Logistic Regression with small noise or few samples[1] Fréjus 2023

Felix Kuchelmeister,
joint work with Sara van de Geer,
ETH Zürich

September 19, 2023

[1] Based on Kuchelmeister and van de Geer [2023].

Default (GPT-3.5)

FE

Give one example of a classification algorithm

Default (GPT-3.5)

FE

Give one example of a classification algorithm

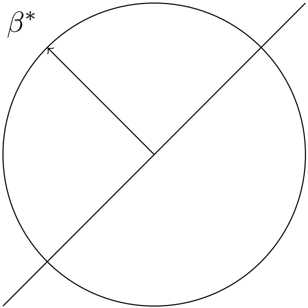
One example of a classification algorithm is the **Logistic Regression** algorithm. Logistic Regression is a supervised learning algorithm used for binary or multi-class classification problems. It is widely used in various fields, including healthcare for disease prediction,

What is logistic regression?

- Data: features $x_i \in \mathbb{R}^p$ and labels $y_i \in \{-1, +1\}$.
- $y_i = sign(x_i^T \beta^* + \sigma\epsilon_i)$, $\|\beta^*\|_2 = 1$, $\sigma > 0$.

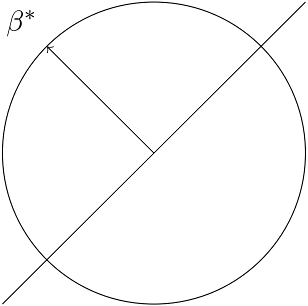
What is logistic regression?

- Data: features $x_i \in \mathbb{R}^p$ and labels $y_i \in \{-1, +1\}$.
- $y_i = sign(x_i^T \beta^* + \sigma \epsilon_i)$, $\|\beta^*\|_2 = 1$, $\sigma > 0$.



What is logistic regression?

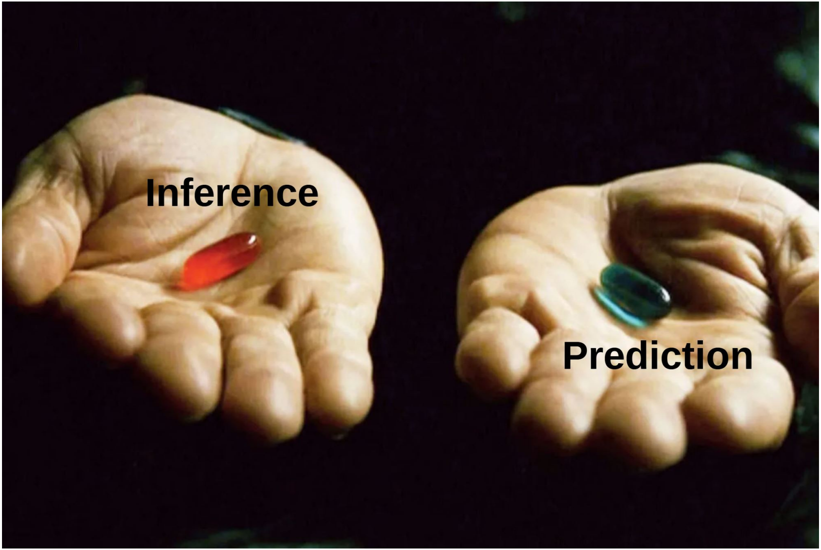
- Data: features $x_i \in \mathbb{R}^p$ and labels $y_i \in \{-1, +1\}$.
- $y_i = sign(x_i^T \beta^* + \sigma\epsilon_i)$, $\|\beta^*\|_2 = 1$, $\sigma > 0$.



Logistic regression is...

$$\arg\min_{\gamma \in \mathbb{R}^p} \sum_{i=1}^{n} \log(1 + \exp(-y_i x_i^T \gamma))$$

What is logistic regression used for?



Problems of logistic regression

- Distribution of estimator $\hat{\gamma}$ difficult to calculate

[2] Hauck Jr and Donner [1977]
[3] Nemes et al. [2009]
[4] Zhao, Sur, and Candes [2022]

Problems of logistic regression

- Distribution of estimator $\hat{\gamma}$ difficult to calculate
- Asymptotically normal.

[2] Hauck Jr and Donner [1977]
[3] Nemes et al. [2009]
[4] Zhao, Sur, and Candes [2022]

Problems of logistic regression

- Distribution of estimator $\hat{\gamma}$ difficult to calculate
- Asymptotically normal.
- Approximation is bad if:

[2] Hauck Jr and Donner [1977]
[3] Nemes et al. [2009]
[4] Zhao, Sur, and Candes [2022]

Problems of logistic regression

- Distribution of estimator $\hat{\gamma}$ difficult to calculate
- Asymptotically normal.
- Approximation is bad if:

σ small[2]

---

[2] Hauck Jr and Donner [1977]
[3] Nemes et al. [2009]
[4] Zhao, Sur, and Candes [2022]

Problems of logistic regression

- Distribution of estimator $\hat{\gamma}$ difficult to calculate
- Asymptotically normal.
- Approximation is bad if:

σ small[2]

n small[3]

---

[2] Hauck Jr and Donner [1977]
[3] Nemes et al. [2009]
[4] Zhao, Sur, and Candes [2022]

Problems of logistic regression

- Distribution of estimator $\hat{\gamma}$ difficult to calculate
- Asymptotically normal.
- Approximation is bad if:

σ small[2]

n small[3]

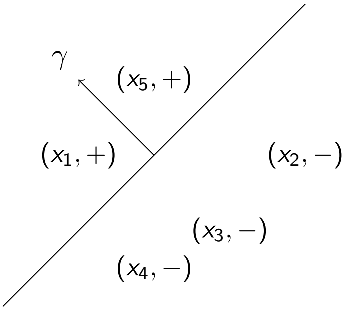
p large[4]

[2] Hauck Jr and Donner [1977]
[3] Nemes et al. [2009]
[4] Zhao, Sur, and Candes [2022]

What's the problem?

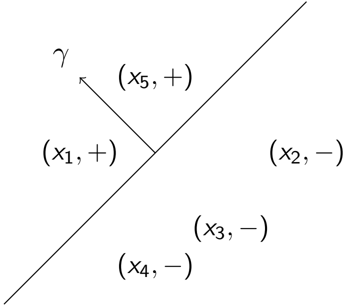
What's the problem?

- Linear separation



What's the problem?

- Linear separation



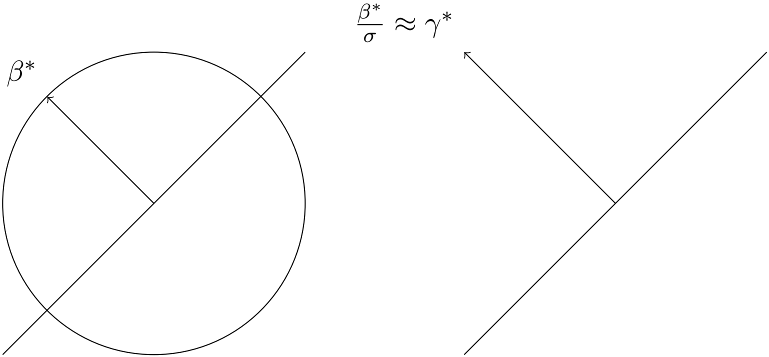
- Monotone likelihood: $\gamma \mapsto \log(1 + \exp(-yx^T\gamma))$.
- $\|\gamma\|_2 \nearrow \infty$ implies Loss $\searrow 0$.

- This is likely, if: $\sigma \approx 0$, $n \ll \infty$, $p \gg 1$.

What's the problem?

"there is an urgent need for new research to provide guidance for supporting sample size considerations for binary logistic regression"
van Smeden et al. [2016]

The model

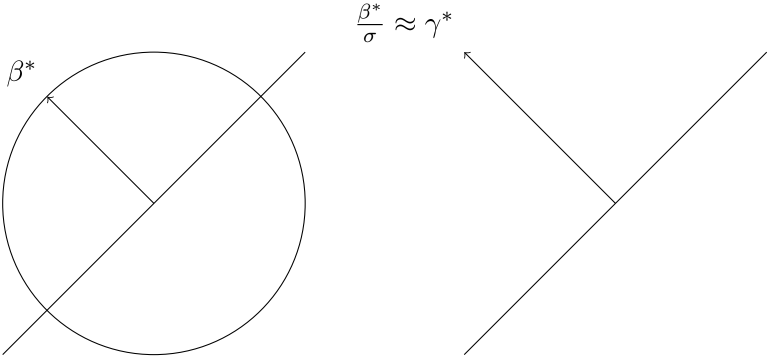


n i.i.d. observations $(x_i, y_i) \in \mathbb{R}^p \times \{-1, +1\}$, where:

$$y_i := sign(x_i^T \beta^* + \sigma \epsilon_i)$$

Parameters $\beta^* \in S^{p-1}$ and $\sigma > 0$ unknown. We assume $(x, \epsilon) \sim \mathcal{N}(0, I_{p+1})$.

The model



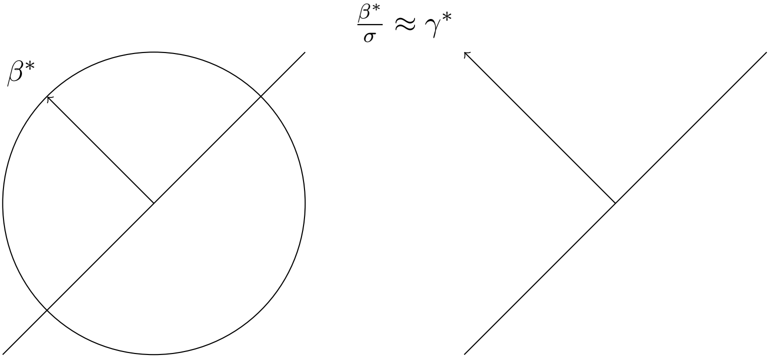
n i.i.d. observations $(x_i, y_i) \in \mathbb{R}^p \times \{-1, +1\}$, where:

$$y_i := sign(x_i^T \beta^* + \sigma \epsilon_i)$$

Parameters $\beta^* \in S^{p-1}$ and $\sigma > 0$ unknown. We assume $(x, \epsilon) \sim \mathcal{N}(0, I_{p+1})$.

$$\gamma^* := \arg\min_{\gamma \in \mathbb{R}^p} \mathbb{E} \log(1 + \exp(-y x^T \gamma))$$

The model



n i.i.d. observations $(x_i, y_i) \in \mathbb{R}^p \times \{-1, +1\}$, where:

$$y_i := sign(x_i^T \beta^* + \sigma \epsilon_i)$$

Parameters $\beta^* \in S^{p-1}$ and $\sigma > 0$ unknown. We assume $(x, \epsilon) \sim \mathcal{N}(0, I_{p+1})$.

$$\gamma^* := \arg\min_{\gamma \in \mathbb{R}^p} \mathbb{E} \log(1 + \exp(-yx^T\gamma))$$

Estimation

$$\gamma^* := \arg\min_{\gamma \in \mathbb{R}^p} \mathbb{E} \log(1 + \exp(-yx^T\gamma))$$

Estimation

$$\gamma^* := \arg\min_{\gamma \in \mathbb{R}^p} \mathbb{E} \log(1 + \exp(-yx^T\gamma))$$

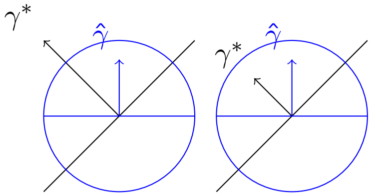
$$\text{“}\hat{\gamma}_\infty := \arg\min_{\gamma \in \mathbb{R}^p} \sum_{i=1}^{n} \log(1 + \exp(-y_i x_i^T \gamma))\text{”}$$

Estimation

$$\gamma^* := \arg\min_{\gamma \in \mathbb{R}^p} \mathbb{E} \log(1 + \exp(-yx^T\gamma))$$

$$\text{“}\hat{\gamma}_\infty := \arg\min_{\gamma \in \mathbb{R}^p} \sum_{i=1}^{n} \log(1 + \exp(-y_i x_i^T \gamma))\text{”}$$

$$\hat{\gamma}_M := \arg\min_{\|\gamma\|_2 \leq M} \sum_{i=1}^{n} \log(1 + \exp(-y_i x_i^T \gamma))$$



Classical asymptotics

E.g. van der Vaart [2000]:

$$\sqrt{n}(\hat{\gamma}_\infty - \gamma^*) \to \mathcal{N}(0, I_{\sigma,\beta^*}^{-1})$$

Classical asymptotics

E.g. van der Vaart [2000]:

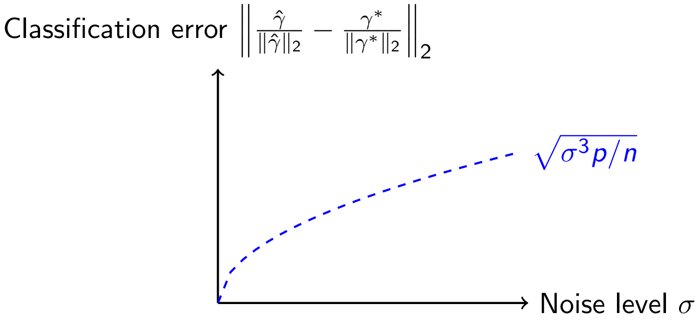
$$\sqrt{n}(\hat{\gamma}_\infty - \gamma^*) \to \mathcal{N}(0, I_{\sigma,\beta^*}^{-1})$$

Gives asymptotic rate ($\sigma \lesssim 1$):

$$\sqrt{\frac{p}{n\sigma}} \lesssim \|\hat{\gamma}_\infty - \gamma^*\|_2 \lesssim \sqrt{\frac{p}{n\sigma^3}}$$

Weird.

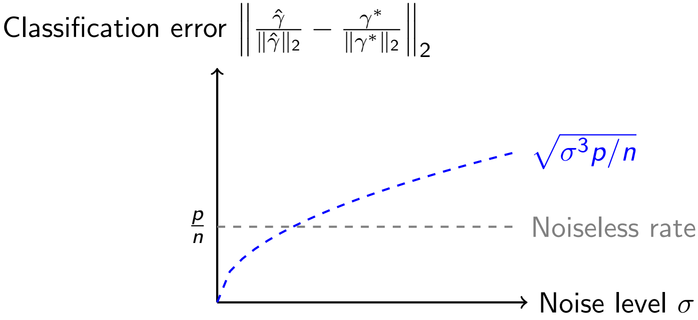
Solution: Treat classification separately



- Asymptotic upper bound[5]: $\sqrt{\sigma^3 p/n}$ if $\sigma \lesssim 1$.

[5] Taking $\|\hat{\gamma}_\infty - \gamma^*\|_2 \sim \sqrt{\frac{p}{n\sigma}}$

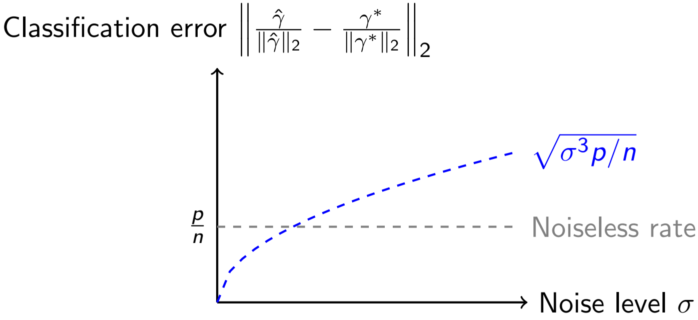
Solution: Treat classification separately



- Asymptotic upper bound[6]: $\sqrt{\sigma^3 p/n}$ if $\sigma \lesssim 1$.
- Finite sample rate p/n if $\sigma = 0$ [Balcan and Long, 2013]

[6]Taking $\|\hat{\gamma}_\infty - \gamma^*\|_2 \sim \sqrt{\frac{p}{n\sigma}}$

Solution: Treat classification separately



- Asymptotic upper bound[6]: $\sqrt{\sigma^3 p/n}$ if $\sigma \lesssim 1$.
- Finite sample rate p/n if $\sigma = 0$ [Balcan and Long, 2013]
- This cannot be the finite sample rate!
  What happens if σ is small?

[6]Taking $\|\hat{\gamma}_\infty - \gamma^*\|_2 \sim \sqrt{\frac{p}{n\sigma}}$

Solution: Large and small noise regime

Noise level $\sigma \sim \frac{1}{\|\gamma^*\|_2}$	Small	Large

Solution: Large and small noise regime

Noise level $\sigma \sim \frac{1}{\|\gamma^*\|_2}$	Small	Large
Classification $\left\| \frac{\hat{\gamma}}{\|\hat{\gamma}\|_2} - \frac{\gamma^*}{\|\gamma^*\|_2} \right\|_2$	Easy	Hard

Solution: Large and small noise regime

Noise level $\sigma \sim \frac{1}{\\|\gamma^*\\|_2}$	Small	Large
Classification $\left\\|\frac{\hat{\gamma}}{\\|\hat{\gamma}\\|_2} - \frac{\gamma^*}{\\|\gamma^*\\|_2}\right\\|_2$	Easy	Hard
Confidence $\lvert\\|\hat{\gamma}\\|_2 - \\|\gamma^*\\|_2\rvert$	Hard	Easy

Solution: Large and small noise regime

Noise level $\sigma \sim \frac{1}{\|\gamma^*\|_2}$	Small	Large
Classification $\left\| \frac{\hat{\gamma}}{\|\hat{\gamma}\|_2} - \frac{\gamma^*}{\|\gamma^*\|_2} \right\|_2$	Easy	Hard
Confidence $\left| \|\hat{\gamma}\|_2 - \|\gamma^*\|_2 \right|$	Hard	Easy

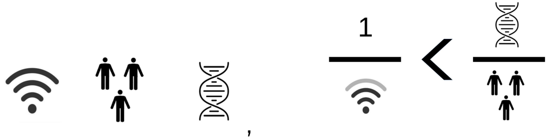
- What is 'small/large'?
- Problems if strong signal, few observations or high dimension


,

Solution: Large and small noise regime

Noise level $\sigma \sim \frac{1}{\|\gamma^*\|_2}$	Small	Large
Classification $\left\| \frac{\hat{\gamma}}{\|\hat{\gamma}\|_2} - \frac{\gamma^*}{\|\gamma^*\|_2} \right\|_2$	Easy	Hard
Confidence $\left| \|\hat{\gamma}\|_2 - \|\gamma^*\|_2 \right|$	Hard	Easy

- What is ‘small/large’?
- Problems if strong signal, few observations or high dimension



- $\sigma \leq \frac{p}{n}$

Main result

Theorem (K & van de Geer, 2023)

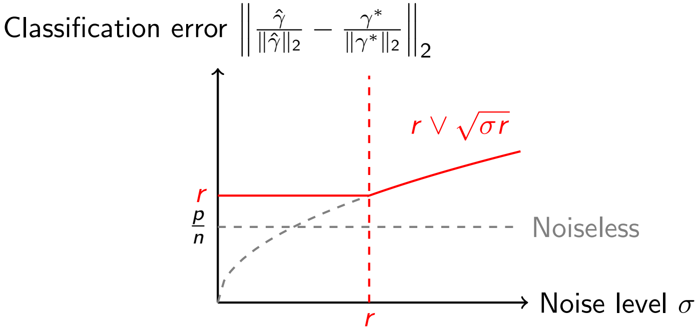
Let $t > 0$ and:

$$r := \frac{p \log n + t}{n} \lesssim 1, \quad M \gtrsim \frac{1}{r}.$$

Then with probability at least $1 - 5\exp(-t)$,

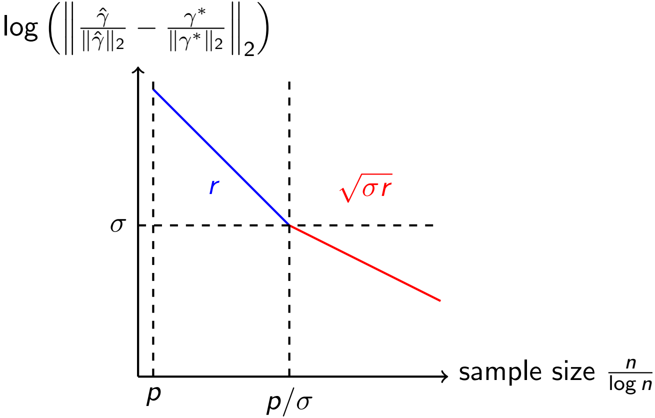
Regime	$\sigma \lesssim r$	$r \lesssim \sigma \lesssim 1$
Classification	$\left\Vert \frac{\hat{\gamma}}{\Vert\hat{\gamma}\Vert} - \frac{\gamma^*}{\Vert\gamma^*\Vert_2} \right\Vert_2 \lesssim r$	$\left\Vert \frac{\hat{\gamma}}{\Vert\hat{\gamma}\Vert} - \frac{\gamma^*}{\Vert\gamma^*\Vert_2} \right\Vert_2 \lesssim \sqrt{\sigma r}$
Confidence	$\Vert\hat{\gamma}\Vert_2 \gtrsim \frac{1}{r}$	$\lvert \Vert\hat{\gamma}\Vert_2 - \Vert\gamma^*\Vert_2 \rvert \lesssim \sqrt{\frac{r}{\sigma^3}}$

Classification error VS noise level

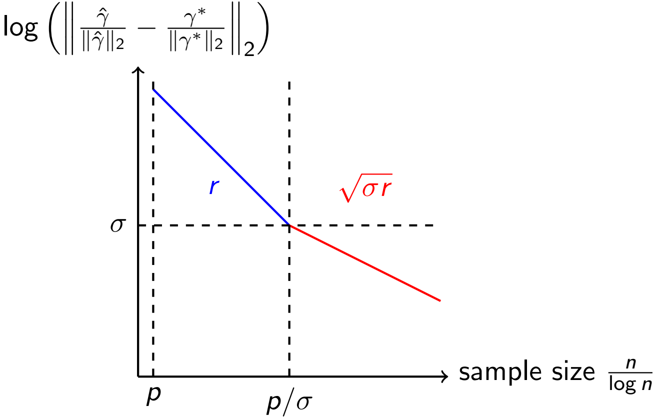


Here $r := \frac{p \log n}{n}$.

Classification error VS sample size



Classification error VS sample size



- Improving performance is “cheaper” for small n!

How do we know which regime occurs?

Recall that $r := \frac{p \log n}{n}$.

Regime	$\sigma \lesssim r$	$r \lesssim \sigma \lesssim 1$
Confidence	$\|\hat{\gamma}\|_2 \gtrsim \frac{1}{r}$	$\left\| \|\hat{\gamma}\|_2 - \|\gamma^*\|_2 \right\| \lesssim \sqrt{\frac{r}{\sigma^3}}$

How do we know which regime occurs?

Recall that $r := \frac{p \log n}{n}$.

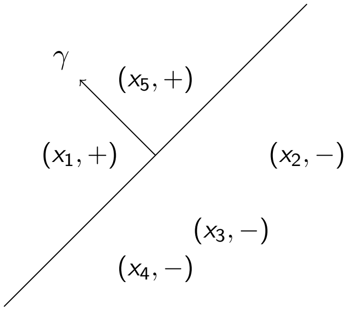
Regime	$\sigma \lesssim r$	$r \lesssim \sigma \lesssim 1$
Confidence	$\|\hat{\gamma}\|_2 \gtrsim \frac{1}{r}$	$\big\lvert \|\hat{\gamma}\|_2 - \|\gamma^*\|_2 \big\rvert \lesssim \sqrt{\frac{r}{\sigma^3}}$

It follows that:

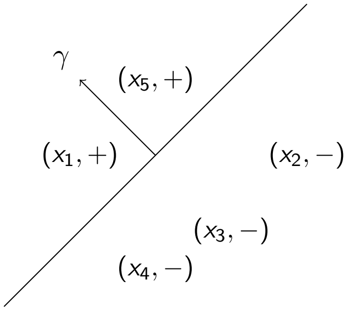
$$\|\hat{\gamma}\|_2 \gtrsim \frac{n}{p \log n} \Rightarrow \text{small noise regime}$$

$$\|\hat{\gamma}\|_2 \lesssim \frac{n}{p \log n} \Rightarrow \text{large noise regime}$$

What can we say if the data is separable?

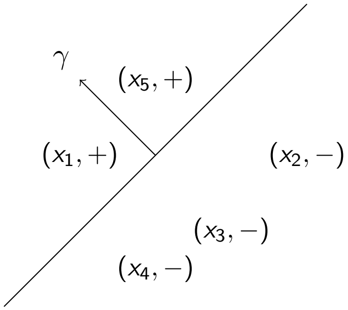


What can we say if the data is separable?



- 'large noise' $\sigma \gtrsim p \log(n)/n \Rightarrow$ not separable

What can we say if the data is separable?



- 'large noise' $\sigma \gtrsim p \log(n)/n \Rightarrow$ not separable
- Separable $\Rightarrow$ not large noise! (whp)
- Same rate as noiseless case (up to $\log n$)

Some ideas of proof: $\sigma \gtrsim r$

- Split loss in two parts, treat separately:

$$\log(1 + \exp(-|x^T\gamma|)) + |x^T\gamma| 1\{yx^T\gamma < 0\}$$

Some ideas of proof: $\sigma \gtrsim r$

- Split loss in two parts, treat separately:

$$\log(1 + \exp(-|x^T\gamma|)) + |x^T\gamma| 1\{yx^T\gamma < 0\}$$

- Quantify distance to γ^* with:

$$d_*(\gamma) := \sqrt{\|\gamma^*\|_2 \left\| \frac{\gamma}{\|\gamma\|_2} - \frac{\gamma^*}{\|\gamma^*\|_2} \right\|_2^2 + \frac{|\|\gamma\|_2 - \|\gamma^*\|_2|^2}{\|\gamma^*\|_2^3}}$$

Some ideas of proof: $\sigma \gtrsim r$

- Split loss in two parts, treat separately:

$$\log(1+\exp(-|x^T\gamma|)) + |x^T\gamma| 1\{yx^T\gamma < 0\}$$

- Quantify distance to γ^* with:

$$d_*(\gamma) := \sqrt{\|\gamma^*\|_2 \left\| \frac{\gamma}{\|\gamma\|_2} - \frac{\gamma^*}{\|\gamma^*\|_2} \right\|_2^2 + \frac{|\|\gamma\|_2 - \|\gamma^*\|_2|^2}{\|\gamma^*\|_2^3}}$$

- Lower bound excess risk with Taylor expansion + convexity

Some ideas of proof: $\sigma \gtrsim r$

- Split loss in two parts, treat separately:

$$\log(1 + \exp(-|x^T\gamma|)) + |x^T\gamma| 1\{yx^T\gamma < 0\}$$

- Quantify distance to γ^* with:

$$d_*(\gamma) := \sqrt{\|\gamma^*\|_2 \left\| \frac{\gamma}{\|\gamma\|_2} - \frac{\gamma^*}{\|\gamma^*\|_2} \right\|_2^2 + \frac{|\|\gamma\|_2 - \|\gamma^*\|_2|^2}{\|\gamma^*\|_2^3}}$$

- Lower bound excess risk with Taylor expansion + convexity
- Upper bound excess risk with empirical process theory
  Bernstein & Bousquet's inequality, localization, peeling

Some ideas of proof: $\sigma \lesssim r$

- Problem: possibly $\|\gamma^*\|_2 \not\leq M$, maybe $P_n l(\gamma^*) < P_n l(\hat{\gamma})$.

Some ideas of proof: $\sigma \lesssim r$

- Problem: possibly $\|\gamma^*\|_2 \not\lesssim M$, maybe $P_n l(\gamma^*) < P_n l(\hat{\gamma})$.
- Exploit linearity of $\|\gamma\|_2 \mapsto |x^T\gamma| 1\{yx^T\gamma < 0\}$,
  compare $\hat{\gamma}$ to $\|\hat{\gamma}\|_2 \frac{\gamma^*}{\|\gamma^*\|_2}$.
  Exploit that $\log(1 + \exp(-|x^T\gamma|))$ is small if $\|\gamma\|_2$ is huge.

Some ideas of proof: $\sigma \lesssim r$

- Problem: possibly $\|\gamma^*\|_2 \not\leq M$, maybe $P_n l(\gamma^*) < P_n l(\hat{\gamma})$.
- Exploit linearity of $\|\gamma\|_2 \mapsto |x^T\gamma| 1\{yx^T\gamma < 0\}$, compare $\hat{\gamma}$ to $\|\hat{\gamma}\|_2 \frac{\gamma^*}{\|\gamma^*\|_2}$.
  Exploit that $\log(1 + \exp(-|x^T\gamma|))$ is small if $\|\gamma\|_2$ is huge.
- New problem: need to control $\|\hat{\gamma}\|_2$.
  Three case distinctions:

$$\|\hat{\gamma}\|_2 \geq M/2 \geq \|\hat{\gamma}\|_2 \geq 6 \geq \|\hat{\gamma}\|_2.$$

Some ideas of proof: $\sigma \lesssim r$

- Problem: possibly $\|\gamma^*\|_2 \not\leq M$, maybe $P_n l(\gamma^*) < P_n l(\hat{\gamma})$.
- Exploit linearity of $\|\gamma\|_2 \mapsto |x^T\gamma| 1\{yx^T\gamma < 0\}$,
  compare $\hat{\gamma}$ to $\|\hat{\gamma}\|_2 \frac{\gamma^*}{\|\gamma^*\|_2}$.
  Exploit that $\log(1 + \exp(-|x^T\gamma|))$ is small if $\|\gamma\|_2$ is huge.
- New problem: need to control $\|\hat{\gamma}\|_2$.
  Three case distinctions:

$$\|\hat{\gamma}\|_2 \geq M/2 \geq \|\hat{\gamma}\|_2 \geq 6 \geq \|\hat{\gamma}\|_2.$$

- Lower bound excess risk using Gaussian tail bounds.

Some ideas of proof: $\sigma \lesssim r$

- Problem: possibly $\|\gamma^*\|_2 \not\leq M$, maybe $P_n l(\gamma^*) < P_n l(\hat{\gamma})$.
- Exploit linearity of $\|\gamma\|_2 \mapsto |x^T\gamma| 1\{yx^T\gamma < 0\}$,
  compare $\hat{\gamma}$ to $\|\hat{\gamma}\|_2 \frac{\gamma^*}{\|\gamma^*\|_2}$.
  Exploit that $\log(1 + \exp(-|x^T\gamma|))$ is small if $\|\gamma\|_2$ is huge.
- New problem: need to control $\|\hat{\gamma}\|_2$.
  Three case distinctions:

$$\|\hat{\gamma}\|_2 \geq M/2 \geq \|\hat{\gamma}\|_2 \geq 6 \geq \|\hat{\gamma}\|_2.$$

- Lower bound excess risk using Gaussian tail bounds.
- Upper bound similar as before (angles - easier).

Final slide

Logistic regression has problems if:

σ small[7]

n small[8]

p large[9]

---

[7] Hauck Jr and Donner [1977]
[8] Nemes et al. [2009]
[9] Zhao, Sur, and Candes [2022]

Final slide

Logistic regression has problems if:



σ small[7] n small[8] p large[9]

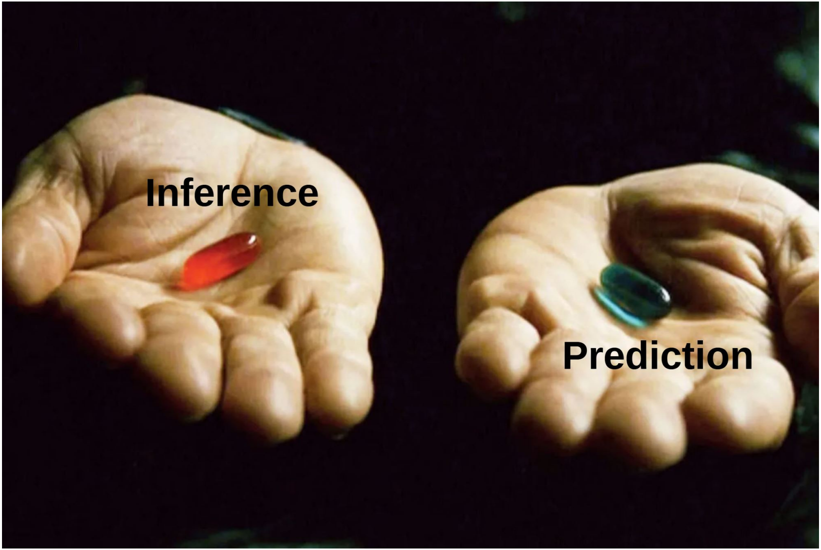
New:

- Fast classification if $\|\hat{\gamma}\|_2 \gtrsim \frac{n}{p \log n}$,

- Parametric rate if $\|\hat{\gamma}\|_2 \lesssim \frac{n}{p \log n}$,

[7] Hauck Jr and Donner [1977]
[8] Nemes et al. [2009]
[9] Zhao, Sur, and Candes [2022]

Merci pour votre attention!



References I

Maria-Florina Balcan and Phil Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316. PMLR, 2013.

Walter W Hauck Jr and Allan Donner. Wald's test as applied to hypotheses in logit analysis. *Journal of the american statistical association*, 72(360a):851–853, 1977.

Felix Kuchelmeister and Sara van de Geer. Finite sample rates for logistic regression with small noise or few samples. *arXiv preprint arXiv:2305.15991*, 2023.

Szilard Nemes, Junmei Miao Jonasson, Anna Genell, and Gunnar Steineck. Bias in odds ratios by logistic regression modelling and sample size. *BMC medical research methodology*, 9:1–5, 2009.

Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

References II

Maarten van Smeden, Joris AH de Groot, Karel GM Moons, Gary S Collins, Douglas G Altman, Marinus JC Eijkemans, and Johannes B Reitsma. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC medical research methodology*, 16:1–12, 2016.

Qian Zhao, Pragya Sur, and Emmanuel J Candes. The asymptotic distribution of the mle in high-dimensional logistic models: Arbitrary covariance. *Bernoulli*, 28(3):1835–1861, 2022.