is to perform empirical risk minimization over a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of i.i.d. pairs in $\mathbb{R}^d \times \{-1, 1\}$. The supervised classification problem with the linear class is thus formulated as

$$\widehat{\theta}_n^{0/1} \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Y_i\langle \theta, X_i\rangle < 0)\,.$$

This problem is not convex, as $\phi_{0/1} : t \mapsto \mathbf{1}(t < 0)$ is not a convex function. As such, it is numerically intractable. To circumvent this issue, one can replace $\phi_{0/1}$ by a convex surrogate ϕ. A common choice is the logistic surrogate $\phi(t) = \log(1 + e^t)$ which leads to the maximum-likelihood estimator (MLE)

$$\widehat{\theta}_n = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \log(1 + e^{-Y_i\langle \theta, X_i\rangle})\,, \tag{7.2}$$

that we extensively studied so far in terms of its performance for probabilistic forecasts.

In this chapter, we are concerned with the classification performance of the MLE as a plug-in linear classifier, and want to bound from above its binary risk $\mathbf{P}(Y\langle \widehat{\theta}_n, X\rangle < 0)$, more precisely its excess risk

$$R(\widehat{\theta}_n) - R^*$$

where $R^* = \inf_{f:\mathbb{R}^d\to\mathbb{R}} \mathbf{P}(Yf(X) < 0)$ is the Bayes risk.

7.1.2 Existing results

Zhang's lemma: slow rate. A seminal result due to Zhang [Zha04] shows that under suitable conditions on the surrogate function ϕ—that the logistic loss satisfies—the excess classification risk is controlled by the excess ϕ-risk. More precisely, Zhang's lemma applied to the logistic loss can be stated as follows. Suppose for clarity that the model is well-specified with parameter θ^*. Then there exists a numerical constant C such that for all θ,

$$R(\theta) - R(\theta^*) \leqslant C\sqrt{L(\theta) - L(\theta^*)}\,, \tag{7.3}$$

where as before, $L(\theta) = \mathbf{E}[\log(1 + \exp(-Y\langle \theta, X\rangle))]$ is the logistic risk. In particular, using previous non-asymptotic results for the logistic excess risk of the MLE, e.g., (1.14)

from [OB21], this yields a classification risk scaling as $\sqrt{d/n}$, which is the typical slow rate provided by Vapnik-Chervonenkis theory for the empirical risk minimizer with respect to the binary loss.

In short, (7.3) ensures that the MLE $\widehat{\theta}_n$ is consistent when used as a plug-in for binary classification, but in terms of quantitative bounds, it significantly degrades the rate of convergence to 0.

We emphasize that compared to the previous section where we discussed the sharpness of the results with respect to the signal strength B, the issue here has to do with the very rate of convergence with respect to the sample size n. In contrast, the results mentioned in the introduction regarding the logistic excess risk (Section 1.2.3), the dependence with respect to the dimension d and the sample size n were sharp (up to logarithmic terms sometimes), and only the dependence with respect to the signal strength left space for improvement.

When using the logistic loss as a convex surrogate for the 0/1 loss, one needs to ensure some guarantees for the statistical performance of the resulting predictor for the specific problem of binary classification.

Our goal here is to improve over the general result obtained from Zhang's lemma [Zha04] in the specific case of logistic regression.

Faster rates under margin and Bernstein conditions. Recall that in the problem of binary classification, the ultimate oracle, called Bayes predictor, admits a closed form in terms of the *regression function* $\zeta(x) = \mathbf{P}(Y = 1 \,|X = x)$, as

$$f_{\text{Bayes}}(X) = 2\zeta(X) - 1\,. \tag{7.4}$$

In short, given the knowledge of the full joint distribution of (X, Y), the Bayes predictor returns the most likely label for each point X. Therefore, if the distribution of (X, Y) is such that the variable $\zeta(X)$ concentrates around $1/2$, the classification problem is intrinsically harder.

Margin (or noise) conditions prevent this phenomenon and quantify how the regression function $\zeta(x) = \mathbf{P}(Y = 1 \,|X = x)$ behaves in the neighborhood of the decision boundary $\{x : \zeta(x) = 1/2\}$. Indeed, for the points x close to this region, x does not provide much information about the outcome Y. The Mammen-Tsybakov condition [MT99, Tsy04] provides a quantitative bound on the mass of the neighborhood of the decision boundary.

Definition 7.1 (Tsybakov soft margin condition)**.** The pair (X, Y) satisfies the Tsybakov noise condition if there exist $A, \alpha > 0$ and $\varepsilon_0 \in (0, 1/2]$ such that for all $\varepsilon \in (0, \varepsilon_0]$,

$$\mathbf{P}(|\zeta(X) - 1/2| \leqslant \varepsilon) \leqslant A\varepsilon^{\alpha}\,. \tag{7.5}$$

In the introduction (Section 1.8), we alluded to the so-called Bernstein condition, which allows for fast rates. The Bernstein condition is a bound on the variance of the loss in terms of the excess risk.

Definition 7.2 (Bernstein condition)**.** The class of predictors $\mathcal{F}$, the loss function ℓ and the distribution P (that of (X, Y)) satisfy the (A, β)-Bernstein condition (with $A > 0$, $\beta \in [0, 1]$) if for all $f \in \mathcal{F}$,

$$\mathbf{E}(\ell(f, Z) - \ell(f^*, Z))^2 \leqslant A\big[L(f) - L(f^*)\big]^{\beta}\,. \tag{7.6}$$

The two important facts are the following: (i) Bernstein's condition (7.6) implies that there are predictors with rates of convergence faster than $n^{-1/2}$, namely $(d/n)^{1/(2-\beta)}$ and (ii) the margin condition with exponent α (7.5) implies the Bernstein condition (7.6) with exponent $\beta = \alpha/(\alpha+1)$. In short, if the margin condition is satisfied with exponent α, that is

$$\mathbf{P}(|\zeta(X) - 1/2| \leqslant \varepsilon) \leqslant A\varepsilon^{\alpha}\,, \tag{7.7}$$

then one can expect to find a predictor $\widehat{f}_n$ with excess risk

$$\mathbf{P}(Y\widehat{f}_n(X) < 0) - R^* \leqslant C\Big(\frac{d}{n}\Big)^{\frac{\alpha+1}{\alpha+2}}\,. \tag{7.8}$$

Consider now a well-specified logit model with parameter $\theta^* = Bu^*$, $u^* \in S^{d-1}$. Then by definition the regression function is

$$\zeta(X) = \sigma(\langle\theta^*, X\rangle) = \frac{1}{1+\exp(-\langle\theta^*, X\rangle)}\,. \tag{7.9}$$

For the sake of clarity, suppose in addition that $X \sim \mathsf{N}(0, I_d)$. Then, we write $\langle\theta^*, X\rangle = BU$ with $U \sim \mathsf{N}(0,1)$. It follows that for all $\varepsilon \in (0, 1/4]$,

$$\mathbf{P}(|\zeta(X) - 1/2| \leqslant \varepsilon) = \mathbf{P}(|\sigma(BU) - 1/2| \leqslant \varepsilon) = \mathbf{P}\Big(|U| \leqslant \frac{\tau(\varepsilon)}{B}\Big) \lesssim \frac{\varepsilon}{B}\,,$$

where

$$\tau(\varepsilon) = \sigma^{-1}(1/2+\varepsilon) = \log\Big(\frac{1+2\varepsilon}{1-2\varepsilon}\Big) \underset{\varepsilon\to 0}{=} O(\varepsilon)\,.$$

This means that the well-specified logistic model with Gaussian design and signal strength B satisfies the margin condition (7.5) with exponent 1 and constant B. Therefore, one can expect to find a linear classifier $\widetilde{\theta}_n$ with excess risk

$$R(\widetilde{\theta}_n) - R(\theta^*) \leqslant C\Big(\frac{d}{n}\Big)^{2/3}\,, \tag{7.10}$$

either in expectation or with high probability.

Let us now summarize the situation regarding the Gaussian, well-specified model. On one hand, Zhang's lemma (7.3) applied to the logistic loss and combined with the logistic risk bound for the MLE from Theorem 3.1 shows that its classification performance satisfies

$$R(\widehat{\theta}_n) - R(\theta^*) \leqslant C\sqrt{\frac{d}{n}}\,. \tag{7.11}$$

On the other hand, this model satisfies Tsybakov's margin condition (7.5) with exponent 1, suggesting a better classification rate

$$R(\widehat{\theta}^{0/1}_n) - R(\theta^*) \leqslant C\Big(\frac{d}{n}\Big)^{2/3}\,,$$

achieved by the ERM with respect to the binary loss for this problem.

It turns out that not only a better rate is possible, but is actually achieved by the MLE. Indeed, Theorem 7.1 in the next section shows that the MLE achieves a classification rate of d/n. We then generalize it to the regular case in Theorem 7.2. Both of these results derive from an improvement of Zhang's lemma using the margin assumption, specifically for the linear class. This is done in Proposition 7.1 for the Gaussian case and in Proposition 7.2 for the regular case. The starting point of their proofs is the following classical representation formula for the excess risk in binary classification, see, e.g., [DGL96, Theorem 2.2].

Lemma 7.1. *Let ζ denote the regression function. For every predictor $f : \mathbb{R}^d \to \{-1, 1\}$, one has*

$$R(f) - R(f^*) = \mathbf{E}\big[|2\zeta(X) - 1|\mathbf{1}\big(f(X)f_{\mathrm{Bayes}}(X) < 0\big)\big]\,.$$

7.2 Sharp rates in the logistic model with Gaussian design

Throughout this section, we always assume that the model is well-specified, meaning that (7.9) holds for a parameter $\theta^* \in \mathbb{R}^d$, and we further assume that $\|\theta^*\| \geqslant e$.

Theorem 7.1 below shows that the MLE, when it exists, is also an optimal linear classifier (up to absolute constant) for the binary loss.

Theorem 7.1. *Grant the assumptions of Theorem 1.1. Then with probability at least $1 - \delta$ the MLE $\widehat{\theta}_n$ exists and satisfies the classification risk bound*

$$R(\widehat{\theta}_n) - R(\theta^*) \leqslant C\frac{d + \log(1/\delta)}{n}\,. \tag{7.12}$$

Theorem 7.1 is a direct consequence of Theorem 3.1 and of Proposition 7.1 below. The latter can be seen as an improvement of Zhang's lemma for the linear class of classifiers, and relates the excess risk in classification to the Euclidean distance between corresponding directions.

Proposition 7.1. *Let $X \sim \mathsf{N}(0, I_d)$ and $u^* = \theta^*/\|\theta^*\|$. Then for any $u \in S^{d-1}$,*

$$R(u) - R(u^*) \leqslant 3.2\|\theta^*\| \cdot \|u - u^*\|^2\,. \tag{7.13}$$

The starting point of the proof of Proposition 7.1 (as well as that of Proposition 7.2 in the next section) is the following classical representation formula for the excess risk in binary classification.

Proof. Recall the notation $R(u)$ that denotes the $0-1$ risk of a *direction* $u \in S^{d-1}$, that is $R(u) = \mathbf{P}(Y\langle u, X\rangle < 0)$. Since the model is well-specified, one has $\zeta(X) = \mathbf{P}(Y = 1|X) = \sigma(\langle\theta^*, X\rangle)$. Let also $\beta = \|\theta^*\|$. Let us sketch the argument before providing the details. We write, using Lemma 7.1 and the fact that by symmetry $|\sigma(t) - 1/2| = \sigma(|t|) - 1/2$,

$$\begin{aligned} R(u) - R(u^*) &= 2\mathbf{E}\big[\big(\sigma(\beta|\langle u^*, X\rangle|) - 1/2\big)\mathbf{1}(\langle u, X\rangle\langle u^*, X\rangle < 0)\big] \\ &\leqslant 2\beta\mathbf{E}\big[|\langle u^*, X\rangle|\mathbf{1}(\langle u, X\rangle\langle u^*, X\rangle < 0)\big] \\ &\leqslant C\beta\mathbf{E}\Big[\exp\Big(-\frac{\langle u^*, X\rangle^2}{C'\|u - u^*\|^2}\Big)|\langle u^*, X\rangle|\Big] \\ &\leqslant C\beta\|u - u^*\|^2\,. \end{aligned} \tag{7.14}$$

It remains to justify (7.14). Conditioning on $\langle u^*, X\rangle$ we get

$$\mathbf{E}\big[|\langle u^*, X\rangle|\mathbf{1}(\langle u, X\rangle\langle u^*, X\rangle < 0)\big] = \mathbf{E}\big[\mathbf{P}\big(\langle u, X\rangle\langle u^*, X\rangle < 0\big|\langle u^*, X\rangle\big) \cdot |\langle u^*, X\rangle|\big]$$

We now decompose u as $u = \langle u, u^*\rangle u^* + \sqrt{1 - \langle u, u^*\rangle^2} w$ with $\langle u^*, w\rangle = 0$. Hence

$$\begin{aligned}
&\mathbf{P}\big(\langle u, X\rangle\langle u^*, X\rangle < 0 \big| \langle u^*, X\rangle\big) \\
&\quad = \mathbf{P}\big(\mathrm{sign}(\langle u^*, X\rangle) \neq \mathrm{sign}(\langle u, u^*\rangle\langle u^*, X\rangle + \sqrt{1 - \langle u, u^*\rangle^2}\langle w, X\rangle) | \langle u^*, X\rangle\big) \\
&\quad \leqslant \mathbf{P}\big(\sqrt{1 - \langle u, u^*\rangle^2}|\langle w, X\rangle| > |\langle u, u^*\rangle\langle u^*, X\rangle| \,\big|\, \langle u^*, X\rangle\big) \\
&\quad = 2Q\left(\frac{|\langle u, u^*\rangle\langle u^*, X\rangle|}{\sqrt{1 - \langle u, u^*\rangle^2}}\right), \qquad (7.15)
\end{aligned}$$

Where $Q(t) = \mathbf{P}(Z > t)$, $Z \sim \mathsf{N}(0,1)$ is the tail of the standard Gaussian distribution. The last equality holds because $\langle w, X\rangle$ is independent of $\langle u^*, X\rangle$. Using that $|\langle u, u^*\rangle| \geqslant 1/2$, $\sqrt{1 - \langle u, u^*\rangle^2} \leqslant \|u - u^*\|$, and the classical inequality $Q(t) \leqslant e^{-t^2/2}/2$ which holds for all $t \geqslant 0$, we deduce that

$$\mathbf{P}\big(\langle u, X\rangle\langle u^*, X\rangle < 0 \big| \langle u^*, X\rangle\big) \leqslant \exp\left(-\frac{\langle u^*, X\rangle^2}{4\|u - u^*\|^2}\right).$$

Now we integrate over $\langle u^*, X\rangle$ and use the fact that if $Z \sim \mathsf{N}(0,1)$, then for all $\alpha > 0$,

$$\begin{aligned}
\mathbf{E}\big[|Z| \exp(-(Z/\alpha)^2)\big] &= \int_{\mathbb{R}} |t| e^{-t^2/\alpha^2} (2\pi)^{-1/2} e^{-t^2/2} \mathrm{d}t \\
&= (2\pi)^{-1/2} \int_{\mathbb{R}} |t| \exp\Big(-\Big(\frac{1}{2} + \frac{1}{\alpha^2}\Big)t^2\Big)\mathrm{d}t \\
&= \sigma \int_{\mathbb{R}} |t| \exp\Big(-\frac{t^2}{2\sigma^2}\Big)(2\pi\sigma^2)^{-1/2}\mathrm{d}t\,, \quad \sigma^2 = \frac{\alpha^2}{2+\alpha^2} \\
&= \sigma^2 \mathbf{E}|Z|\,.
\end{aligned}$$

Then by standard computations, $\mathbf{E}|Z| = \sqrt{2/\pi}$. Using this identity with $\alpha = 2\|u - u^*\|$ and combining everything, we deduce that

$$\begin{aligned}
R(u) - R(u^*) &\leqslant 2\beta \mathbf{E}\left[|\langle u^*, X\rangle| \exp\left(-\frac{\langle u^*, X\rangle^2}{4\|u - u^*\|^2}\right)\right] \\
&\leqslant 2\beta \frac{\alpha^2}{2+\alpha^2}\sqrt{\frac{2}{\pi}} \leqslant 4\sqrt{\frac{2}{\pi}}\,\beta\|u - u^*\|^2\,. \qquad \blacksquare
\end{aligned}$$

7.3 Near-optimal rates for non-Gaussian designs

The results from the previous section were specific to the Gaussian design setting. In this section, we extend the previous results to the case where X is not Gaussian. More precisely, we assume that X is isotropic ($\mathbf{E}XX^\top = I_d$) and that it satisfies Assumption 5.1 from Chapter 5, with $K \geqslant e$. We further assume that X satisfies a standard margin assumption, which is a slight strengthening of Assumption 5.2.

Assumption 7.1. There exists $c > 0$ such that for all $t > 0$, $\mathbf{P}(|\langle u^*, X\rangle| \leqslant t) \leqslant ct$.

Note that in Chapter 6, this assumption was only needed for t bounded away from 0, namely $t \geqslant 1/B$. Assumption 7.1 above is closer in spirit to the standard margin condition from Definition 7.1. Assumption 7.1 is naturally satisfied *in all directions* when

X is Gaussian, as for all $u \in S^{d-1}$, $\langle u, X\rangle \sim \mathsf{N}(0,1)$, and thus admits an upper-bounded density.

Theorem 7.2 below extends Theorem 7.1 to non-Gaussian, but regular designs, with the strengthened one-dimensional margin assumption discussed above. It is a direct consequence of Proposition 7.2 and Theorem 6.1.

Theorem 7.2. *Let X be a K-sub-exponential, isotropic random vector satisfying Assumption 1.5 with constant c and Assumption 1.3 with constant c and scale $1/B$. There are some constants $c_1, c_2 > 0$, depending only on c and K such that for every $\delta \in (0,1)$, if $n \geqslant c_1 B \log^4(B)(d + \log(1/\delta))$, then with probability at least $1-\delta$, the MLE $\widehat{\theta}_n$ exists and satisfies*

$$R(\widehat{\theta}_n) - R(\theta^*) \leqslant c_2 K^2 \frac{d + \log(1/\delta)}{n} \log^2\left(\frac{Bn}{d}\right).$$

Proposition 7.2. *Let X be isotropic, K-sub-exponential and satisfying Assumption 7.1 with constant c. Then, for all $u \in S^{d-1}$ such that $\|u - u^*\| \leqslant 1/2$,*

$$R(u) - R(u^*) \leqslant 10 c K^2 \|\theta^*\| \cdot \|u - u^*\|^2 \log^2\left(\frac{1}{\|u - u^*\|}\right).$$

Proof. We let $\beta = \|\theta^*\|$ as before and start again from the fact that

$$R(u) - R(u^*) \leqslant 2\beta \mathbf{E}\big[|\langle u^*, X\rangle| \mathbf{1}(\langle u, X\rangle \langle u^*, X\rangle < 0)\big]. \tag{7.16}$$

In addition, as $\langle u, X\rangle = \langle u^*, X\rangle + \langle u - u^*, X\rangle$, if $\langle u, X\rangle$ and $\langle u^*, X\rangle$ have opposite sign, then necessarily $|\langle u^*, X\rangle| \leqslant |\langle u - u^*, X\rangle|$. Hence,

$$\mathbf{E}\big[|\langle u^*, X\rangle| \mathbf{1}(\langle u, X\rangle \langle u^*, X\rangle < 0)\big] \leqslant \mathbf{E}\big[|\langle u - u^*, X\rangle| \mathbf{1}(|\langle u^*, X\rangle| \leqslant |\langle u - u^*, X\rangle|)\big]. \tag{7.17}$$

Let $\lambda > 0$ be a parameter to tune later. We condition on the event

$$E = \big\{|\langle u - u^*, X\rangle| \leqslant \lambda \|u - u^*\|\big\}, \tag{7.18}$$

which for λ large enough has high probability. We write

$$\begin{aligned}
&\mathbf{E}\big[|\langle u - u^*, X\rangle| \cdot \mathbf{1}(|\langle u^*, X\rangle| \leqslant |\langle u - u^*, X\rangle|)\big] \\
&\quad = \mathbf{E}\big[|\langle u - u^*, X\rangle| \cdot \mathbf{1}(|\langle u^*, X\rangle| \leqslant |\langle u - u^*, X\rangle|) \cdot \mathbf{1}(|\langle u - u^*, X\rangle| \leqslant \lambda \|u - u^*\|)\big] \\
&\quad + \mathbf{E}\big[|\langle u - u^*, X\rangle| \cdot \mathbf{1}(|\langle u^*, X\rangle| \leqslant |\langle u - u^*, X\rangle|) \cdot \mathbf{1}(|\langle u - u^*, X\rangle| > \lambda \|u - u^*\|)\big]
\end{aligned} \tag{7.19}$$

For the first expectation in the right hand side above, we have

$$\begin{aligned}
&\mathbf{E}\big[|\langle u - u^*, X\rangle| \mathbf{1}(|\langle u^*, X\rangle| \leqslant |\langle u - u^*, X\rangle|) \mathbf{1}(|\langle u - u^*, X\rangle| \leqslant \lambda \|u - u^*\|)\big] \\
&\qquad \leqslant \lambda \|u - u^*\| \mathbf{P}(|\langle u^*, X\rangle| \leqslant \lambda \|u - u^*\|) \\
&\qquad \leqslant c \lambda^2 \|u - u^*\|^2.
\end{aligned} \tag{7.20}$$

On the complement of the event E (7.18), we drop the factor $\mathbf{1}(|\langle u^*, X\rangle| \leqslant |\langle u - u^*, X\rangle|)$ and only keep $\mathbf{1}_{E^c}$

$$\begin{aligned}
&\mathbf{E}\big[|\langle u - u^*, X\rangle| \mathbf{1}(|\langle u^*, X\rangle| \leqslant |\langle u - u^*, X\rangle|) \mathbf{1}(|\langle u - u^*, X\rangle| > \lambda \|u - u^*\|)\big] \\
&\qquad \leqslant \mathbf{E}\big[|\langle u - u^*, X\rangle| \mathbf{1}(|\langle u - u^*, X\rangle| \geqslant \lambda \|u - u^*\|)\big] \\
&\qquad = \|u - u^*\| \mathbf{E}\big[|\langle w, X\rangle| \mathbf{1}(|\langle w, X\rangle| \geqslant \lambda)\big],
\end{aligned}$$

where we let $w = (u - u^*)/\|u - u^*\|$. By Cauchy-Schwarz inequality and since $\langle w, X\rangle$ is K-sub-exponential,

$$\mathbf{E}\big[|\langle w, X\rangle| \mathbf{1}(|\langle w, X\rangle| \geqslant \lambda)\big] \leqslant \sqrt{\mathbf{E}\langle w, X\rangle^2}\sqrt{\mathbf{P}(|\langle w, X\rangle| \geqslant \lambda)} \leqslant \exp(-\lambda/(2K))$$

We then let $\lambda = 2K \log(1/\|u - u^*\|)$, so that

$$\mathbf{E}\big[|\langle u - u^*, X\rangle| \mathbf{1}(|\langle u^*, X\rangle| \leqslant |\langle u - u^*, X\rangle|)\big] \leqslant \|u - u^*\|^2 .$$

Combining with (7.20), we obtain

$$\mathbf{E}\big[|\langle u - u^*, X\rangle| \mathbf{1}(|\langle u^*, X\rangle| \leqslant |\langle u - u^*, X\rangle|)\big] \leqslant \|u - u^*\|^2 \left[1 + 4cK^2 \log^2\left(\frac{1}{\|u - u^*\|}\right)\right].$$

We then plug this in (7.17) and finally, in (7.16). The claim follows. ∎

7.4 Proofs of Theorems 7.1 and 7.2

Our results on the performance of the MLE in the previous chapters were stated in terms of bounds on its excess logistic risk. These derived from the localization argument of Lemma 2.1, which allowed us to bound the distance to θ^* with respect to the norm induced by the Fisher information H. Given the structure of H, this also gives a bound on the estimation error of the norm, and more importantly for the matter at hand here, a bound on the estimation error of the direction of θ^*. This follows from the third point of Lemma 8.4, reproduced below for completeness.

Lemma 7.2 (Chapter 3, Lemma 8.4)**.** *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 $u = \theta/\|\theta\|$,*

$$\|u - u^*\| \leqslant \frac{\sqrt{2}r}{(1-r)B} .$$

Proof of Theorem 7.1 By Theorem 3.1 for all $\delta \in (0, 1)$, if $n \geqslant CB(d + \log(1/\delta))$, then with probability at least $1 - \delta$, the MLE $\widehat{\theta}_n$ exists and is localized as

$$\|\widehat{\theta}_n - \theta^*\|_H \leqslant C\sqrt{\frac{d + \log(1/\delta)}{n}} .$$

By Lemma 7.2, this shows that the direction of θ^* is estimated accurately in Euclidean norm, in the sense that on the same event,

$$\|\widehat{u}_n - u^*\| \leqslant C'\sqrt{\frac{d + \log(1/\delta)}{Bn}} . \tag{7.21}$$

Therefore, on the event where the MLE is localized, and hence (7.21) holds, we deduce from Proposition 7.1 that

$$R(u) - R(u^*) \leqslant C\|\theta^*\|\frac{d + \log(1/\delta)}{Bn} \leqslant C\,\frac{d + \log(1/\delta)}{n} .$$

Proof of Theorem 7.2. By Theorem 6.1, there exist some constants $c_1, c_2 > 0$, depending only on c and K such that for every $\delta \in (0,1)$, if $n \geqslant c_1 B \log^4(B)(d + \log(1/\delta))$, then with probability at least $1 - \delta$, the MLE $\widehat{\theta}_n$ exists and

$$\|\widehat{\theta}_n - \theta^*\|_H \leqslant c_2 \log(B) \sqrt{\frac{d + \log(1/\delta)}{n}}\,. \tag{7.22}$$

Let $\widehat{u}_n = \widehat{\theta}_n / \|\widehat{\theta}_n\|$. By Lemma 7.2, on the same event,

$$\|\widehat{u}_n - u^*\| \leqslant c_3 \log(B) \sqrt{\frac{d + \log(1/\delta)}{Bn}}\,.$$

To conclude, the function $t \mapsto t \log(1/t)$ increases on $(0, 1/e]$. In addition, using that $B \leqslant n/(c_1 \log^4(B)(d + \log(1/\delta)))$, we further bound

$$\frac{Bn}{(d + \log(1/\delta)) \log^2(B)} \leqslant \frac{n^2}{c_1 (d + \log(1/\delta))^2 \log^6(B)} \leqslant \frac{1}{c_1} \Big(\frac{n}{d}\Big)^2\,,$$

which by Proposition 7.2 finishes the proof.

Chapter 8

Technical results

8.1 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 8.1 (ψ_α-norm)**.** Let $\alpha > 0$. If X is a real random variable, its *ψ_α-norm* is defined as

$$\|X\|_{\psi_\alpha} = \sup_{p \geqslant 2} \left\{ \frac{2^{1/\alpha} e \|X\|_p}{p^{1/\alpha}} \right\} \in [0, +\infty] \,, \tag{8.1}$$

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 (8.1) ensures that (i) if $\mathbf{E}[X^2] = 1$, 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 8.2 (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

$$\mathbf{E} \exp(\lambda X) \leqslant \exp\left(\frac{\sigma^2 \lambda^2}{2(1 - \lambda K)} \right) . \tag{8.2}$$

Recall that X is said to be centered if $\mathbf{E}[X] = 0$. The basic properties of sub-gamma and sub-exponential variables are gathered in the following lemma:

Lemma 8.1. *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*

$$\mathbf{P}\big(|X| \geqslant K t^{1/\alpha}\big) \leqslant e^{-2t} \,. \tag{8.3}$$

2. *If X is (σ^2, K)-sub-gamma, then for every $t \geqslant 0$, one has*

$$\mathbf{P}(X \geqslant \sigma\sqrt{2t} + Kt) \leqslant e^{-t} \,. \tag{8.4}$$

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*

$$\mathbf{E}[|X|^p] \leqslant \sigma^2 K^{p-2} p!/2\,, \tag{8.5}$$

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 8.1 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 3.1.

Proof. For the first point, for any $p \geqslant 2$, Markov's inequality implies that

$$\mathbf{P}(|X| \geqslant e\|X\|_p) = \mathbf{P}\big(|X|^p \geqslant e^p\|X\|_p^p\big) \leqslant \frac{\mathbf{E}[|X|^p]}{e^p\|X\|_p^p} = e^{-p}\,.$$

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 $\mathbf{P}(X \geqslant t) \leqslant e^{-\lambda t}\mathbf{E}e^{\lambda X}$ and optimizing over $\lambda \geqslant 0$.

The third point follows from the definition and the fact that, by independence,

$$\mathbf{E}\big[e^{\lambda(X_1+\cdots+X_n)}\big] = \mathbf{E}[e^{\lambda X_1}]\cdots\mathbf{E}[e^{\lambda X_n}]\,.$$

We now turn to the fourth point. For every $\lambda \in [0, 1/K)$, one has

$$\begin{aligned}\mathbf{E}e^{\lambda X} &\leqslant 1 + \lambda\mathbf{E}[X] + \sum_{p\geqslant 2}\frac{\lambda^p\,\mathbf{E}[|X|^p]}{p!} \leqslant 1 + \frac{\sigma^2\lambda^2}{2}\sum_{p\geqslant 2}\frac{\lambda^{p-2}K^{p-2}p!}{p!}\\ &= 1 + \frac{\sigma^2\lambda^2}{2(1-\lambda K)} \leqslant \exp\left(\frac{\sigma^2\lambda^2}{2(1-\lambda K)}\right).\end{aligned}$$

For the fifth point, we first note that, as $\mathbf{E}[e^{|X|/(2K)}] \leqslant \mathbf{E}[e^{X/(2K)}] + \mathbf{E}[e^{-X/(2K)}] < \infty$, by dominated convergence the function $\phi : \lambda \mapsto \log\mathbf{E}[e^{\lambda X}]$ is well-defined and twice continuously differentiable over $(-1/(2K), 1/(2K))$, with $\phi(0) = 0$, $\phi'(0) = \mathbf{E}[X]$ and $\phi''(0) = \mathrm{Var}(X)$. Hence, $\phi(\lambda) = \mathbf{E}[X]\lambda + \mathrm{Var}(X)\lambda^2/2 + o(\lambda^2)$ 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 $\mathbf{E}[X] = 0$ and $\mathrm{Var}(X) \leqslant \sigma^2$. Next, in order to bound $\|X\|_{\psi_1}$, we apply the sub-gamma condition (8.2) to $\lambda = 1/(\sigma \vee 2K)$, which gives:

$$\begin{aligned}\mathbf{E}\left[\exp\left(\frac{|X|}{\sigma\vee 2K}\right)\mathbf{1}(X \geqslant 0)\right] &\leqslant \mathbf{E}\left[\exp\left(\frac{X}{\sigma\vee 2K}\right)\right]\\ &\leqslant \mathbf{E}\left[\exp\left(\frac{\sigma^2/\sigma^2}{2(1-K/(2K))}\right)\right] = e\,.\end{aligned}$$

Applying the same inequality to $-X$ and summing gives:

$$\mathbf{E}\Big[\exp\Big(\frac{|X|}{\sigma\vee 2K}\Big)\Big]\leqslant 2e\,.$$

Now, a simple analysis of function shows that $e^u-eu\geqslant 0$ for any $u\geqslant 0$, hence (applying this to u/p) $\big(\frac{eu}{p}\big)^p\leqslant e^u$. Hence, for any $p\geqslant 3$, one has

$$\mathbf{E}\Big[\Big(\frac{e|X|}{p(\sigma\vee 2K)}\Big)^p\Big]\leqslant \mathbf{E}\Big[\exp\Big(\frac{|X|}{\sigma\vee 2K}\Big)\Big]\leqslant 2e\,,$$

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 H o lder's inequality:

$$\begin{aligned}\mathbf{E}[|X|^p]&=\mathbf{E}[|X|^{2(1-1/r)}|X|^{p-2+2/r}]\\&\leqslant \mathbf{E}[X^2]^{1-1/r}\mathbf{E}[X^{(p-2)r+2}]^{1/r}\\&\leqslant \sigma^{2-2/r}\|X\|_{(p-2)r+2}^{[(p-2)r+2]/r}\\&\leqslant \sigma^{2-2/r}\Big[\frac{[(p-2)r+2]K}{2e}\Big]^{p-2+2/r}\\&=\sigma^2\Big(\frac{Kr}{2}\Big)^{p-2}\Big(\frac{r}{2}\Big)^{2/r}\Big(\frac{K}{\sigma}\Big)^{2/r}\Big(\frac{p-2+2/r}{e}\Big)^{p-2+2/r}.\end{aligned}$$

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

$$\sqrt{2\pi p}\Big(\frac{p}{e}\Big)^p\leqslant p!\leqslant p^p\,. \tag{8.6}$$

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

$$\begin{aligned}\mathbf{E}[|X|^p]&\leqslant\sigma^2\big(K\log(K/\sigma)\big)^{p-2}e^{1+1/e}p!/(6\sqrt{\pi})\\&\leqslant\sigma^2\Big(K\log\Big(\frac{K}{\sigma}\Big)\Big)^{p-2}p!/2\,,\end{aligned}\tag{8.7}$$

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

$$\mathbf{E}[|X|^p]\leqslant\Big(\frac{Kp}{2e}\Big)^p\leqslant\Big(\frac{K}{2}\Big)^p p!=\frac{1}{2}\frac{K^2}{2}\Big(\frac{K}{2}\Big)^{p-2}p!\,,\tag{8.8}$$

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 8.2. *Let X be a random variable such that $\mathbf{E}[X^2] \leqslant \sigma^2$ and $X \leqslant b$ almost surely, for some $\sigma^2 > 0$ and $b > 0$. Then*

1. *$X - \mathbf{E}[X]$ is $(\sigma^2, b/3)$-sub-gamma.*
2. *For all $\lambda \in [0, b^{-1}]$, $\log \mathbf{E} e^{\lambda X} \leqslant \lambda \mathbf{E}[X] + \sigma^2/b^2$.*

Proof. By homogeneity we assume that $b = 1$. Let $X' = X - \mathbf{E}[X]$. Using Bennett's inequality [BLM13, Theorem 2.9], one has, for all $\lambda > 0$,

$$\log \mathbf{E} e^{\lambda X'} \leqslant \sigma^2 \phi(\lambda)\,, \quad \phi(\lambda) = e^\lambda - \lambda - 1\,. \tag{8.9}$$

Moreover, for every $\lambda \in [0, 1/3)$

$$\phi(\lambda) = \sum_{k \geqslant 2} \frac{\lambda^k}{k!} = \frac{\lambda^2}{2} \sum_{k \geqslant 0} \frac{\lambda^k}{(k+2)!/2} \leqslant \frac{\lambda^2}{2} \sum_{k \geqslant 0} \frac{\lambda^k}{3^k} = \frac{\lambda^2}{2(1 - \lambda/3)}\,,$$

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 (8.9) and use that $\phi(\lambda) \leqslant \phi(1) = e - 2 \leqslant 1$ for all $\lambda \in [0, 1]$. ∎

8.2 Polar coordinates and spherical caps

In this section we gather some results on geometric facts that are used throughout Chapters 3 to 7.

8.2.1 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 8.3. *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*

$$\frac{\|u - u^*\|}{\sqrt{2}} \leqslant \frac{\|\theta_\perp\|}{\|\theta\|} = \sqrt{1 - \langle u, u^* \rangle^2} \leqslant \|u - u^*\|\,. \tag{8.10}$$

2. *If $\|u - u^*\| \leqslant 1$, then $\|\theta\|/2 \leqslant \langle u^*, \theta \rangle \leqslant \|\theta\|$.*
3. *One has*

$$|\langle u^*, \theta - \theta^* \rangle| \leqslant \big| \|\theta\| - \|\theta^*\| \big| + \|\theta^*\| \cdot \frac{\|u - u^*\|^2}{2}\,. \tag{8.11}$$

4. *One has*

$$\big|\|\theta\| - \|\theta^*\|\big| \leqslant |\langle u^*, \theta - \theta^*\rangle| + \frac{\|\theta_\perp\|^2}{\|\theta\| + \|\theta^*\|}\,. \tag{8.12}$$

Proof. We start with the first point. By orthogonality,

$$\begin{aligned}\|\theta_\perp\|^2 &= \|\theta\|^2 - \langle u^*, \theta\rangle^2 = \|\theta\|^2\big[1 - \langle u^*, u\rangle^2\big] \\ &= \|\theta\|^2\big[1 - \langle u^*, u\rangle\big]\big[1 + \langle u^*, u\rangle\big] = \frac{1}{2}\|\theta\|^2\|u - u^*\|^2\big[1 + \langle u^*, u\rangle\big]\,.\end{aligned} \tag{8.13}$$

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 (8.13) 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. ∎

We also use repeatedly the following result, which controls the behavior of the norm and direction within H-ellipsoids.

Lemma 8.4. *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

$$\frac{(\langle \theta, u^* \rangle - B)^2}{B^3} + \frac{\|\theta - \langle \theta, u^* \rangle u^*\|^2}{B} \leqslant \frac{r^2}{B} \,. \tag{8.14}$$

For the upper bound in the first point, remark first that $\theta = (1+r)\theta^*$ satisfies (8.14) 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 (8.14), $(\langle \theta, u^* \rangle - B)^2 \leqslant B(r^2 - \alpha^2)$. 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 (8.14), 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 (8.14),

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

The proof is concluded by Point 1. ∎

8.2.2 Spherical caps

In Section 3.5, we defined spherical caps through their angles as

$$\mathsf{C}(u, \varepsilon) = \left\{ v \in S^{d-1},\ \langle u, v \rangle \geqslant 0,\ |\sin(u, v)| \leqslant \varepsilon \right\} , \tag{8.15}$$

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

$$\widetilde{\mathsf{C}}(u, r) = \left\{ v \in S^{d-1},\ \|u - v\| \leqslant r \right\} . \tag{8.16}$$

The following result provides a formal statement of this equivalence.

Fact 8.1. *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*

$$\mathsf{C}(u, \varepsilon/\sqrt{2}) \subset \widetilde{\mathsf{C}}(u, \varepsilon) \subset \mathsf{C}(u, \varepsilon) \,. \tag{8.17}$$

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, (8.17) follows from the first point of Lemma 8.3. ∎

Chapter 9

Conclusion and future work

9.1 Conclusion

In this thesis, we investigated several aspects of maximum-likelihood estimation il logistic regression: existence of the MLE, its performance with respect to the logarithmic loss and its performance as a binary classifier. We emphasized in particular the critical role of the signal strength B, both for the question of existence of the MLE, and for its performance.

In the ideal setting of a well-specified logistic model with a Gaussian design, we proved that the MLE undergoes a sharp "phase transition" around the critical sample size Bd: if $n \ll Bd$, then the MLE exists with probability exponentially small in d, and if $n \gg Bd$, then with probability $1 - e^{-n/B}$ the MLE exists and satisfies the optimal risk bound from asymptotic theory. We thereby answer a question from [HM24] regarding the estimation error of the direction of the true parameter in logistic regression.

We then identified necessary and sufficient conditions on the design for the MLE to exhibit a near-Gaussian behavior, at least in the well-specified case, and then proved optimal guarantees for a misspecified model under these regularity conditions.

We also proved the first $O(d/n)$ bounds on the classification excess risk of (the plug-in of) the MLE. This significantly improved over the $\sqrt{d/n}$ obtained by combining Zhang's lemma with a bound on the excess logistic risk, and more interestingly, it is better than the $(d/n)^{2/3}$ suggested by the fact that the model satisfies the margin condition with exponent 1.

9.2 Future work

9.2.1 Direction-dependent estimation

In this thesis, our focus was on the predictive performance of the MLE, either in terms of logistic loss for probability assignments, or binary loss for classification.

A possible future direction would be to derive sharper bounds for the estimation error of θ^* in general directions, namely estimate $\langle v, \theta^* \rangle$ for an arbitrary v on the unit sphere. This can naturally be done by the plug-in estimator $\langle v, \widehat{\theta}_n \rangle$, but bounding its estimation error requires a refinement compared to the analysis regarding the prediction performance of the MLE. Of course, a particular case of interest is when v is a coordinate direction (a vector in the canonical basis of $\mathbb{R}^d$). In this case, deriving sharp bounds on the estimation error of the coordinates θ_j^* directly serves as a building block to provide non-asymptotic

confidence intervals for these coefficients. In turn, they give rise to non-asymptotic tests on their nullity, providing improved guarantees for reliable finite-sample inference in the logistic model.

A closely related question is that of the estimation of the signal strength itself. This problem also has practical implications for reliable inference, as the MLE is known to overestimate the effects of the covariates on the outcome when the dimension is not negligible compared to the sample size [SC19]. As a byproduct of our analysis, we obtained a bound on $|\|\widehat{\theta}_n\| - \|\theta^*\||$, which is sub-optimal with respect to B. Namely, we obtained that with probability at least $1-\delta$,

$$\big|\|\widehat{\theta}_n\| - \|\theta^*\|\big| \lesssim \sqrt{B^3 \frac{d + \log(1/\delta)}{n}}\,.$$

The limiting distribution of $\sqrt{n}(\widehat{\theta}_n - \theta^*)$ suggests that the dependence with respect to δ is sharp, but the fact that B^3 multiplies the dimension is clearly pessimistic. It would be interesting to obtain a finite-sample bound $O\big(\sqrt{(Bd + B^3\log(1/\delta))/n}\big)$, as it is what one can expect from the convergence in distribution of $\sqrt{n}(\widehat{\theta}_n - \theta^*)$.

9.2.2 High-dimensional regime

All the results in this thesis deal with the unregularized MLE, which is canonical in some sense. The strong non-existence result of Chapter 4 (Theorem 4.2) shows that, except for the special case of a bounded signal, the regime where it is reasonable to perform logistic regression is that where $n \gg d$. An interesting direction would be to analyze logistic regression in the framework of high-dimensional statistics, that is, the Lasso logistic regression, or ℓ_1-regularized MLE, defined for a penalty parameter $\lambda \geqslant 0$ as

$$\widehat{\theta}_n^{\lambda} = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-Y_i\langle\theta, X_i\rangle}\right) + \lambda\|\theta\|_1 \right\}.$$

This would allow in particular to handle the regime $n \asymp d$, where if the signal is strong, there is a high probability that the MLE does not exist. A possible direction would be to investigate the high dimensional regime, where $n < d$, under an additional sparsity assumption on the true signal θ^*, namely that only s coordinates of θ^* are non-zero, for some $s < d$.

In this setting, the ℓ_1 regularization would untie the existence of the estimator from the geometry of the dataset, and allow to derive risk bounds that adapt to the level of sparsity, as is the case with the Lasso linear regression. The goal would still be to estimate the conditional probability $\mathbf{P}(Y = 1\,|X) = \sigma(\langle\theta^*, X\rangle)$ (in the well-specified case, considered as a first step), and the performance of $\widehat{\theta}_n^{\lambda}$ would still be assessed as before by its excess logistic risk $L(\widehat{\theta}_n^{\lambda}) - L(\theta^*)$.

Chapter 10

Introduction en Français

Perspective historique et importance pratique. La régression logistique est un modèle classique décrivant la dépendance d'une variable binaire à des caractéristiques multivariées. Sa description simple la rend très populaire dans une grande variété d'applications en sciences expérimentales et sociales. Un exemple typique se trouve dans la recherche biomédicale : comment peut-on prédire si un patient développera une certaine maladie (e.g., le cancer du poumon) à partir de marqueurs génétiques et d'autres caractéristiques cliniques de cette personne ? La régression logistique (ou le modèle logistique) est une manière d'aborder un tel problème.

Le modèle logit a été introduit pour la première fois par Berkson [Ber44] dans le contexte d'études cliniques sur les dosages médicamenteux. D'un point de vue biomédical, l'usage du modèle *probit* (dans lequel la fonction de lien est la fonction de répartition de la loi normale) était plus répandu, mais Berkson a soutenu que le modèle logistique avait l'avantage d'être plus interprétable et plus facile à traiter numériquement. Il proposa de modéliser la dépendance d'une variable $Y = \pm 1$ à une covariée réelle X (en notation moderne) via sa distribution de probabilité conditionnelle :

$$\mathbf{P}(Y = 1|X = x) = \frac{1}{1 + e^{-(\theta_0+\theta_1 x)}} = \sigma(\theta_0 + \theta_1 x)\,, \tag{10.1}$$

où θ_0 et θ_1 sont les paramètres inconnus à estimer, et

$$\sigma(t) = \frac{e^t}{1 + e^t} = \frac{1}{1 + e^{-t}}\,, \quad t \in \mathbb{R}\,, \tag{10.2}$$

est appelée la fonction *sigmoïde*.

Berkson met en avant la pertinence de la fonction logit, qui est l'inverse de la sigmoïde (10.2),

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right), \tag{10.3}$$

qui, dans le modèle logistique (10.1), transforme le logarithme du rapport de vraisemblance $p/(1-p)$ en une fonction affine de la variable explicative x :

$$\text{logit}(p) = \theta_0 + \theta_1 x\,, \quad p = \mathbf{P}(Y = 1|X = x)\,. \tag{10.4}$$

Il propose donc la procédure suivante : étant données des observations $(x_1, y_1), \ldots, (x_n, y_n)$, on calcule les rapports de cotes correspondants à l'aide de la fonction logit (10.3), puis on ajuste le modèle linéaire (10.4) en utilisant la méthode classique des moindres carrés.

Bien que Berkson ait correctement introduit le modèle logistique (univarié) (10.1), il n'a pas préconisé l'utilisation de la méthode générale d'estimation par maximum de vraisemblance, qui était pourtant connue depuis de nombreuses années dans le contexte de l'estimation paramétrique dans les modèles statistiques.

Quelques années plus tard, Cox formalise la régression logistique dans des termes plus proches de la formulation moderne dans [Cox58]. Signalons les principales différences avec les travaux antérieurs de Berkson. Premièrement, Cox considère le modèle logit dans un but plus général, contrairement à la description de Berkson qui se concentre sur les expériences biologiques. Deuxièmement, Cox met en avant le caractère plus naturel de l'estimation par maximum de vraisemblance dans le modèle logistique. Enfin, Cox considère également des variables explicatives multivariées[1] $x \in \mathbb{R}^d$, pour une certaine dimension $d \geqslant 1$.

Description moderne. Dans sa forme moderne générale, le modèle logistique peut être décrit de la manière suivante. Étant donnée une dimension $d \geqslant 1$, le *modèle logistique* est la famille de lois conditionnelles sur la variable $y \in \{-1, 1\}$ sachant les covariables représentées par le vecteur $x \in \mathbb{R}^d$, définie par :

$$\mathcal{P}_{\text{logit}} = \left\{ p_\theta : \theta \in \mathbb{R}^d \right\}, \quad \text{où} \quad p_\theta(y|x) = \sigma(y\langle \theta, x \rangle), \tag{10.5}$$

pour tout $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$, où l'on pose

$$\sigma(s) = \frac{e^s}{e^s + 1} = \frac{1}{1 + e^{-s}}$$

pour $s \in \mathbb{R}$, la fonction sigmoïde, et où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire usuel sur $\mathbb{R}^d$. On dit qu'un couple aléatoire (X, Y) sur $\mathbb{R}^d \times \{-1, 1\}$ suit le modèle logistique si la loi conditionnelle de Y sachant X appartient à $\mathcal{P}_{\text{logit}}$.

Historiquement, le modèle logit a été étudié sous l'hypothèse que les données étaient générées selon le modèle, comme c'est souvent le cas en estimation paramétrique. Plus précisément, le cadre statistique était le suivant : le statisticien observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ dans $\mathbb{R}^d \times -1, 1$ supposés indépendants et de même loi inconnue P, et cette loi est telle que le modèle est bien spécifié, c'est-à-dire qu'il existe $\theta^* \in \mathbb{R}^d$ tel que pour tout i,

$$\mathbf{P}(Y_i = 1 \,|X_i) = \sigma(\langle \theta^*, X_i \rangle). \tag{10.6}$$

Le statisticien ajuste alors le modèle par maximum de vraisemblance, c'est-à-dire en calculant

$$\widehat{\theta}_n = \arg\max_{\theta \in \mathbb{R}^d} \prod_{i=1}^n p_\theta(Y_i|X_i) = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \log\left(1 + e^{-Y_i \langle \theta, X_i \rangle}\right), \tag{10.7}$$

dans le but de faire de l'inférence sur le paramètre θ^*, notamment pour tester si certains coefficients sont nuls. Ce point de vue est motivé par le besoin de tester si une variable $x^{(j)}$ donnée a un effet statistiquement significatif sur la variable binaire y—par exemple, si un gène donné influence la probabilité de développer une certaine maladie.

Cependant, cette approche présente une limite importante : elle suppose que le modèle est *bien spécifié*, c'est-à-dire que la loi réelle des données appartient au modèle, ou de manière équivalente que (10.6) est vérifiée.

[1] Techniquement, Cox traite les cas $d = 1$ et $d = 2$ mais le passage au modèle en dimension deux se généralise facilement et pose les bases du modèle général.

Au-delà de l'inférence, la régression logistique est également attrayante comme méthode de prédiction, où l'objectif n'est plus de tester la significativité de certaines variables, mais d'estimer précisément la probabilité conditionnelle $\mathbf{P}(Y = 1|X = x)$ pour tout x. Cette perspective prédictive est l'objet principal de cette thèse. En particulier, nous considérons aussi le cadre général de l'apprentissage statistique, où le modèle n'est en général pas supposé bien spécifié. Dans ce contexte, la régression logistique sert à approximer la vraie loi conditionnelle de Y sachant X par une loi appartenant au modèle logit.

10.1 Questions principales

Dans cette thèse, nous étudions les performances prédictives de l'estimateur du maximum de vraisemblance (MLE, pour *maximum likelihood estimator* en anglais), dans l'esprit de l'apprentissage statistique. Nous nous intéressons donc aux deux questions suivantes :

1. *Existence* : Quand le MLE existe-t-il ?

2. *Performance* : Lorsque le MLE existe, quelle est sa précision ?

Pour rendre ces deux questions précises, une discussion s'impose.

Commençons par clarifier la signification géométrique de l'existence (et de l'unicité) du MLE ; nous renvoyons à [AA84] ainsi qu'à l'introduction de [CS20] pour une discussion intéressante sur ce point, accompagnée de références approfondies. L'unicité du MLE est en réalité une question assez directe : dès que les points $X_1, \dots, X_n$ engendrent $\mathbb{R}^d$ (ce qui est vrai avec forte probabilité lorsque $n \gtrsim d$, sous des hypothèses raisonnables sur X), la deuxième fonction dans (10.7) que le MLE minimise est strictement convexe sur $\mathbb{R}^d$, et admet donc au plus un minimiseur. La question de l'existence du MLE a un contenu géométrique plus riche. Supposons encore pour simplifier que $X_1, \dots, X_n$ engendrent $\mathbb{R}^d$, de sorte que pour tout $\theta \neq 0$, il existe $i \in \{1, \dots, n\}$ tel que $\langle \theta, X_i \rangle \neq 0$. Alors, *le MLE existe si et seulement si les données ne sont pas linéairement séparables*, ce qui signifie qu'il n'existe pas de $\theta \neq 0$ tel que

$$\{X_i : 1 \leqslant i \leqslant n, Y_i = 1\} \subset \mathcal{H}_\theta^+ = \{x \in \mathbb{R}^d : \langle \theta, x \rangle \geqslant 0\}$$

et

$$\{X_i : 1 \leqslant i \leqslant n, Y_i = -1\} \subset \mathcal{H}_\theta^- = \{x \in \mathbb{R}^d : \langle \theta, x \rangle \leqslant 0\},$$

ou, de manière plus concise, s'il n'existe pas de $\theta \neq 0$ tel que $Y_i \langle \theta, X_i \rangle \geqslant 0$ pour tout $i = 1, \dots, n$. En effet, si un tel θ existe, alors la deuxième fonction de (10.7) évaluée en $t\theta$ reste majorée lorsque $t \to +\infty$; or une fonction strictement convexe qui admet un minimiseur global diverge à l'infini, donc la fonction objectif n'admet pas de minimiseur global. Réciproquement, si un tel θ n'existe pas, de simples arguments de compacité montrent que la fonction dans (10.7) diverge à l'infini et est continue, donc elle admet un minimiseur global.

Deuxièmement, pour évaluer les performances de l'estimateur du maximum de vraisemblance (MLE), il convient de spécifier une notion de précision. Dans ce travail, nous nous concentrons principalement sur les performances prédictives du MLE, mesurées par son risque sous la perte logistique. Plus précisément, nous considérons le problème consistant à assigner des probabilités aux valeurs possibles ± 1 de Y, connaissant le vecteur de covariables associé X. Chaque paramètre $\theta \in \mathbb{R}^d$ induit une loi conditionnelle p_θ définie

en (10.5). On peut alors définir la perte logistique ℓ (au point $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$) et le risque L de θ respectivement par :

$$\ell(\theta, (x, y)) = -\log p_\theta(y|x) = \log(1 + e^{-y\langle\theta,x\rangle}) \quad \text{et} \quad L(\theta) = \mathbf{E}[\ell(\theta, (X, Y))]\,. \tag{10.8}$$

Ainsi, la perte logistique correspond à l'opposée de la log-vraisemblance (ou perte logarithmique) pour le modèle logistique. La perte logarithmique est un critère classique pour évaluer la qualité de prévisions probabilistes : elle favorise des prédictions bien calibrées en pénalisant à la fois les probabilités trop confiantes et pas assez confiantes. En particulier, assigner une probabilité nulle à une étiquette y qui se réalise conduit à une perte infinie. De plus, ce critère est intimement lié au MLE, qui correspond au minimiseur du *risque empirique* $\widehat{L}_n : \mathbb{R}^d \to \mathbb{R}$ sous la perte logistique, défini par :

$$\widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta, (X_i, Y_i)) = \frac{1}{n}\sum_{i=1}^n \log(1 + e^{-Y_i\langle\theta,X_i\rangle})\,. \tag{10.9}$$

Avec ces définitions à disposition, on peut mesurer la performance prédictive du MLE via son excès de risque sous la perte logistique, à savoir $L(\widehat{\theta}_n) - L(\theta^*)$, où $\theta^* \in \arg\min_{\theta\in\mathbb{R}^d} L(\theta)$ (en supposant cet ensemble non vide). Nous donnons dans les Chapitres 3 et 6 des bornes supérieures non asymptotiques valables avec forte probabilité pour cette quantité.

La perte logistique intervient également de manière naturelle dans la théorie de l'apprentissage statistique [BBL05, Kol11, Bac24], comme relaxation convexe de l'erreur de classification [Zha04, BJM06]. En tant que telle, nous étudierons également (dans le Chapitre 7) les performances en classification du MLE utilisé comme classifieur par plug-in, visant à prédire correctement l'étiquette Y d'un point X. Par classifieur plug-in, on entend des prédicteurs de la forme $f_\theta(x) = \text{sign}(\langle\theta, x\rangle)$, où $\text{sign}(t) = \mathbf{1}(t \geqslant 0) - \mathbf{1}(t < 0)$ pour tout réel t.

Dans ce cadre, la performance est évaluée par le risque de classification, défini pour tout classifieur $f : \mathbb{R}^d \to \mathbb{R}$ comme $R(f) = \mathbf{P}(Yf(X) < 0)$. Nous mesurerons donc les performances en classification binaire de $\widehat{\theta}_n$ via l'excès de risque de classification de $f_{\widehat{\theta}_n}$ par rapport au meilleur classifieur possible, à savoir :

$$R(\widehat{\theta}_n) - \inf_f R(f) = \mathbf{P}(Y\langle\widehat{\theta}_n, X\rangle < 0) - \inf_f \mathbf{P}(Yf(X) < 0)\,,$$

où l'infimum est pris sur l'ensemble des fonctions mesurables de $\mathbb{R}^d$ dans $\mathbb{R}$.

Troisièmement, l'existence du MLE dépend du jeu de données et constitue donc un événement aléatoire, de la même manière que l'excès de risque $L(\widehat{\theta}_n) - L(\theta^*)$ est une variable aléatoire. Ainsi, l'existence et la précision du MLE dépendent de la loi jointe P de (X, Y). Pour donner un sens précis aux questions précédentes, il faut donc spécifier quelles distributions P sont considérées. Remarquons que la loi jointe P est caractérisée par (a) la loi marginale P_X de X, appelée *design*, ou "plan d'expérience", et (b) la loi conditionnelle $P_{Y|X}$ de Y sachant X.

Nous considérons dans ce mémoire trois cadres différents de généralité croissante, selon les hypothèses respectives sur P_X et $P_{Y|X}$, mais pour fixer les idées et faciliter la comparaison avec les résultats précédents, nous commencerons dans cette introduction par le plus simple :

(a) Le design suit une loi gaussienne : $X \sim \mathsf{N}(0, \Sigma)$ pour une matrice Σ symétrique définie positive. Par invariance du problème par transformation linéaire inversible de X, on peut supposer sans perte de généralité que $\Sigma = I_d$, ce que nous ferons par la suite.

(b) Le modèle est *bien spécifié*, au sens où la loi conditionnelle $P_{Y|X}$ appartient au modèle logistique $\mathcal{P}_{\text{logit}}$. Autrement dit, il existe $\theta^* \in \mathbb{R}^d$ tel que $\mathbf{P}(Y = 1|X) = \sigma(\langle \theta^*, X \rangle)$.

Outre son caractère naturel, l'attrait de ce cadre réside dans le fait que le problème ne dépend que d'un petit nombre de paramètres : la taille d'échantillon n, la dimension des données d, la probabilité $1 - \delta$ avec laquelle les garanties s'appliquent, et surtout la *force du signal* (ou rapport signal-bruit, ou température inverse) $B = \max\{e, \|\theta^*\|\}$, où $\|\cdot\|$ désigne la norme euclidienne.

Il est utile de commenter le rôle de la dimension d et de la force du signal B. Intuitivement, deux effets distincts peuvent rendre les données linéairement séparables. D'une part, plus la dimension d est grande, plus il existe de degrés de liberté pour séparer linéairement les données. D'autre part, un effet provient de la force du signal : plus le signal B est fort, plus les étiquettes Y_i ont tendance à avoir le même signe que $\langle \theta^*, X_i \rangle$—et donc, plus il est probable que les données soient séparées par θ^* ou par une direction « proche ». Comme nous le verrons, les effets de « dimensionnalité » et de « force du signal » interagissent entre eux. On note également qu'intuitivement, un signal plus fort rend le problème de *classification* (prédire la valeur de Y et minimiser le taux d'erreurs) plus facile. Cela revient à dire que plus B est grand, plus l'erreur d'estimation de la *direction* $u^* = \theta^*/\|\theta^*\|$ de θ^* devrait être petite. En revanche, lorsqu'il est confronté à un signal fort, il est connu que le MLE (voir, par exemple, [CS20, SC19] et les références citées) a tendance à sous-estimer l'incertitude sur les étiquettes, c'est-à-dire à fournir des probabilités conditionnelles trop confiantes pour Y sachant X. C'est notamment le cas lorsque les données sont presque linéairement séparées, auquel cas le MLE prédit des probabilités proches de 0 ou 1. Ainsi, pour le problème d'*estimation de densité conditionnelle* que nous considérons, un signal plus fort peut dégrader les performances du MLE. Cela devrait se manifester par le fait que la *norme* du MLE (par opposition à sa direction) peut être très différente de celle de θ^*, de sorte que l'erreur d'estimation globale de θ^* peut être plus grande.

Pour résumer, nous cherchons à établir des garanties explicites et *non asymptotiques* d'existence et de précision du MLE, en fonction des paramètres pertinents B, d, n, δ—idéalement, dans une situation générale où ces paramètres peuvent prendre des valeurs arbitraires. Notre objectif est double : obtenir d'une part la dépendance optimale en tous les paramètres dans le cas d'un design gaussien et d'un modèle bien spécifié ; et d'autre part, examiner dans quelle mesure ces résultats s'étendent à des distributions plus générales. Enfin, nous étudierons les performances en classification du classifieur plug-in associé au MLE, qui, comme mentionné ci-dessus, est étroitement lié à l'estimation de la direction de θ^*.

10.2 Résultats existants

Avant de décrire nos contributions, nous passons en revue les résultats connus relatifs aux questions que nous abordons. En tant que méthode statistique de base, la régression logistique a été largement étudiée dans la littérature, nous nous concentrons donc sur les résultats les plus directement pertinents pour notre cadre. À nouveau, dans un souci de comparaison, nous nous focaliserons principalement sur le cas d'un design gaussien et d'un modèle bien spécifié, bien que nous discutions aussi des extensions possibles.

10.2.1 Asymptotique classique

Le comportement du MLE est bien compris dans le cadre de l'asymptotique paramétrique classique [LCY00, vdV98]. Dans ce cadre, la loi P est fixée (et donc, la dimension d et la force du signal B aussi), tandis que la taille de l'échantillon n tend vers l'infini. Dans ce cas, lorsque $n \to \infty$, le MLE $\widehat{\theta}_n$ existe avec une probabilité tendant vers 1, converge vers θ^* à vitesse $1/\sqrt{n}$, et est asymptotiquement normal, avec une covariance asymptotique donnée par l'inverse de la matrice d'information de Fisher [vdV98, §5.2–5.6]. Cela implique que l'excès de risque converge vers 0 à vitesse $1/n$, et plus précisément que

$$2n\big\{L(\widehat{\theta}_n) - L(\theta^*)\big\} \xrightarrow{\text{(d)}} \chi^2(d)\,, \tag{10.10}$$

où $\xrightarrow{\text{(d)}}$ désigne la convergence en loi, et $\chi^2(d)$ la loi du χ^2 à d degrés de liberté. En combinant ceci avec une borne sur la queue supérieure de la loi du χ^2, on obtient : pour tout $d \geqslant 1$, $\theta^* \in \mathbb{R}^d$, et $\delta \in (0,1)$,

$$\liminf_{n\to\infty} \mathbf{P}\Big(L(\widehat{\theta}_n) - L(\theta^*) \leqslant \frac{d + 2\log(1/\delta)}{n}\Big) \geqslant 1 - \delta\,, \tag{10.11}$$

avec la convention que $L(\widehat{\theta}_n) - L(\theta^*) = +\infty$ si les données sont linéairement séparables. Notons que la convergence (10.10) ne vaut que dans le cas bien spécifié, et que dans le cas mal spécifié, l'excès de risque normalisé $2n\{L(\widehat{\theta}_n) - L(\theta^*)\}$ converge vers une autre loi limite dépendant de la loi P de (X, Y) ; voir [vdV98, Example 5.25 p. 55] et, par exemple, les introductions de [OB21, MG22] pour des discussions complémentaires sur ce point.

Du côté positif, la garantie en forte probabilité (10.11) est précise, compte tenu de la convergence en loi (10.10) de l'excès de risque. En revanche, il convient de noter que cette garantie est purement asymptotique : elle est valable lorsque $n \to \infty$ alors que tous les autres paramètres du problème sont fixés. Cela ne permet pas de traiter le régime moderne de grande dimension, où la dimension d peut être grande et éventuellement comparable à n. De plus, cette garantie ne précise pas quelle doit être la taille de l'échantillon n (en fonction de B, d, δ) pour que le comportement asymptotique (10.11) soit valide—en particulier, elle ne fournit aucune information sur la taille d'échantillon requise pour que le MLE existe.

10.2.2 Asymptotique en grande dimension

Plusieurs des limites de la théorie asymptotique classique peuvent être levées en considérant un autre cadre asymptotique, à savoir le « régime asymptotique en grande dimension », où $d, n \to \infty$ tout en ayant d/n qui converge vers une constante fixée. Ce cadre a suscité un intérêt considérable en statistique au cours de la dernière décennie (voir, par exemple, [EK18b, Mon18] et les références citées pour un aperçu partiel de cette ligne de recherche). L'intérêt de ce cadre est qu'il permet de capturer les effets liés à la grande dimension, puisque la dimension n'est plus négligeable comparée à la taille de l'échantillon.

La question de l'existence du MLE dans ce régime de grande dimension asymptotique a été abordée dans les travaux fondateurs de Candès et Sur [CS20], prolongeant un résultat antérieur de Cover [Cov65] dans le cas sans signal où $\theta^* = 0$. Plus précisément, le résultat principal de Candès et Sur [CS20, Théorèmes 1–2] peut être énoncé comme

suit[2] : il existe une fonction $h : \mathbb{R}^+ \to (0,1)$ telle que le résultat suivant ait lieu. Fixons $\beta \in \mathbb{R}^+$ et $\gamma \in (0,1)$, et supposons que $d = d_n \to \infty$ lorsque $n \to \infty$, avec $d/n \to \gamma$. Si $X \sim \mathsf{N}(0, I_d)$ et $\mathbf{P}(Y = 1|X) = \sigma(\langle \theta^*, X \rangle)$, avec $\theta^* = \theta^*_d \in \mathbb{R}^d$ tel que $\|\theta^*\| = \beta$, et que les données sont composées de n copies i.i.d. de (X, Y), alors

$$\lim_{n\to\infty} \mathbf{P}(\text{MLE existe}) = \begin{cases} 0 & \text{si } \gamma > h(\beta) \\ 1 & \text{si } \gamma < h(\beta)\,. \end{cases} \tag{10.12}$$

De plus, la quantité $h(\beta)$ est définie comme l'infimum de l'espérance d'une famille explicite de variables aléatoires dépendant de β (voir eq. (2.4) dans [CS20]), et la courbe de la fonction h est tracée numériquement dans cet article.

Les conditions (10.12) fournissent une caractérisation précise de l'existence du MLE dans le régime asymptotique de grande dimension, et en particulier, établissent une transition de phase nette pour cette propriété, dépendant de la valeur du rapport d'aspect $\gamma = \lim d/n$.

10.2.3 Garanties non asymptotiques

Nous présentons ici les garanties non asymptotiques disponibles pour le MLE en régression logistique dans la littérature, en nous concentrant sur celles qui sont les plus pertinentes pour notre cadre.

Tout d'abord, il découle des résultats de [CLL20] (en combinant en particulier les Théorèmes 1 et 8) qu'il existe une constante $c > 0$ telle que, si $n \geqslant e^{cB} d$, alors avec probabilité au moins $1 - 2e^{-d/c}$, le MLE $\widehat{\theta}_n$ existe et satisfait

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant \frac{e^{cB} d}{n}\,. \tag{10.13}$$

Ce résultat est entièrement explicite et présente une dépendance optimale en la taille d'échantillon n et la dimension d ; la probabilité $1 - e^{-d/c}$ avec laquelle la borne $O_B(d/n)$ est valable est elle aussi optimale, à la lumière des résultats asymptotiques (10.10) et (10.11). En revanche, la dépendance en la force du signal B est exponentielle, ce qui s'avère fortement sous-optimal dans le cas d'un design gaussien. En réalité, la borne (10.13) s'applique dans un cadre plus général, où le modèle peut être mal spécifié et où le design est seulement supposé sous-gaussien. Comme nous le verrons ci-dessous, une dépendance exponentielle en la norme est inévitable si l'on suppose seulement que la distribution du design est sous-gaussienne.

Jusqu'à récemment, les garanties non asymptotiques les plus précises pour le MLE en régression logistique avec un design gaussien étaient dues à Ostrovskii et Bach [OB21]. Plus précisément, en combinant le Théorème 4.2 et la Proposition D.1 de [OB21], on obtient qu'il existe une constante $c > 0$ telle que, pour tout $\delta \leqslant 1/2$, si $n \geqslant c \log^4(B) B^8 d \log(1/\delta)$, alors avec probabilité au moins $1 - \delta$, le MLE existe et satisfait

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant \frac{B^3 d \log(1/\delta)}{n}\,. \tag{10.14}$$

Comme la borne (10.13), ce résultat présente une dépendance optimale à la dimension d et à la taille d'échantillon n ; et bien que la borne implique un terme de déviation $d \log(1/\delta)$

[2] En réalité, le Théorème 1 de [CS20] traite du cas de la régression logistique avec un terme constant (biais), tandis que le Théorème 2 concerne la régression logistique sans terme constant, que nous discutons ici.

proportionnel à la dimension (ce qui est sous-optimal pour petits δ), il pourrait être affiné en un terme additif $d+\log(1/\delta)$ avec des modifications mineures de la preuve dans [OB21]. Ce résultat améliore considérablement la borne générale (10.13) dans le cas d'un design gaussien, en remplaçant la dépendance exponentielle en la norme B par une dépendance polynomiale. Il convient également de noter que le résultat de [OB21] est valable dans le cas général mal spécifié, et constitue alors la meilleure garantie disponible dans la littérature. Cela étant dit, comme nous le verrons plus loin, la dépendance polynomiale en B dans la condition d'existence du MLE ainsi que dans la borne de risque peut être améliorée. Par exemple, dans le cas bien spécifié, la borne (10.14) est supérieure au risque asymptotique (10.11) d'un facteur B^3, ce qui suggère des améliorations possibles.

Récemment, alors que l'article [CLM24] était en préparation, deux travaux supplémentaires [KvdG23, HM23] ont apporté des contributions importantes à l'étude de la régression logistique en design gaussien, avec un accent particulier sur la dépendance à la force du signal B. Le plus proche de notre cadre est celui de Kuchelmeister et van de Geer [KvdG23], qui étudient le MLE pour la régression logistique en design gaussien, en supposant toutefois que la loi conditionnelle de Y sachant X suit un modèle probit plutôt qu'un modèle logit. En dépit de différences techniques réelles entre les modèles probit et logit, cela reste qualitativement proche du modèle logit bien spécifié. Avec une notion naturelle de force du signal B dans le modèle probit (l'inverse du paramètre de bruit σ dans leur travail), le Théorème 2.1.1 de [KvdG23] affirme que, pour une constante absolue c, si $n \geqslant cB(d\log n + \log(1/\delta))$, alors le MLE existe et satisfait

$$\left\| \frac{\widehat{\theta}_n}{\|\widehat{\theta}_n\|} - \frac{\theta^*}{\|\theta^*\|} \right\| \leqslant c\sqrt{\frac{d\log n + \log(1/\delta)}{Bn}}, \ \left| \|\widehat{\theta}_n\| - \|\theta^*\| \right| \leqslant cB^{3/2}\sqrt{\frac{d\log n + \log(1/\delta)}{n}}. \tag{10.15}$$

Bien que les bornes (10.15) contrôlent les erreurs d'estimation de la norme et de la direction du paramètre, on peut également les reformuler en termes d'excès de risque logistique :

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c' \, \frac{d\log n + \log(1/\delta)}{n} \tag{10.16}$$

pour une certaine constante $c' > 0$. Cette garantie correspond au risque asymptotique (10.11) à un facteur $\log n$ près, et comme nous le verrons ci-dessous, la condition d'existence du MLE donnée dans [KvdG23] est aussi presque optimale à des facteurs logarithmiques près. Notons également que d'autres résultats sur la séparation linéaire dans des contextes plus généraux ont été obtenus par Kuchelmeister [Kuc24].

Hsu et Mazumdar [HM23] considèrent le problème d'estimation de la direction du paramètre, $\theta^*/\|\theta^*\|$ (ce qui suffit pour la classification, c'est-à-dire la prédiction de la valeur la plus probable de Y connaissant X, contrairement à l'estimation de probabilités conditionnelles), là encore avec un accent mis sur la dépendance à la force du signal B. Comme [KvdG23], ils considèrent le cas d'un design gaussien, mais supposent cette fois que les données suivent un modèle logit plutôt qu'un modèle probit. Ils considèrent notamment des estimateurs différents du MLE pour la régression logistique, en particulier le minimiseur du risque empirique de classification. Ils établissent des bornes supérieures sur l'erreur d'estimation du même ordre que la première borne de (10.15), encore une fois avec des facteurs logarithmiques en n. Ils établissent également des bornes inférieures minimax sur l'erreur d'estimation de $\theta^*/\|\theta^*\|$, qui montrent que la borne supérieure précédente est optimale à des facteurs logarithmiques près. Ils posent aussi explicitement la question de savoir si le MLE atteint ces bornes optimales.

Bien que ces résultats représentent des avancées décisives, ils laissent ouvertes certaines questions importantes. Premièrement, les garanties font apparaître des facteurs logarithmiques supplémentaires en la taille d'échantillon, qui sont vraisemblablement sous-optimaux mais semblent difficiles à éviter dans les analyses de [KvdG23] et [HM23], laissant un écart entre bornes supérieures et inférieures. Même si les facteurs logarithmiques représentent une forme bénigne de sous-optimalité, la régression logistique avec un design gaussien est sans doute un problème suffisamment élémentaire pour justifier la recherche de résultats précis. Deuxièmement, et de manière peut-être plus importante encore, ces résultats sont spécifiques au cas d'un design gaussien et d'un modèle bien spécifié, ce qui soulève la question du comportement du MLE pour des designs plus généraux ou dans le cas de modèles mal spécifiés.

Dans tous les résultats discutés ci-dessus, la dépendance en la dimension et en la taille d'échantillon est essentiellement optimale (à des facteurs logarithmiques près), et la sous-optimalité provient de la dépendance en la force du signal B, dans le régime à fort signal. C'est pourquoi, dans cette thèse, nous nous concentrons sur la dépendance en B, en particulier lorsque ce dernier est grand. C'est en effet dans ce régime que les prédictions probabilistes du MLE deviennent peu fiables, car la force du signal elle-même est difficile à estimer.

10.2.4 Classification binaire

Jusqu'à présent, nous avons discuté des résultats établis sur les performances du MLE en termes de perte logistique, qui constitue une mesure naturelle de la qualité pour l'estimation des probabilités conditionnelles. Comme mentionné plus haut, la perte logistique est également un substitut convexe naturel à la perte binaire en classification supervisée. Elle satisfait à la définition de *calibration* introduite par [BJM06], généralisant les résultats importants de [Zha04], et entre ainsi dans le cadre du célèbre lemme de Zhang. Ce résultat permet de majorer l'erreur de classification d'un prédicteur en fonction de son excès de risque pour la perte convexe. Dans le cas de la régression logistique, sous l'hypothèse que le modèle est bien spécifié et pour un design gaussien, on obtient ainsi la borne suivante :

$$R(\theta) - R(\theta^*) \leqslant C\sqrt{L(\theta) - L(\theta^*)}\,, \tag{10.17}$$

où $R(\theta)$ désigne le risque de classification du prédicteur linéaire associé au vecteur θ, c'est-à-dire $R(\theta) = \mathbf{P}(Y\langle\theta, X\rangle < 0)$, et C est une constante universelle. En utilisant les résultats non asymptotiques sur l'excès de risque logistique du MLE, par exemple (10.14) de [OB21], on en déduit un risque de classification qui décroît comme $\sqrt{d/n}$, ce qui correspond à la vitesse lente typique donnée par la théorie de Vapnik-Chervonenkis. En résumé, (10.17) garantit que le MLE $\widehat{\theta}_n$ est consistant lorsqu'il est utilisé comme prédicteur en classification binaire, mais en termes de bornes quantitatives, ce résultat dégrade significativement les vitesses obtenues pour l'excès de risque logistique présentés dans la section précédente.

Nous soulignons que, contrairement à la section précédente où nous discutions de la précision des résultats vis-à-vis de la force du signal B, le problème ici concerne bien la vitesse de convergence en la taille de l'échantillon n ; alors que dans la section précédente, les dépendances en la dimension d et en la taille d'échantillon n étaient optimales (à des facteurs logarithmiques près), seule la dépendance en la force du signal laissait une marge d'amélioration.

10.3 Résumé des contributions

Nous sommes maintenant en mesure de fournir une vue d'ensemble de haut niveau de nos résultats principaux ; nous renvoyons aux chapitres correspondants pour les énoncés précis et les commentaires supplémentaires.

Design gaussien, modèle bien spécifié. Tout d'abord, dans le cas d'un design gaussien et d'un modèle logit bien spécifié, le Théorème 3.1 fournit des garanties optimales (à des constantes absolues près) pour l'existence et la précision du MLE. Plus précisément, il existe une constante universelle c telle que : pour tout $\delta \leqslant 1/2$, si $n \leqslant c^{-1}B(d + \log(1/\delta))$, alors

$$\mathbf{P}(\text{MLE existe}) \leqslant 1 - \delta\,. \tag{10.18}$$

En revanche, si $n \geqslant cB(d + \log(1/\delta))$, alors avec probabilité *au moins* $1 - \delta$, le MLE existe et satisfait

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c\,\frac{d + \log(1/\delta)}{n}\,. \tag{10.19}$$

Cela supprime un facteur $\log n$ de la borne (10.16) déduite du travail de [KvdG23] dans le cas d'un modèle probit, et répond par l'affirmative (après traduction de cette borne de risque en une borne sur l'erreur d'estimation) à une question de [HM23] concernant l'optimalité du MLE.

En résumé, ce résultat fournit des conditions nécessaires et suffisantes sur la taille d'échantillon n (à des constantes numériques près) pour que le MLE existe avec forte probabilité, et montre que dans le régime où le MLE existe, il atteint de manière non asymptotique le même risque que celui prédit par le comportement asymptotique (10.11) pour B, d, δ fixes et $n \to \infty$.

Ce résultat implique en particulier que, si $n \gg Bd$, alors le MLE existe avec probabilité au moins $1 - \exp\big(-\frac{n}{c'B}\big)$ pour une certaine constante c', et que cette estimée est optimale. Cela fournit une version quantitative de la convergence vers 1 dans la transition de phase (10.12) de Candès et Sur [CS20]. En revanche, dans le régime où $n \ll Bd$, le Théorème 3.1 montre seulement que la probabilité d'existence du MLE est bornée par une constante (disons, $1/2$), plutôt que de converger vers 0 comme dans la transition de phase (10.12). Nous complétons donc le Théorème 3.1 par un résultat de non-existence du MLE (Théorème 4.2), qui affirme que si $n \ll Bd/\kappa$ pour un certain $\kappa \geqslant 1$, alors

$$\mathbf{P}(\text{MLE existe}) \leqslant c\exp\big(-\max\big\{\kappa\sqrt{d}, \kappa^2 d/B^2\big\}/c\big) \tag{10.20}$$

pour une certaine constante c. Cela peut être vu comme une version quantitative de la convergence vers 0 dans la transition de phase (10.12) de [CS20].

Design régulier, modèle bien spécifié. Les résultats précédents sont spécifiques au cas d'un design gaussien, que l'on peut considérer comme le cas le plus favorable. Cela soulève la question naturelle suivante : quelles propriétés de la loi gaussienne sont responsables du comportement du MLE décrit précédemment ? De manière équivalente, pour quels designs le MLE se comporte-t-il de manière similaire au cas gaussien, au moins dans le cas bien spécifié ?

Une intuition naturelle serait qu'un design à queues légères conduise à un comportement similaire au design gaussien, ce qui est effectivement le cas en régression linéaire.

Cependant, ceci est loin d'être vrai en régression logistique : comme mentionné précédemment, si le design est seulement supposé sous-gaussien (comme dans [CLL20]), alors une dépendance exponentielle en la norme est inévitable.

Dans le chapitre 5, nous identifions des hypothèses appropriées sur le design conduisant à un comportement proche de celui du cas gaussien. En plus des queues légères (Hypothèse 5.1), les hypothèses incluent une condition sur le comportement des projections linéaires unidimensionnelles du design au voisinage de 0 (Hypothèse 5.2), qui est liée aux conditions de marge classiques dans la littérature sur la classification [MT99, Tsy04]. Cependant, comme montré dans la Proposition 5.1, une autre hypothèse est nécessaire pour obtenir un comportement proche de celui du cas gaussien (en un sens approprié) ; cette condition non standard porte sur les projections linéaires *bidimensionnelles* du design, plutôt que simplement ses marginales unidimensionnelles. Par analogie avec la condition de marge unidimensionnelle classique, nous appelons cette condition « hypothèse de marge bidimensionnelle ».

Sous ces hypothèses de régularité sur le design mais toujours dans le cas bien spécifié, le Théorème 6.1 montre que le MLE se comporte de manière similaire au cas gaussien, à des facteurs poly-logarithmiques près en la norme B. Plus précisément, pour une certaine constante c (dépendant des constantes des conditions de régularité), si $n \geqslant c\log^4(B)B(d+\log(1/\delta))$, alors avec probabilité $1-\delta$, le MLE existe et satisfait

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c\log^4(B)\,\frac{d+\log(1/\delta)}{n}\,. \tag{10.21}$$

Design régulier, modèle mal spécifié. Enfin, nous abordons le cadre le plus général, dans lequel aucune hypothèse n'est faite sur la loi conditionnelle de Y sachant X ; en particulier, nous ne supposons plus qu'elle appartient au modèle logit. Cela étant dit, comme discuté précédemment, il est toujours possible de définir le minimiseur θ^* du risque logistique, et de s'intéresser à l'excès de risque $L(\widehat{\theta}_n) - L(\theta^*)$ du MLE, qui correspond au minimiseur du risque empirique (ERM pour l'acronyme anglais) associé à la perte logistique. Cela correspond au problème de l'apprentissage statistique avec la perte logistique.

Comme discuté dans la Section 10.2, dans de nombreux régimes d'intérêt, la meilleure garantie disponible dans la littérature pour ce problème est celle de [OB21], à savoir la borne (10.14) sur le risque excédentaire en $B^3 d\log(1/\delta)/n$, lorsque le design est gaussien mais que le modèle peut être mal spécifié. Le Théorème 6.2 (Chapitre 6) améliore ces garanties de la manière suivante : il montre que si le design est régulier (au sens déjà défini) et $n \geqslant c\log^4(B)(Bd+B^2\log(1/\delta))$, alors avec probabilité au moins $1-\delta$, le MLE existe et satisfait

$$L(\widehat{\theta}_n) - L(\theta^*) \leqslant c\log^4(B)\frac{d+B\log(1/\delta)}{n}\,; \tag{10.22}$$

ici comme dans (10.21), c est une constante dépendant des constantes intervenant dans les conditions de régularité sur X. Par exemple, pour un δ constant et $B \lesssim d$, cela supprime un facteur d'environ B^3 par rapport à la meilleure garantie précédente (10.14) pour l'apprentissage statistique avec la perte logistique.

Nos garanties dans le cas mal spécifié présentent une dépendance plus forte en la norme B que dans le cas bien spécifié ; en particulier, les termes de « déviation »(ceux qui dépendent de la probabilité d'échec δ) intervenant à la fois dans la condition d'existence du MLE et dans la borne sur le risque excédentaire sont amplifiés d'un facteur B. Cela soulève la question de savoir si cet écart est essentiel ou bien un artefact de l'analyse.

Il s'avère que même dans le cas d'un design gaussien, la garantie (10.22) est optimale (à des facteurs polylog(B) près) dans le cadre général mal spécifié, à la fois pour la condition d'existence du MLE et pour la borne sur son excès de risque, comme le montre le Théorème 6.2.

Classification binaire. On se place ici dans le cadre de la classification binaire, discuté en section 10.2.4. La vitesse lente en $\sqrt{d/n}$ que donne le lemme de Zhang (combiné aux résultats sur l'excès de risque logistique du MLE) est pessimiste pour deux raisons. Premièrement, comme discuté précédemment, les résultats permettant de contrôler l'excès de risque impliquent entre autres choses une majoration de l'erreur d'estimation de la direction du paramètre θ^*. Cette erreur est naturellement liée à celle de classification binaire, ce qui suggère de meilleures vitesses atteignables, notamment par le MLE. Deuxièmement, dans le cas d'un modèle gaussien bien spécifié, on peut montrer que la condition de marge de Tsybakov [MT99, Tsy04] est vérifiée avec exposant 1. Ceci suggère une meilleure vitesse pour la classification, d'ordre $(d/n)^{2/3}$.

Nous montrons dans le chapitre 7 que le MLE atteint en fait la vitesse optimale d/n dans le cas gaussien, et quasiment optimale $\frac{d}{n}\log^2(\frac{n}{d})$ dans le cas régulier.

Lois satisfaisant les hypothèses de régularité. Bien que l'identification des conditions sur le design assurant un comportement du MLE similaire au cas gaussien puisse être une question intéressante en soi, ces hypothèses générales ne constituent une véritable généralisation du cas gaussien que si l'on peut exhiber d'autres exemples pertinents de lois les satisfaisant.

Pour illustrer ces conditions, nous considérons dans le chapitre 5 deux familles de lois : les lois log-concaves et les mesures produit. Le cas log-concave est globalement similaire au cas gaussien (qu'il contient comme cas particulier), en ce sens que les hypothèses de régularité sont vérifiées pour toute valeur du paramètre $\theta^* \in \mathbb{R}^d$ — c'est-à-dire pour toute direction $u^* = \theta^*/\|\theta^*\|$ et toute intensité de signal $B = \max\{e, \|\theta^*\|\}$. En revanche, le cas des mesures produit est plus subtil, car les conditions de régularité sont fortement sensibles à la direction du paramètre u^*. En effet, nous montrons que pour des designs à coordonnées i.i.d. (un exemple canonique étant le *design de Bernoulli*, avec coordonnées i.i.d. uniformes sur $\{-1, 1\}$), selon la direction du paramètre, le MLE peut se comporter comme dans le cas gaussien soit uniquement pour une intensité de signal triviale (constante) $B = O(1)$, soit jusqu'à une grande intensité de signal $B = O(\sqrt{d})$.

10.4 Travaux connexes supplémentaires

Nous présentons à présent d'autres travaux antérieurs pertinents sur la régression logistique, au-delà des résultats discutés dans la Section 10.2.

Régression logistique comme apprentissage statistique convexe. La régression logistique est un cas particulier d'apprentissage statistique convexe (ou d'optimisation stochastique convexe), ce qui permet de bénéficier des garanties issues de ce cadre. Par exemple, un argument classique de convergence uniforme basé sur la propriété de Lipschitz de la perte logistique implique une borne de risque excédentaire en $B\sqrt{d/n}$ pour l'ERM sur une boule de rayon $O(B)$. Cette borne supérieure présente un taux de convergence lent en $n^{-1/2}$ lorsque $n \to \infty$, contrairement au taux asymptotique réel en n^{-1}.

Pour améliorer ce taux lent, il faut renforcer la simple convexité en posant des hypothèses supplémentaires sur la courbure de la perte ou du risque. Une notion classique de courbure en optimisation est la forte convexité [BV04, Bub15, Bac24] de la perte, mais la perte logistique n'est pas fortement convexe par rapport au paramètre de régression θ, car elle ne varie que dans une seule direction. Une notion plus appropriée est celle de "concavité exponentielle" (exp-concavité), qui provient de l'apprentissage en ligne [Vov98, CBL06]. En exploitant cette propriété de la perte logistique, il est montré dans [PZ23] (voir aussi [Meh17, DVW21]) que l'ERM contraint à une boule de rayon $O(B)$ atteint un risque excédentaire d'au plus $O(de^{cBR}/n)$, où $R > 0$ est tel que $\|X\| \leqslant R$ presque sûrement. Dans le cas isotrope où $\mathbf{E}[XX^\top] = I_d$, on a $R \geqslant \mathbf{E}[\|X\|^2]^{1/2} = \sqrt{d}$, de sorte que la garantie précédente croît au mieux comme $de^{cB\sqrt{d}}/n$, avec une dépendance exponentielle en la norme B et la racine carrée de la dimension d. Cela reflète la très faible courbure déterministe de la perte logistique.

Une autre propriété pertinente de la perte logistique est la (pseudo-)auto-concordance (c'est-à-dire une borne sur la dérivée troisième de la perte en fonction de la dérivée seconde), mise en avant par [Bac10] et utilisée pour analyser la régression logistique dans une série de travaux [Bac10, Bac14, BM13, OB21], les résultats les plus précis dans cette direction étant ceux d'Ostrovskii et Bach [OB21] discutés dans la Section 1.2.

Enfin, une condition classique pour obtenir des taux rapides pour l'ERM en théorie de l'apprentissage statistique est une borne sur la variance des différences de pertes en fonction du risque excédentaire [Mas07, Kol11], connue sous le nom de condition de Bernstein [BM06]. Des garanties générales pour l'ERM sous une perte convexe et lipschitzienne sont obtenues dans [ACL19] en utilisant cette condition. Ces résultats sont raffinés dans le travail [CLL20] grâce à une version locale de cette condition ; nous avons discuté l'application de leurs résultats à la régression logistique dans la Section 1.2.

Asymptotique en grande dimension. Comme discuté dans la Section 1.2, Candès et Sur [CS20] ont caractérisé la transition de phase pour l'existence du MLE dans le cas bien spécifié avec un design gaussien, dans le régime asymptotique de grande dimension où $d/n \to \gamma \in (0, 1)$ et $\beta = \|\theta^*\| \in \mathbb{R}^+$ est fixé. Ce résultat sur l'existence est complété dans [SC19] par une étude du comportement du MLE sous les mêmes hypothèses et dans le même régime asymptotique ; en particulier, il y est montré que la loi jointe des vrais et estimés coefficients converge vers une certaine distribution. Ces résultats ont été généralisés, entre autres, aux matrices de covariance arbitraires du design [ZSC22], à la régression logistique ridge [SAH19], à des modèles binaires plus généraux [TPT20], à la régression logistique multinomiale [TB24] et aux données manquantes [VM24].

Designs adversariels, estimateurs impropres et robustes. Notre objectif dans ce travail est de caractériser les performances du MLE dans des situations "régulières", c'est-à-dire lorsque la loi du design satisfait certaines conditions assurant un comportement proche du cas gaussien. Une perspective différente mais complémentaire consiste à considérer les performances du MLE (ou d'autres estimateurs) pour la régression logistique sous des lois de design adverses.

Comme on peut s'y attendre, les performances du MLE se détériorent considérablement dans le cas de lois de design adverses. En particulier, un résultat de [HKL14] pour l'apprentissage statistique avec la perte logistique implique que, lorsqu'aucune hypothèse n'est faite sur le design si ce n'est une borne presque sûre $\|X\| \leqslant R$, alors le MLE (ou tout estimateur "propre" qui retourne une densité conditionnelle appartenant au modèle logis-

tique) ne peut obtenir un excès de risque en moyenne (par rapport à une boule de rayon B) meilleur que $O(BR/\sqrt{n})$, tant que $n \leqslant e^{cBR}$. Cette dépendance exponentielle à la norme du paramètre peut être contournée en recourant à des estimateurs "impropres", c'est-à-dire des estimateurs qui retournent des densités conditionnelles ne relevant pas du modèle logistique ; cela inclut l'agrégation bayésienne [KN05, FKL+18, QRZ24] ou des estimateurs ajustés tenant compte de l'incertitude via des "étiquettes virtuelles" [MG22, JGR20]. Nous renvoyons aussi à [Vij21, vdHZCB23] pour d'autres procédures atteignant des garanties précises avec grande probabilité, mais au prix d'un coût computationnel élevé.

Une direction connexe est celle de la régression logistique robuste. Dans [CLL20], des bornes de risque avec grande probabilité sont établies pour des estimateurs basés sur des médianes de moyennes, lorsque le design X peut être à queue lourde. De plus, des estimateurs atteignant des garanties quasi-optimales en distance de Hellinger ont été proposés dans [BC24]. D'un point de vue statistique, ces estimateurs sont plus robustes que le MLE ; en revanche, leur coût computationnel semble exponentiel en la dimension, ce qui les rend moins utilisables en pratique.

Bibliography

[AA84] Adelin Albert and John A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[AC10] Jean-Yves Audibert and Olivier Catoni. Linear regression through PAC-Bayesian truncation. *Preprint arXiv:1010.0072*, 2010.

[AC11] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 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. *The Annals of Statistics*, 47(4):2117–2144, 2019.

[Ada08] Radosław Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

[ALMT14] Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[AT07] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

[AZ24] Pedro Abdalla and Nikita Zhivotovskiy. Covariance estimation: optimal dimension-free guarantees for adversarial corruption and heavy tails. *Journal of the European Mathematical Society*, 2024.

[Bac10] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[Bac14] Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

[Bac24] Francis Bach. *Learning Theory from First Principles*. MIT Press (forthcoming), 2024.

[BB08] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems 20*, pages 161–168, 2008.

[BBL05] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[BC24] Yannick Baraud and Juntong Chen. Robust estimation of a regression function in exponential families. *Journal of Statistical Planning and Inference*, 233:106167, 2024.

[Ber44] Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365, 1944.

[BJM06] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[BL19] Sergey G. Bobkov and Michel Ledoux. One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 2019.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[BM06] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and 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. *Comptes Rendus Mathématique*, 334(6):495–500, 2002.

[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.

[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, Cambridge, New York, USA, 2006.

[CG17] Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *Preprint arXiv:1712.02747*, 2017.

[CLL20] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with Lipschitz and convex loss functions. *Probability Theory and Related Fields*, 176:897–940, 2020.

[CLM24] Hugo Chardon, Matthieu Lerasle, and Jaouad Mourtada. Finite-sample performance of the maximum likelihood estimator in logistic regression. *arXiv preprint arXiv: 2411.02137*, 2024.

[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.

[Cox58] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242, 1958.

[CS20] Emmanuel J. Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 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. *Journal of Machine Learning Research*, 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. *Probability Theory and Related Fields*, 170(1):95–175, 2018.

[EK18b] Noureddine El Karoui. Random matrices and high-dimensional statistics: beyond covariance matrices. In *Proceedings of the International Congress of Mathematicians*, volume 4, pages 2875–2894, Rio de Janeiro, 2018.

[Fel68] William Feller. On the Berry-Esseen theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 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 *Proceedings of the 31st Conference on Learning Theory*, pages 167–208, 2018.

[Giu18] Ilaria Giulini. Robust dimension-free Gram operator estimates. *Bernoulli*, 24(4B):3864–3923, 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 *Proceedings of the 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.

[HM24] Daniel Hsu and Arya Mazumdar. On the sample complexity of parameter estimation in logistic regression with normal design. In *Thirty-Seventh Annual Conference on Learning Theory*, 2024.

[JGR20] Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. In *Proceedings of the 33rd Conference on Learning Theory*, pages 2085–2108. PMLR, 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 of Probability and its Applications*, 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. *The Annals of Probability*, 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.

[LV07] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

[Mas07] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, 2007.

[McA99] David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*, pages 164–170, 1999.

[Meh17] Nishant Mehta. Fast rates with high probability in exp-concave statistical learning. In *Proceedings of the 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. *Journal of Machine Learning Research*, 23(31):1–49, 2022.

[Mon18] Andrea Montanari. Mean field asymptotics in high-dimensional statistics: From exact results to efficient algorithms. In *Proceedings of the 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. *The Annals of Statistics*, 50(4):2157–2178, 2022.

[MT99] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[MZ25] Arshak Minasyan and Nikita Zhivotovskiy. Statistically optimal robust mean and covariance estimation for anisotropic Gaussians. *Mathematical Statistics and Learning*, 2025.

[OB21] Dmitrii Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 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, 1st edition, 2010.

[Oli16] Roberto I. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3):1175–1194, 2016.

[OR24] Roberto I. Oliveira and Zoraida F. Rico. Improved covariance estimation: optimal robustness and sub-gaussian guarantees under heavy tails. *The Annals of Statistics*, 52(5):1953–1977, 2024.

[PZ23] Nikita Puchkin and Nikita Zhivotovskiy. Exploring local norms in exp-concave statistical learning. In *Proceedings of the 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. *Proceedings of the National Academy of Sciences*, 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.

[Tal96] Michel Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103, 1996.

[Tal21] Michel Talagrand. *Upper and lower bounds for stochastic processes: Decomposition Theorems*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 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*, volume 36, pages 70892–70925, 2024.

[TPT20] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Sharp asymptotics and optimal performance for inference in binary models. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 3739–3749, 2020.

[Tro12] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[Tsy04] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[Tyu12] I. S. Tyurin. A refinement of the remainder in the Lyapunov theorem. *Theory of Probability and its Applications*, 56(4):693–696, 2012.

[Vap00] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[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 Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 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.

[Ver12] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, pages 210–268. Cambridge University Press, Cambridge, 2012.

[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. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

[Zha04] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

[Zhi24] Nikita Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, 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.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Théorie non asymptotique pour l'estimation par maximum de vraisemblance en régression logistique

Mots clés : Apprentissage statistique, Régression logistique, Grande dimension, Processus empiriques, Inégalités de concentration.

Résumé : Cette thèse étudie plusieurs aspects de la régression logistique. Ce modèle classique décrit la dépendance probabiliste d'une variable binaire à des covariables multivariées. Nous nous intéressons à l'existence et aux performances de l'estimateur du maximum de vraisemblance (EMV) dans ce modèle. L'existence de l'EMV est équivalente à l'absence de séparation linéaire des données. C'est donc une question purement géométrique. Les performances de l'EMV sont principalement mesurées par son excès de risque logistique, qui quantifie la précision des probabilités conditionnelles associées.
Nous établissons des garanties précises dans trois situations de généralité croissante ; avec un accent particulier sur le rôle de la force du signal. Nous commençons par le cas stylisé d'un design gaussien et d'un modèle bien spécifié. Dans ce cadre, nous montrons que l'EMV subit une transition de phase abrupte autour d'une taille critique d'échantillon. Ce seuil dépend de l'intensité du signal et de la dimension des données. En dessous de ce seuil, avec forte probabilité, l'EMV n'existe pas . Au-dessus, il existe avec forte probabilité et atteint la même borne de risque (optimale) que dans le régime asymptotique.
Nous généralisons ensuite cette analyse de deux façons. D'abord, nous considérons des designs non gaussiens, tout en conservant un modèle bien spécifié. Nous identifions des conditions nécessaires et suffisantes pour que l'EMV présente un comportement proche du cas gaussien. En particulier, nous introduisons une nouvelle *condition de marge bidimensionnelle*.
Ensuite, nous permettons au modèle d'être mal spécifié. La régression logistique est alors considérée comme un problème d'apprentissage statistique et le paramètre optimal défini comme le minimiseur du risque logistique en population. Dans ce cadre, les conditions d'existence du MLE et les bornes de risque associées sont dégradées par rapport au cas bien spécifié. Nous montrons toutefois que ces dégradations sont inévitables.
Notre approche repose sur un argument de localisation convexe. Elle utilise des inégalités de concentration précises pour des vecteurs aléatoires et des bornes uniformes pour une collection de matrices aléatoires.
Enfin, nous étudions les performances de l'EMV pour la classification. Nous améliorons la vitesse lente issue du lemme de Zhang, et montrons que l'EMV atteint la vitesse paramétrique optimale de classification.

Title : Finite-sample theory for maximum-likelihood estimation in logistic regression

Keywords : Statistical learninng, Logistic regression, High dimension, Empirical processes, Concentration inequalities

Abstract : This thesis studies several aspects of logistic regression, a classical model to describe the probabilistic dependence of a binary outcome on multivariate covariates. We study the maximum-likelihood estimator (MLE) in this model, investigating its existence and performance. The existence of the MLE is equivalent to the dataset not being linearly separated and is as such a purely geometric question. Performance is mostly measured by the excess logistic risk of the MLE, which measures the accuracy of its probabilistic forecasts,. We provide sharp guarantees in three cases of increasing order of generality, with a special emphasis on the signal strength, starting with the stylized case of a Gaussian design and a well-specified logistic model. We establish that the MLE undergoes a sharp phase transition around a critical sample size that depends on the signal strength and the dimension of the data. Below this threshold, with high probability, the MLE does not exist. Conversely, when the sample size exceeds this critical threshold, with high probability, the MLE exists and satisfies the same (optimal) risk bound as in the asymptotic regime.
Our approach is based on a convex localization argument and involves sharp deviation inequalities for random vectors and uniform bounds for a collection of random matrices.
When then generalize this in two ways. First, by considering non Gaussian designs (still with a well-specified model). We identify necessary and sufficient conditions for the MLE to exhibit a near-Gaussian behavior, and in particular a new *two-dimensional margin condition*, Second, we allow the model to be misspecified, and consider logistic regression as a statistical learning problem, where the optimal parameter is defined as the minimizer of the population logistic risk. In this setting, the condition on the sample size for the MLE to exist and the risk bound it satisfies are degraded compared to the well-specified case, but in a way that we prove to be unavoidable. Finally, we investigate the classification performance of the MLE, and improve the slow rate derived from Zhang's lemma to the optimal parametric rate.

Institut Polytechnique de Paris
91120 Palaiseau, France