On the sample complexity of parameter estimation in logistic regression with normal design

Daniel Hsu (Columbia)
Arya Mazumdar (UCSD)

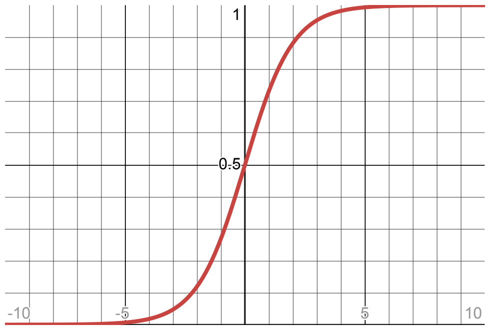
COLT 2024

Data model for noisy binary classification

- $(X, Y) \sim P_w$: covariate vector $X \sim \mathrm{N}(0, I_d)$; binary response Y

$$\Pr_w(Y = y \mid X = x) = \frac{1}{1 + e^{-y\langle x, w\rangle}}, \qquad \forall x \in \mathbb{R}^d, y \in \{-1,1\}$$

"logistic regression"



- Parameter $w \in \mathbb{R}^d$
- **Estimation goal**: Given i.i.d. sample from $P_{w^\star}$ ($w^\star$ unknown), construct estimate $\hat{w}$ such that

$$\|\hat{w} - w^\star\| \le \epsilon$$

How large should the sample size be?

Clues from classical asymptotic theory?

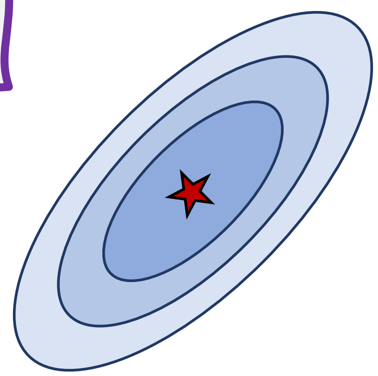
- Maximum likelihood estimator given data $(x_i, y_i)_{i=1}^n$

$$\hat{w}_{\text{mle}} = \arg\min_w \sum_{i=1}^n \ln\left(1 + e^{-y_i \langle x_i, w \rangle}\right)$$

MLE may not exist!

- Asymptotically (as $n \to \infty$),

$$\sqrt{n}\left(\hat{w}_{\text{mle}} - w^\star\right) \xrightarrow{\text{dist.}} \mathrm{N}(0, \mathcal{J}(w^\star)^{-1})$$



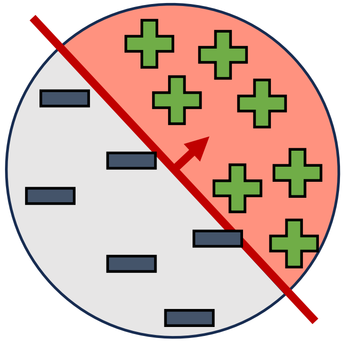
- Very roughly:

$$\mathbb{E}\|\hat{w}_{\text{mle}} - w^\star\| \to \sqrt{d/n}$$

Dependence on $\|w^\star\|$?

- **"Conclusion":** sample complexity is d/ϵ^2 ???

Learning half-spaces



- As $\|w^\star\| \to \infty$, response Y is determined by X:
$$Y = \text{sign}(\langle X, \theta^\star \rangle)$$
where $\theta^\star = w^\star / \|w^\star\|$
- PAC learning homogeneous half-spaces under $X \sim \text{Uniform}(S^{d-1})$
- **Long (1995, 2003):** sample complexity is d/ϵ (cf. d/ϵ^2)
  - … to guarantee **classification error rate** $\leq \epsilon$
  - … which is proportional to **parameter error** $\|\hat{\theta} - \theta^\star\|$
- AKA "1-bit compressed sensing"

Question

- What is the role of $\|w^\star\|$?
  - Fix $\|w^\star\| = \beta$, and only consider estimating $\theta^\star = w^\star/\|w^\star\| \in S^{d-1}$
  - β is akin to **signal-to-noise ratio**, also called **inverse temperature**

$$\Pr_{\beta\theta^\star}(Y = 1 \mid X = x) = \frac{1}{1 + \exp(-\beta\langle x, \theta^\star\rangle)}$$

  - No signal ($\beta = 0$): hopeless
  - No noise ($\beta = \infty$): PAC learning half-spaces
- **Revised goal**: Given i.i.d. sample from $P_{\beta\theta^\star}$ for some unknown $\theta^\star$, construct estimate $\hat{\theta}$ such that

$$\|\hat{\theta} - \theta^\star\| \le \epsilon$$

Main result

- Sample complexity* to ensure $\|\hat{\theta} - \theta^\star\| \leq \epsilon$ (in expectation or w.h.p.):

$$n^*(d, \epsilon, \beta) \asymp \begin{cases} \dfrac{d}{\beta^2 \epsilon^2} & \text{if } \beta \lesssim 1 & \text{"high temperature"} \\ \dfrac{d}{\beta \epsilon^2} & \text{if } 1 \lesssim \beta \lesssim 1/\epsilon & \text{"moderate temperature"} \\ \dfrac{d}{\epsilon} & \text{if } 1/\epsilon \lesssim \beta & \text{"low temperature"} \end{cases}$$

*up to logarithmic factors in d and $1/\epsilon$

Logistic loss

- Logistic loss (i.e., negative log-likelihood of $\mathrm{Ber}(\beta\langle x, \theta\rangle)$ on (x, y)):

$$\ell(\theta; x, y) = \ln\left(1 + e^{-\beta y\langle x,\theta\rangle}\right)$$

- Excess risk with logistic loss:

$$\mathbb{E}[\ell(\theta; X, Y) - \ell(\theta^\star; X, Y)] = \mathbb{E}\left[\mathrm{KL}\left(\mathrm{Ber}(\beta\langle X, \theta^\star\rangle) \| \mathrm{Ber}(\beta\langle X, \theta\rangle)\right)\right]$$

- Normal design $\longrightarrow$ very good estimates of expected KL divergence

Sample complexity lower bound

- To use Fano's inequality, suffices to prove good upper bound on

$$\mathbb{E}\left[\mathrm{KL}\left(\mathrm{Ber}(\beta\langle X, \theta^\star\rangle) \| \mathrm{Ber}(\beta\langle X, \theta\rangle)\right)\right]$$

  - High temp ($\beta \lesssim 1$): textbook exercise

$$n^*(d, \epsilon, \beta) \gtrsim \frac{d}{\beta^2 \epsilon^2}$$

  - Moderate temp ($1 \lesssim \beta \lesssim 1/\epsilon$): not well-known?

$$n^*(d, \epsilon, \beta) \gtrsim \frac{d}{\beta \epsilon^2}$$

  - Low temp ($1/\epsilon \lesssim \beta$): unclear how to get tight bound with Fano

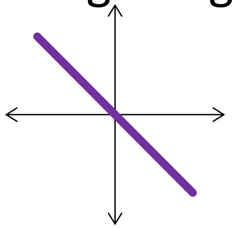
Instead, extend Long's 1995 lower bound for $\beta = \infty$ to all $\beta \gtrsim 1/\epsilon$

Sample complexity upper bound

- Three different estimators, depending on temperature
  - High temp ($\beta \lesssim 1$): minimize average **linear loss** (Servedio, 1999: "Average" algorithm)

Also: Plan & Vershynin (2012)

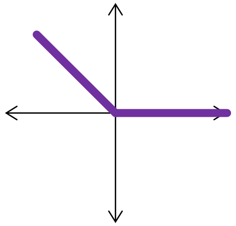
$$\hat{\theta} = \arg\min_{\theta \in S^{d-1}} \sum_{i=1}^{n} -y_i \langle x_i, \theta \rangle$$



  - Moderate or low temp ($1 \lesssim \beta$): minimize average **ReLU loss**

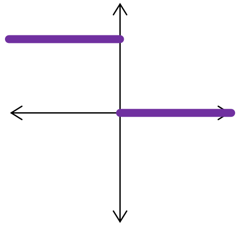
Inspired by Kuchelmeister & van de Geer (2023)

$$\hat{\theta} = \arg\min_{\theta \in S^{d-1}} \sum_{i=1}^{n} \max\{0, -y_i \langle x_i, \theta \rangle\}$$



  - Low temp ($1/\epsilon \lesssim \beta$): minimize average **0-1 loss**

$$\hat{\theta} = \arg\min_{\theta \in S^{d-1}} \sum_{i=1}^{n} 1\{y_i \langle x_i, \theta \rangle \leq 0\}$$



What we couldn't get to work

- Minimize average **logistic loss** (i.e., MLE)
  - Taylor-expand the estimation error (Portnoy, 1988; He & Shao, 2000; …)
  - Use self-concordance of logistic loss (Bach, 2010; Ostrovskii & Bach, 2021)
  - Our attempts gave suboptimal dependence on β

😓

Recap and open problems

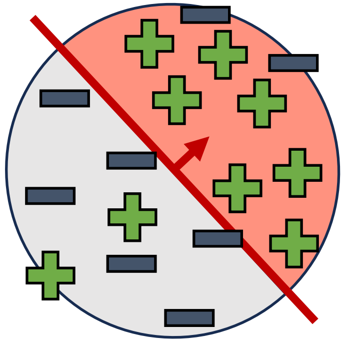
- Two "change points" in sample complexity for logistic regression

$$n^*(d, \epsilon, \beta) \asymp \begin{cases} \dfrac{d}{\beta^2 \epsilon^2} & \text{if } \beta \lesssim 1 & \text{"high temperature"} \\ \dfrac{d}{\beta \epsilon^2} & \text{if } 1 \lesssim \beta \lesssim 1/\epsilon & \text{"moderate temperature"} \\ \dfrac{d}{\epsilon} & \text{if } 1/\epsilon \lesssim \beta & \text{"low temperature"} \end{cases}$$

- Q: Efficient algorithms? MLE? Estimation of $||w^\star||$?

Thank you!

Learning noisy half-spaces



- <u>(Distribution-free) agnostic PAC learning half-spaces</u>
- **VC theory:** To ensure $\leq \epsilon$ **excess classification error rate**

$$\mathrm{err}(\hat{\theta}) - \mathrm{err}(\theta^\star) \leq \epsilon$$

  sample complexity is at most $d(1/\epsilon + \mathrm{err}(\theta^\star)/\epsilon^2)$ (up to logs)
- But we want guarantee about **parameter error** $\|\hat{\theta} - \theta^\star\|$
  - Can relate in low temp ($1/\epsilon \lesssim \beta$) regime, but unclear otherwise
  - **Useful fact:** $\mathrm{err}(\theta^\star) \asymp 1/\beta$ when $\beta \gtrsim 1$

Bregman divergence

- Bernoulli distribution $\mathrm{Ber}(\eta)$ has "mean parameter" $g'(\eta) = \frac{1}{1+e^{-\eta}}$;
  - $g(\eta) = \ln(1 + e^{\eta})$ is log partition function; g' is its derivative
- KL between Bernoulli distributions as Bergman divergence:

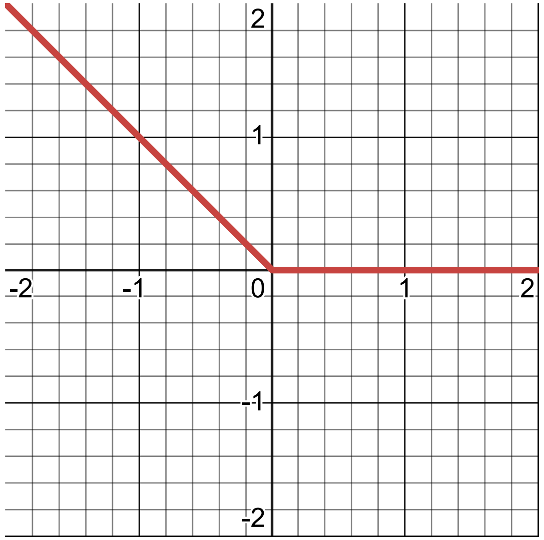
$$\mathrm{KL}\big(\mathrm{Ber}(\eta^\star)\|\mathrm{Ber}(\eta)\big) = g(\eta) - g(\eta^\star) - g'(\eta^\star)(\eta - \eta^\star)$$

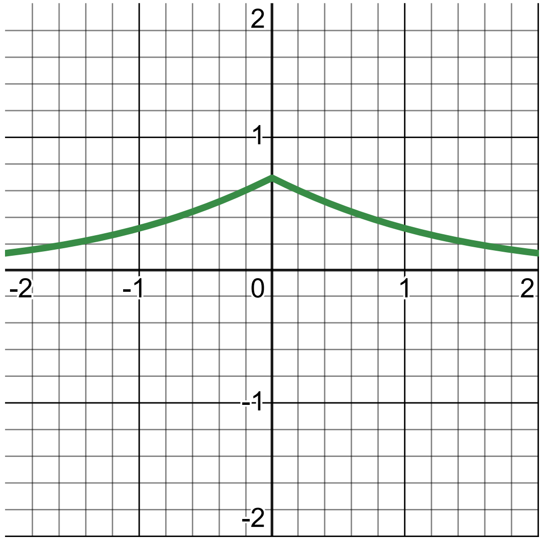
- When $\eta^\star = \beta\langle X, \theta^\star\rangle$ and $\eta = \beta\langle X, \theta\rangle$ and $X \sim \mathrm{N}(0, I_d)$:

$$\mathbb{E}\big[\mathrm{KL}\big(\mathrm{Ber}(\beta\langle X, \theta^\star\rangle)\|\mathrm{Ber}(\beta\langle X, \theta\rangle)\big)\big]$$
$$= \beta\ \mathbb{E}[g'(\beta\langle X, \theta^\star\rangle)\langle X, \theta - \theta^\star\rangle]$$

ReLU loss

- Nice observation of Kuchelmeister and van de Geer (2023):
$$\ln\left(1 + e^{-\beta y\langle x,\theta\rangle}\right) = \mathrm{ReLU}(-\beta y\langle x,\theta\rangle) + \ln\left(1 + e^{-\beta|\langle x,\theta\rangle|}\right)$$
- (Scaled) excess risk with ReLU loss = excess risk with logistic loss
  - Uses spherical symmetry of $\mathrm{N}(0, I_d)$
  - **Caveat:** optimization over the sphere





Adaptivity

- **If β is unknown:** suffices to coarsely distinguish "high temp" ($\beta \lesssim 1$) and "medium or low temp" ($1 \lesssim \beta$) regimes
  - Estimate classification error rate of $\theta^\star$
  - Can use training error rate of ERM (with zero-one loss) on dataset of size d/ϵ
  - Based on outcome, decide whether to use linear loss or ReLU loss on full data