université PARIS-SACLAY

Mathématiques Orsay

MASTER'S DEGREE IN MATHEMATICS - M2 MATHEMATICS OF RANDOMNESS

Finite sample behaviour of the maximum likelihood estimator in the Poisson model under Gaussian design

A non-asymptotic tour of the Poisson generalized linear model

Master's Thesis

submitted in partial fulfillment of the requirements for the degree of

Master's Degree in Mathematics - M2 Mathematics of randomness

AUTHOR

Antoine Roche

SUPERVISOR

Prof. Matthieu Lerasle

Laboratoire de Mathématiques d'Orsay

Université Paris-Saclay

2026

Summary

This thesis is an expository account of a recent line of work on the non-asymptotic behaviour of the maximum likelihood estimator (MLE) in Poisson regression with a standard Gaussian design, based on the preprint Finite sample behavior of the maximum likelihood estimator in the Poisson model under Gaussian design by M. Lerasle and A. Roche. Two questions organise the presentation. First, when does the MLE exist? The answer is a sharp phase transition at $n \asymp d$: below the threshold the estimator fails to exist almost surely, above it it exists with high probability, and, in sharp contrast with logistic regression, the location of the threshold does not depend on the signal strength $\|\beta_*\|$. Second, when it exists, how good is it? Here the answer involves a much larger sample size threshold, exponential in $\|\beta_*\|^2$, and this mismatch between the two thresholds creates an intermediate regime, specific to the Poisson model, in which the MLE provably exists but is not yet known to attain the asymptotic d/n rate. A separate result gives a localisation bound valid throughout this intermediate regime. The aim of the present text is pedagogical rather than encyclopaedic: complete proofs are deliberately not reproduced. Instead we isolate the four mechanisms that actually drive the results; exponential tilting of a log-normal weight, convex localisation, conic integral geometry, and PAC-Bayesian control of heavy-tailed empirical processes. Technical constants are suppressed throughout, and the reader is referred to the preprint for full arguments.

Acknowledgements

First and foremost, I would like to thank Matthieu, my supervisor. This work emerged from his initial idea and came to fruition over time during our discussions. His availability, insights, deep bibliographic knowledge and humor made working with him a pleasure, and I am excited to continue working with him and Guillaume Lecué during my PhD starting this fall.
I would also like to thank my mother and sister, with whom I spent most of this summer while writing this thesis, and who helped me get my mind off work. Last but not least, I'd like to mention my cats: Iggy, whose purrs and gentle rubs had a definite impact on this first work, and Réglisse, my best friend I'll never forget.

To my friends and my family.

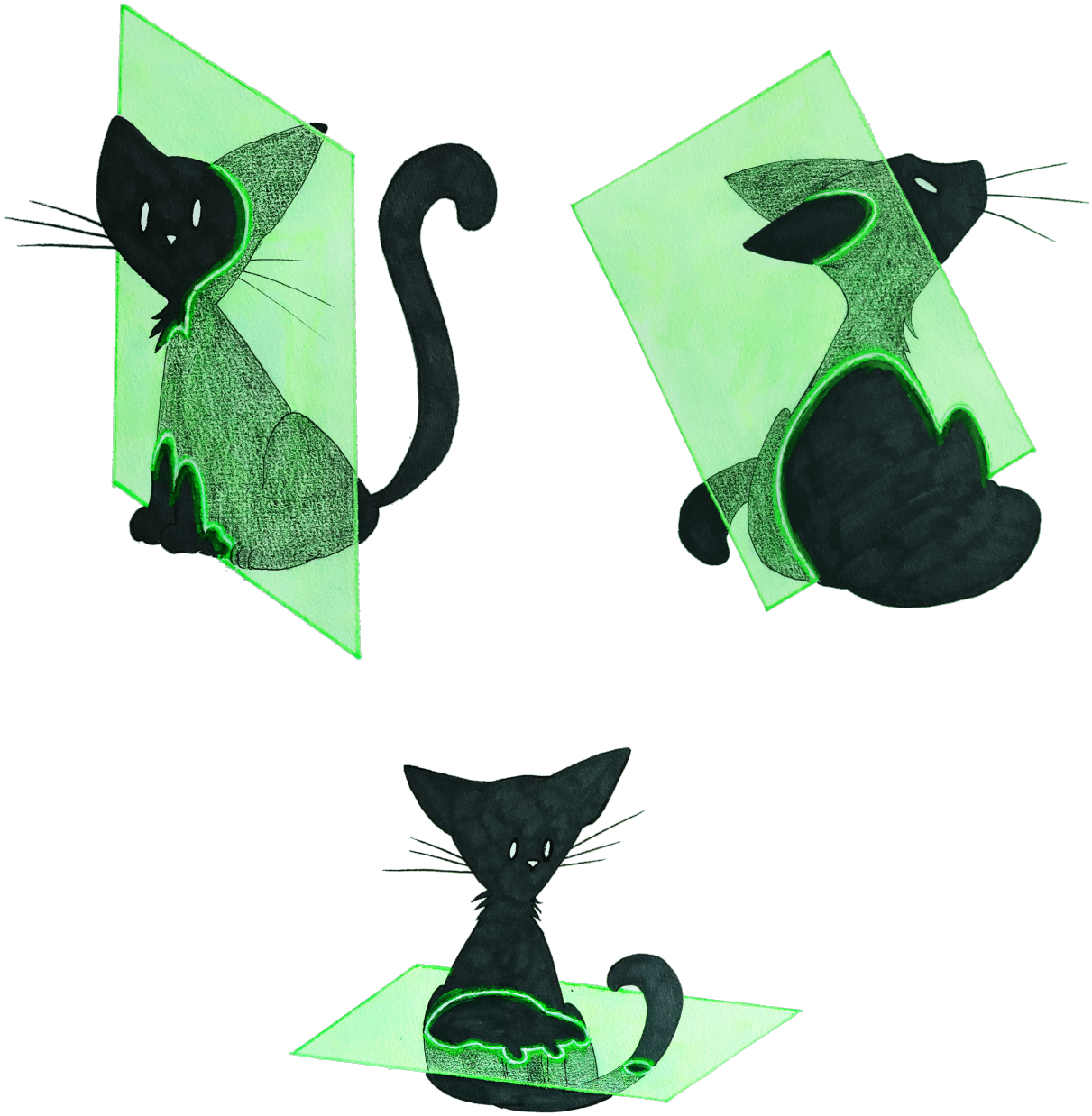

Figure 1: Conic section cats drawn by my sister.

Contents

Acknowledgements **iii**

1 Introduction **1**
1.1 Count data and the Poisson model . . . 1
1.2 The two questions . . . 1
1.3 Why the signal strength must be a free parameter . . . 2
1.4 What is specific to Poisson regression . . . 2
1.5 Contributions of the paper, and outline . . . 3

2 Setting, notation, and the asymptotic benchmark **4**
2.1 The model . . . 4
2.2 Loss, risk, and the estimator . . . 4
2.3 Gradient and Hessian . . . 5
2.4 Notation . . . 5
2.5 The asymptotic benchmark . . . 6
2.6 Related work . . . 6

3 Why the Poisson model is of interest **7**
3.1 The weights are log-normal, hence have no Laplace transform . . . 7
3.2 Exponential tilting: why the model is explosive . . . 7
3.3 What is left to work with . . . 8

4 Main results **9**
4.1 Existence of the MLE . . . 9
4.2 The large sample regime . . . 10
4.3 The intermediate regime . . . 11
4.4 The three regimes, and the gap that remains . . . 12

5 The machinery **12**
5.1 Convex localisation . . . 12
5.2 PAC-Bayes . . . 14

6 A guided tour of the proofs **16**
6.1 Existence: random cones . . . 16
6.2 Lower bound on the empirical Hessian . . . 18
6.3 Upper bound on the empirical Hessian . . . 19
6.4 Upper bound on the empirical gradient . . . 20
6.5 Assembling the pieces . . . 21

7 Numerical experiments **21**
7.1 The existence transition . . . 22
7.2 The risk threshold . . . 25
7.3 Wilks' theorem . . . 26
7.4 What the simulations do not show . . . 26

8 Which assumptions are structural? **27**
8.1 The design . . . 27
8.2 Well-specification . . . 28
8.3 Summary . . . 29

9 Conclusion and open problems **29**

A Notation **30**

B Two computations in dimension one **31**
B.1 The tilting integral 31
B.2 The proportion of zero counts 31

C AI usage disclosure **31**

1 Introduction

1.1 Count data and the Poisson model

Regression models for count data are among the oldest and most widely used tools in applied statistics. When the response Y takes integer values: a number of insurance claims, of particles detected, of citations received, of patents created by a given corporation, the Gaussian linear model is structurally wrong, as it puts mass on negative values and it imposes a variance that does not depend on the mean. Indeed, count data typically exhibit exactly the opposite behaviour, for instance, a company that develops hundreds of new patents per year will tend to see a bigger variation in the number of patents created yearly than a company that only develops a few. In this sense, the variance depends directly on the mean. Poisson regression, one of the founding examples of the theory of generalised linear models (GLMs) [19, 7], resolves both difficulties by positing

$$Y \mid X \sim \mathcal{P}\left(\exp(\langle \beta_*, X \rangle)\right),$$

so that the conditional mean and the conditional variance both equal $\exp(\langle \beta_*, X \rangle)$, and the parameter β_* acts on the log-scale. The model is used routinely in econometrics, epidemiology, engineering reliability and the social sciences.

The classical theory of this model is asymptotic and, in this sense, complete. If the dimension d and the parameter β_* are held fixed while the sample size n tends to infinity, then under mild regularity conditions on the model the MLE $\widehat{\beta}_n$ exists with probability tending to one, is consistent, is asymptotically normal with covariance the inverse Fisher information [27, 8], and its excess risk satisfies Wilks' theorem: $2n(L(\hat{\beta}_n) - L(\beta_*))$ converges in distribution to a $\chi^2(d)$ distribution. In particular the excess risk decays like d/n.

What this theory does not say is how large n must be, as a function of the different parameters of the problem, for these statements to be usable. That is the main question the present work addresses.

1.2 The two questions

Throughout, $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ are i.i.d. observations from the model, with a standard Gaussian design $X_i \sim \mathcal{N}(0, I_d)$, and $\widehat{\beta}_n$ denotes the maximiser of the likelihood, equivalently the minimiser of the empirical negative log-likelihood. Concretely, the aim of the paper can be broken down into three simple questions, the third one emerging naturally from these first two,

1. When does the MLE exist?
2. How large should n be, as a function of the other parameters of the problem, for the excess risk to be of the asymptotic order d/n with high probability?

It turns out that, informally, the MLE in the Poisson model exists with high probability whenever $n \gtrsim d$, while the asymptotic risk bound is attained for $n \gtrsim T(\|\beta_*\|, d)$ where T is a much larger threshold that depends on B, d. In consequence, there exists an "intermediate" regime $d \lesssim n \lesssim T(\|\beta_*\|, d)$ where the MLE exists with high probability, yet, the excess risk does not satisfy the asymptotic risk bound of order d/n. This leads to the third question,

3. What can we say about the MLE in the intermediate regime?

We emphasise once again that in the non-asymptotic setting, the questions are to be understood in a high probability sense, and not in an almost sure sense. While it is straightforward to see why the questions regarding non-asymptotic excess risk bounds are interesting, for instance from a practitioner standpoint (at least in their nature, because in our paper the numerical constants are too big for the results to present a practical interest); the existence question highlights a

genuine potential failure in a non-penalized setting. Indeed, the empirical negative log-likelihood of the Poisson model is convex but not always coercive, and when it is not, it simply has no minimiser. Santos Silva and Tenreyro [24] exhibited configurations of data in dimension one for which this happens and their simple setting already hints at the more general phenomenon in higher dimension, where it becomes a problem of random convex geometry. We introduce tools regarding phase transitions in convex programs studied by Amelunxen, Lotz, McCoy and Tropp [1] to solve this problem.

We now motivate why $\|\beta_*\|$ is treated as a key parameter throughout.

1.3 Why the signal strength must be a free parameter

A recurring feature of the results below is that the quantity

$$B = \max(2, \|\beta_*\|)$$

is treated as a third free parameter, on the same footing as n and d, with no restriction on its order relative to the other two. This deserves a word of justification, because much of the non-asymptotic literature either normalises $\|\beta_*\|$ away or treats it as a bounded constant absorbed into universal constants. In GLMs the signal strength is not a nuisance: it changes the geometry of the problem. We mention three examples, all from the linear models literature to emphasise the point. We refer the interested reader to the preprint by Lerasle and Roche [17] for a more elaborate discussion of each of the following examples, and more.

- In logistic regression, the celebrated result of Candès and Sur [9] shows that in the proportional asymptotics $d/n \to \gamma$, the MLE exists if and only if γ lies below a threshold $h(\|\beta_*\|)$, and h is decreasing. A strong signal makes the labels more predictable, hence makes linear separation, and therefore non-existence, more likely.
- In the probit model, Kuchelmeister and van de Geer [16] show that the signal-to-noise ratio determines which of two distinct estimation regimes one is in, with genuinely different rates, and that classification and SNR estimation trade off against each other.
- In Poisson regression, the signal strength enters the risk thresholds exponentially, through the factor $e^{B^2/2}$ (Theorems 4.2 and 4.3), while leaving the existence threshold untouched at constant order.

Normalising $\|\beta_*\|$ therefore does not simplify the problem but instead hides it, as B can be understood as a complexity parameter in our setting, much like the dimension.

1.4 What is specific to Poisson regression

The short answer is that the Poisson likelihood has explosive curvature. The empirical Hessian of the loss at a parameter β is

$$\nabla^2 \widehat{L}_n(\beta) \;=\; \frac{1}{n}\sum_{i=1}^{n} e^{\langle \beta,\, X_i\rangle} X_i X_i^\top, \tag{1}$$

a weighted sample covariance matrix whose weights $e^{\langle \beta,\, X_i\rangle}$ are log-normal under a Gaussian design. Log-normal random variables have no finite Laplace transform: $\mathbb{E}[\exp(s\, e^{Bg})] = +\infty$ for every $s > 0$ when $g \sim \mathcal{N}(0,1)$. Hence, standard concentration tools, which rely fundamentally on Laplace transform methods do not apply here directly (Bernstein, sub-Gaussian concentration, Orlicz norms for exponential families, etc.). Moreover, many approaches for similar problems from the finite sample GLM literature that rely on empirical processes, self-concordance in the sense of Ostrovskii and Bach [22], or even the general finite-sample machinery of Spokoiny [25]

do not apply here either as they all assume at least a local Laplace transform for the gradient and/or Hessian of the loss (once again, we refer to the associated preprint [17] for more discussion on this point). For instance, as mentioned by Ostrovskii and Bach in their paper, even in the case of a bounded design, the self-concordance approach cannot be applied in the case of Poisson regression.
Table 1 summarises what one might call the three corners of the GLM landscape and shows where Poisson differ from the logistic and linear models.

	Linear	Logistic	Poisson
Hessian weight	1	$\sigma'(\langle \beta, X \rangle) \in (0, \frac{1}{4}]$	$e^{\langle \beta, X \rangle}$
Effect of $\|\beta_*\|$ on curvature	flat	shrinking	explosive
Laplace transform (Gaussian design)	yes	yes	**no**
Existence threshold	$n \gtrsim d$	$n \gtrsim Bd$	$n \gtrsim d$
Risk threshold	$n \gtrsim d$	$n \gtrsim Bd$	$n \gtrsim e^{\frac{B^2}{2}(1+\varepsilon^{-1})} d^{1+\varepsilon}$

Table 1: The three corners. In linear regression the Hessian does not depend on β at all; in logistic regression the weights are bounded and are in fact small when the signal is strong; in Poisson regression they are unbounded and grow with the signal. Only in Poisson regression do the two thresholds fail to coincide.

The mismatch recorded in the last two rows of table 1 is one of the central phenomenon occuring in the model. Existence is a question of convex geometry; risk is a question of concentration. There is no a priori reason for the two to require the same sample size, and in logistic regression the fact that they do is, from this point of view, a coincidence. Since in Poisson regression the existence and risk thresholds do not match, the gap between them forms an intermediate regime in which the estimator exists with high probability but does not satisfy an asymptotic-type risk guarantee.

1.5 Contributions of the paper, and outline

The preprint [17] establishes three results, which we restate and discuss in Section 4:

- a sharp non-asymptotic existence threshold at $n \asymp d$, with almost sure non-existence below and high-probability existence above (Theorem 4.1);
- a sample size condition, exponential in B^2, above which the MLE attains the asymptotic excess risk $\lesssim (d+z)/n$ (Theorem 4.2);
- a localisation bound valid in the intermediate regime $n \gtrsim d^{1+\varepsilon}$, together with the excess risk bound it implies (Theorem 4.3).

Along the way, two PAC-Bayesian inequalities of independent interest are proved, for sub-Gamma random vectors and sub-Gamma random matrices, extending results of Gupta–Lee–Price and of Zhivotovskiy, respectively [14, 29].
The rest of this thesis is organised as follows. Section 2 fixes the model, the loss and the notation, and recalls what the asymptotic theory gives in greater precision. Section 3 explains in detail why the Poisson model resists the standard machinery; this section is the one to read if one wants to understand where the exponential factors come from. Section 4 states and discusses the three main results. Section 5 presents the two workhorses: the convex localisation lemma and the PAC-Bayes method, which strongly structure the proofs of the risk bounds results. Section 6 is a guided tour of the proofs, with the technical steps replaced by heuristics wherever possible. Section 7 is the most concrete section of the thesis, presenting numerical experiments

that directly benchmark the theoretical results we obtained, it aims to conclude on the overview of the paper. Section 8 discusses which assumptions are structural and which are conveniences. Section 9 collects open problems and directions for future works. The interested reader can find the full proofs and many more details throughout the original paper by Lerasle and Roche [17].

2 Setting, notation, and the asymptotic benchmark

2.1 The model

Fix a dimension $d \geqslant 1$. The Poisson model is the family of conditional distributions

$$\mathcal{P}_{\text{poisson}} = \{p_\beta : \beta \in \mathbb{R}^d\}, \quad p_\beta(y \mid x) = e^{-e^{\langle \beta,\, x\rangle}} \frac{e^{y\langle \beta,\, x\rangle}}{y!}, \quad (x, y) \in \mathbb{R}^d \times \mathbb{N}. \tag{2}$$

We say that (X, Y) follows the Poisson model, or that the model is well specified, if the conditional law of Y given X belongs to this family, i.e. if there exists $\beta_* \in \mathbb{R}^d$ with $Y \mid X \sim \mathcal{P}(\exp(\langle \beta_*,\, X\rangle))$. Throughout, the design is standard Gaussian, $X \sim \mathcal{N}(0, I_d)$.

Remark (Isotropy is free). Assuming $\mathrm{Var}(X) = I_d$ rather than a general $\Sigma \succ 0$ costs nothing in this setting. Writing $\langle \beta,\, X\rangle = \langle \Sigma^{1/2}\beta,\, \Sigma^{-1/2}X\rangle$ shows that the model with design $X \sim \mathcal{N}(0, \Sigma)$ and parameter β is exactly the model with design $\Sigma^{-1/2}X \sim \mathcal{N}(0, I_d)$ and parameter $\Sigma^{1/2}\beta$. Since no regularisation is used anywhere, every result below transfers verbatim, with the signal strength read as $\left\|\Sigma^{1/2}\beta_*\right\|$. It is worth noting that the signal strength is, in a sense, the signal strength "as seen by the design".

2.2 Loss, risk, and the estimator

Dropping the term $\log y!$, which does not depend on β, the negative log-likelihood of a single observation $(x, y) \in \mathbb{R}^d \times \mathbb{N}$ at $\beta \in \mathbb{R}^d$ is $\ell(y, \langle \beta,\, x\rangle)$ with

$$\ell(y, t) \;=\; e^t - y\, t. \tag{3}$$

Two elementary properties of ℓ will matter enormously:

- $t \mapsto \ell(y, t)$ is strictly convex for every $y \geqslant 0$;
- it is coercive (i.e. $\to +\infty$ as $|t| \to \infty$) if and only if $y > 0$. For $y = 0$ one simply has $\ell(0, t) = e^t$, which is still convex, but tends to 0 as $t \to -\infty$.

The second point is the entire content of the existence problem: observations with a zero count contribute a term that offers no resistance to sending the parameter off to infinity in a suitable direction, potentially precluding existence of the MLE. We will return to this in Section 6.1.
The population risk and its empirical counterpart are defined in the usual way by

$$L(\beta) = \mathbb{E}\big[\ell(Y, \langle \beta,\, X\rangle)\big], \quad \widehat{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Big[e^{\langle \beta,\, X_i\rangle} - Y_i\langle \beta,\, X_i\rangle\Big], \tag{4}$$

and the MLE is $\widehat{\beta}_n = \arg\min_{\beta \in \mathbb{R}^d} \widehat{L}_n(\beta)$, assuming a minimiser exists. As β_* is the population minimiser, the object of interest is the excess risk

$$L(\widehat{\beta}_n) - L(\beta_*) \;\geqslant\; 0.$$

2.3 Gradient and Hessian

Differentiating (4),

$$\nabla \widehat{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\big(e^{\langle \beta,\, X_i\rangle} - Y_i\big)X_i, \tag{5}$$

$$\nabla^2 \widehat{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} e^{\langle \beta,\, X_i\rangle} X_i X_i^\top. \tag{6}$$

Note that the Hessian does not involve the responses at all: it is purely a function of the design. This is an obvious fact but it has large consequences on the assumptions. Moreover, the modelling assumptions on Y enters at exactly one place, namely the control of (5), as the responses $(Y_i)_{1\leqslant i\leqslant n}$ do enter the gradient.
Under the Gaussian design, the population Hessian has a closed form which we will use constantly.

Lemma 2.1 (Population Hessian)**.** *For $X \sim \mathcal{N}(0, I_d)$ and any $\beta \in \mathbb{R}^d$, writing $u = \beta/\|\beta\|$,*

$$\nabla^2 L(\beta) \;=\; \mathbb{E}\big[e^{\langle \beta,\, X\rangle} X X^\top\big] \;=\; e^{\|\beta\|^2/2}\Big(\|\beta\|^2\, uu^\top + I_d\Big). \tag{7}$$

Sketch. By rotational invariance of $\mathcal{N}(0, I_d)$ in the hyperplane $u^\perp$, the matrix must be of the form $\lambda_1 uu^\top + \lambda_2(I_d - uu^\top)$. The two eigenvalues are one-dimensional Gaussian integrals: with $g \sim \mathcal{N}(0,1)$ and $b = \|\beta\|$, $\lambda_1 = \mathbb{E}[e^{bg}g^2] = e^{b^2/2}(1+b^2)$ and, using independence of coordinates, $\lambda_2 = \mathbb{E}[e^{bg}]\,\mathbb{E}[g^2] = e^{b^2/2}$. □

Formula (7) already hints at a first difficulty. In the direction of the signal the curvature is of order $B^2 e^{B^2/2}$, in every orthogonal direction it is of order $e^{B^2/2}$, and both explode with the signal strength. This anisotropy is why the natural geometry of the problem is not Euclidean but the one induced by the matrix

$$H \;:=\; e^{B^2/2}\big(B^2 u_* u_*^\top + I_d\big), \quad u_* := \beta_*/\|\beta_*\|, \tag{8}$$

which is a proxy for $\nabla^2 L(\beta_*)$, equivalent to it up to universal constants. We mention that this approach of using the geometry induced by the variance/curvature of the problem is very standard in arguments of this form, for instance, even though, as we mentioned, the general methodology of Spokoiny [25] cannot be applied there, using the geometry induced by the Hessian (or the covariance of the gradient under misspecification) is a central idea throughout that paper.

2.4 Notation

n and d denote the sample size and dimension; $B := \max(2, \|\beta_*\|)$ is the signal strength (the truncation at 2 is a convenience that avoids degeneracies when $\|\beta_*\|$ is small); S^{d-1} is the unit sphere of $\mathbb{R}^d$. $a \gtrsim b$ means $a \geqslant Cb$ for a universal constant C, and $a \asymp b$ means $a \gtrsim b \gtrsim a$. For (symmetric) matrices, $\preceq$ is the Loewner order, that is $A \preceq B$ whenever $B - A$ is positive semidefinite (we write $0 \preceq B - A$), $\succeq$ is the reverse order. For $A \succeq 0$ we write $\|x\|_A^2 = \langle Ax,\, x\rangle$ and denote by $B_A(x,r)$, $S_A(x,r)$ the corresponding ball and sphere. $\|A\|$ is the operator norm, $\mathrm{Tr}(A)$ the trace, and $\langle A,\, B\rangle_F = \mathrm{Tr}(A^\top B)$ the Frobenius inner product. Finally $z > 0$ is always a confidence parameter, statements holding with probability at least $1 - c\,e^{-z}$. A table of notation is collected in Appendix A.

2.5 The asymptotic benchmark

To know what a finite-sample bound should aim for, recall what happens when d, β_* and the law of (X, Y) are fixed and $n \to \infty$. Under mild regularity assumptions, the MLE exists with probability tending to one, is consistent, and $\sqrt{n}(\widehat{\beta}_n - \beta_*)$ converges to a centered Gaussian with covariance matrix the inverse Fisher information. Expanding the risk to second order around β_* and using this convergence yields Wilks' theorem,

$$2n\big(L(\widehat{\beta}_n) - L(\beta_*)\big) \xrightarrow{(d)} \chi^2(d). \tag{9}$$

Combining (9) with a standard deviation bound for the chi-square distribution [3] gives, for any $\eta > 0$,

$$\liminf_{n\to\infty} \mathbb{P}\left(L(\widehat{\beta}_n) - L(\beta_*) \leqslant \frac{d + 2\log(1/\eta)}{n}\right) \geqslant 1 - \eta. \tag{10}$$

This gives us our benchmark. And justifies why we call a bound of the form $L(\widehat{\beta}_n) - L(\beta_*) \lesssim (d+z)/n$ an "asymptotic-type excess risk bound", as it is exactly the finite-sample analogue of (10), and by (9) it cannot be improved in order. Question 2 (from Section 1.2) thus emerges naturally: how large must n be for it to hold?
It is worth stressing what (10) does not provide. It is a statement about a limit; it holds for fixed d and fixed β_*; and the rate of convergence in (9), which is what would tell us how large n must be, is not part of the statement and in general depends on d and $\|\beta_*\|$ in an unspecified way. Turning (10) into a usable statement is the whole objective of Theorem 4.2.

2.6 Related work

We briefly situate the results. This subsection can be skipped on first reading and is roughly a shortened version of the literature review of our preprint [17].

Asymptotics and econometrics. Beyond the classical theory (van der Vaart [27]; Le Cam and Yang [8]), Poisson regression has been studied intensively in econometrics, often under misspecification. Gouriéroux, Monfort and Trognon [13] showed that the Poisson quasi-MLE remains consistent and asymptotically normal in a much wider setting, provided only that the conditional mean is correctly specified. This robustness property explains the popularity of the "Poisson pseudo-maximum-likelihood" estimator in trade economics. Giles and Feng [12], and Chen and Giles [10], computed bias and MSE expansions up to $O(\text{poly}(n^{-1}))$ and proposed almost-unbiased corrections; these results are asymptotic in nature.

Existence. Santos Silva and Tenreyro gave explicit data configurations for which the Poisson MLE fails to exist [24]. Koriyama and Bellec [15] characterised existence for unregularised M-estimators in single-index models in terms of a separation condition, and connected the question to the conic phase transitions of Amelunxen, Lotz, McCoy and Tropp [1], their result holds in some specific high-dimensional asymptotics, but emcompasses much more than just Poisson regression. Candès and Sur [9] determined the exact high-dimensional asymptotic phase transition for logistic regression.

Finite-sample GLM theory. Chardon, Lerasle and Mourtada [11] obtained sharp non-asymptotic results for logistic regression, tracking $\|\beta_*\|$ explicitly; their work is the closest analogue of the one reviewed here, and the natural point of comparison. Ostrovskii and Bach [22] developed a self-concordance-based finite-sample analysis of M-estimators; Spokoiny [25] developed a general theory allowing misspecification. Both require exponential moment conditions on the gradient that fail for the Poisson model, even under a bounded design (see Section 3). Kuchelmeister and van de Geer [16] studied the probit model with the signal to noise ratio as a free parameter.

PAC-Bayes. The method goes back to McAllester [18] and was developed extensively by Catoni and co-authors [5, 6]; it has become a standard tool for dimension-free matrix and vector concentration through the work of Oliveira [23], Mourtada [20] and Zhivotovskiy [29]. The two inequalities proved in the preprint extend Gupta, Lee and Price [14] (vectors) and Zhivotovskiy [29] (matrices).

3 Why the Poisson model is of interest

This section contains no theorems. Its purpose is to explain in an reader-friendly way, where the Poisson model is unique (compared to logistic regression and linear regression for instance). And where the core of the problem lies. Hopefully, this section serves two purposes: making the results more digestible, and giving a teaser of the obstacles to overcome.

3.1 The weights are log-normal, hence have no Laplace transform

An important part of the difficulty is contained in the random variable

$$W = e^{\langle \beta,\, X\rangle}, \quad X \sim \mathcal{N}(0, I_d),$$

which, for $\beta \neq 0$, writing $b = \|\beta\|$ and $g = \langle \beta/b,\, X\rangle \sim \mathcal{N}(0,1)$, is the log-normal random variable e^{bg}. Log-normal random variables have finite moments of all orders:

$$\mathbb{E}\big[e^{kbg}\big] = e^{k^2 b^2/2} \quad (k \in \mathbb{N}).$$

However these moments grow like e^{k^2} rather than like $k!$ (which is the rate of growth of moments of sub-exponential random variables), consequently

$$\mathbb{E}\big[\exp(s\, e^{bg})\big] = +\infty \quad \text{for every } s > 0.$$

This single fact rules out, in one stroke, every argument whose first step is "bound the moment generating function of the summand", as mentioned in the introduction. Concretely:

- **Bernstein / Chernoff for the Hessian.** The summands $e^{\langle \beta,\, X_i\rangle}\langle v,\, X_i\rangle^2$ in (6) are not sub-exponential in any direction v, so no Chernoff bound can be applied to them directly.
- **Self-concordance (Ostrovskii–Bach [22]).** The Poisson loss is self-concordant in the sense of Nesterov and Nemirovski [21, 2]. Indeed $\partial_t^3 \ell = \partial_t^2 \ell = e^t$. But the finite-sample theory of Ostrovskii and Bach additionally requires the gradient and Hessian of the loss to be sub-Gaussian, and, conditionally on $(X_1, \dots, X_n)$, the gradient involves centered Poisson random variables which are not sub-Gaussian (the mean is never zero in this model).

This is why the non-asymptotic literature on Poisson regression is surprisingly sparse.

3.2 Exponential tilting: why the model is explosive

Consider the empirical Hessian (6), with $\|\beta_*\| = B$ and look at the curvature in the direction of the signal,

$$u_*^\top \nabla^2 \widehat{L}_n(\beta_*)\, u_* \;=\; \frac{1}{n}\sum_{i=1}^{n} e^{Bg_i} g_i^2, \quad g_i = \langle u_*,\, X_i\rangle \sim \mathcal{N}(0,1).$$

Its expectation is $\mathbb{E}[e^{Bg} g^2] = e^{B^2/2}(1 + B^2)$, in agreement with Lemma 2.1. The question is: where does this expectation come from? Write

$$\mathbb{E}\big[e^{Bg} g^2\big] = \int_{\mathbb{R}} e^{Bx} x^2\, \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx = e^{B^2/2} \int_{\mathbb{R}} x^2\, \frac{e^{-(x-B)^2/2}}{\sqrt{2\pi}}\, dx.$$

The integrand, after tilting, is a Gaussian density centered at B. In other words: the expectation of the exponential weight is entirely carried by the event

$$\{\, g \approx B \,\}, \quad \text{whose probability is of order } e^{-B^2/2}. \tag{11}$$

Observations contribute essentially nothing, except the rare ones for which $\langle u_*, X_i \rangle$ happens to be of order B; those contribute a weight of order e^{B^2}, and the balance $e^{-B^2/2} \cdot e^{B^2} = e^{B^2/2}$ reproduces the mean.
Two consequences follow immediately, and they are exactly the two exponential thresholds of Theorem 4.2. These examples motivate why B, the signal strength can be seen as a complexity parameter, much like the dimension d. Indeed, a larger B makes the Hessian harder to estimate, thus, a larger number of samples are needed.

Lower bound: $n \gtrsim e^{B^2/2}$. For the empirical Hessian to be at least of the correct order, at least one of the n observations must realise the rare event (11). Since that event has probability $\approx e^{-B^2/2}$, one needs

$$n \gtrsim e^{B^2/2}$$

merely to see it happen once, and somewhat more to see it happen often enough that the count concentrates. There is nothing one can do about this: it is a statement about the model, not about the proof technique. If $n \ll e^{B^2/2}$ then, with high probability, the sample simply does not contain the observations that carry the curvature, and the empirical Hessian cannot be guaranteed to be lower bounded (up to a numerical constant) by the population one with high probability.

Upper bound: $n \gtrsim e^{B^2}$. To bound the empirical Hessian from above by its mean one must smooth out the same rare events, which requires controlling the second moment,

$$\mathbb{E}\big[e^{2Bg} g^4\big] \asymp e^{2B^2} B^4.$$

A crude fluctuation heuristic, comparing the standard deviation of the average to its mean (this is the mechanism behind Bernstein's inequality), gives

$$\frac{\sqrt{\mathrm{Var}\,/n}}{\mathrm{mean}} \asymp \frac{e^{B^2}/\sqrt{n}}{e^{B^2/2}} = \sqrt{\frac{e^{B^2}}{n}} \ll 1 \iff n \gg e^{B^2},$$

which is a strictly stronger requirement than the previous one. The asymmetry in this heavy-tailed average is interesting: it is easier to guarantee that sufficiently many large values have been seen than to guarantee that potentially absurdly large values have been sufficiently smoothed out by the averaging process. In a sense, this is a discussion about the necessary scales for the law of large numbers to "kick in" from above and below.

Remark. Note that these are heuristics about a one-dimensional quantity, while the theorems concern a $d \times d$ matrix uniformly over a set of parameters. In consequence the actual proofs are much more delicate. However, the exponential factors in the final statements are exactly the ones produced by the two displays above; the empirical process machinery contributes the polynomial factors in d, which is natural for an inequality about a $d \times d$ matrix: d potential directions have to be controlled in our case.

3.3 What is left to work with

If Laplace transforms are unavailable, what tools remain? The preprint uses two, and the division of labour between them is worth stating explicitly, because it is the main methodological lesson of the work.

(i) **For lower bounds on the Hessian: bounded functions.** One may always discard mass. Since $e^{\langle \beta,\, X_i\rangle}\langle v,\, X_i\rangle^2 \geqslant c\,\mathbf{1}_{A_i}$ for a suitable event A_i and constant c, a lower bound on the Hessian reduces to a lower bound on an empirical process of indicator functions. Indicators are bounded, and for bounded classes of finite VC dimension the whole classical toolbox, such as Vapnik–Chervonenkis bounds [28], Talagrand's [26] and Bousquet's [4] inequalities applies without any moment assumption. This is how Lemma 6.1 and Lemma 6.2 are proven.

(ii) **For upper bounds: truncation plus PAC-Bayes.** Upper bounds cannot discard mass, so one truncates instead: on the event $\Omega_b = \{\max_i \langle u_*,\, X_i\rangle \leqslant \sqrt{2\log n + 2z}\}$, which has probability at least $1 - e^{-z}$ due to the Gaussian tails, the weights are bounded by $M = \exp(B\sqrt{2\log n + 2z})$, and the truncated variables do have a Laplace transform, with sub-Gamma parameters $(\sigma^2, b) \approx (e^{2B^2}, M)$. One then needs a concentration inequality that is uniform over a sphere or over a subspace and that degrades gracefully in the sub-Gamma scale parameter: this is what the PAC-Bayes inequalities of Section 5.2 provide.

The interplay between M and the target is what produces the peculiar-looking ε-dependent conditions of the theorems: the truncation level M grows like $e^{B\sqrt{2\log n}}$, i.e. slower than any power of n but not by much. Making the truncation error negligible amounts to solving an inequality of the type $\exp(B\sqrt{2\log n} - \log n) \leqslant \exp(-B^2/2)$, and the elementary bound $(a-b)^2 \geqslant \frac{a^2}{1+\varepsilon} - \frac{b^2}{\varepsilon}$ used to linearise it is exactly what introduces the free parameter ε and the factor d^{ε}.

4 Main results

We now state the three theorems, in each case followed by a discussion of what they mean. Universal constants are denoted $C_1, C_2, \ldots$ and their values, which are explicit but enormous, are omitted; the confidence parameter is $z > 0$.

4.1 Existence of the MLE

Theorem 4.1 (Existence)**.** *Assume $X \sim \mathcal{N}(0, I_d)$ and that the model is well specified.*

(a) If $n < d$, the MLE does not exist almost surely.

(b) There is a universal constant C_1 such that, for every $z > 0$, if

$$n \;\geqslant\; C_1\,(d \vee z), \tag{12}$$

then the MLE exists with probability at least $1 - 6e^{-z}$.

Thus $n \asymp d$ is both necessary and sufficient. Several comments are in order.

The transition is very sharp. Compare with logistic regression, where the analogous threshold is at $n \asymp Bd$, and where below the threshold the probability of non-existence decreases exponentially in d (which is very fast) but does not drop abruptly [11, Theorem 2]. In the Poisson model the threshold for high-probability existence coincides, up to constants, with the threshold below which non-existence is deterministic. There is no window in which non-existence is merely likely: one passes from probability zero to high-probability existence within a constant factor.

The signal strength has disappeared. This is the most striking feature, and it is worth understanding why. Recall that the loss term of observation i is coercive if and only if $Y_i > 0$. If there exists b_* such that $b_* \neq 0$ is orthogonal to every covariate X_i with $Y_i > 0$ and satisfies $\langle b_*,\, X_i\rangle \leqslant 0$ for every i with $Y_i = 0$, we say that b_* Poisson separates the data. In this case

moving along the ray tb_* keeps the coercive terms constant and does not increase the other terms, so $\widehat{L}_n$ has no minimiser. Conversely, if no such b_* exists, $\widehat{L}_n$ is coercive and a minimiser exists by compactness and continuity. So existence is governed by the number

$$|Z| \;=\; \#\{i : Y_i = 0\}$$

of non-coercive terms, and, conditionally on $|Z|$, the existence condition turns out to be

$$n \;\gtrsim\; d + |Z|. \tag{13}$$

Read (13) as: the sample size must be large enough to defeat two obstructions, the dimension (which produces separating directions deterministically when $n < d$) and the lack of coercivity (which is caused by too many zero counts).
Now, $|Z|$ is Binomial$(n, p(b))$ with

$$p(b) \;=\; \mathbb{E}\big[\mathbb{P}(Y = 0 \mid X)\big] \;=\; \mathbb{E}\big[\exp(-e^{bg})\big], \quad g \sim \mathcal{N}(0,1),\; b = \|\beta_*\|\,,$$

which is strictly increasing in b, with $p(0) = 1/e \approx 0.37$ and $p(b) \to 1/2$ as $b \to \infty$. A stronger signal therefore does produce more zero counts and does make existence harder, but only by a constant factor, since p stays in $[1/e, 1/2]$, and $|Z|$ concentrates around $np(b) \asymp n$. Substituting into (13) gives $n(1-p) \gtrsim d$, i.e. $n \gtrsim d$, with the signal strength absorbed into the constant. Contrast this with logistic regression, where B enters the threshold linearly: there, a strong signal makes the labels align with $\langle \beta_*, X_i \rangle$, which directly increases the probability that a separating hyperplane exists.

Existence is geometry, not statistics. The proof, sketched in Section 6.1, does not estimate anything: it computes the probability that a uniformly distributed random subspace of dimension $d-1$ intersects a given convex cone non-trivially, using the approximate kinematic formula of Amelunxen et al. [1]. Nothing about the accuracy of $\widehat{\beta}_n$ is involved. This is precisely why the threshold of Theorem 4.1 has no reason to coincide with the one in Theorem 4.2, and, in the Poisson model, it does not. We emphasise thus once again that the case of logistic regression, where both thresholds match, is the exception rather than the rule.

4.2 The large sample regime

Theorem 4.2 (Asymptotic-type risk bound)**.** *There are universal constants C_2, C_3 such that the following holds. If there exists $\varepsilon \in (0,1)$ with*

$$n \;\geqslant\; C_2 \exp\!\left(\frac{B^2}{2}\big(1+\varepsilon^{-1}\big)\right)\Big((d+z)\,e^{4B\sqrt{z}}\Big)^{1+\varepsilon}, \tag{14}$$

then, with probability at least $1 - 6e^{-z}$, the MLE exists and satisfies

$$L(\widehat{\beta}_n) - L(\beta_*) \;\leqslant\; C_3\,\frac{d+z}{n}. \tag{15}$$

The conclusion (15) is exactly the finite-sample form of the asymptotic benchmark (10), and by Wilks' theorem it is optimal in order. Everything of interest is therefore in the hypothesis (14).

Reading the condition. Ignore for a moment the confidence parameter, i.e. take $z \asymp 1$ (so that it gets absorbed by the constants). Then (14) reads

$$n \;\gtrsim\; e^{\frac{B^2}{2}(1+\varepsilon^{-1})}\, d^{1+\varepsilon},$$

and one is free to choose $\varepsilon \in (0,1)$ to make the right-hand side as small as possible. This exhibits a genuine trade-off between the two parameters: taking ε close to 1 costs a factor d but keeps the exponential factor at e^{B^2}; taking ε small keeps the polynomial factor close to d but blows up the exponential to $e^{B^2/2\varepsilon}$.

Where the exponentials come from. They come from Section 3.2, and nowhere else. To apply the localisation machinery one needs a uniform lower bound $\nabla^2 \widehat{L}_n(\beta) \succeq c_0 H$ near β_*, and by the tilting heuristic this requires $n \gtrsim e^{B^2/2}$; one also needs an upper bound at β_* in order to control the variance of the gradient, and that requires $n \gtrsim e^{B^2}$. The factor $e^{4B\sqrt{z}}$ is the price of the truncation level, and d^ε is the price of linearising the truncation inequality (making the last term in a Bernstein inequality negligible compared to the first two terms). It is important to be clear that the exponential factor is structural: it is a property of the Poisson model with a Gaussian design, not an artefact. One may reasonably conjecture that the correct threshold is $e^{B^2} d$ rather than $e^{\frac{B^2}{2}(1+\varepsilon^{-1})/2} d^{1+\varepsilon}$, but no condition free of e^{cB^2} can suffice for an excess-risk bound of asymptotic order.

Interpretation. A useful way to read (14) is as the sample size at which the law of large numbers takes effect for the exponential weights. Below it, the empirical Hessian is not a faithful picture of the population Hessian: it is either too small (missing the rare heavy observations) or wildly inflated (dominated by one of them). The MLE is then still perfectly well defined. Indeed, by Theorem 4.1 it exists as soon as $n \gtrsim d$, but the second-order Taylor picture that underlies (10) is not yet valid.

4.3 The intermediate regime

Theorem 4.3 (Localisation in the intermediate regime)**.** *Let $\varepsilon \in (0,1)$ and set $\delta = \frac{2\varepsilon}{2+\varepsilon} \leqslant 1$. There exist a universal constant C_4 and a constant K_ε depending only on ε such that, whenever*

$$n \;\geqslant\; K_\varepsilon \, d^{\,1+\varepsilon} \, e^{8z}, \tag{16}$$

one has, with probability at least $1 - 7e^{-z}$,

$$\big\|\widehat{\beta}_n - \beta_*\big\| \;\leqslant\; C_4 \, e^{B^2/2\delta} \sqrt{\frac{d+z}{n}}, \tag{17}$$

and consequently

$$L(\widehat{\beta}_n) - L(\beta_*) \leqslant f(B,\delta) \frac{d+z}{n}, \quad f(B,\delta) = 4\kappa^4 e^{2B^2/\delta} \exp\big(2\kappa^2 e^{B^2/\delta}\big), \tag{18}$$

for a universal constant κ.

The signal strength has moved from the hypothesis to the conclusion. This is one of the main structural points. Condition (16) does not involve B at all. What one buys with a small sample size is therefore not a weaker probability statement but a weaker conclusion: the localisation radius (17) carries the factor $e^{B^2/2\delta}$, which is exactly the ignorance about the curvature that one has not paid to resolve. This result on the intermediate regime is therefore only interesting in the case where B can be of arbitrary order compared to d and n. Indeed, if B is treated as a bounded constant, the problem simplifies and (17) is the familiar $\big\|\widehat{\beta}_n - \beta_*\big\| \lesssim \sqrt{(d+z)/n}$ with (18) that recovers the conclusions of Theorem 4.2, the dependence on B of the sample size being absorbed by universal constants.

Reading of (18). The excess risk bound is obtained from the localisation bound in a somewhat crude way: one bounds $\nabla^2 L$ deterministically and uniformly over the Euclidean ball of radius (17), and since $\nabla^2 L(\beta)$ grows like $e^{\|\beta\|^2/2}$, a ball of radius $R^{1/2}$ produces a factor $e^{R/2}$. With $R \asymp e^{B^2/\delta}$ this yields the doubly exponential $f(B,\delta)$. So (18) has the correct rate in n and d but a constant that is astronomically bad in B, however, this is the bound one obtains if the MLE behaves adversarially, i.e. is on the boundary of the localisation radius. The

informative statement of Theorem 4.3 is really the localisation bound (17). Moreover, any theoretical guarantee on the MLE in this intermediate regime (or even in the large n regime) is still a new result in the literature.

4.4 The three regimes, and the gap that remains

Putting the three theorems together gives the picture in Figure 2 (for a fixed $\varepsilon > 0$).

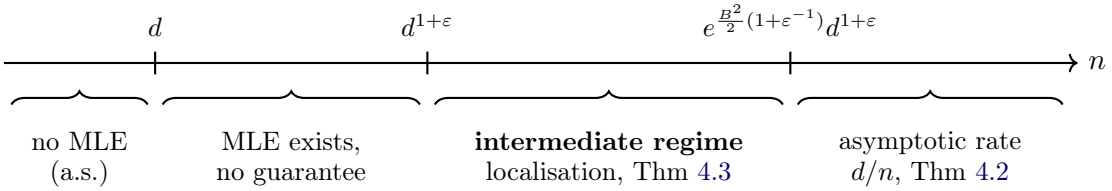

Figure 2: The three sample size regimes define four zones on the axis, d and B are fixed and the confidence parameter z is of constant order. In logistic regression the second and third zones are empty: the existence threshold and the asymptotic risk threshold coincide at $n \asymp Bd$.

Two gaps are visible. The first, between $d^{1+\varepsilon}$ and $e^{\frac{B^2}{2}(1+\varepsilon^{-1})/2}d^{1+\varepsilon}$, is the intermediate regime. It is a genuine feature of the model: because the existence question is geometric and the risk question is a concentration question, the two thresholds have no reason to agree, and here they differ by a factor that can be made arbitrarily large by increasing B. Theorem 4.3 fills it with a weaker but non-vacuous guarantee.
The second gap, between d and $d^{1+\varepsilon}$, is almost certainly an artefact, but one we were unable to remove. It persists even when B is a bounded constant, in which case all exponential factors are absorbed into universal constants and one is left with: the MLE exists for $n \gtrsim d$, but performance guarantees only start at $n \gtrsim d^{1+\varepsilon}$ for every $\varepsilon > 0$. Closing this gap, that is, obtaining a risk guarantee that is fully linear in d, is stated in the paper as an open problem. As can be seen in the axis, this would also lead to the full picture of the unpenalized Gaussian Poisson MLE in terms of existence and risk.

5 The machinery

Two general tools carry the two risk theorems. The first, convex localisation, is deterministic and reduces everything to two random quantities. The second, the PAC-Bayes method, is what allows those two quantities to be controlled in the absence of Laplace transforms. We present both in some detail, because they are transferable well beyond the Poisson model. Moreover, while the localisation approach is very standard, the PAC-Bayes type inequalities we develop and are of independent interest. When they apply, these inequalities give very clean deviation bounds.

5.1 Convex localisation

The following lemma is the backbone of the paper. It is entirely deterministic; the only structural assumptions are that $\widehat{L}_n$ is convex, that $\widehat{L}_n$ and L are twice continuously differentiable, and that β_* minimises L.

Lemma 5.1 (Convex localisation)**.** *Suppose there exist a positive definite matrix $H \in \mathbb{R}^{d\times d}$ and reals $r_0, c_0, c_1, \nu > 0$ such that:*

(i) $\nabla^2 \widehat{L}_n(\beta) \succeq c_0 H$ *for every* β *with* $\|\beta - \beta_*\|_H \leqslant r_0$ *(empirical curvature from below);*

(ii) $\left\|\nabla \widehat{L}_n(\beta_*)\right\|_{H^{-1}} \leqslant \nu$ *(gradient at the truth);*

(iii) $\nabla^2 L(\beta) \preceq c_1 H$ *for every* β *with* $\|\beta - \beta_*\|_H \leqslant r_0$ *(population curvature from above).*

If moreover $\nu < c_0 r_0/2$*, then* $\widehat{L}_n$ *admits a unique minimiser* $\widehat{\beta}_n$*, and*

$$\left\|\widehat{\beta}_n - \beta_*\right\|_H \;\leqslant\; \frac{2\nu}{c_0}, \quad L(\widehat{\beta}_n) - L(\beta_*) \;\leqslant\; \frac{2\nu^2 c_1}{c_0^2}. \tag{19}$$

Proof. The argument is short enough to give in full, and it is instructive.
Step 1: the loss increases on a sphere. Let $r \in (2\nu/c_0, r_0]$ and take β with $\|\beta - \beta_*\|_H = r$. A second-order Taylor expansion with the Lagrange remainder, followed by Cauchy–Schwarz in the H-geometry and assumption (i), gives

$$\widehat{L}_n(\beta) - \widehat{L}_n(\beta_*) \;\geqslant\; \langle \nabla \widehat{L}_n(\beta_*), \beta - \beta_* \rangle + \tfrac{1}{2} \inf_{\beta' \in B_H(\beta_*, r)} \|\beta - \beta_*\|^2_{\nabla^2 \widehat{L}_n(\beta')} \;\geqslant\; -\nu r + \frac{c_0 r^2}{2} \;>\; 0$$

whenever $\nu < c_0 r/2$. So $\widehat{L}_n$ is strictly larger than $\widehat{L}_n(\beta_*)$ everywhere on the H-sphere of radius r.
Step 2: convexity propagates this to the whole space. Let β' satisfy $\|\beta' - \beta_*\|_H = r' \geqslant r$ and set $t = r/r' \in (0, 1]$, so that $\beta = (1-t)\beta_* + t\beta'$ lies on the sphere. Convexity of $\widehat{L}_n$ gives $(1-t)\widehat{L}_n(\beta_*) + t\widehat{L}_n(\beta') \geqslant \widehat{L}_n(\beta) > \widehat{L}_n(\beta_*)$, whence $\widehat{L}_n(\beta') > \widehat{L}_n(\beta_*)$. Therefore $\inf_{\mathbb{R}^d} \widehat{L}_n = \inf_{B_H(\beta_*, r)} \widehat{L}_n$, and a minimiser exists by compactness and continuity. Strict convexity on the ball (implied by (i)) gives uniqueness, and the minimiser lies in $B_H(\beta_*, r)$. Letting $r \downarrow 2\nu/c_0$ gives the first half of (19).
Step 3: from localisation to risk. Since $\nabla L(\beta_*) = 0$, a Taylor expansion and assumption (iii) give $L(\beta) - L(\beta_*) \leqslant \frac{c_1}{2} \|\beta - \beta_*\|^2_H$ on the ball, and it remains to plug in the localisation bound. □

What the lemma buys. Lemma 5.1 is very interesting because it reduces the entire problem of localizing the MLE and bounding its excess risk to that of controlling two (three taking into account the deterministic upper bound on the population Hessian) relevant quantities. Existence and accuracy both follow from the single inequality $\nu < c_0 r_0/2$: informally, the gradient at the truth must be small compared with the curvature times the radius on which that curvature is guaranteed. To apply it one must:

(a) choose H, ideally with $H \asymp \nabla^2 L(\beta_*)$, which by Lemma 2.1 leads to the definition (8);

(b) determine the largest radius r_0 on which the population Hessian is still comparable to H, this step is purely deterministic, is similar to self-concordance computations, and it gives $r_0 \asymp e^{B^2/4}$ in the H-geometry;

(c) prove a high-probability uniform lower bound on $\nabla^2 \widehat{L}_n$ over $B_H(\beta_*, r_0)$;

(d) prove a high-probability upper bound on $\left\|\nabla \widehat{L}_n(\beta_*)\right\|_{H^{-1}}$.

Steps (c) and (d) are the probabilistic content, and they occupy Sections 6.2 and 6.4 below. Step (d) itself requires an upper bound on the empirical Hessian at β_*. This is not because the localisation lemma asks for one, but because the empirical Hessian is the conditional covariance of the empirical gradient; this is the role of Lemma 6.3. Moreover, note that the choice $H \asymp \nabla^2 L(\beta_*)$ in step (a) is motivated by the LLN. The heuristic being that if by step (b), we obtain that for all $\beta \in B_H(\beta_*, r)$, $\nabla^2 L(\beta) \asymp \nabla^2 L(\beta_*)$ for some $r > 0$ and we want to guarantee that on this same set $\nabla^2 \hat{L}_n(\beta) \succeq c_0 H$ and $\nabla^2 L(\beta_*) \preceq c_1 H$. Then for n large enough we have by the law of large numbers and step (b),

$$c_0 H \preceq \nabla^2 \hat{L}_n(\beta) \asymp \nabla^2 L(\beta) \asymp \nabla^2 L(\beta_*) \preceq c_1 H.$$

This is of course just a heuristic, but intuitively, this approach forces our hand for the choice of H in the large sample size regime, where the comparisons above do make sense.

Why $r_0 \asymp e^{B^2/4}$. It is worth doing this small computation, because it explains the form of the localisation radius. The smallest eigenvalue of H is $e^{B^2/2}$, so an H-ball of radius $re^{B^2/4}$ is contained in a Euclidean ball of radius r; moreover, in the signal direction the constraint is tighter, $|\langle \beta - \beta_*, u_* \rangle| \leqslant r/B$. By Lemma 2.1 the population Hessian at β involves the factor $e^{\|\beta\|^2/2}$, and

$$\|\beta\|^2 = \|\beta_*\|^2 + 2\langle \beta - \beta_*, \beta_* \rangle + \|\beta - \beta_*\|^2 \ \leqslant \ B^2 + 2B \cdot \frac{r}{B} + r^2 = B^2 + 2r + r^2,$$

so for $r \leqslant 1/2$ the exponent changes by $O(1)$ and $\nabla^2 L(\beta) \asymp \nabla^2 L(\beta_*)$. This is the content of Lemma B.3 of the preprint, which gives $\nabla^2 L(\beta) \preceq 45H$ on $B_H(\beta_*, \frac{1}{2}e^{B^2/4})$. The localisation set is a very anisotropic region in consequence: thin (of width $\asymp 1/B$) in the signal direction and of constant width in the orthogonal ones, and it is precisely the region on which the Poisson loss behaves like a quadratic (bounded above and below by H and the Taylor expansion is uniformly accurate there).

The modification used in the intermediate regime. In the intermediate regime the programme above fails at step (c): one cannot guarantee $\nabla^2 \widehat{L}_n \succeq c_0 H$ on the whole of $B_H(\beta_*, r_0)$, because that requires $n \gtrsim e^{B^2/2}$ by Section 3.2. The paper's answer is to change the geometry rather than the core of the argument:

- replace the sharp lower bound $\nabla^2 \widehat{L}_n \succeq c_0 H$ by the much weaker $\nabla^2 \widehat{L}_n \succeq c_0 I_d$, which throws away all dependence on B but, crucially, holds uniformly over all of $\mathbb{R}^d$, so that no radius restriction is needed;
- compensate by measuring the gradient in the stronger norm $\|\cdot\|_{H_0^{-1}}$, where

$$H_0 := e^{B^2/\delta}\, I_d, \quad \delta = \tfrac{2\varepsilon}{2+\varepsilon}, \tag{20}$$

  satisfies $H_0^{-1} \preceq H^{-1}$ (often by a lot).

Running Step 1 of the proof of Lemma 5.1 with these two ingredients, one in the Euclidean geometry, the other in the H_0 geometry, gives localisation as soon as $\nu < \frac{c_0}{2} \frac{\|\beta - \beta_*\|^2}{\|\beta - \beta_*\|_{H_0}}$, i.e. as soon as the Euclidean radius exceeds $\kappa\, e^{B^2/2\delta} \sqrt{(d+z)/n}$. This is exactly (17). The mismatch between the two geometries is what produces the factor $e^{B^2/2\delta}$: it can be seen as the price of not knowing the curvature precisely.

5.2 PAC-Bayes

In this section, we discuss the PAC-Bayes approach in obtaining deviation bounds as PAC-Bayes type inequalities form the technical core of the paper. We also present two new PAC-Bayes type inequalities that can be seen as extensions of methods proposed by Gupta et al. [14] (for random vectors) and Zhivotovskiy [29] (for matrices). In a few words, the PAC-Bayes method produces uniform deviation bounds by smoothing: instead of controlling $\sup_\theta f(X, \theta)$ directly, one controls $\mathbb{E}_{\omega \sim \rho_\theta}[f(X, \omega)]$ for a family of probability measures ρ_θ ("posteriors") simultaneously, which "smooth out" $f(X)$. The cost of such a method is a divergence term measuring how far each ρ_θ is from a fixed reference measure π (the "prior"). The choice of the family $(\rho_\theta)_\theta$ is entirely free (up to mild technical considerations), and that freedom is where the power of the method lies. We first present the generic PAC-Bayes inequality, which, at a glance, may not seem that interesting, yet, many of the more interesting PAC-Bayes type inequalities emerge directly from this lemma once specific priors and posterior families are plugged in.
Before stating the lemma, we clarify that Θ is a set that parametrizes the family of distributions (of ω) $(\rho_\theta)_{\theta \in \Theta}$ and that π is just another potential probability distribution for the random variable

ω. Finally, for X, ω independent random variables, we denote by $\mathbb{E}_{\rho_\theta}[f(X,\omega)] = \mathbb{E}[f(X,\omega) \mid X]$, i.e. the smoothed version of the process where only $\omega \sim \rho_\theta$ is integrated out (note that this is still a random variable that depends only on X after integrating ω out). $\mathbb{E}_X$ denotes the same thing but with X integrated out.

Lemma 5.2 (Generic PAC-Bayes inequality)**.** *Let X and ω be independent, π a prior on the range of ω, and $(\rho_\theta)_{\theta\in\Theta}$ a family of posteriors, each absolutely continuous with respect to π with density γ_θ. For any $z > 0$ and $s > 0$, with probability at least $1 - e^{-z}$, simultaneously for all $\theta \in \Theta$,*

$$\mathbb{E}_{\rho_\theta}\big[f(X,\omega)\big] \;\leqslant\; \frac{1}{s}\Big(\mathbb{E}_{\rho_\theta}\big[\log \mathbb{E}_X[e^{sf(X,\omega)}]\big] + \mathbb{E}_{\rho_\theta}\big[\log \gamma_\theta(\omega)\big] + z\Big). \tag{21}$$

Sketch. The change-of-measure with Jensen gives $\mathbb{E}_{\rho_\theta}[g] \leqslant \log \mathbb{E}_\pi[e^g] + \mathbb{E}_{\rho_\theta}[\log \gamma_\theta]$ for any g whose integral makes sense. We apply it to $g_X(\omega) = sf(X,\omega) - \log \mathbb{E}_X[e^{sf(X,\omega)}]$, which is constructed precisely so that $\mathbb{E}_X[e^{g_X(\omega)}] = 1$ for each fixed ω. Then Markov's inequality applied to $\log \mathbb{E}_\pi[e^{g_X}]$, together with Fubini, gives the claim. □

The two structural features to retain are: (1) the bound is uniform in θ at no extra cost, this is what replaces a union bound or a chaining argument; (2) it involves the Laplace transform of f only through ρ_θ-averages, so if the Laplace transform is finite only in a neighbourhood of the origin it suffices to keep the posteriors concentrated near the origin. The tension between "concentrated near the origin, to keep the Laplace term finite" and "spread out, to keep the divergence term small" is the core of the method, and it is exactly what is optimised in the proofs of the two theorems below.

Sub-Gamma vectors. The first inequality of the preprint generalises a result of Gupta, Lee and Price [14] by allowing an arbitrary seminorm to measure closeness to zero. Note that in our case, the theorem is applied to a "true" norm, however, the proof only uses the triangle inequality and the fact that $\|\lambda x\| = |\lambda| \, \|x\|$ for a scalar λ, hence we give the result for the slightly more general setting of an arbitrary seminorm.

Definition 5.3. *Let $\|\cdot\|_\star$ be a seminorm on $\mathbb{R}^d$ and $\Sigma \succ 0$. A centered random vector X is $(\star, \Sigma)$ sub-Gamma if*

$$\mathbb{E}\big[e^{\langle u,\, X\rangle}\big] \leqslant \exp\left(\frac{\|u\|_\Sigma^2}{2}\right) \quad \text{for every } u \text{ with } \|u\|_\star \leqslant 1.$$

Theorem 5.4 (PAC-Bayes for sub-Gamma vectors)**.** *Let X be $(\star, \Sigma)$ sub-Gamma and let $w_d = \mathbb{E}\big[\|g\|_\star\big]$, $\Delta_d = \sup_{u\in S^{d-1}} \|u\|_\star$ for $g \sim \mathcal{N}(0, I_d)$ be the Gaussian width and diameter of the dual ball. Then for any $z > 0$, with probability at least $1 - e^{-z}$,*

$$\|X\| \;\lesssim\; \sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{\|\Sigma\|\,(\log 2 + z)} + w_d\sqrt{\log 2 + z} + \Delta_d(\log 2 + z).$$

The first two terms are what one would get from a genuine sub-Gaussian vector; the last two quantify the degradation caused by the fact that the Laplace transform is only controlled in a $\|\cdot\|_\star$-neighbourhood of the origin. The proof takes ρ_θ to be a Gaussian $\mathcal{N}(\theta, \tau I_d)$ restricted to $\{\|\omega - \theta\|_\star \leqslant r\}$, chooses $r = 2w_d\sqrt{\tau}$ so that Markov's inequality keeps the normalising constant above $1/2$, and then optimises in (s, τ) across four regimes according to which of $\mathrm{Tr}\,\Sigma$, $\|\Sigma\|\,(\log 2 + z)$, $w_d^2(\log 2 + z)$ dominates.

Sub-Gamma matrices. The second inequality generalises a result of Zhivotovskiy [29] on sums of independent matrices, and is the tool used to control the empirical Hessian uniformly over a subspace.

Definition 5.5. *A centered random matrix X is (σ^2, b) sub-Gamma if for all unit vectors u, v and all $|s| < 1/b$, $\mathbb{E}[\exp(s\langle X,\, uv^\top\rangle_F)] \leqslant \exp(s^2\sigma^2)$.*

Theorem 5.6 (PAC-Bayes for sub-Gamma matrices)**.** *Let X be (σ^2, b) sub-Gamma, let $\Gamma_U, \Gamma_V \succ 0$ with associated ellipsoids $\mathcal{E}_U, \mathcal{E}_V$, and write $r(\Gamma) = \operatorname{Tr}(\Gamma)/\|\Gamma\|$ for the effective rank. Then with probability at least $1 - e^{-z}$,*

$$\sup_{u\in\mathcal{E}_U,\, v\in\mathcal{E}_V} u^\top X v \;\lesssim\; \sqrt{\|\Gamma_U\|\,\|\Gamma_V\|}\Big(\sigma\,\mathcal{C} \vee b\,\mathcal{C}^2\Big), \quad \mathcal{C} = \sqrt{r(\Gamma_U) + r(\Gamma_V) + 2(2\log 2 + z)}.$$

The proof applies Lemma 5.2 to $f(x, \omega) = \langle x,\, \omega\rangle_F$ with $\omega = UV^\top$ built from two independent Gaussian vectors, restricted to balls, and optimises. Two features matter for the application: the bound is dimension-free, depending on the effective ranks rather than on d; and the sub-Gamma scale b enters only in the second, quadratic term, which is what allows the truncation level M of Section 3 to be traded against n.

Remark (Control over a subspace). A small but essential corollary: if the sub-Gamma control is only available for (u, v) in a strict subspace $E \subsetneq \mathbb{R}^d$, one applies the theorem to $P^\top X P$ where P is the orthogonal projector on E, using $\langle X,\, Puv^\top P^\top\rangle_F = \langle P^\top X P,\, uv^\top\rangle_F$. This is exactly what is needed here, because the centered empirical Hessian behaves quite differently along u_* (where the exponential weight and the quadratic factor are driven by the same Gaussian) and along $u_*^\perp$ (where they are independent).

6 A guided tour of the proofs

We now go through the arguments. The aim is not completeness, the preprint runs to fifty and so pages, but to make each step feel natural. Constants are dropped without warning.

6.1 Existence: random cones

Step 1: a deterministic characterisation. Say that $b \in \mathbb{R}^d \setminus \{0\}$ Poisson separates the data if

$$\forall i \leqslant n : \quad \begin{cases} \langle X_i,\, b\rangle = 0 & \text{if } Y_i > 0, \\ \langle X_i,\, b\rangle \leqslant 0 & \text{if } Y_i = 0. \end{cases} \tag{22}$$

Specialising a result of Koriyama and Bellec [15] on unregularised M-estimators in single-index models, one obtains that the MLE fails to exist if and only if such a b exists. The intuition is exactly the coercivity discussion of Section 2: along the ray tb, the terms with $Y_i > 0$ are frozen (their argument $\langle X_i,\, tb\rangle$ is identically zero) and the terms with $Y_i = 0$ are non-increasing, so the loss cannot grow; a strictly convex function with a global minimiser must diverge at infinity in every direction, so there is no minimiser.
Note also that if $n < d$ the covariates cannot span $\mathbb{R}^d$, so any nonzero b orthogonal to all of them separates: non-existence is then deterministic. This is part (a) of Theorem 4.1, and it is what makes the transition so sharp.

Step 2: reformulation as a cone intersection. Assume $n \geqslant d$ and let $\mathbf{X}$ be the $n \times d$ design matrix. By rotational invariance of $\mathcal{N}(0, I_d)$ we may assume that β_* is aligned with the first coordinate, so that Y_i depends only on X_{i1}; the independence of the Gaussian coordinates then gives $(Y_i, X_{i1}) \perp\!\!\!\perp (X_{i2}, \ldots, X_{id})$. Writing $b = \sum_j b_j e_j$ and $p_i = \sum_{j\geqslant 2} b_j X_{ij}$, condition (22) becomes a membership condition for $p = (p_1, \ldots, p_n)$ in the random cone

$$C(u, y) = \Big\{p \in \mathbb{R}^n : \exists t \in \mathbb{R},\ \forall i,\ tu_i + p_i = 0 \text{ if } y_i > 0,\quad tu_i + p_i \leqslant 0 \text{ if } y_i = 0\Big\},$$

evaluated at $u = \mathbf{X}e_1$. Hence

$$\text{MLE fails} \iff \operatorname{Span}(\mathbf{X}e_2, \dots, \mathbf{X}e_d) \cap C(\mathbf{X}e_1, y) \neq \{0\}.$$

Conditionally on $(\mathbf{X}e_1, y)$, the left-hand subspace is, almost surely, a $(d-1)$-dimensional subspace of $\mathbb{R}^n$ drawn uniformly at random: this is where Gaussianity is used a second time, while C is a fixed convex cone.

Step 3: the approximate kinematic formula. For a convex cone $C \subset \mathbb{R}^n$ define its statistical dimension

$$\delta(C) = \mathbb{E}\big[\|\Pi_C(g)\|^2\big], \quad g \sim \mathcal{N}(0, I_n),$$

where Π_C is the Euclidean projection onto C. This is the right notion of "effective dimension" for a cone: it coincides with the dimension for linear subspaces, is monotone under inclusion, additive under products, invariant under isometries, and satisfies $\delta(C) + \delta(C^\circ) = n$ if we write C° the polar cone of C. The approximate kinematic formula of Amelunxen, Lotz, McCoy and Tropp [1] states, roughly, that a randomly rotated cone and a random subspace drawn uniformly intersect non-trivially essentially exactly when the sum of their statistical dimensions exceeds the ambient dimension, with a transition window of width $\sqrt{n}$. Precisely, if $m \geqslant \delta(C) + \lambda$, then a uniformly random $(n-m)$-dimensional subspace L_{n-m} satisfies

$$\mathbb{P}\big(L_{n-m} \cap C \neq \{0\}\big) \;\leqslant\; 4\exp\!\left(-\frac{\lambda^2}{8(\omega^2(C) + \lambda)}\right).$$

Where $\omega(C)$ is a function of $(\delta(C), no)$ called the transition window. Here the subspace has dimension $d-1$, so $m = n - d + 1$, and the condition to check is

$$n - d + 1 \;\geqslant\; \delta(C) + \lambda. \tag{23}$$

Step 4: computing $\delta(C)$. Split the indices into $Z = \{i : y_i = 0\}$ and $P = \{i : y_i > 0\}$. Up to a permutation, C is contained in $\mathbb{R}u_P \times \big(\mathbb{R}u_Z + \mathbb{R}_-^{|Z|}\big)$, so by the properties of δ listed above,

$$\frac{|Z|}{2} \;\leqslant\; \delta(C) \;\leqslant\; 1 + |Z|.$$

The picture is transparent: the statistical dimension of the obstruction is the number of zero counts. Since $|Z| \sim \text{Binomial}(n, p)$ with $p \in [\frac{1}{e}, \frac{1}{2}]$, Hoeffding's inequality gives $\delta(C) = np + O(\sqrt{nz})$, and (23) becomes

$$n(1-p) - \sqrt{2nz} \;\gtrsim\; d + \lambda.$$

Because $1 - p$ is bounded below by the universal constant $\frac{e-1}{2e} > 0$, one may take $\lambda \asymp n$, which makes the exponential bound of Step 3 as small as $4e^{-z}$ provided $n \gtrsim z$, and the condition reduces to $n \gtrsim d \vee z$. This proves Theorem 4.1.

Remark. The whole of the signal strength dependence is buried in $p = p(\|\beta_*\|)$, and it survives only as the constant $1-p \in [\frac{1}{2}, 1-\frac{1}{2e}]$. Had the model been such that $p(b) \to 1$ as $b \to \infty$, i.e. had a strong signal produced almost exclusively zero counts, the constant would have degenerated and B would have entered the threshold, as it does in logistic regression. It does not, because $\mathbb{E}[\exp(-e^{bg})] \to \frac{1}{2}$: for a strong signal, half the observations have $\langle \beta_*, X_i \rangle \ll 0$ and give a zero count, but the other half have $\langle \beta_*, X_i \rangle \gg 0$ and give an enormous, very coercive count. We note that this is strongly due to the symmetry of the Gaussian distribution, so a natural question that emerges would concern the behaviour of the threshold for different distributions. Unfortunately, the results of Amelunxen et al. require some strong properties on the distribution of $\mathbf{X}$, for instance, that the random subspace generated by $\operatorname{Span}(\mathbf{X}e_2, \dots, \mathbf{X}e_d)$ is uniformly drawn among $d-1$ dimensional subspaces. Hence, this proof scheme cannot be transferred to other, more

general distributions and the very interesting question of existence in a more general setting stays open. Note however that the almost sure non-existence of the MLE whenever $n < d$ is guaranteed in all cases as this is a purely deterministic statement due to the Poisson separation condition being satisfied by any vector $u \in \mathrm{Span}(X_1, \dots, X_n)^\perp$.

6.2 Lower bound on the empirical Hessian

Lemma 6.1 (Large sample regime)**.** *With probability at least* $1-e^{-z}$*, for every* $\beta \in B_H(\beta_*, \frac{1}{2}e^{B^2/4})$*,*

$$\nabla^2 \widehat{L}_n(\beta) \;\succeq\; c\, H,$$

provided $n \gtrsim e^{B^2/2} B^2 (Bd + z)$*.*

The proof is the technical heart of the paper. Its structure is as follows.

Slicing. Fix $v \in S^{d-1}$ and β in the ball, with $u = \beta / \|\beta\|$. One must show

$$\frac{1}{n} \sum_{i=1}^{n} e^{\langle \beta,\, X_i \rangle} \langle v,\, X_i \rangle^2 \;\gtrsim\; e^{B^2/2} \big(1 + B^2 \langle u_*,\, v \rangle^2\big).$$

Discard everything except the observations lying in the tilting region identified in Section 3.2, namely $\langle u,\, X_i \rangle \in [B - \gamma, B]$ with $\gamma \asymp 1$, and slice that region into the intervals

$$I_k = \Big[B - \tfrac{k+1}{B},\; B - \tfrac{k}{B}\Big], \quad k = 0, \dots, k_0 - 1,\; k_0 = \gamma B.$$

The width $1/B$ is not arbitrary: it is chosen so that the exponential weight $e^{\|\beta\| \langle u,\, X_i \rangle}$ varies by at most a constant factor across a slice. Indeed on I_k one gets $\langle \beta,\, X_i \rangle \geqslant B^2 - k - 5$. Discarding also the observations with $|\langle v,\, X_i \rangle|$ too small (below a threshold $\eta/4$, with $\eta = \max(1, B|\langle u,\, v \rangle|)$ calibrated to the two eigenvalues of H), one arrives at a lower bound by an empirical process of indicator functions:

$$v^\top \nabla^2 \widehat{L}_n(\beta) v \;\gtrsim\; \frac{\eta_*^2}{n} \sum_{k=0}^{k_0 - 1} \sum_{i=1}^{n} e^{B^2 - k}\, \mathbf{1}\big\{\langle u,\, X_i \rangle \in I_k,\; |\langle v,\, X_i \rangle| > \eta/4\big\}. \tag{24}$$

With $\eta_* = \max(1, B|\langle u_*,\, v \rangle|)$. This is the reduction announced in Section 3: all the unboundedness has been moved into the deterministic coefficients e^{B^2-k}, and what remains random is a bounded empirical process.

The mean is already of the right order. Write $p_{k,u,v}$ for the probability of the event in (24) and split the process into its mean and its fluctuation. Two elementary computations handle the mean.

- $p_{k,u,v} \gtrsim \mathbb{P}(g \in I_k)$: decomposing $v = \langle u,\, v \rangle u + \alpha w$ with $w \perp u$, the variables $\langle u,\, X \rangle$ and $\langle w,\, X \rangle$ are independent standard Gaussians, so the two constraints decouple, and one checks in both regimes ($|\langle u,\, v \rangle|$ large or small) that the result holds.
- $\sum_{k < k_0} e^{B^2-k} \mathbb{P}(g \in I_k) \asymp e^{B^2/2}$: this is the tilting computation of Section 3.2 written as a Riemann sum. On I_k one has $e^{B^2-k-1} \leqslant e^{Bx} \leqslant e^{B^2-k}$, so the sum is sandwiched between $e^{B^2/2} \int_{B-k_0/B}^{B} \varphi(x - B)dx$ and e times it.

Together these give $\sum_k e^{B^2-k} \eta_*^2 p_{k,u,v} \gtrsim v^\top H v$, i.e. the deterministic part of (24) is already what we want.

The fluctuations are negligible. It remains to show that the centered process is small compared with its mean, *uniformly in* $u, v \in S^{d-1}$ *and simultaneously for all* k. The class of sets $\{x : \langle u, x\rangle \in [a, b],\ |\langle v, x\rangle| > \nu\}$ is a union of two intersections of three half-spaces, so by the VC-dimension bounds of van der Vaart and Wellner [28] it is a VC class of dimension $O(d)$. Since all the functions are indicators with variance at most $\mathbb{P}(g \in I_k)$, Bousquet's [4] version of Talagrand's [26] inequality together with a standard bound on the expected supremum gives, for n large enough,

$$\sup_{u,v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \big(\mathbf{1}\{\cdots\} - p_{k,u,v} \big) \right| \ \leqslant\ \theta\, \mathbb{P}(g \in I_k)$$

for a small constant θ, with a union bound over the $k_0 \leqslant B$ slices. The sample size required is governed by the smallest probability involved, $\mathbb{P}(g \in I_k) \asymp B^{-1} e^{-B^2/2}$: the classical VC condition $n \gtrsim d \log(1/p)/p$ then reads $n \gtrsim e^{B^2/2} B^2 (Bd + z)$. Once again, the exponential factor is the reciprocal of the probability of the tilting event.

Lemma 6.2 (Intermediate regime)**.** *With probability at least* $1 - 2e^{-z}$*, for every* $\beta \in \mathbb{R}^d$*,*

$$\nabla^2 \widehat{L}_n(\beta) \ \succeq\ c\, I_d, \quad c > \tfrac{1}{100},$$

provided $n \gtrsim d \vee z$*.*

This is straightforward: for any β with $u = \beta / \|\beta\|$,

$$v^\top \nabla^2 \widehat{L}_n(\beta) v = \frac{1}{n} \sum_i e^{\langle \beta, X_i \rangle} \langle v,\, X_i \rangle^2 \ \geqslant\ \frac{1}{n} \sum_i \mathbb{1}_{\{\langle u, X_i\rangle \geqslant 1,\ |\langle v, X_i\rangle| \geqslant 1\}},$$

because on that event the exponential weight is at least 1 and the quadratic factor is at least 1. All the B-dependence has been thrown away, but in exchange the bound is uniform over the whole parameter space, no localisation set is needed, and one only has to lower bound the empirical measure of a VC class of sets, each of probability at least $\mathbb{P}(\mathcal{N}(0,1) \geqslant 1)^2 \geqslant 0.025$ (the last inequality needs a small correlation argument, since $\langle u, X\rangle$ and $\langle v, X\rangle$ are not independent). A standard uniform lower-tail bound for VC classes [3] then gives the result as soon as $n \gtrsim d$.

Remark. It is worth pausing on the fact that the lower bound is the easy direction here. This inverts the usual intuition, and the reason is structural: lower-bounding an average of non-negative terms allows one to discard data, which turns an unbounded problem into a bounded one for free. Upper bounds enjoy no such luxury.

6.3 Upper bound on the empirical Hessian

Lemma 6.3 (Large sample regime)**.** *Let* $\varepsilon \in (0, 1)$*. With probability at least* $1 - 3e^{-z}$*,* $\nabla^2 \widehat{L}_n(\beta_*) \preceq C\, H$*, provided* $n \geqslant \exp\big(\frac{B^2}{2}(1 + \varepsilon^{-1})\big)\big((d + z) e^{4B\sqrt{z}}\big)^{1+\varepsilon}$*.*

Reduction to two directions. Because H has exactly two eigenspaces, $\mathbb{R}u_*$ with eigenvalue $(B^2+1)e^{B^2/2}$ and $u_*^\perp$ with eigenvalue $e^{B^2/2}$, and because cross terms are handled by $2ab \leqslant a^2 + b^2$, the matrix inequality reduces to the two scalar bounds

$$u_*^\top \nabla^2 \widehat{L}_n(\beta_*) u_* \lesssim B^2 e^{B^2/2} \quad \text{and} \quad \sup_{v \in S^{d-1} \cap u_*^\perp} v^\top \nabla^2 \widehat{L}_n(\beta_*) v \lesssim e^{B^2/2}.$$

For the first term, the Gaussian appearing in the exponential weight is the same as the one being squared, so one is averaging $e^{B g_i} g_i^2$, which is a single one-dimensional, very heavy-tailed average, with no dependence on the dimension. In the second, $\langle u_*, X\rangle$ and $\langle v, X\rangle$ are independent, so the weight decouples from the quadratic factor, but the bound must hold uniformly over a sphere of dimension $d - 2$: this is where the dimension d enters (and it is the only source of d).

Truncation. Both bounds start by conditioning on

$$\Omega_b = \Big\{\max_i \langle u_*, X_i\rangle \leqslant \sqrt{2\log n + 2z}\Big\}, \quad \mathbb{P}(\Omega_b) \geqslant 1 - e^{-z},$$

on which all weights are at most $M = \exp\big(B\sqrt{2\log n + 2z}\big)$. This is the step that manufactures a Laplace transform where none existed previously.

Direction u_*: Bernstein. On Ω_b the truncated variables are sub-Gamma with variance proxy $\mathbb{E}[e^{2Bg}g^4] \asymp B^4 e^{2B^2}$ and scale M, so Bernstein's inequality gives

$$\frac{1}{n}\sum_i e^{Bg_i} g_i^2 \;\lesssim\; B^2 e^{B^2/2} + \sqrt{\frac{B^4 e^{2B^2} z}{n}} + \frac{Mz}{n},$$

and the second and third terms are dominated by the first (expectation term) precisely when $n \gtrsim e^{B^2}$: note that this is the second heuristic of Section 3.2.

Orthogonal subspace: PAC-Bayes for matrices. For $u, v \perp u_*$ and $k \geqslant 2$, independence and Cauchy–Schwarz give

$$\mathbb{E}\Big[e^{kB\langle u_*, X\rangle}\big|\langle u, X\rangle\langle v, X\rangle\big|^k\Big] \;\leqslant\; M^{k-2}\,\mathbb{E}[e^{2Bg}]\,\mathbb{E}[g^{2k}] \;\leqslant\; (2M)^{k-2}\,4e^{2B^2}\,k!,$$

i.e. the truncated centered Hessian is (σ^2, b) sub-Gamma with $\sigma^2 \asymp e^{2B^2}/n$ and $b \asymp M/n$. Theorem 5.6, applied on the subspace $u_*^\perp$ as in the Remark after 5.5, then yields

$$\sup_{u,v\in S^{d-1}\cap u_*^\perp} u^\top \nabla^2 \widehat{L}_n(\beta_*) v \;\lesssim\; \underbrace{e^{B^2/2}}_{\text{mean}} + \underbrace{e^{B^2}\sqrt{\frac{d+z}{n}}}_{\text{sub-Gaussian part}} + \underbrace{M\,\frac{d+z}{n}}_{\text{truncation part}}. \tag{25}$$

The second term is dominated by the first iff $n \gtrsim e^{B^2}(d+z)$. The third is where the parameter ε is born: one needs $\exp\big(-\frac{B^2}{2} + B\sqrt{2\log n + 2z} - \log n\big) \lesssim 1$, i.e.

$$\exp\left(-\Big(\tfrac{B}{\sqrt{2}} - \sqrt{\log n}\Big)^2 + B\sqrt{2z}\right) \lesssim 1,$$

and one converts this, via $(a-b)^2 \geqslant \frac{a^2}{1+\varepsilon} - \frac{b^2}{\varepsilon}$ into the sample size condition of the lemma. The reader now has a complete account of every factor in (14).

Intermediate regime. When n is too small for (25) to give $e^{B^2/2}$, one settles for a cruder target: the same argument, with the same truncation and the same optimisation but aiming at the isotropic matrix $H_0 = e^{B^2/\delta} I_d$ instead of H, gives $\nabla^2 \widehat{L}_n(\beta_*) \preceq C e^{B^2/\delta} I_d$ under the B-free condition $n \gtrsim_\varepsilon d^{1+\varepsilon} e^{8z}$. The relation $\frac{1}{\delta} = \frac{1}{2} + \frac{1}{\varepsilon}$ is exactly what makes the optimization work: it is the value of δ for which the inflation $e^{B^2/\delta - B^2/2} = e^{B^2/\varepsilon}$ absorbs the truncation error.

6.4 Upper bound on the empirical gradient

Lemma 6.4. *Let $M \in \{H, H_0\}$ according to the regime. With probability at least $1 - 5e^{-z}$,*

$$\big\|\nabla \widehat{L}_n(\beta_*)\big\|_{M^{-1}} \;\lesssim\; \sqrt{\frac{d+z}{n}},$$

under the corresponding sample size conditions: (14) for $M = H$, (16) for $M = H_0$.

The proof is where the responses finally enter, and where the well-specification assumption is used. Conditionally on the design, $\nabla \widehat{L}_n(\beta_*) = \frac{1}{n}\sum_i (e^{\langle \beta_*, X_i\rangle} - Y_i)X_i$ is a sum of independent centered terms, and the elementary Poisson bound $\mathbb{E}[e^{s(Y-\lambda)}] \leqslant e^{\lambda s^2}$ for $|s| \leqslant 1$ gives, for any v,

$$\mathbb{E}\Big[\exp\big(\langle v,\, M^{-1/2}\nabla \widehat{L}_n(\beta_*)\rangle\big)\,\Big|\,\mathcal{X}\Big] \;\leqslant\; \exp\left(\frac{\|v\|^2_{M^{-1/2}\nabla^2 \widehat{L}_n(\beta_*)M^{-1/2}}}{n}\right),$$

valid as long as $\frac{1}{n}\max_i |\langle M^{-1/2}v,\, X_i\rangle| \leqslant 1$. In other words, $M^{-1/2}\nabla \widehat{L}_n(\beta_*)$ is $(\star, \Sigma)$ sub-Gamma in the sense of Definition 5.3, with

$$\|v\|_\star = \frac{1}{n}\max_i \big|\langle X_i,\, M^{-1/2}v\rangle\big|, \quad \Sigma = \frac{2}{n}M^{-1/2}\nabla^2 \widehat{L}_n(\beta_*)M^{-1/2}.$$

Remark (Why the Hessian upper bound was needed). The covariance proxy Σ is, up to the factor $2/n$, the empirical Hessian: in a well-specified model the variance of the score equals the Fisher information. It is the reason why controlling the empirical Hessian from above is indispensable even though the localisation lemma never asks for it.
Note that the pseudo-norm $\|\cdot\|_\star$ is actually a norm conditionally on the data $\mathcal{X}$, to the best of our knowledge, the use of such a norm that differs clearly from an ellipsoid norm to measure closeness to zero in a PAC-Bayes setting is novel.

Applying Theorem 5.4 then requires four quantities. The two leading ones are immediate from the Hessian upper bound: on that event $\Sigma \preceq \frac{C}{n} I_d$, hence $\mathrm{Tr}(\Sigma) \lesssim d/n$ and $\|\Sigma\| \lesssim 1/n$, contributing $\sqrt{d/n} + \sqrt{z/n}$, which is the announced rate. The remaining two are the Gaussian width and diameter of the dual ball, and they are of a smaller order: since $\big\|M^{-1/2}\big\| \leqslant 1$, both are controlled by $\max_i \|X_i\|$, a standard Gaussian maximum computation gives $\mathbb{E}[\max_i \|X_i\|] \lesssim \sqrt{d} + \sqrt{\log n}$, hence those terms are contributions of order $\frac{\sqrt{d\log n}}{n}\sqrt{z}$ and $\frac{\sqrt{d}+\sqrt{\log n}+\sqrt{z}}{n}$, both negligible compared to $\sqrt{(d+z)/n}$ whenever n is larger than an absolute constant.

6.5 Assembling the pieces

We can now provide an overview of the three main theorems.

- **Theorem 4.2.** Take H as in (8). Lemma 6.1 gives assumption (i) of Lemma 5.1 with c_0 a constant and $r_0 \asymp e^{B^2/4}$; the deterministic comparison of the population Hessian gives (iii) with $c_1 \leqslant 45$; Lemma 6.3 feeds Lemma 6.4, which gives (ii) with $\nu \asymp \sqrt{(d+z)/n}$. The condition $\nu < c_0 r_0/2$ becomes $\sqrt{(d+z)/n} \lesssim e^{B^2/4}$, which is far weaker than (14) and is therefore automatically satisfied. Conclusion (19) is (15).

- **Theorem 4.3.** Same programme in the modified geometry of Section 5.1: Lemma 6.2 for the curvature (uniformly on $\mathbb{R}^d$, hence no radius constraint), Lemma 6.4 with $M = H_0$ for the gradient. Localisation follows in the Euclidean ball of radius $\kappa e^{B^2/2\delta}\sqrt{(d+z)/n}$, and the risk bound follows from the deterministic bound $\nabla^2 L(\beta) \preceq 2Re^{R/2} I_d$ on that ball, with $R \asymp e^{B^2/\delta}$.

- **Theorem 4.1.** Independent of all the above: pure conic geometry, as in Section 6.1. (One could also deduce existence from Lemma 5.1, but only under the far stronger sample size conditions of the other two theorems, which is exactly the point of treating existence separately.)

7 Numerical experiments

Note on the assumptions: In Section 8, we will discuss potential relaxations of the assumptions to non-Gaussian settings, however, this entire section relies on the assumption of a

Gaussian design, both for closed form expressions (see (26)) but more importantly for the basic computational feasibility, as sampling from arbitrary distributions (other than Gaussians) in high dimensions is a difficult problem. Second, $\|\beta_*\|$ is really $\|\beta_*\|$ in these simulations, and not $B = \max(2, \|\beta_*\|)$, this choice was natural in the simulation setting as defining $B \geqslant 2$ was purely for technical reasons, to avoid pathologies in the population Hessian during proofs.

The thresholds of Theorem 4.1 and Theorem 4.2 are stated up to universal constants that are far too large to be meaningful numerically, but their *scaling* in n, d and B can be tested directly in moderate dimension. This section reports such an experiment. The code is publicly available on GitHub. Two features make the design of the experiment very clean.

Existence is decidable exactly. Suppose we are given a dataset $((Y_1, X_1), \dots, (Y_n, X_n))$. Denote by $X_P \in \mathbb{R}^{|P| \times d}$ (where $|P|$ is the number of positive counts $Y_i > 0$) the matrix whose rows are the $X_i's$ associated with the $Y_i's > 0$. Write the same thing for X_Z, the matrix associated with the $Y_i's = 0$.
By the characterisation (22), the MLE fails to exist if and only if some $b \neq 0$ Poisson separates the sample. The equality constraints say that $b \in \mathbb{R}^d$ lies in $\ker(X_P)$, where $P = \{i : Y_i > 0\}$; writing $b = Vc$ for V an orthonormal basis of that kernel, the question becomes whether there is $c \neq 0$ with $Mc \leqslant 0$ (coordinate-wise), $M = X_Z V$. This is settled by a small linear program, so no thresholding or convergence heuristic is involved: each sample is classified exactly (up to potential measure-zero events).

The excess risk is available in closed form. Under $X \sim \mathcal{N}(0, I_d)$ both terms of the population risk are Gaussian integrals,

$$L(\beta) = \mathbb{E}\big[e^{\langle \beta,\, X\rangle}\big] - \mathbb{E}\big[Y\langle \beta,\, X\rangle\big] = e^{\|\beta\|^2/2} - e^{\|\beta_*\|^2/2}\langle \beta,\, \beta_*\rangle,$$

the second by the tilting identity of Appendix B. Writing $\mu = \beta - \beta_*$ and $s = \langle \mu,\, \beta_*\rangle + \|\mu\|^2/2$ this rearranges into the numerically stable form

$$L(\beta) - L(\beta_*) = e^{\|\beta_*\|^2/2}\Big[\big(e^s - s - 1\big) + \tfrac{1}{2}\,\|\mu\|^2\Big], \tag{26}$$

in which both bracketed terms are non-negative. So the excess risk is evaluated exactly, with no Monte-Carlo error and no cancellation: the only randomness in the plots below is that of the sample itself.
The MLE is computed by damped Newton with backtracking, which converges in under ten iterations here (the loss is self-concordant). One caveat is worth recording: when the MLE does not exist the loss is flat along the separating ray, so the gradient underflows while the iterate drifts to infinity, and the optimiser reports spurious convergence at a large meaningless β. Convergence of the optimiser is therefore *not* a test of existence; we always run the linear program first.

7.1 The existence transition

What is being plotted. Recall the approximate kinematic condition as given in Section (23): the probability of existence of the MLE is predicted to sharply increase in λ whenever

$$n \geqslant \delta(C) + d - 1 + \lambda.$$

Hence, by transition, we mean the precise moment of this predicted switch to high probability existence, that is, (reformulating in terms of n/d):

$$\frac{n}{d} = 1 + \frac{\delta(C) - 1}{d} \asymp 1 + \frac{\delta(C)}{d}, \quad \text{(setting } \lambda = 0\text{)}.$$

For each fixed pair $(d, \|\beta_*\|)$ and each value of the ratio $r = n/d$, we draw 300 independent samples of size $n = \lceil rd \rfloor$ and record the fraction for which the MLE exists, as decided by the linear program. Figure 3 plots that fraction against r. At finite d the resulting curve is a smooth increasing function of r, not a step. To quantify this idea of transition, write $r_q(d)$ for the ratio at which the estimated probability crosses the level q (obtained by linear interpolation on the grid). The measured values at $\|\beta_*\| = 2$ are:

d	10	20	40	80	160
$r_{0.05}$	1.070	1.103	1.147	1.178	1.210
$r_{0.50}$	1.241	1.290	1.301	1.298	1.300
$r_{0.95}$	1.686	1.586	1.489	1.440	1.396
width $r_{0.95} - r_{0.05}$	0.616	0.484	0.343	0.262	0.187

Two things are visible. The midpoint $r_{0.50}$ is stable in d, settling at 1.30 from $d = 40$ onwards: this midpoint is a function of the ratio n/d alone, which is the content of $n \asymp d$. And the window over which the probability climbs from 0.05 to 0.95 shrinks steadily, roughly like $d^{-1/2}$. The $d^{-1/2}$ rate is what one expects. Indeed, omitting constants, and using deviation bounds heuristics ($\alpha \in [1/2, 1]$ to get the exact $\delta(C)$ between $|Z|/2$ and $|Z|$):

$$n - d + 1 = \delta(C) = \alpha|Z| \asymp \alpha(np \pm \sqrt{n})$$

we find

$$-\frac{1}{1-\alpha p}\frac{1}{\sqrt{d}} \leqslant \frac{n}{d} - \frac{1}{1-\alpha p} \leqslant \frac{1}{1-\alpha p}\frac{1}{\sqrt{d}}.$$

The transition is driven by the number of zero counts $|Z|$, whose binomial fluctuations are of size $\sqrt{n} \asymp \sqrt{d}$, i.e. of relative size $d^{-1/2}$ once divided by d.

Where the transition sits. The location can be predicted from the proof itself. By Step 3 of Section 6.1, existence holds when $(d-1) + \delta(C) < n$, and Step 4 brackets the statistical dimension of the obstruction as $|Z|/2 \leqslant \delta(C) \leqslant 1 + |Z|$. With $|Z| \approx np$ the two ends of the bracket predict

$$r_c = \frac{1}{1-p/2} \qquad \text{and} \qquad r_c = \frac{1}{1-p}$$

respectively. These are far apart: 1.22 against 1.58 at $\|\beta_*\| = 0$. Experimentally, the simulation decides between them in this setting. At $d = 160$:

$\|\beta_*\|$	0	1	2	3	4
p	0.367	0.381	0.416	0.433	0.446
$1/(1-p/2)$	1.224	1.235	1.263	1.277	1.287
measured $r_{0.50}$	1.226	1.251	1.301	1.325	1.343
$1/(1-p)$	1.579	1.615	1.713	1.765	1.807

The measured midpoint (proxy for the transition) tracks the *lower* end of the bracket, with a small positive bias that grows slowly with $\|\beta_*\|$. In other words $\delta(C) \approx |Z|/2$: the statistical dimension of the obstruction is governed by the orthant $\mathbb{R}_-^{|Z|}$, whose statistical dimension is exactly $|Z|/2$. The upper end $1/(1-p)$ is not the transition at all: it is the ratio beyond which $|P| \geqslant d$, so that X_P spans $\mathbb{R}^d$ and existence becomes automatic for trivial reasons. That mechanism is far too crude to explain the observed behaviour at $d = 80$, $\|\beta_*\| = 2$ and $r = 1.6$ the MLE exists in every one of 400 samples while $|P| \geqslant d$ holds in only 22% of them. Existence below $1/(1-p)$ is genuinely a consequence of the negativity constraints on the zero-count observations, that is, of the conic geometry, and not of a rank condition.

This also refines the discussion after Theorem 4.1: the signal strength enters only through p, and since p ranges over $[1/e, 1/2]$, $1/(1-p)$ (proxy for the transition up to a constant α) can only ever move within $[1.22, 1.33]$, whatever $\|\beta_*\|$ may be. The measured spread of the midpoint, 1.226 to 1.343 as $\|\beta_*\|$ goes from 0 to 4, occupies almost exactly that interval. The independence of the threshold from B is therefore not approximate but bounded a priori, in sharp contrast with the factor B of logistic regression.
Figure 3 plots the estimated $\mathbb{P}(\text{MLE exists})$ against n/d, for $d \in \{10, 40, 160\}$ and $\|\beta_*\| \in \{0, 2, 4\}$ (300 samples per point).

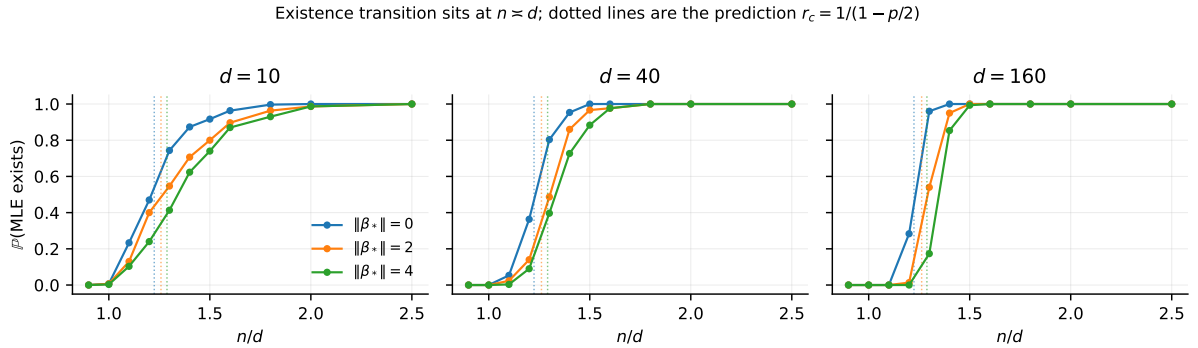

Figure 3: Probability that the MLE exists, against n/d. Dotted lines mark the lower bound of the predicted location $1/(1-\hat{2}p)$. The transition is a function of n/d alone, sharpens with d, and is only mildly displaced by the signal strength.

Where z appears. Theorem 4.1 carries a confidence parameter z which has no counterpart in the simulation, since the simulation estimates the probability itself rather than a bound on it. The two are reconciled by treating z as a free knob and optimising it. Given n and d, the conclusion $\mathbb{P}(\text{fails}) \leqslant 6e^{-z}$ is available for every z satisfying $n \geqslant C_1(d \vee z)$, so the best choice is the largest admissible one, $z = n/C_1$, legitimate as soon as $n \geqslant C_1 d$. The theorem is therefore equivalent to the single statement

$$n \geqslant C_1 d \qquad \Longrightarrow \qquad \mathbb{P}(\text{MLE fails to exist}) \;\leqslant\; 6\, e^{-n/C_1}, \tag{27}$$

and in the coordinates of Figure 3, at fixed $r = n/d$, the right-hand side reads $6e^{-rd/C_1}$: the failure probability should decay *exponentially in the dimension* at any fixed ratio above threshold. This is precisely the sharpening of the curves with d, and it can be checked quantitatively. At $\|\beta_*\| = 2$, estimating the failure probability from 4000 samples per point and regressing its logarithm on d:

	$d = 10$	20	40	80	120	160	fitted rate
$r = 1.4$	.290	.245	.171	.098	.064	.037	$0.0134\, d$
$r = 1.5$	.182	.115	.038	.009	.003	$<.001$	$0.0376\, d$
$r = 1.6$	.106	.040	.007	.0002	$<.001$	$<.001$	$0.0859\, d$

The decay is exponential in d, as (27) requires. It is interesting to see that the rate is not proportional to r with a single constant, as the crude form (27) would suggest: it steepens sharply as r moves away from $r_c \approx 1.30$. This is not a contradiction as (27) is a bound, valid for all admissible r, and it is not necessarily tight for all values of r.
Finally, the constants. Tracking them through the proof of Theorem 4.1 gives $C_1 = 2/c_8 \leqslant 13$ for the dimension part of the condition and $8/c_8^2 \leqslant 321$, respectively $c_{10} \leqslant 367$, for the part

involving z. The dimension condition therefore guarantees nothing until $n \geqslant 13d$, whereas the MLE in fact exists with probability indistinguishable from one already at $n = 2d$; and a bound of the form $\mathbb{P}(\text{fails}) \leqslant 0.05$ requires $z = \log(6/0.05) \approx 4.8$, hence $n \geqslant 1760$ irrespective of d. The guarantee is thus quite conservative.

7.2 The risk threshold

For the risk we fix $d = 5$ and record the scaled excess risk $n(L(\widehat{\beta}_n) - L(\beta_*))/d$, computed from (26). Wilks' theorem (9) predicts that this settles at $1/2$, since $2n(L(\widehat{\beta}_n) - L(\beta_*)) \xrightarrow{d} \chi^2(d)$ has mean d. This is what Figure 4 shows, and the point of interest is *when*: the curves are ordered by $\|\beta_*\|$ and the sample size at which each reaches its plateau grows very fast with the signal strength.

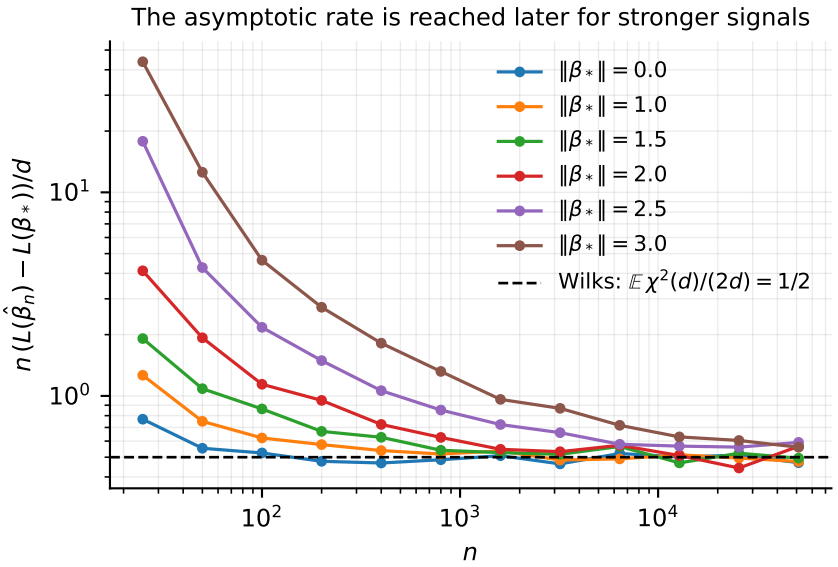

Figure 4: Scaled excess risk against n, at $d = 5$. All curves converge to the Wilks level $1/2$, but the sample size required to do so grows sharply with $\|\beta_*\|$.

To quantify this, let $n_*(B)$ be the smallest n at which the mean scaled excess risk drops below 0.6, obtained by log-linear interpolation on a geometric grid. The measured values are:

$\|\beta_*\|$	1.0	1.5	2.0	2.5	3.0
n_*	113	336	1082	4837	27067

Regressing $\log n_*$ on B^2 gives

$$\log n_* \;\approx\; 4.20 + 0.68\, B^2, \qquad R^2 = 0.997.$$

The fit confirms the shape predicted by Section 3.2: the threshold is exponential in B^2, and strikingly so, with 99.7% of the variation in $\log n_*$ explained by B^2 alone over more than two orders of magnitude in n_*. The fitted exponent 0.68 falls between the two heuristics of Section 3.2: the Hessian lower bound, which needs only $n \gtrsim e^{B^2/2}$ (slope $1/2$), and the Hessian upper bound, which needs $n \gtrsim e^{B^2}$ (slope 1). Both reference slopes are drawn in Figure 5. That the measured exponent is strictly larger than $1/2$ is consistent with the upper bound being the binding constraint, as the proof of Theorem 4.2 suggests; we would not read more into the precise value 0.68, since the level 0.6 used to define n_* is arbitrary. Moreover, simulations only concern the modest range of B that is computationally feasible.

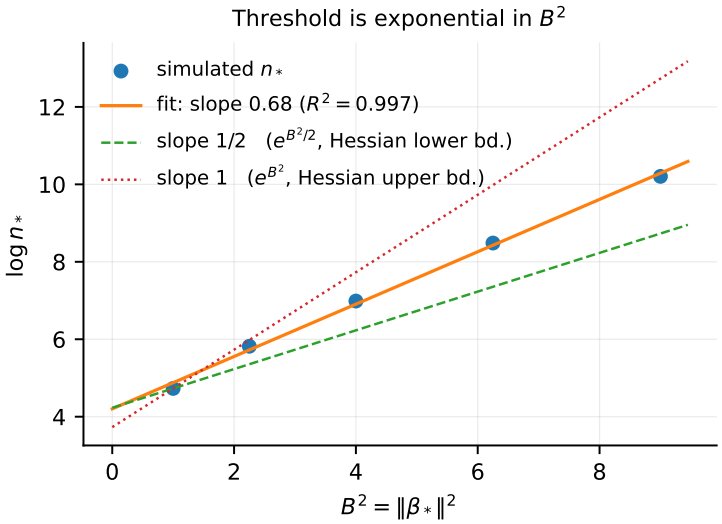

Figure 5: $\log n_*$ against B^2, with the two reference slopes of Section 3.2. The exponential-in-B^2 form fits essentially perfectly; the exponent lies between the lower- and upper-bound heuristics.

7.3 Wilks' theorem

Finally, Figure 6 checks the benchmark itself, and serves more as a sanity check than as a truly interesting plot: once n is past the threshold, the histogram of $2n(L(\widehat{\beta}_n) - L(\beta_*))$ matches $\chi^2(d)$, with sample means within 2% of d and Kolmogorov–Smirnov p-values comfortably above 0.05 in every configuration tested, the null is not rejected (we cannot reject that both samples come from the same distribution). This confirms that the constant $1/2$ used as the plateau level above is the right target, and that (15) cannot be improved in order. It is the only end-to-end validation that the closed-form excess risk, the Newton fit, and the sampler are all correct at once. Indeed, if an additional factor of 2 were present in the tilting algebra, Figure 4 would still show a clean plateau, but at 0.25, and nothing else would catch it.

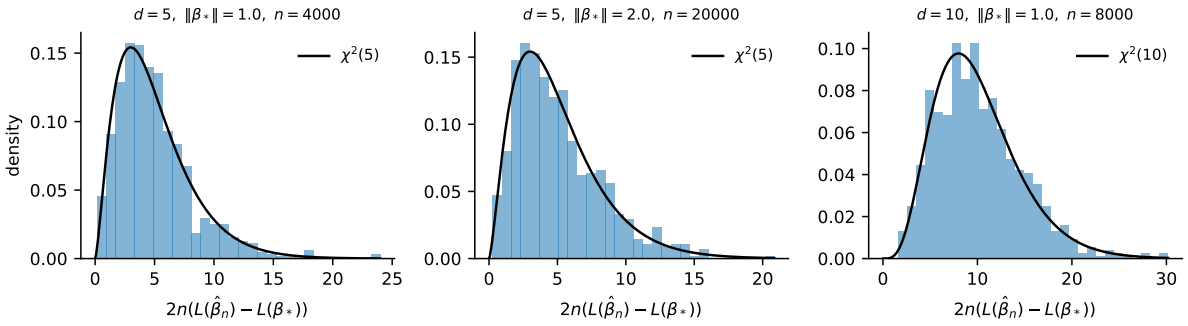

Figure 6: Empirical law of $2n(L(\widehat{\beta}_n) - L(\beta_*))$ against the $\chi^2(d)$ density, for samples past the risk threshold.

7.4 What the simulations do not show

Two caveats. First, all of this concerns the *sufficiency* direction: the experiments locate the sample size at which the MLE starts to behave well; they do not prove that it behaves badly below, even though they do provide evidence in this direction. They are therefore consistent with, but do not establish, the matching lower bound raised as open problem **Q2**. Second,

the accessible range of B is small, since n_* grows like e^{cB^2} and already exceeds 2.5×10^4 at $\|\beta_*\| = 3$; the intermediate regime, where the gap between the two thresholds is large, is out of computational reach for exactly the reason that makes it interesting.

8 Which assumptions are structural?

All the results rest on two hypotheses: an isotropic Gaussian design and a well-specified model. It is worth asking, for each, whether it is a genuine feature of the problem or a convenience. The short answer is that the risk bounds rest on much weaker foundations than their statements suggest, that well-specification is by far the least essential of the two, and that the existence result is the fragile one.

8.1 The design

Recall first (Remark 2.1) that isotropy is not an assumption at all: any $\mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$ reduces to it, with B read as $\max(2, \|\Sigma^{1/2}\beta_*\|)$. The real question is Gaussianity.

Gaussianity is nearly maximal. The natural candidate generalisation is an isotropic sub-Gaussian design. One would normally regard this as a mild weakening. Here it is almost the largest possible class, for a reason specific to the Poisson model: the population risk itself,

$$L(\beta) = \mathbb{E}\big[e^{\langle \beta,\, X\rangle}\big] - \mathbb{E}\big[Y\langle \beta,\, X\rangle\big],$$

is finite for every $\beta \in \mathbb{R}^d$ only if the marginal moment generating function $\psi_u(t) = \mathbb{E}[e^{t\langle u,\, X\rangle}]$ is finite for every $t \in \mathbb{R}$ and every direction u. If the design has, say, merely sub-Gamma marginals, then $\psi_u(t) = +\infty$ beyond some finite t, and the risk is identically $+\infty$ outside a ball: the estimation problem is not even well posed for signal strengths above a design-determined threshold. Since the whole point of the paper is to let B be arbitrary, very light tails are forced by the model, not chosen for convenience. This is an interesting feature of the Poisson model, as it has no analogue in linear or logistic regression, where the loss is at most quadratic. To truly get a result akin to a general one, considering regressors with ψ_α tails for $\alpha > 1$ and study how the risk bounds vary would be an interesting question. Solving the problem, however, would probably be similar in approach to what we did in the Gaussian case.

Where Gaussianity actually enters the risk bounds. The architecture of the argument is distribution-free: Lemma 5.1 is deterministic, the empirical process arguments use only VC bounds [28] and Talagrand's inequality [26], and the two PAC-Bayes inequalities are stated for arbitrary sub-Gamma vectors and matrices. Gaussianity is used in exactly three computations.

(D1) **The closed form of the population Hessian** (Lemma 2.1), which dictates the choice of the proxy H and the deterministic comparison $\nabla^2 L(\beta) \asymp \nabla^2 L(\beta_*)$ on the localisation set. For a general isotropic design one would simply take $H \asymp \nabla^2 L(\beta_*)$; the statements would then be relative to the true population Hessian and not necessarily explicit in B anymore. This is a potential loss of information but not an obstruction. The comparison itself is a self-concordance-type statement in the sense of Nesterov–Nemirovski and would have to be re-derived for each specific case.

(D2) **The identification of the tilting scale.** The ubiquitous $e^{B^2/2}$ is always $\psi(B) = \mathbb{E}[e^{B\langle u,\, X\rangle}]$, and the rare event $\langle u_*,\, X\rangle \asymp B$ is the corresponding tilting point. This part should survive under regularity assumptions on the design, with ψ being a natural upper bound for the design's MGF if we assume sub-Gaussian tails. What is more delicate is the small-ball argument $p_{k,u,v} \gtrsim \mathbb{P}(g \in I_k)$ of Section 6.2, which uses that $\langle u,\, X\rangle$ and $\langle w,\, X\rangle$

are independent for $u \perp w$, a property typical of the Gaussian among isotropic laws. For a general design one would have to assume a joint small-ball condition of the form

$$\mathbb{P}\big(\langle u,\, X\rangle \in I_k,\ |\langle v,\, X\rangle| > \eta/4\big) \ \geqslant\ c\,\mathbb{P}\big(\langle u,\, X\rangle \in I_k\big).$$

Identifying natural classes of distributions satisfying such a condition is an open problem; the analogous question for logistic regression was addressed by Chardon, Lerasle and Mourtada [11] through a notion of "regular" distributions.

(D3) **Decoupling across $\mathbb{R}u_* \oplus u_*^{\perp}$.** This occurs in the moment computations of Section 6.3 (for instance). A similar argument can still be used by applying Cauchy–Schwarz to the whole product instead of using independence. It gives the same conclusions with $\psi(2B) = e^{2B^2}$ replaced by $\psi(4B)^{1/2} = e^{4B^2}$, i.e. a worse exponent in the sample size conditions but no change in the structure.

Existence is a different matter. Gaussianity is used twice in Section 6.1, and neither use is cosmetic. Rotational invariance is what allows the reduction to the first coordinate; and the approximate kinematic formula requires the random subspace $\mathrm{Span}(\mathbf{X}e_2, \dots, \mathbf{X}e_d)$ to be uniformly distributed among $(d-1)$-dimensional subspaces of $\mathbb{R}^n$, conditionally on $(\mathbf{X}e_1, y)$. Both are immediate for a standard Gaussian design and both fail in general. Since rotational invariance is central to the conic geometry, extending Theorem 4.1 beyond the Gaussian case would likely require a genuinely different argument, not a refinement of this one.

8.2 Well-specification

Here the situation is much more comfortable, and the reason is the observation made after (6): the responses do not appear in the Hessian. Consequently Lemmas 6.1, 6.2 and 6.3 are statements about the design and about β_* alone, and the modelling assumption enters at exactly one point, the control of the gradient (Lemma 6.4). Only two properties of the conditional law of Y given X are used there:

(M1) **Correct specification of the conditional mean:** $\mathbb{E}[Y \mid X] = \exp(\langle \beta_*,\, X\rangle)$ almost surely. This is what makes $\nabla L(\beta_*) = 0$, so that β_* really is the population minimiser, and the empirical gradient centered.

(M2) **Conditional sub-Gamma behaviour with variance proportional to the mean:** there exist $c_1, c_2 > 0$ with

$$\mathbb{E}\big[\exp\big(s(Y - \mathbb{E}[Y \mid X])\big) \mid X\big] \leqslant \exp\big(c_2 s^2 e^{\langle \beta_*,\, X\rangle}\big), \quad |s| \leqslant 1/c_1.$$

For $Y \sim \mathcal{P}(\lambda)$ this holds with $c_1 = c_2 = 1$.

Under (M1)–(M2) the proofs go through unchanged with the covariance proxy inflated by c_2, so Theorems 4.2 and 4.3 hold with the risk bounds and sample size conditions multiplied by a factor depending only on (c_1, c_2); the rates are unchanged. The natural setting for the risk bounds is thus the quasi-likelihood one: correct conditional mean, plus a variance that is at most proportional to the mean.

The price of overdispersion. It is instructive to see what happens for the standard misspecified model of the econometrics literature, [13]

$$Y \mid \xi, X \sim \mathcal{P}\big(\exp(\langle \beta_*,\, X\rangle + \xi)\big), \quad \xi \perp\!\!\!\perp X,\ \mathbb{E}[e^{\xi}] = 1,$$

in which ξ absorbs explanatory variables not included in X. Assumption (M1) still holds, so β_* is still the population minimiser. But the conditional variance (conditional on $\mathcal{X}$) is now

$e^{\langle \beta_*, X\rangle} + \mathrm{Var}(e^{\xi})e^{2\langle \beta_*, X\rangle}$: the variance proxy acquires a quadratic term, and the conditional covariance of the gradient involves

$$\frac{1}{n}\sum_{i=1}^{n} e^{2\langle \beta_*, X_i\rangle} X_i X_i^\top = \nabla^2 \widehat{L}_n(2\beta_*).$$

Since the Hessian upper bounds are statements about the design alone, they can simply be applied at $2\beta_*$ instead of β_*, at the cost of replacing B by $2B$ everywhere: the conclusions survive with $e^{B^2/2}$ replaced by e^{2B^2} in (14). This is a substantial degradation, but a quantitative one: it makes the price of multiplicative overdispersion explicit, which offers more information than what the asymptotic theory on Poisson regression does.

Existence again. Even here well-specification is barely used: the characterisation (22) is a statement about the loss and the observed data, not about the model. Inspecting the proof, only two facts are needed: that Y is conditionally independent of X given $\langle \beta_*, X\rangle$ (which preserves the reduction to the first coordinate, and which holds in particular for the overdispersed model above), and that $p = \mathbb{P}(Y = 0)$ is bounded away from 1. Any bound $p \leqslant 1 - c$ would do, with the constant C_1 depending on c.

8.3 Summary

We summarise here all of the assumptions made to help the reader determine at a glance what is really needed and what can be relaxed.

Assumption	Used for	Status
Isotropy $\Sigma = I_d$	everything	free (reparametrisation)
Gaussian marginals	risk bounds	near-maximal here; partly relaxable (ψ_α designs)
Independence on $u_*^\perp$	moment bounds	convenience ($e^{2B^2} \to e^{4B^2}$)
Small-ball condition	Hessian lower bound	needed; open for which laws
Rotational invariance	existence	**structural** for this proof
Well-specified $Y \mid X$	gradient bound only	relaxable to (M1)–(M2)

Table 2: What is structural and what is not.

9 Conclusion and open problems

The final picture that emerges is (we think) an interesting one, we use this section to provide a conclusion on the work and present remaining open problems that are related.
In a generalised linear model, two very different things can go wrong at small sample sizes. The estimator can fail to exist, which is a question of convex geometry: it happens when the data admit a direction along which the empirical loss refuses to grow, and whether such a direction exists is a matter of how a random cone meets a random subspace. Or the estimator can exist but be unreliable, which is a question of concentration: it happens when the empirical gradient and Hessian have not yet settled near their expectations. There is no reason for these two failure modes to disappear at the same sample size. In logistic regression they happen to, at $n \asymp Bd$. In Poisson regression they do not: existence is secured at $n \asymp d$, with the signal strength entering only through a constant, whereas reliability requires $n \gtrsim e^{\frac{B^2}{2}(1+\varepsilon^{-1})} d^{1+\varepsilon}$, because the exponential weights concentrate only once the sample is large enough to contain many realisations of an event of probability $\asymp e^{-B^2/2}$ (and the variance must be negligible compared to the expectation for the upper bound which introduces the even bigger $n \gtrsim e^{B^2}$). The interval between the two thresholds is a real regime, not a gap in the analysis, and Theorem 4.3 shows that something, a localisation bound degraded by a factor $e^{B^2/2\delta}$, can still be said there.

We close with the questions we find to be the most natural to ask now.

Q1. Remove the d^{ε}. Even for bounded B, there remains a gap between the existence threshold $n \gtrsim d$ and the risk threshold $n \gtrsim d^{1+\varepsilon}$. The factor comes from linearising the truncation inequality, so we are convinced that it is an artefact of the method rather than of the model. A sharper treatment of the truncation, perhaps a peeling argument over the levels of $\max_i \langle u_*, X_i \rangle$ instead of a single threshold, might remove it. We state this explicitly as an open problem.

Q2. Matching lower bounds. The paper gives heuristic reasons why e^{cB^2} must appear (Section 3.2), but no formal lower bound is proven showing that the MLE genuinely fails to achieve the d/n rate when $n \ll e^{B^2/2} d$. Establishing such a statement, or an information-theoretic lower bound for any estimator in the intermediate regime, would complete the picture and would answer the more subtle question about whether the intermediate regime is a deficiency of the MLE or of the model. Perhaps there exists a genuinely better estimator than the MLE in Poisson regression, if so, it could be potentially very interesting for practitioners.

Q3. Beyond the Gaussian design. Two sub-questions, both seemingly of high difficulty: describing the classes of isotropic designs that are light-tailed enough and that satisfy the joint small-ball condition (D2); and proving an existence phase transition without rotational invariance, perhaps the hardest open question that remains, this would probably require novel theory for phase transitions in random convex programs.

Q4. Regularization. Everything here concerns the unregularized MLE. Since non-existence is caused by directions along which the loss is flat, even a vanishing ridge penalty restores existence unconditionally. An interesting question is quantitative: what is the optimal penalty level as a function of (n, d, B) in the intermediate regime; does it improves the risk bounds ?

A Notation

Symbol	Meaning
n, d	sample size, dimension
β_*, B	true parameter; $B = \max(2, \|\beta_*\|)$, the signal strength
u_*	$\beta_* / \|\beta_*\|$, the signal direction
z	confidence parameter; statements hold w.p. $\geqslant 1 - ce^{-z}$
$\ell(y, t)$	$e^t - yt$, the Poisson loss
L, $\widehat{L}_n$	population and empirical risk
$\widehat{\beta}_n$	maximum likelihood estimator
H	$e^{B^2/2}(B^2 u_* u_*^\top + I_d)$, proxy for $\nabla^2 L(\beta_*)$
H_0	$e^{B^2/\delta} I_d$, isotropic proxy used in the intermediate regime
ε, δ	free parameter in $(0, 1)$; $\delta = 2\varepsilon/(2+\varepsilon)$
$\|Z\|$	number of zero counts, $\#\{i : Y_i = 0\}$
$\delta(C)$	statistical dimension of a cone C
w_d, Δ_d	Gaussian width and diameter of a dual ball
$a \gtrsim b$, $a \asymp b$	inequality / equivalence up to universal constants
$\preceq$	Loewner order

B Two computations in dimension one

Both computations below are elementary but they carry most of the intuition, so we record them in full.

B.1 The tilting integral

Let $g \sim \mathcal{N}(0,1)$ and $b > 0$. Completing the square,

$$\mathbb{E}\big[e^{bg} h(g)\big] = \int_{\mathbb{R}} h(x)\, \frac{e^{bx - x^2/2}}{\sqrt{2\pi}}\, dx = e^{b^2/2} \int_{\mathbb{R}} h(x)\, \frac{e^{-(x-b)^2/2}}{\sqrt{2\pi}}\, dx = e^{b^2/2}\, \mathbb{E}\big[h(g+b)\big].$$

Taking $h = 1$ gives $\mathbb{E}[e^{bg}] = e^{b^2/2}$; taking $h(x) = x^2$ gives $\mathbb{E}[e^{bg} g^2] = e^{b^2/2}(1 + b^2)$, which is Lemma 2.1 in dimension one. The formula also makes the localisation of the mass explicit: after tilting, the relevant Gaussian is centered at b, and

$$\mathbb{P}(g \in [b-1, b]) \;\asymp\; \frac{e^{-b^2/2}}{b},$$

so the event carrying the expectation has probability of order $e^{-b^2/2}$ up to a polynomial factor. Everything exponential in the paper is a consequence of this display.
For the same reason,

$$\mathrm{Var}\big(e^{bg} g^2\big) \;\leqslant\; \mathbb{E}\big[e^{2bg} g^4\big] = e^{2b^2}\mathbb{E}[(g+2b)^4] \;\asymp\; b^4 e^{2b^2},$$

so the ratio of standard deviation to mean for a single observation is of order $e^{b^2/2}$, and for an average of n of order $e^{b^2/2}/\sqrt{n}$: the two-sided control requires $n \gtrsim e^{b^2}$, as claimed in Section 3.2.

B.2 The proportion of zero counts

Let $b = \|\beta_*\|$ and, as in Section 4,

$$p(b) = \mathbb{P}(Y = 0) = \mathbb{E}\big[e^{-e^{bg}}\big], \quad g \sim \mathcal{N}(0,1).$$

By symmetry of g, $p(b) = \mathbb{E}\big[\frac{1}{2}\big(e^{-e^{bg}} + e^{-e^{-bg}}\big)\big]$, and since $t \mapsto e^{-e^t} + e^{-e^{-t}}$ is increasing on $\mathbb{R}_+$, p is increasing in b. At the endpoints, $p(0) = e^{-1} \approx 0.368$, while as $b \to \infty$ the variable e^{bg} tends to 0 on $\{g < 0\}$ and to $+\infty$ on $\{g > 0\}$, so $p(b) \to \frac{1}{2}$ by dominated convergence. Crude bounds suffice for the proof:

$$\frac{1}{2e} \;\leqslant\; p(b) \;\leqslant\; \frac{e+1}{2e} \approx 0.684.$$

The upper bound is what matters: it keeps $1 - p$ bounded below by the universal constant $(e-1)/(2e)$, and this is precisely why the signal strength does not enter the existence threshold.

C AI usage disclosure

With the goal of total transparency and scientific honesty, we disclose here the impact AI, more precisely, Anthropic's model Claude Opus 5 had on this thesis.
Aside from finding typos, and improving the overall flow of the sentences, LLMs were used to write the Python code for the simulations in Section 7. No other scientific content was produced by them.

References

[1] D. Amelunxen, M. Lotz, M. B. McCoy and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. Information and Inference 3(3):224–294, 2014.

[2] F. Bach. Self-concordant analysis for logistic regression. arXiv:0910.4627, 2009.

[3] S. Boucheron, G. Lugosi and P. Massart. Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, 2013.

[4] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. C. R. Math. Acad. Sci. Paris 334(6):495–500, 2002.

[5] O. Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. IMS Lecture Notes Monograph Series 56, 2007.

[6] O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. arXiv:1712.02747, 2017.

[7] A. C. Cameron and P. K. Trivedi. Regression Analysis of Count Data. Cambridge University Press, 2013.

[8] L. M. L. Le Cam and G. L. Yang. Asymptotics in Statistics: Some Basic Concepts. Springer Series in Statistics. Springer-Verlag, 1990.

[9] E. J. Candès and P. Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. Annals of Statistics 48(1):27–42, 2020.

[10] Q. Chen and D. Giles. Finite-sample properties of the maximum likelihood estimator for the Poisson regression model with random covariates. Comm. Statist. Theory Methods 40:1000–1014, 2011.

[11] H. Chardon, M. Lerasle and J. Mourtada. Finite-sample performance of the maximum likelihood estimator in logistic regression. arXiv:2411.02137.

[12] D. Giles and H. Feng. Almost unbiased estimation of the Poisson regression model. Department of Economics, University of Victoria, Econometrics Working Papers, Jan. 2009.

[13] C. Gourieroux, A. Monfort and A. Trognon. Pseudo maximum likelihood methods: applications to Poisson models. Econometrica 52(3):701–720, 1984.

[14] S. Gupta, J. C. H. Lee and E. Price. High-dimensional location estimation via norm concentration for subgamma vectors. ICML, 2023.

[15] T. Koriyama and P. C. Bellec. Phase transitions for the existence of unregularized M-estimators in single index models. arXiv:2501.03163, 2025.

[16] F. Kuchelmeister and S. van de Geer. Finite sample rates for logistic regression with small noise or few samples. Sankhya A, 2024.

[17] M. Lerasle and A. Roche. Finite sample behavior of the maximum likelihood estimator in the Poisson model under Gaussian design. Preprint, 2026. (the article reviewed in this thesis)

[18] D. A. McAllester. PAC-Bayesian stochastic model selection. Machine Learning 51(1):5–21, 2003.

[19] P. McCullagh and J. A. Nelder. Generalized Linear Models, 2nd ed. Chapman and Hall/CRC, 1989.

[20] J. Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. Annals of Statistics 50(4):2157–2178, 2022.

[21] Y. Nesterov and A. Nemirovskii. Interior-Point Polynomial Algorithms in Convex Programming. SIAM, 1994.

[22] D. M. Ostrovskii and F. Bach. Finite-sample analysis of M-estimators using self-concordance. Electronic Journal of Statistics 15(1):326–391, 2021.

[23] R. I. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. Probab. Theory Related Fields 166(3):1175–1194, 2016.

[24] J. M. C. Santos Silva and S. Tenreyro. On the existence of the maximum likelihood estimates in Poisson regression. Economics Letters 107(2):310–312, 2010.

[25] V. Spokoiny. Parametric estimation. Finite sample theory. Annals of Statistics 40(6), 2012.

[26] M. Talagrand. Upper and Lower Bounds for Stochastic Processes. Springer, 2014.

[27] A. W. van der Vaart. Asymptotic Statistics. Cambridge University Press, 1998.

[28] A. W. van der Vaart and J. A. Wellner. A note on bounds for VC dimensions. IMS Collections 5:103–107, 2009.

[29] N. Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. Electronic Journal of Probability 29:1–28, 2024.