Data subsampling for Poisson regression with pth-root-link

Han Cheng Lie (University of Potsdam), Alexander Munteanu (TU Dortmund)
December 10-15, 2024 | NeurIPS, Vancouver, BC, Canada

pth-root-link Poisson regression problem:
given $X \in \mathbb{R}^{n\times d}$ with *row* vectors $x_i = (1, x_i^{(1)}, \ldots, x_i^{(d-1)})$, $Y \in \mathbb{N}_0^n$, $p \in \{1, 2\}$,

$$\text{find} \quad \beta^* \in \operatorname{argmin}_{\beta \in D(0)} \sum_{i=1}^{n} (x_i\beta)^p - py \log(x_i\beta) + \log(y!),$$

where $D(\eta) := \{\beta \mid \forall i \in [n]\colon x_i\beta > \eta\}$.

Link functions: canonical log-link intractable in our setting [Molina et al., 2018], so consider popular alternatives [Cochran, 1940]:

- ID-link ($p = 1$)
- square-root-link ($p = 2$)

Our Goal: reduce instance size n by subsampling. Preserve a $(1+\varepsilon)$-approximation. Hereby save computational resources such as

- data storage
- runtime
- energy
- etc.

Data Subsampling

Sensitivity sampling framework: [Langberg, Schulman, 2010]

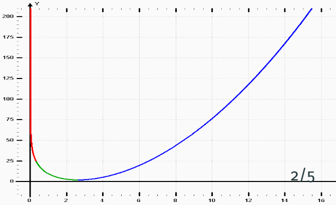
- sample proportional to sensitivity scores (relative contribution of single data points)
- main complexity parameters: VC dimension Δ, total sensitivity $\mathfrak{S}$
- sample size $m \in \tilde{O}(\Delta\mathfrak{S}/\varepsilon^2)$ yields $(1 \pm \varepsilon)$-approximation

VC dimension bounds:

- $O(d^2)$ (complexity of evaluating the loss [Anthony, Bartlett, 2002])
- $O(d \log(n) \log(y_{\max})/\varepsilon) \subseteq \tilde{O}(d/\varepsilon)$ (grouping and rounding technique [Munteanu et al., 2018, 2022])

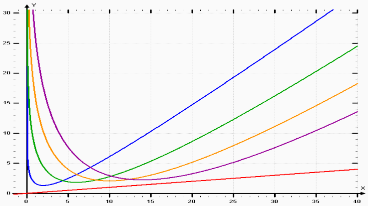
Bounding the sensitivity: pth-root-link requires to handle three intervals:

1. large $x_i\beta \geq y_i^{1/p}$ (relate to the ℓ_p-norm $(x_i\beta)^p$)
2. medium $\eta < x_i\beta < y_i^{1/p}$ (uniform sampling ✓)
3. small $0 < x_i\beta \leq \eta$ (domain shift)



Handling large $x\beta \geq y^{1/p}$

Bounds on the (individual) loss $g_y(x\beta)$:



- $(x\beta)^p \geq g_y(x\beta) \geq \frac{(x\beta - y^{1/p})^p}{\lambda}$
- $\lambda = 1$ for $p = 2$ ✓
- but $\lambda \in \Theta\left(\sqrt{\frac{y}{\log(y)}}\right)$ required for $p = 1$

Novel complexity parameter ρ:

- ρ-complexity quantifies balance between upper and lower bound:

$$\sup_{\beta \in \mathbb{R}^d} \frac{\sum_{j=1}^n |x_j\beta|^p}{\sum_{j=1}^n \left|x_j\beta - y_j^{1/p}\right|^p} \leq \rho$$

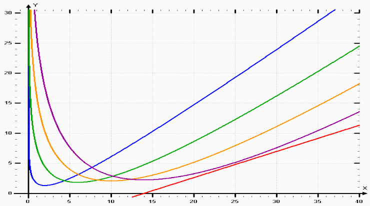
- natural interpretation w.r.t. the Poisson model and optimization

Bounding the total sensitivity for all $x_i\beta > \eta$:

$$\mathfrak{S} \in \begin{cases} O\left(\rho d\sqrt{y_{\max}/\log(y_{\max})} + \log\log(1/\eta)\right), & \text{for } p = 1 \\ O\left(\rho d + \log\left(y_{\max}\right) + \log\log(1/\eta)\right), & \text{for } p = 2. \end{cases}$$

Bounds on the (individual) loss $g_y(x\beta)$:

- $(x\beta)^p \geq g_y(x\beta) \geq \frac{(x\beta - y^{1/p})^p}{\lambda}$
- $\lambda = 1$ for $p = 2$ ✓
- but $\lambda \in \Theta\left(\sqrt{\frac{y}{\log(y)}}\right)$ required for $p = 1$



Novel complexity parameter ρ:

- ρ-complexity quantifies balance between upper and lower bound:

$$\sup_{\beta \in \mathbb{R}^d} \frac{\sum_{j=1}^{n} |x_j\beta|^p}{\sum_{j=1}^{n} |x_j\beta - y_j^{1/p}|^p} \leq \rho$$

- natural interpretation w.r.t. the Poisson model and optimization

Bounding the total sensitivity for all $x_i\beta > \eta$:

$$\mathfrak{S} \in \begin{cases} O\left(\rho d \sqrt{y_{\max}/\log(y_{\max})} + \log\log(1/\eta)\right), & \text{for } p = 1 \\ O\left(\rho d + \log\left(y_{\max}\right) + \log\log(1/\eta)\right), & \text{for } p = 2. \end{cases}$$

Handling small $x_i\beta \leq \eta$

Domain shift:

- Problem: cannot bound the sensitivity for contributions close to zero due to asymptote
- *domain shift* avoids this issue by optimizing over $\beta \in D(\eta) \subseteq D(0)$
- all $\beta \in D(\eta)$ satisfy $\forall i \in [n]: x_i\beta > \eta$

Optimization over $D(\eta)$:

- there exists a $(1+\varepsilon)$-approximate solution in $D(\varepsilon)$
- sensitivity sampling preserves the loss up to another $(1+\varepsilon)$ factor

$\Rightarrow$ we can find $\tilde{\beta} \in D(\varepsilon)$ evaluated on the subsample that satisfies

$$f(X\tilde{\beta}) \leq (1+\varepsilon)\, f(X\beta^*),\ \text{where } \beta^* \in \text{argmin}_{\beta\in D(0)} f(X\beta).$$

Optimization requires the extreme points $\mathcal{E}$ on the convex hull:

- Worst case $|\mathcal{E}| = n$
- Smoothed complexity: $\mathbb{E}\left[|\mathcal{E}|\right] \in O\left(\frac{\log^{1.5d-1}(n)}{\sigma^d} + \log^{d-1}(n)\right)$ [Damerow, 2006]
- ε-kernel approximation: $O(\frac{1}{\varepsilon}^{(d-1)/2})$ [Chan, 2004, Blum, Har-Peled, Raichel, 2019]

Limitations

General lower bounds:

- $\Omega(n)$ against (weighted) subsets of data
- Information theoretic $\Omega(n/\log(n))$ against any data reduction

Dependence on parameters:

- For $p = 1$: $\lambda \in \Theta\left(\sqrt{y_{\max}/\log(y_{\max})}\right)$ via novel bounds on the Lambert W_0 function improving over [Roig-Solvas, Sznaier, 2022]
- linear dependence on ρ and λ but d^2 from VC dimension $\times$ sensitivity
- $\tilde{\Theta}(d)$ likely to suffice [Munteanu, Omlor, 2024]

Domain shift and the choice of p:

- Domain shift fails to preserve $(1+\varepsilon)$-approximation for $p \geq 3$
- Indicates that other techniques needed, if even possible