A Computationally Efficient Method for Learning Exponential Family Distributions

Abhin Shah, Devavrat Shah, Gregory W. Wornell

Massachusetts Institute of Technology

Motivation

Exponential Family Distributions

x_1	x_2	x_3	x_4	$\cdots$	x_{n-1}	x_n
81	7	?	44		3	6
11	?	33	91		66	63
55	77	55	97		?	97
35	1	40	17		68	?

Joint distribution → Conditional distribution

Exponential Family Distributions

$$f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}^T\boldsymbol{\phi}(\mathbf{x}) + \beta(\mathbf{x})\right)$$

$$f_{\mathbf{x}}(\mathbf{x};\Theta) \propto \exp\left(\left\langle \Theta, \Phi(\mathbf{x}) \right\rangle\right)$$

We focus on — *Minimal and truncated* — exponential family

- No known method that is both computationally and statistically efficient for learning Θ

Maximum Likelihood Estimator

MLE

$$\min -\frac{1}{n}\sum_{t=1}^{n}\left\langle \Theta, \Phi(\mathbf{x}^{(t)})\right\rangle + \log \int_{\mathbf{x}\in\mathcal{X}} \exp\left(\left\langle \Theta, \Phi(\mathbf{x})\right\rangle\right) d\mathbf{x}.$$

1. Consistency
2. Asymptotic normality
3. Asymptotic efficiency
4. Computationally hard

We provide a computationally efficient proxy for the MLE.

Loss function

1. A novel, computationally tractable loss function —

$$\mathcal{L}_n(\Theta) = \frac{1}{n}\sum_{t=1}^{n} \exp\Big(- \Big\langle \Theta, \Phi(\mathbf{x}^{(t)}) \Big\rangle \Big)$$

- Minimizing the population version of $\mathcal{L}_n(\Theta)$ is equivalent to a MLE of $f_{\mathbf{x}}(\mathbf{x}; \Theta^* - \Theta)$.

Main Results

Consistency and Normality

$$\hat{\Theta}_n \in \arg\min_{\Theta \in \Lambda} \mathcal{L}_n(\Theta).$$

- $\hat{\Theta}_n$ is asymptotically consistent!
- Under some mild conditions, $\hat{\Theta}_n$ is asymptotically normal!

$\downarrow$

- Even though MLE is intractable, this M-estimation is tractable.
- However, unlike MLE, this is not asymptotically efficient.

Main Results

Finite-sample guarantees

- Structure recovery with $\Omega(\text{poly}(k_1 k_2))$ samples.

- Our framework can capture various constraints on the natural parameters including sparse, low-rank, sparse-plus-low-rank

  1. Θ is a sparse matrix
  2. Θ is a low-rank matrix
  3. Θ is a sparse-plus-low-rank matrix

Open question

Can computational and asymptotic efficiency be achieved by a single estimator for this class of exponential family?

Please visit our poster - Thank you!