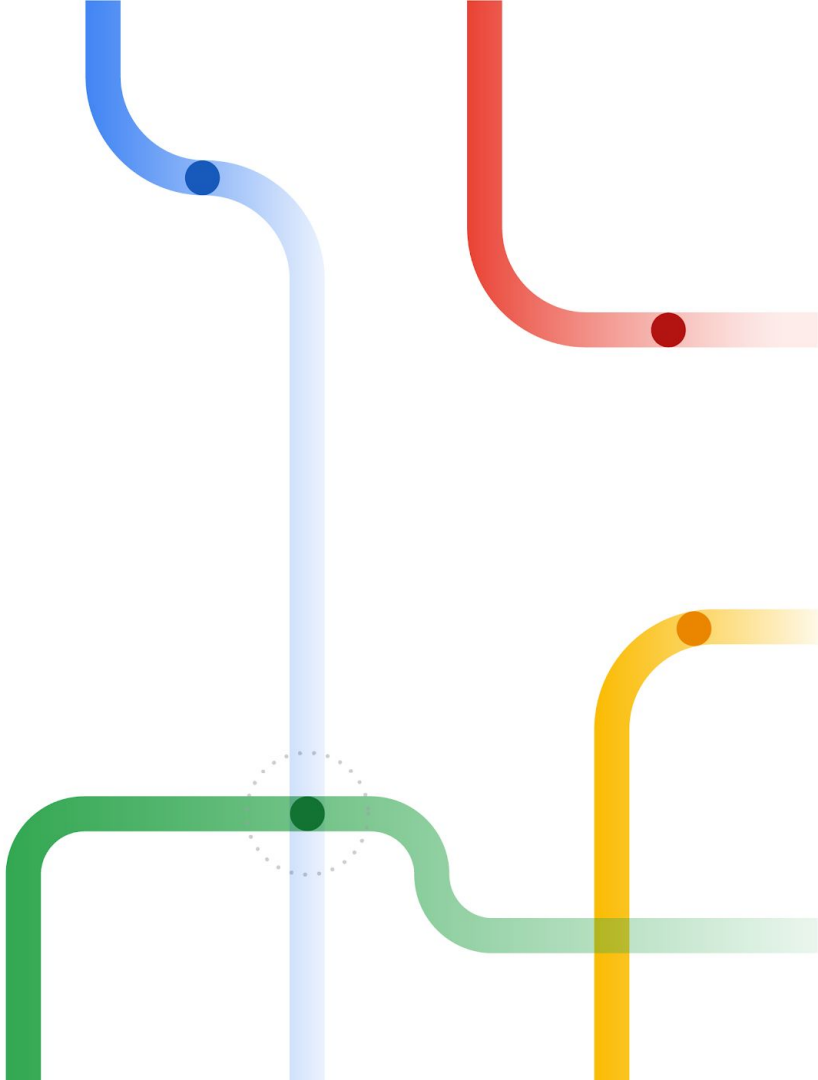
On the benefits of maximum likelihood estimation in regression and forecasting

Rajat Sen

Joint work with Pranjal Awasthi, Abhimanyu Das and Ananda Theertha Suresh

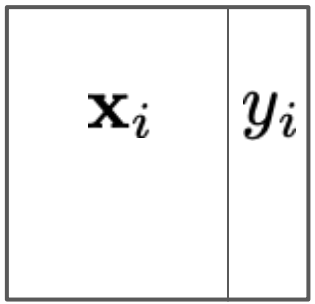
Google Research

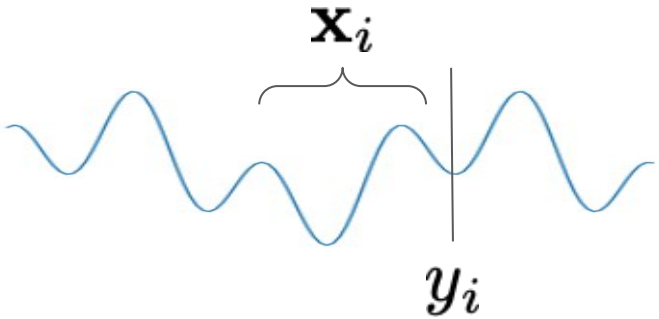


Regression & Forecasting in Practice

Regression & Forecasting in Practice

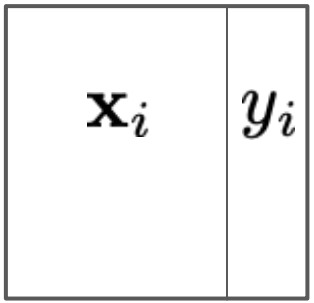
- A training dataset $(\mathbf{X}, y^n)$

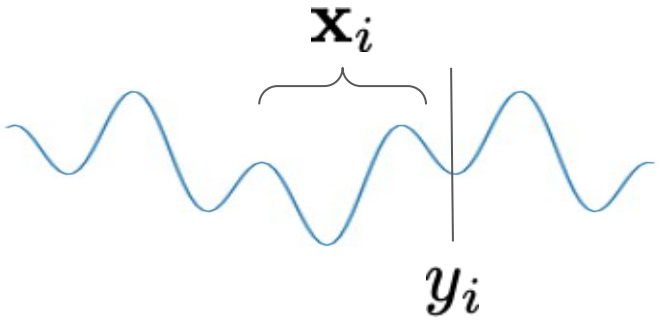




Regression & Forecasting in Practice

- A training dataset $(\mathbf{X}, y^n)$



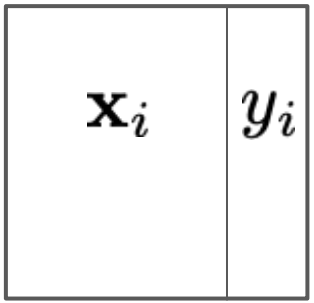


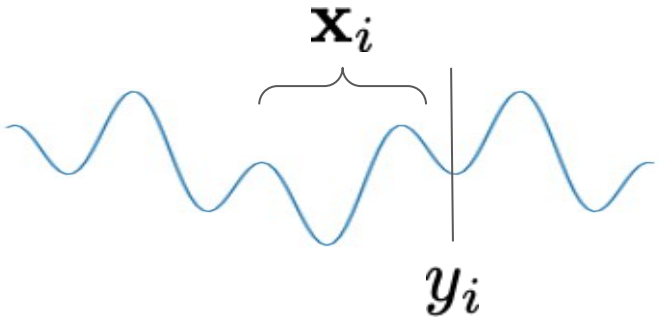
Objective

$$\text{argmin}_{h \in \mathcal{H}} \mathbb{E}_{(\mathbf{x},y) \in \mathcal{D}} \left[\ell(y, h(\mathbf{x})) \right]$$

Regression & Forecasting in Practice

- A training dataset $(\mathbf{X}, y^n)$





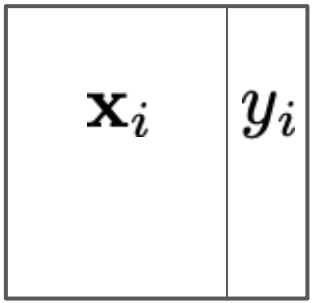
Test Set

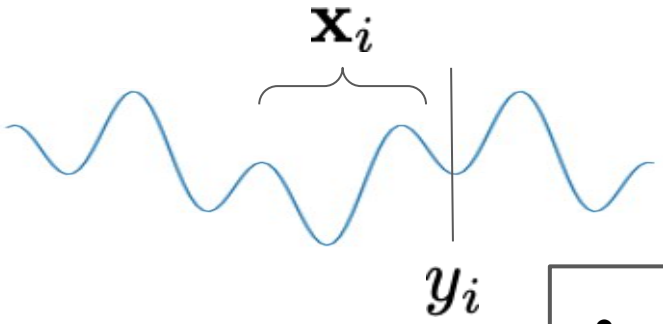
Objective

$$\text{argmin}_{h \in \mathcal{H}} \mathbb{E}_{(\mathbf{x},y) \in \mathcal{D}} \left[\ell(y, h(\mathbf{x})) \right]$$

Regression & Forecasting in Practice

- A training dataset $(\mathbf{X}, y^n)$





Test Set

Objective

$$\text{argmin}_{h \in \mathcal{H}} \mathbb{E}_{(\mathbf{x},y) \in \mathcal{D}} \left[\ell(y, h(\mathbf{x})) \right]$$

- MSE
- MAE
- MAPE
- RE
- MSLE...

Approach 1: Target Metric Optimization

- Use target metric as loss and perform Empirical Risk Minimization

$$\hat{h} = \text{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(\mathbf{x}_i))$$

Approach 1: Target Metric Optimization

- Use target metric as loss and perform Empirical Risk Minimization

$$\hat{h} = \text{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(\mathbf{x}_i))$$

- Need to train a different model for every target metric
- (PRO) Well understood for many function classes and works reasonably well in practice for many datasets
- (CON) It is provably sub-optimal in some settings like heavy tailed response or covariate [Lugosi et al. 2019]

Lugosi G, Mendelson S. Mean estimation and regression under heavy-tailed distributions: A survey. Foundations of Computational Mathematics. 2019 Oct;19(5):1145-90.

Approach 2: MLE & Post-hoc inference

- Assume that the data is generated from a distribution from a known parametric family of distributions
- Minimize empirical negative log-likelihood over the family of distributions agnostic to the target metric

$$\hat{\boldsymbol{\theta}}_{mle} = \text{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} -\log\left(p(y_i|\mathbf{x}_i;\boldsymbol{\theta})\right)$$

Approach 2: MLE & Post-hoc inference

- During inference predict the statistic from the MLE distribution that minimizes the target metric

$$\hat{h}(\boldsymbol{x}) = \text{argmin}_{\hat{y}} \mathbb{E}_{y \sim p(y|\boldsymbol{x}; \hat{\boldsymbol{\theta}}_{mle})} \left[\ell(y, \hat{y}) \right]$$

Approach 2: MLE & Post-hoc inference

- During inference predict the statistic from the MLE distribution that minimizes the target metric

$$\hat{h}(\boldsymbol{x}) = \text{argmin}_{\hat{y}} \mathbb{E}_{y \sim p(y|\boldsymbol{x}; \hat{\boldsymbol{\theta}}_{mle})} \left[\ell(y, \hat{y}) \right]$$

- For target metric MSE it is the mean from the MLE
- For MAE it is the median
- Can be computed efficiently for a family of relative errors [Gneiting 11]

$$\left| 1 - \left(\frac{y}{\hat{y}} \right)^{\beta} \right| \rightarrow \text{Median of} \sim y^{\beta} p(y|\mathbf{x}; \hat{\boldsymbol{\theta}}_{mle})$$

Tilmann Gneiting. Making and evaluating point forecasts. Journal of the American Statistical Association, 106(494):746–762, 2011.

Approach 2: MLE & Post-hoc inference

- During inference predict the statistic from the MLE distribution that minimizes the target metric

$$\hat{h}(\boldsymbol{x}) = \text{argmin}_{\hat{y}} \mathbb{E}_{y \sim p(y|\boldsymbol{x}; \hat{\boldsymbol{\theta}}_{mle})} [\ell(y, \hat{y})]$$

- (PRO) Only one model can serve many target metrics by adapting the inference
- (PRO) Probabilistic forecasting for free
- (PRO) Can capture inductive biases or domain knowledge by choosing appropriate family like zero inflated negative-binomial for sparse count data
- (CON) Choice of MLE might have a large impact on performance

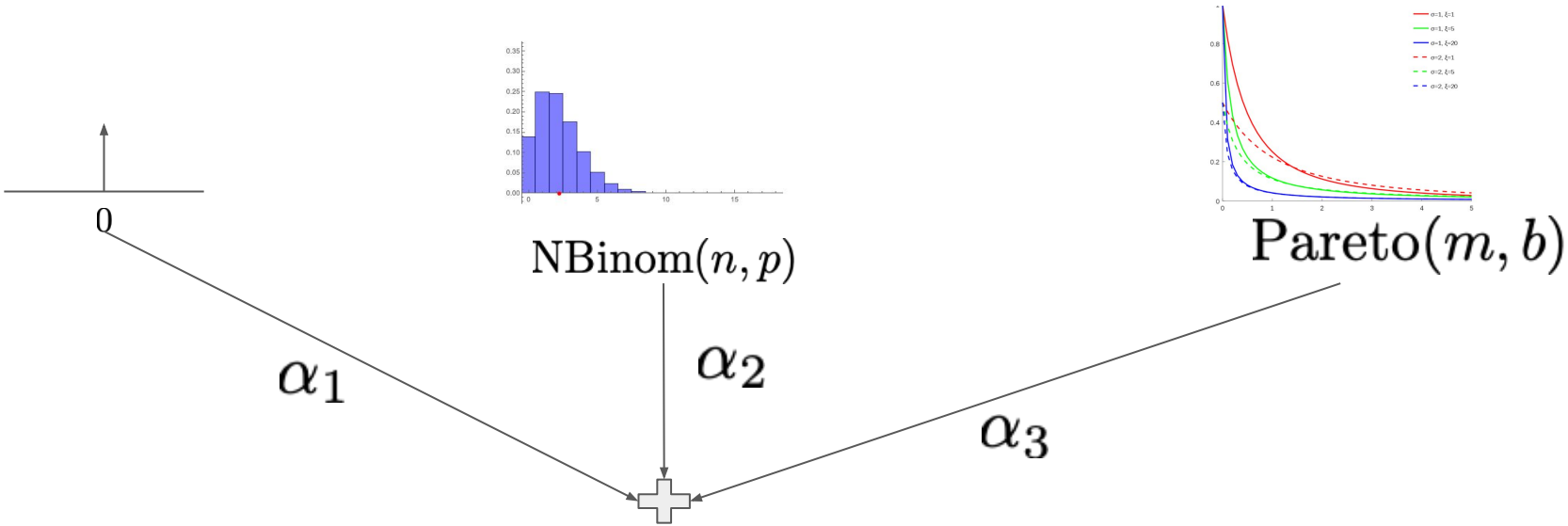
Approach 2: MLE & Post-hoc inference

- During inference predict the statistic from the MLE distribution that minimizes the target metric

$$\hat{h}(\boldsymbol{x}) = \text{argmin}_{\hat{y}} \mathbb{E}_{y \sim p(y|\boldsymbol{x}; \hat{\boldsymbol{\theta}}_{mle})} [\ell(y, \hat{y})]$$

- (PRO) Only one model can serve many target metrics by adapting the inference
- (PRO) Probabilistic forecasting for free
- (PRO) Can capture inductive biases or domain knowledge by choosing appropriate family like zero inflated negative-binomial for sparse count data
- (CON) Choice of MLE might have a large impact on performance

Choice of Likelihood



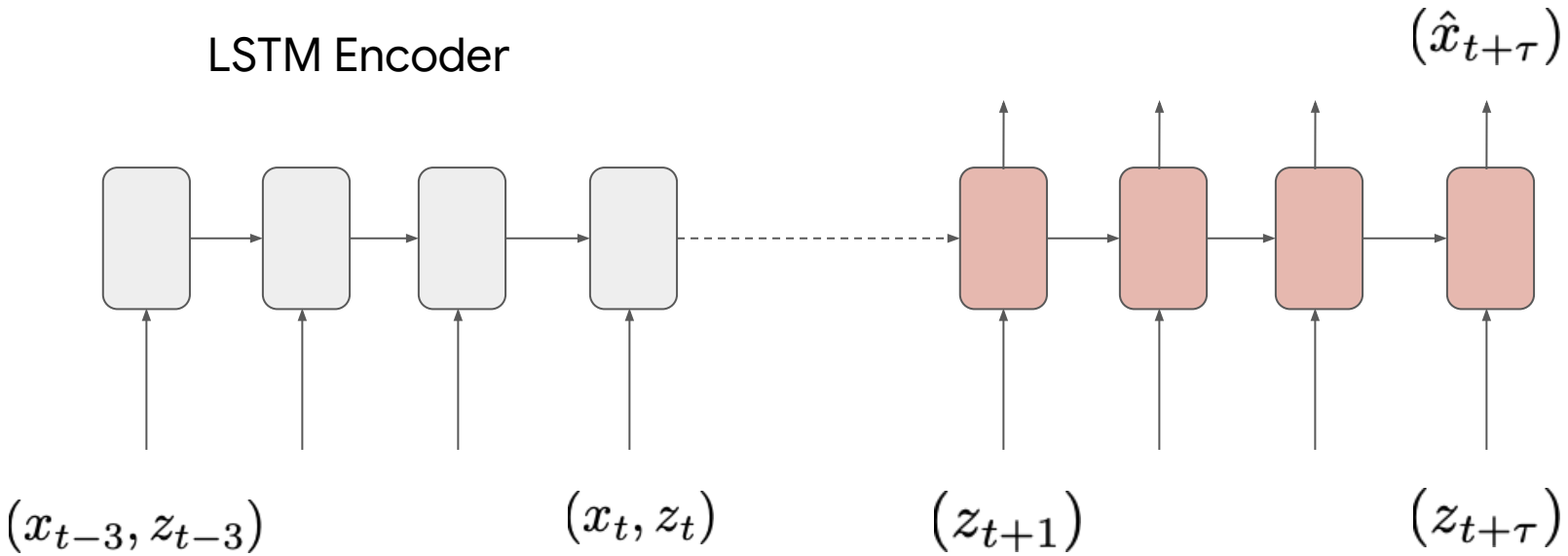
Some experimental results....

Architecture for Time-Series Models

TMO

LSTM Decoder

LSTM Encoder

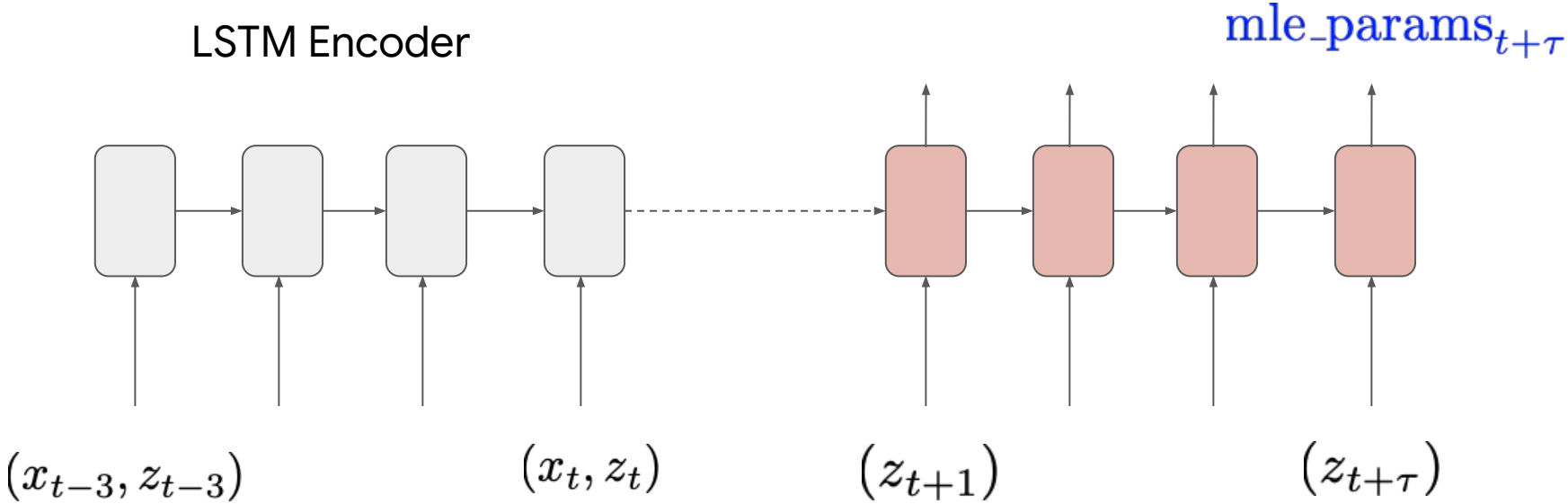


Architecture for Time-Series Models

MLE with post-hoc Inference

LSTM Decoder

LSTM Encoder



Empirical Results for different Target Metrics

Model	Favorita			M5		
	MAPE	WAPE	RMSE	MAPE	WAPE	RMSE
TMO(MSE)	0.6121±0.0075	0.2891±0.0023	175.3782±0.8235	0.5045±0.004	0.2839±0.0008	7.507±0.023
TMO(MAE)	0.3983±0.0012	0.2258±0.0006	**161.4919**±0.4748	0.4452±0.0005	**0.266**±0.0001	**7.0503**±0.0094
TMO(MAPE)	0.3199±0.0011	0.2528±0.0016	192.3823±1.3871	0.3892±0.0001	0.3143±0.0007	11.3799±0.1965
TMO(Huber)	0.432±0.0033	0.2366±0.0018	164.7006±0.7178	0.4722±0.0007	0.269±0.0002	7.093±0.0133
MLE(ZNBP)	**0.3139**±0.0011	**0.2238**±0.0009	164.6521±1.5185	**0.3864**±0.0001	0.2677±0.0002	7.2133±0.0152

Regression Datasets

Model	Bicycle Share			Gas Turbine		
	MAPE	WAPE	RMSE	MAPE	WAPE	RMSE
TMO(MSE)	0.2503±0.0008	0.1421±0.0003	878.5815±1.3059	0.8884±0.0118	0.3496±0.0041	1.5628±0.0071
TMO(MAE)	0.2594±0.0011	0.1436±0.0003	901.1357±1.4943	**0.774**±0.0054	**0.3389**±0.0019	1.5789±0.0067
TMO(MAPE)	0.2382±0.0012	0.1469±0.0012	899.9163±4.8219	0.8108±0.0009	0.8189±0.001	3.0573±0.0019
TMO(Huber)	0.2536±0.0011	0.1414±0.0004	889.1173±1.9654	0.902±0.0128	0.3598±0.0049	1.5992±0.0082
MLE(ZNBP)	**0.1969**±0.0018	**0.1235**±0.001	**767.4368**±7.1274	0.9877±0.0019	**0.3379**±0.0004	**1.4547**±0.0054

Model	p10QL	p90QL
TMO (Quantile)	0.0973±0.0002	0.0628±0.0019
MLE(ZNBP)	**0.0788**±0.0008	**0.0536**±0.0007

In Theory: Can MLE be better than TMO? Is MLE competitive with any estimator?

MLE is competitive with any estimator

- We prove a general result that shows the MLE estimator is competitive with any estimator for a target metric in finite samples under some assumptions
- The competitive ratio depends on the size of the family of distribution (or the size of its cover). This is different from the competitive result in Acharya et al 2016. Also does not follow from classical MVUE results.
- We show that our assumptions are valid for a class of convex likelihood families for least square regression

Acharya J, Das H, Orlitsky A, Suresh AT. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. InInternational Conference on Machine Learning 2017 Jul 17 (pp. 11-21). PMLR.

Application: Poisson regression with Identity Link

- We consider fixed design Poisson regression with the identity link function

$$p(y_i = k|\mathbf{x}_i;\boldsymbol{\theta}^*) = \frac{\mu_i^k e^{-\mu_i}}{k!} \text{ where } \mu_i = \langle \boldsymbol{\theta}^*, \mathbf{x}_i \rangle$$

- We will consider the square loss as our target metric for simplicity

Application: Poisson regression with Identity Link

- We consider fixed design Poisson regression with the identity link function

$$p(y_i = k|\mathbf{x}_i;\boldsymbol{\theta}^*) = \frac{\mu_i^k e^{-\mu_i}}{k!} \text{ where } \mu_i = \langle \boldsymbol{\theta}^*, \mathbf{x}_i \rangle$$

- We find a (complicated) estimator based on [Lugosi and Mendelson, 19] with guarantee

$$\mathcal{E}(\boldsymbol{\theta}_{\text{est}}) \leq c \cdot \|\boldsymbol{\theta}^*\|^2 \cdot \lambda_{\max}(\Sigma) \left(\frac{d + \log(\frac{1}{\delta})}{n} \right).$$

- Our result allows us to show that the simple MLE based estimator is competitive with the above upto log(n)/n factors

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. Foundations of Computational Mathematics, 19(5):1145–1190, 2019a.

Application: Poisson regression with Identity Link

- We consider fixed design Poisson regression with the identity link function

$$p(y_i = k|\mathbf{x}_i;\boldsymbol{\theta}^*) = \frac{\mu_i^k e^{-\mu_i}}{k!} \text{ where } \mu_i = \langle \boldsymbol{\theta}^*, \mathbf{x}_i \rangle$$

- We find a (complicated) estimator based on [Lugosi and Mendelson, 19] with guarantee

$$\mathcal{E}(\boldsymbol{\theta}_{\text{est}}) \leq c \cdot \|\boldsymbol{\theta}^*\|^2 \cdot \lambda_{\max}(\Sigma) \left(\frac{d + \log(\frac{1}{\delta})}{n} \right)$$

- Our result allows us to show that the simple MLE bas… …petitive with the above upto log(n)/n factors

Better than TMO guarantee!

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-taile… …rvey. Foundations of Computational Mathematics, 19(5):1145–1190, 2019a.

Application: Pareto Regression

- We consider the Pareto regression setting where the scale parameter is relevant for the regression
- By virtue of our results we show that MLE based estimator is competitive with a complicated heavy tailed regression estimator [Hsu and Sabato, 16]
- This shows that we are better than TMO which will fail to have sub-gaussian type guarantees for heavy tailed data

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. Journal of Machine Learning Research, 17(18):1–40, 2016.

Thank you! Questions?