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

Hugo Chardon (CREST, ENSAE)

Joint work with Matthieu Lerasle and Jaouad Mourtada (CREST, ENSAE)

Stat Math Appli 2023, Fréjus,
September 19, 2023.

Contents

Setting

Estimating conditional probabilities

- **Binary outcome** $y \in \{-1, 1\}$; covariates $x \in \mathbb{R}^d$.
- Random pair $Z = (X, Y) \sim P$ on $\mathbb{R}^d \times \{-1, 1\}$, distribution P **unknown**.

Estimating conditional probabilities

- **Binary outcome** $y \in \{-1, 1\}$; covariates $x \in \mathbb{R}^d$.
- Random pair $Z = (X, Y) \sim P$ on $\mathbb{R}^d \times \{-1, 1\}$, distribution P **unknown**.

Definition (well-specified logit model)

$$\mathbb{P}(Y = 1|X) = \sigma(\langle \theta^*, X \rangle), \quad \theta^* \in \mathbb{R}^d \tag{1}$$

where

$$\sigma : t \mapsto \frac{1}{1 + e^{-t}}$$

is the logistic (or sigmoid) function.

Estimating conditional probabilities

- **Binary outcome** $y \in \{-1, 1\}$; covariates $x \in \mathbb{R}^d$.
- Random pair $Z = (X, Y) \sim P$ on $\mathbb{R}^d \times \{-1, 1\}$, distribution P **unknown**.

Definition (well-specified logit model)

$$\mathbb{P}(Y = 1|X) = \sigma(\langle \theta^*, X \rangle), \quad \theta^* \in \mathbb{R}^d \tag{1}$$

where

$$\sigma : t \mapsto \frac{1}{1 + e^{-t}}$$

is the logistic (or sigmoid) function.

- **Goal: estimate** conditional probability $\mathbb{P}(Y = 1|X = x)$ through θ^* with the logarithmic loss

$$L(\theta) = \mathbb{E}\ell(\theta, Z) = \mathbb{E}\left[\log\left(1 + \exp(-Y\langle \theta, X \rangle)\right)\right].$$

Empirical risk minimizer

Logistic Regression: fitting the best logit model when given a random i.i.d. sample $Z_1, \ldots, Z_n \sim P$:

- Empirical risk corresponding to the logarithmic loss

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(-Y_i \langle \theta, X_i \rangle)\right).$$

- We study the empirical risk minimizer (ERM)

$$\hat{\theta}_n = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}}\, L_n(\theta).$$

Empirical risk minimizer

Logistic Regression: fitting the best logit model when given a random i.i.d. sample $Z_1, \dots, Z_n \sim P$:

- Empirical risk corresponding to the logarithmic loss

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(-Y_i \langle \theta, X_i \rangle)\right).$$

- We study the empirical risk minimizer (ERM)

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} L_n(\theta).$$

- Also the maximum likelihood estimator.

Overview of existing results

Asymptotic normality of the MLE, fast rate excess risk

Wilks' theorem: In the well-specified setting, for **fixed** d and $\theta^* \in \mathbb{R}^d$,

$$\limsup_{n\to\infty} \mathbb{P}\left(2n\big(L(\hat{\theta}_n) - L(\theta^*)\big) \geqslant 3(d+t)\right) \leqslant 1 - e^{-t}. \tag{2}$$

Asymptotic normality of the MLE, fast rate excess risk

Wilks' theorem: In the well-specified setting, for **fixed** d and $\theta^* \in \mathbb{R}^d$,

$$\limsup_{n\to\infty} \mathbb{P}\left(2n\big(L(\hat{\theta}_n) - L(\theta^*)\big) \geqslant 3(d+t)\right) \leqslant 1 - e^{-t}. \qquad (2)$$

Optimal rates.

Asymptotic: requires a fixed d and $n \to \infty$, and hides the dependency on θ^*: what happens when $\|\theta^*\| \gg 1$?

How to reach these ideal bounds ?

Question

Minimal sample size and distributional assumptions for

$$L(\hat{\theta}_n) - L(\theta^*) \leqslant C \, \frac{d+t}{n} \tag{3}$$

to hold w.p. $1 - e^{-t}$ (with C an absolute constant) ?

How to reach these ideal bounds ?

Question

Minimal sample size and distributional assumptions for

$$L(\hat{\theta}_n) - L(\theta^*) \leqslant C \, \frac{d+t}{n} \tag{3}$$

to hold w.p. $1 - e^{-t}$ (with C an absolute constant) ?

- The signal strength $B = \|\theta^*\|_\Sigma$ is a critical parameter : if $n \lesssim Bd$, the MLE a.s. **does not exist** (Candès and Sur '20).
- Otherwise it exists, but is $n \gtrsim B(d+t)$ enough to guarantee a bound like (3) ?
- No dependency on B in (3) is crucial.

Previous work on finite sample rates

- (Chinot et al. '20) If $n \gtrsim B^6 d$,

$$L(\hat{\theta}_n) - L(\theta^*) \lesssim B^5 \frac{d}{n} \text{ w.p. } 1 - e^{-d}. \tag{4}$$

Local assumptions (Bernstein condition),

broader scope (general ERM),

Previous work on finite sample rates

- (Chinot et al. '20) If $n \gtrsim B^6 d$,

$$L(\hat{\theta}_n) - L(\theta^*) \lesssim B^5 \frac{d}{n} \text{ w.p. } 1 - e^{-d}. \tag{4}$$

Local assumptions (Bernstein condition),

broader scope (general ERM),

Entanglement of confidence level and dimension,

sub-optimal bounds in the high signal-to-noise ratio (SNR).

Previous work on finite sample rates

- (Chinot et al. '20) If $n \gtrsim B^6 d$,

$$L(\hat{\theta}_n) - L(\theta^*) \lesssim B^5 \frac{d}{n} \text{ w.p. } 1 - e^{-d}. \tag{4}$$

Local assumptions (Bernstein condition),
broader scope (general ERM),
Entanglement of confidence level and dimension,
sub-optimal bounds in the high signal-to-noise ratio (SNR).

- (Ostrovskii and Bach, '21) If

$$n \gtrsim \log^8(B) B^8 d\, t \tag{5}$$

$$L(\hat{\theta}_n) - L(\theta^*) \lesssim B^3 \frac{d\, t}{n} \text{ w.p. } 1 - e^{-t}. \tag{6}$$

All confidence levels $t \geqslant 0$.

Previous work on finite sample rates

- (Chinot et al. '20) If $n \gtrsim B^6 d$,

$$L(\hat{\theta}_n) - L(\theta^*) \lesssim B^5 \frac{d}{n} \text{ w.p. } 1 - e^{-d}. \tag{4}$$

Local assumptions (Bernstein condition),
broader scope (general ERM),
Entanglement of confidence level and dimension,
sub-optimal bounds in the high signal-to-noise ratio (SNR).

- (Ostrovskii and Bach, '21) If

$$n \gtrsim \log^8(B) B^8 d\, t \tag{5}$$

$$L(\hat{\theta}_n) - L(\theta^*) \lesssim B^3 \frac{d\, t}{n} \text{ w.p. } 1 - e^{-t}. \tag{6}$$

All confidence levels $t \geqslant 0$.
Many assumptions, although local.
wrong dependency on B.

Recent works on the topic

- $\sim$ 3 months ago : (van de Geer and Kuchelmeister '23) consider the logistic regression with a *probit* model.

- (Hsu and Mazumdar '23) consider the well-specified logit model with gaussian covariates and compute the sample size required to estimate the direction of θ^*.

Main result

Main result: Gaussian design in the well specified model

Theorem (C., Lerasle, Mourtada)

If $X \sim \mathcal{N}(0, \Sigma)$ and the model is well-specified, if $n \geqslant CB(d+t)$, then w.p. $1 - e^{-t}$,

$$L(\hat{\theta}_n) - L(\theta^*) \leqslant C\,\frac{d+t}{n}. \tag{7}$$

Tight dependencies on B, d and t (match asymptotic theory).

Sharp transition from non existence of the MLE ($n \lesssim Bd$) to existence **with optimal behavior**.

Extensions

Robustness to misspecification

No modelling assumption on $Y|X$. Define

$$\theta^* = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ L(\theta) = \mathbb{E} \log(1 + \exp(-Y\langle \theta, X \rangle)) \right\}$$

as in statistical learning (well defined because L is strictly convex).

Robustness to misspecification

No modelling assumption on $Y|X$. Define

$$\theta^* = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ L(\theta) = \mathbb{E} \log(1 + \exp(-Y\langle \theta, X \rangle)) \right\}$$

as in statistical learning (well defined because L is strictly convex).

Theorem (C., Lerasle, Mourtada)

If $X \sim N(0, \Sigma)$ and **without any assumption** *on $Y|X$, if $n \geqslant C\, B(d + B^2 t)$, then w.p. $1 - e^{-t}$*

$$L(\hat{\theta}_n) - L(\theta^*) \leqslant C \log^4(B) \frac{d + B^2 t}{n}. \tag{8}$$

Robustness to misspecification

No modelling assumption on $Y|X$. Define

$$\theta^* = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ L(\theta) = \mathbb{E} \log(1 + \exp(-Y \langle \theta, X \rangle)) \right\}$$

as in statistical learning (well defined because L is strictly convex).

Theorem (C., Lerasle, Mourtada)

If $X \sim N(0, \Sigma)$ and **without any assumption** *on $Y|X$, if $n \geqslant C\, B(d + B^2 t)$, then w.p. $1 - e^{-t}$*

$$L(\hat{\theta}_n) - L(\theta^*) \leqslant C \log^4(B) \frac{d + B^2 t}{n}. \tag{8}$$

Does not match the well-specified setting bound but significantly improve existing results.

No assumption whatsoever on the link between X and Y.

Design relaxation in the well-specified setting

The Gaussian design assumption is not necessary !

Theorem (C., Lerasle, Mourtada)

In the well-specified model, for more general designs (technical conditions), if $n \geqslant C\, B(d \log B + t)$*, w.p.* $1 - e^{-t}$*,*

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

Proofs Ideas

Outline

Unified framework for the different proofs: localize $\hat{\theta}_n$ by controlling

$$L_n(\theta) - L_n(\theta^*) = \langle \nabla L_n(\theta^*), \theta - \theta^* \rangle + \left\| H_n\big(\tilde{\theta}\big)^{1/2} (\theta - \theta^*) \right\|^2$$

locally by

- bounding from above the gradient at θ^*,
- bounding from below the Hessians around θ^* uniformly.

Deviations of the gradient

Let $\tilde{g}$ denote the suitably rescaled gradient.

- We want to control

$$\left\| H^{-1/2} \nabla L_n(\theta^*) \right\| = \sup_{v \in S^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \langle v, \tilde{g}_i \rangle.$$

  Unbounded but sub-gaussian empirical process.

- Vanilla sub-Gaussian deviations: w.p. $1 - e^{-t}$,

$$\left\| H^{-1/2} \nabla L_n(\theta^*) \right\|^2 \lesssim B^3 \frac{d+t}{n}. \tag{9}$$

- Replace sub-Gaussian norm by **variance**. No distinction leads to bad dependencies on B.

Deviations of the gradient

- Weak variance → Talagrand type inequality. $(X_{i,t})_{t \in T, i \in [n]}$ a bounded centered process ($|X_t| \leqslant b$ a.s.),

$$Z = \sup_{t \in T} \frac{1}{n} \left| \sum_{i=1}^{n} X_{i,t} \right| \quad \sigma^2 = \sup_{t \in T} \mathbb{E} X_t^2,$$

  w.p. $1 - e^{-t}$

$$Z \lesssim \mathbb{E} Z + \sigma \sqrt{t} + bt.$$

  We need a version for **unbounded** processes.
- B^3 from the worst direction. "Super Bernstein" with Sub-Gaussian or sub-exponential norms ? No, leads to a residual B^3 in the second order term.
- Key is **sub-gamma** bounds !

Bounding from below empirical Hessians

- The Hessian does not depend on the conditional distribution of $Y|X$.
- Control the **uniform** lower tail of a collection of random matrices:

$$\begin{aligned}\inf_{\theta\in\Theta}\lambda_{\min}(H^{-1/2}H_n(\theta)H^{-1/2}) &= \inf_{(\theta,v)\in\Theta\times S^{d-1}}\left\langle H^{-1/2}H_n(\theta)H^{-1/2}v,v\right\rangle\\ &= \inf_{(\theta,v)}\frac{1}{n}\sum_{i=1}^{n}\sigma'(\langle\theta,X_i\rangle)\langle v,H^{-1/2}X_i\rangle^2.\end{aligned}$$

- For a single matrix: lower bounds from (Oliveira '16) and (Zhivotovskiy '21). Additional technical difficulty due to the uniformity over Θ and the non linearity of σ' . We adapt the PAC- Bayesian approach.

Conclusion

- **Take home message:** In the well specified setting, as soon as the maximum likelihood estimator exists, it satisfies the optimal bound known from the asymptotic theory !
- A nearly-optimal result still holds in the case of a misspecified model.
- This remains true with much more general designs.

Thank you!

Design relaxation

Exemple of sufficient design conditions. Denote by $V = \langle v, X \rangle$ the projection of X in the direction $v \in S^{d-1}$ and f_V its density. Similarly $f_{U,V}$ the joint density of $\langle u, X \rangle$ and $\langle v, X \rangle$.

- (Sub-exponential design.) For all $v \in S^{d-1}$, $\|\langle v, X \rangle\|_{\psi_1} \leqslant K$,
- (Bounded densities of the one-dimensional marginals.) $\exists M > m > 0$ s.t. $\forall v \in S^{d-1}$,

$$\forall t \in [-1, 1],\ f_V(t) \geqslant m; \quad \forall t \in \mathbb{R},\ f_V(t) \leqslant M. \tag{10}$$

- (dim 2 marginals) $u^* =$ direction of θ^*. $\exists M_2 > m_2 > 0$ s.t. $\forall v \in S^{d-1}$,

$$\forall (s, t) \in [-1, 1]^2,\ f_{U^*,V}(s, t) \geqslant m_2$$
$$\forall (s, t) \in \mathbb{R}^2,\ f_{U^*,V}(s, t) \leqslant M_2.$$