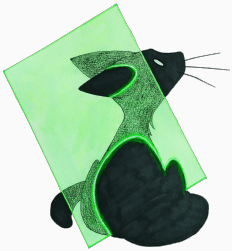
Finite sample behaviour of the MLE in the Poisson model under Gaussian design

M2 thesis – Mathematics of Randomness

Antoine Roche

2026

Laboratoire de Mathématiques d'Orsay — Supervisor: Matthieu Lerasle



Count data and the Poisson GLM

Setting in which responses are counts: insurance claims, detected particles, citations, patents.

The Gaussian linear model is structurally wrong here: it puts mass on negative values, and it forces a variance that does not depend on the mean whereas count data typically show the opposite.

Poisson regression, one of the founding examples of GLMs:

$$Y \mid X \sim \mathcal{P}\big(e^{\langle \beta_*, X\rangle}\big), \quad X \sim \mathcal{N}(0, I_d) \quad \text{(Gaussian)}.$$

so conditional mean = conditional variance $= e^{\langle \beta_*, X\rangle}$, and β_* acts on the log-scale.

Naturally, one wants to estimate β_*, as in logistic regression, we look at the conditional probabilities:

$$p_\beta(y \mid x) = e^{-e^{\langle \beta,\, x\rangle}} \frac{\exp\big(y\langle \beta,\, x\rangle\big)}{y!}$$

The loss is simply the negative log-likelihood (up to $\log y!$):

$$\ell(y, t) = e^t - yt, \quad \hat{L}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(Y_i, \langle \beta, X_i \rangle\big).$$

Assuming existence (and uniqueness here), we define the MLE as the minimiser of the log-likelihood, i.e. the empirical loss

$$\hat{\beta}_n = \underset{\beta \in \mathbb{R}^d}{\arg\min}\, \hat{L}_n(\beta).$$

$\rightarrow$ A natural question is to ask for theoretical guarantees regarding the performance of the MLE. In this case, in a non-asymptotic setting.

The asymptotic setting and benchmark

In the standard asymptotic setting (fix d and β_*, let $n \to \infty$), the theory is essentially complete: Under mild regularity conditions $\hat{\beta}_n$ exists w.p. $\to 1$, is consistent, asymptotically normal, and satisfies Wilks' theorem

$$2n\big(L(\hat{\beta}_n) - L(\beta_*)\big) \xrightarrow{(d)} \chi^2(d),$$

Combining this with Slutsky gives a rate of convergence of order d/n of the excess risk to zero. Moreover, for every $\eta > 0$ (simple deviation bound),

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

This is our benchmark i.e. what we can hope for: a bound $L(\hat{\beta}_n) - L(\beta_*) \lesssim d/n$ that holds with high probability would be the finite-sample analogue, and by Wilks it cannot be improved in order.

The limitations of the asymptotic setting

What the classical asymptotic theory does *not* say:

- Regarding the risk bound: how large n must be as a function of d and of β_* for any of this to be usable.
- Regarding the existence: how large n must be as a function of d and of β_* to guarantee that the MLE exists with high probability.

$\rightarrow$ Two (first) questions naturally ensue.

The three questions

1. When does the MLE exist? (How large must n be)

   $t \mapsto \ell(y, t)$ is strictly convex, but coercive *if and only if* $y > 0$. Zero counts contribute a term that offers no resistance to sending β off to infinity. Non-existence is a genuine failure mode (Santos Silva–Tenreyro, 2010).

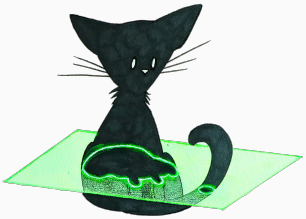
2. How large must n be for the excess risk to reach the asymptotic order d/n with high probability?

The two answers turn out *not* to match: the MLE exists with high probability before it satisfies an asymptotic-type risk bound (and the gap between the two thresholds can be made arbitrarily large). This leads to the third question:

3. What can be said in the intermediate regime in between?

Everything below is non-asymptotic and in high probability, never almost sure.

What makes the Poisson model interesting?



Why the Poisson model is worth studying

This is a legitimate question.

$\rightarrow$ Non-asymptotic results have already been obtained in many GLMs, for instance Ostrovskii and Bach [7]: assuming a self-concordant loss and sub-Gaussian design, gradient and Hessian for the model, they obtain that whenever $n \gtrsim \rho\, d^2 \log(d)$ (roughly)

$$L(\hat{\beta}_n) - L(\beta_*) \lesssim \frac{d}{n},$$

where ρ can hide (potentially very large) dependences on the signal strength $\|\beta_*\|$.

$\rightarrow$ The Poisson loss *is* self-concordant but even assuming Gaussian covariates, the gradient and Hessian have tails that are much too heavy to be sub-Gaussian random variables.

$\rightarrow$ Similarly, other methods, even generic ones [9] rely on some moment assumptions (gradient at least sub-Gamma in some neighborhood of β_* for instance).

$\rightarrow$ No results in the literature (as far as we know that study Poisson regression in a non-asymptotic setting).

Why the Poisson model is worth studying: 3 models, 3 Hessians, 3 behaviours

Chardon, Lerasle and Mourtada [3] carry out the same programme in the logistic case, keeping $B = \max(2, \|\beta_*\|)$ as a free parameter. If we consider Linear regression alongside logistic and Poisson regression. Writing their empirical Hessians $\frac{1}{n}\sum_{i=1}^{n} w_i X_i X_i^\top$ we obtain the following table.

	Linear	Logistic	Poisson
Hessian weight (w_i)	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	$n \gtrsim d$	$n \gtrsim Bd$	$n \gtrsim e^{\frac{B^2}{2}(1+\varepsilon^{-1})} d^{1+\varepsilon}$

$\rightarrow$ 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.

The signal strength is a free parameter

$B = \max(2, \|\beta_*\|)$ is treated on the same footing as n and d: no restriction on its order.

Much of the non-asymptotic literature normalises $\|\beta_*\|$ away. In GLMs this hides the problem rather than simplifying it: the signal strength changes the geometry:

- **logistic** (Candès–Sur): in proportional asymptotics the MLE exists iff d/n is below a threshold $h(\|\beta_*\|)$, with h *decreasing*: a strong signal makes separation, hence non-existence, more likely;
- **probit** (Kuchelmeister–van de Geer): the SNR decides which of two estimation regimes, with different rates, one is in;
- **Poisson**: B enters the risk thresholds exponentially, while leaving the existence threshold at constant order.

B is best read as a complexity parameter, much like the dimension. Why is that ? Let's look at the curvature.

Quantities of interest

Gradient and Hessian of the empirical risk:

$$\nabla \hat{L}_n(\beta) = \frac{1}{n}\sum_i \big(e^{\langle \beta, X_i\rangle} - Y_i\big) X_i, \qquad \nabla^2 \hat{L}_n(\beta) = \frac{1}{n}\sum_i e^{\langle \beta, X_i\rangle} X_i X_i^\top.$$

The Hessian is purely a function of the design. Curvature in the signal direction: $u_*^\top \nabla^2 \hat{L}_n(\beta_*) u_* = \frac{1}{n}\sum_i e^{Bg_i} g_i^2$, with $\mathbb{E}[e^{Bg} g^2] = e^{B^2/2}(1 + B^2)$. Where does this expectation come from?

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

After tilting, the Gaussian is recentred at B. So the expectation is carried entirely by the rare event

$$\{\, g \approx B \,\}, \qquad \mathbb{P} \approx e^{-B^2/2}, \qquad \text{on which the weight is } \approx e^{B^2}.$$

Most observations contribute essentially nothing; a few rare ones carry all the curvature.

Two thresholds, from the same picture

What this implies: For the proof (we will detail why later), one has to estimate $\nabla^2 \hat{L}_n(\beta)$, more precisely, to lower bound it with high probability by its expectation $\nabla^2 L(\beta_*)$. Necessary scales for the LLN to "kick in" (highlights why B can be seen as a complexity parameter).

Lower bound: $n \gtrsim e^{B^2/2}$

To see the rare event *at all*, one needs $n \gtrsim e^{B^2/2}$ and somewhat more for the count to concentrate.

Below that, the sample simply does not contain the observations that carry the curvature. This is a property of the model, not of the proof.

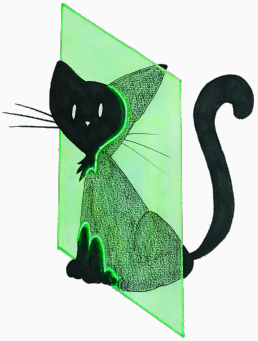
Upper bound: $n \gtrsim e^{B^2}$

To smooth the same rare events out from above one needs the variance to be negligible: $\mathbb{E}[e^{2Bg} g^4] \asymp B^4 e^{2B^2}$, so

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

An asymmetry worth noting: it is easier to guarantee that enough large values have been *seen* than that absurdly large ones have been *averaged out*. A statement about the scales at which the LLN kicks in, from below and from above.

Main results



Theorem 1 — existence

Let $X \sim \mathcal{N}(0, I_d)$, model well specified.

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

(b) If $n \geqslant C_1(d \vee z)$, the MLE exists with probability at least $1 - 6e^{-z}$.

So $n \asymp d$ is necessary and sufficient. Two comments:

- The transition is very sharp. The threshold for high-probability existence coincides, up to constants, with the one below which non-existence is *deterministic*. There is no window where non-existence is merely likely. (In logistic regression the threshold is at $n \asymp Bd$, with an exponential but not abrupt decay below.)
- The signal strength has disappeared.

Why B disappears

Existence is governed by the number of non-coercive terms, $|Z| = \#\{i : Y_i = 0\}$, through

$$n \gtrsim d + |Z|.$$

Two obstructions: the dimension (deterministic when $n < d$), and the lack of coercivity (too many zero counts).

Now $|Z| \sim \mathrm{Bin}(n, p(b))$ with $p(b) = \mathbb{E}[\exp(-e^{bg})]$, increasing in $b = \|\beta_*\|$, and

$$p(0) = e^{-1} \approx 0.37, \qquad p(b) \xrightarrow[b\to\infty]{} \tfrac{1}{2}.$$

So $1 - p$ stays bounded below by a universal constant and B is absorbed into the constant. Then sharp concentration: $|Z| \asymp np(b) \asymp n$.

Why: 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. Symmetry of the Gaussian saves us. In logistic regression a strong signal aligns the labels with $\langle \beta_*, X_i \rangle$ and directly helps separation.

Theorem 2 — the large sample regime

If for some $\varepsilon \in (0,1)$

$$n \;\geqslant\; C_2 \exp\Big(\tfrac{B^2}{2}\big(1+\varepsilon^{-1}\big)\Big)\Big((d+z)\,e^{4B\sqrt{z}}\Big)^{1+\varepsilon},$$

then with probability $\geqslant 1-6e^{-z}$ the MLE exists and $L(\hat{\beta}_n) - L(\beta_*) \;\leqslant\; C_3\,\dfrac{d+z}{n}$.

The conclusion is exactly the finite-sample form of the Wilks benchmark, optimal in order. Everything of interest is in the hypothesis. For $z \asymp 1$ it reads

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

and one may optimise $\varepsilon \in (0,1)$: a genuine trade-off between an extra factor d^{ε} and an inflation of the exponential from e^{B^2} up to $e^{B^2/2\varepsilon}$.

Where each factor comes from

- $e^{B^2/2}$: the uniform lower bound $\nabla^2 \hat{L}_n \succcurlyeq c_0 H$ near β_* one must see the tilting event;
- e^{B^2}: the upper bound at β_*, needed to control the variance of the gradient;
- $e^{4B\sqrt{z}}$: the price of the truncation level;
- d^{ε}: the price of linearising the truncation inequality.

The exponential factor is structural: a property of the Poisson model under Gaussian design, not an artefact. One may conjecture the right threshold is $e^{B^2} d$; no condition free of e^{cB^2} can suffice.

How to read it: this is the sample size at which the law of large numbers takes effect for the exponential weights. Below it the empirical Hessian is either too small (the heavy observations are missing) or wildly inflated (dominated by one of them) the MLE is perfectly well defined, but the second-order Taylor picture is not yet valid.

Theorem 3 — the intermediate regime

Let $\varepsilon \in (0,1)$, $\delta = \frac{2\varepsilon}{2+\varepsilon}$. If $n \geqslant K_\varepsilon\, d^{1+\varepsilon} e^{8z}$, then with probability $\geqslant 1 - 7e^{-z}$,

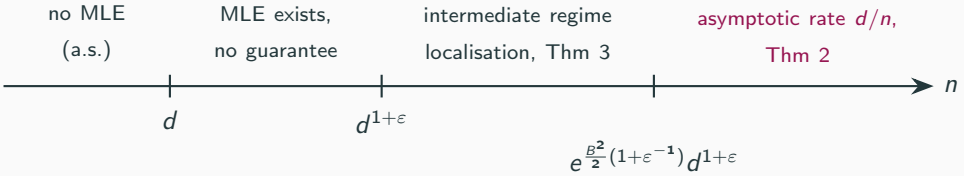
$$\left\| \hat{\beta}_n - \beta_* \right\| \;\leqslant\; C_4\, e^{B^2/2\delta} \sqrt{\frac{d+z}{n}},$$

and consequently $L(\hat{\beta}_n) - L(\beta_*) \leqslant f(B,\delta)\,\frac{d+z}{n}$.

The signal strength has moved from the hypothesis to the conclusion. The sample size condition is free of B; what one does not pay for is resolved curvature, and that ignorance shows up in the localisation radius.

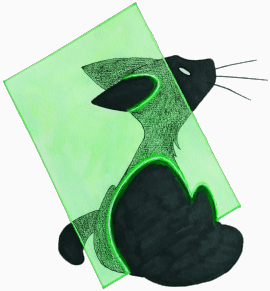
The risk version is obtained crudely, by bounding $\nabla^2 L$ over a Euclidean ball of radius $\asymp e^{B^2/2\delta}$, so f is doubly exponential in B: correct rate in n, d, astronomical constant. The informative statement is the localisation bound.

The three regimes



- The third zone is a real feature of the model: geometry and concentration have no reason to agree, and here they differ by a factor that grows with B. Theorem 3 fills it with a weaker but non-vacuous guarantee.
- The second zone (d to $d^{1+\varepsilon}$) persists even for bounded B and is almost certainly an artefact of the method we could not remove it.
- In logistic regression both middle zones are empty.

Machinery and proofs



Convex localisation

Deterministic lemma
Suppose there are $H \succ 0$ and $r_0, c_0, c_1, \nu > 0$ with

(i) $\nabla^2 \hat{L}_n(\beta) \succcurlyeq c_0 H$ on $\{\|\beta - \beta_*\|_H \leqslant r_0\}$;

(ii) $\left\|\nabla \hat{L}_n(\beta_*)\right\|_{H^{-1}} \leqslant \nu$;

(iii) $\nabla^2 L(\beta) \preccurlyeq c_1 H$ on the same ball.

If $\nu < c_0 r_0/2$, then $\hat{\beta}_n$ exists, is unique, and $\left\|\hat{\beta}_n - \beta_*\right\|_H \leqslant 2\nu/c_0$, $L(\hat{\beta}_n) - L(\beta_*) \leqslant 2\nu^2 c_1/c_0^2$.

Proof in three lines: the loss strictly increases on an H-sphere of radius r; convexity propagates this to all of $\mathbb{R}^d$; then Taylor. Everything reduces to: the gradient at the truth must be small compared with the curvature times the radius on which that curvature is guaranteed. Concretely:

(a),(b) choose $H \asymp \nabla^2 L(\beta_*)$, then find the radius: deterministic, self-concordance-like, gives $r_0 \asymp e^{B^2/4}$;

(c),(d) lower bound $\nabla^2 \hat{L}_n$ uniformly, upper bound $\left\|\nabla \hat{L}_n(\beta_*)\right\|_{H^{-1}}$: the probabilistic content.

Existence: random cones

Step 1. $b \neq 0$ Poisson separates the data if $\langle X_i, b\rangle = 0$ whenever $Y_i > 0$ and $\langle X_i, b\rangle \leqslant 0$ whenever $Y_i = 0$. Then along the ray tb the coercive terms are frozen and the others do not increase: the MLE fails to exist *iff* such a b exists (Koriyama–Bellec). If $n < d$ any b orthogonal to the design works this is part (a).

Step 2. By rotational invariance, align β_* with e_1; then (Y_i, X_{i1}) is independent of the remaining coordinates, and non-existence becomes

$$\mathrm{Span}(Xe_2, \ldots, Xe_d) \;\cap\; \mathcal{C}(Xe_1, y) \;\neq\; \{0\},$$

a uniformly random $(d-1)$-dimensional subspace meeting a fixed convex cone.

Step 3–4. The approximate kinematic formula (Amelunxen–Lotz–McCoy–Tropp): such an intersection is trivial as soon as the statistical dimensions leave room, $n - d + 1 \geqslant \delta(\mathcal{C}) + \lambda$. Here $|Z|/2 \leqslant \delta(\mathcal{C}) \leqslant 1 + |Z|$, and $|Z| \asymp np$ with $p \in [\frac{1}{e}, \frac{1}{2}]$, giving $n \gtrsim d \vee z$. Nothing is estimated anywhere: this is pure convex geometry.

Lower bound on the Hessian: discard mass

If Laplace transforms are unavailable, what is left? For a lower bound one may always throw data away:

$$e^{\langle \beta, X_i \rangle} \langle v, X_i \rangle^2 \;\geqslant\; c\, 1_{A_i},$$

which turns an unbounded problem into an empirical process of indicators and for bounded VC classes the classical toolbox (VC bounds, Talagrand, Bousquet) applies with no moment assumption at all.

Concretely: keep only the observations in the tilting region $\langle u, X_i \rangle \in [B - \gamma, B]$, sliced into intervals of width $1/B$ (so that the weight varies by at most a constant factor across a slice). The mean of the resulting process is *already* of the right order it is the tilting computation written as a Riemann sum and the fluctuations are controlled uniformly over $u, v \in S^{d-1}$.

The sample size is dictated by the smallest probability involved, $\mathbb{P}(g \in I_k) \asymp B^{-1} e^{-B^2/2}$: the classical VC condition $n \gtrsim d \log(1/p)/p$ reproduces exactly the exponential factor $e^{B^2/2}$.

Upper bound: truncation, then PAC-Bayes

Upper bounds cannot discard mass so truncate instead. On $\widehat{\Omega} = \{\max_i \langle S^{-1} X_i \rangle \leqslant \sqrt{2 \log n + 2z}\}$, of probability $\geqslant 1 - e^{-z}$, the weights are bounded by

$$M = \exp\big(B\sqrt{2 \log n + 2z}\big),$$

and the truncated variables do have a Laplace transform, sub-Gamma with parameters $\approx (e^{2B^2}, M)$. This is the step that manufactures an mgf where none existed.

One then needs a concentration inequality that is uniform over a sphere or a subspace and degrades gracefully in the sub-Gamma scale this is what PAC-Bayes provides:

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

Since M grows slower than any power of n but not by much balancing the last term is exactly what gives birth to the free parameter ε.

PAC-Bayes, and two new inequalities

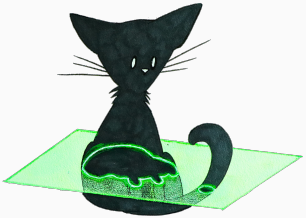
The idea. Instead of controlling $\sup_\theta f(X,\theta)$, control $\mathbb{E}_{\omega\sim\rho_\theta}[f(X,\omega)]$ *simultaneously* for a whole family of posteriors, at the cost of a divergence to a fixed prior. Uniformity comes for free: no union bound, no chaining. The tension posteriors concentrated enough for the Laplace term to be finite, spread out enough for the divergence to be small is what gets optimised.

Two inequalities of independent interest are proved:

- sub-Gamma vectors, extending Gupta–Lee–Price: closeness to zero is measured in an *arbitrary seminorm*, which is what allows the pseudo-norm $\|v\|_\star = \frac{1}{n}\max_i |\langle X_i, M^{-1/2}v\rangle|$ coming from the Poisson score;
- sub-Gamma matrices, extending Zhivotovskiy: *dimension-free*, driven by effective ranks, with the sub-Gamma scale entering only the quadratic term which is precisely what lets the truncation level M be traded against n.

Well-specification enters here, and only here: the covariance of the score *is* the Fisher information, which is why an upper bound on the Hessian is indispensable even though the localisation lemma never asks for one.

Numerical experiments



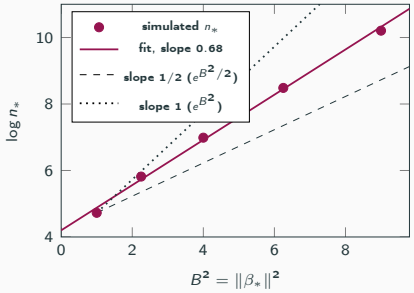
Risk: the threshold is exponential in B^2

$d = 5$; $n_*(B)$ = smallest n at which the scaled excess risk $n(L(\hat{\beta}_n) - L(\beta_*))/d$ drops below 0.6 (Wilks level: 1/2). Excess risk computed in closed form no Monte-Carlo error.

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

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

The fitted slope sits between the two heuristics, 1/2 and 1, consistent with the Hessian *upper* bound being the binding constraint.



Caveat: this locates *sufficiency* only, and B is limited to a small range the intermediate regime is out of computational reach for exactly the reason that makes it interesting.

Conclusion

Which assumptions are structural?

Assumption	Used for	Status
Isotropy $\Sigma = I_d$	everything	free (reparametrisation)
Gaussian marginals	risk bounds	near-maximal here; partly relaxable (ψ_α)
Independence on $S^{-1} \perp$	moment bounds	convenience ($e^{2B^2} \to e^{4B^2}$)
Small-ball condition	Hessian lower bd.	needed; open for which laws
Rotational invariance	existence	structural for this proof
Well-specified $Y \mid X$	gradient bound only	relaxable to correct mean + sub-Gamma

- Gaussianity is nearly maximal, not a convenience: $L(\beta)$ is finite for all β only if the marginal mgf is finite everywhere. With sub-Gamma marginals the risk is $+\infty$ outside a ball the problem is not even well posed for large B. No analogue in linear or logistic regression, where the loss is at most quadratic.
- Well-specification is the least essential: responses never enter the Hessian. Under the usual overdispersed model the conclusions survive with B replaced by $2B$: an explicit price for overdispersion, which the asymptotic theory does not give.

Conclusion and open problems

Two very different things can go wrong at small sample sizes, and no reason for them to disappear together: existence is convex geometry, reliability is concentration. Logistic regression is the coincidence; Poisson is the general case.

Q1 Remove the d^{ε}. A gap $d \to d^{1+\varepsilon}$ persists even for bounded B. It comes from linearising the truncation inequality; a peeling argument over the levels of $\max_i \langle S_{\cdot}^{-1} X_i \rangle$ might remove it.

Q2 Matching lower bounds. Does the MLE genuinely fail below $e^{B^2/2} d$? Or is the intermediate regime a deficiency of the *model* rather than of the estimator i.e. is there a better estimator than the MLE here?

Q3 Beyond the Gaussian design. Which light-tailed isotropic laws satisfy the joint small-ball condition; and existence without rotational invariance likely requiring new theory for random convex programs.

Q4 Regularisation. A vanishing ridge penalty restores existence unconditionally; what is the optimal level as a function of (n, d, B)?

Thank you for your attention.

References i

[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] 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.

[3] H. Chardon, M. Lerasle and J. Mourtada.
Finite-sample performance of the maximum likelihood estimator in logistic regression.
arXiv:2411.02137, 2024.

References ii

[4] S. Gupta, J. C. H. Lee and E. Price.

High-dimensional location estimation via norm concentration for subgamma vectors.

ICML, 2023.

[5] T. Koriyama and P. C. Bellec.

Phase transitions for the existence of unregularized M-estimators in single index models.

arXiv:2501.03163, 2025.

[6] M. Lerasle and A. Roche.

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

Preprint, 2026.

References iii

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

[8] 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.

[9] V. Spokoiny.
Parametric estimation. Finite sample theory.
Annals of Statistics, 40(6):2877–2909, 2012.

[10] 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.