A Computationally Efficient Method for Learning Exponential Family Distributions

Abhin Shah, Devavrat Shah, Gregory Wornell

Exponential Family

• An exponential family is a set of parametric probability distributions with probability densities of the following canonical form:

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

where $\mathbf{x} \in \mathcal{X}$ is a realization of the random vector $\mathbf{x}$, $\boldsymbol{\theta} \in \mathbb{R}^k$ is the natural parameter, $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^k$ is the natural statistic, k denotes the number of parameters, and β is the log base function.

• Motivated by the kind of constraints on the natural parameters we focus on, an equivalent representation of $f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})$ is:

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

where $\Theta = [\Theta_{ijl}] \in \mathbb{R}^{k_1\times k_2\times k_3}$ is the natural parameter, $\Phi = [\Phi_{ijl}] : \mathcal{X} \to \mathbb{R}^{k_1\times k_2\times k_3}$ is the natural statistic, $k_1 \times k_2 \times k_3 - 1 = k$, and $\langle\langle \Theta, \Phi(\mathbf{x})\rangle\rangle$ denotes the tensor inner product, i.e., the sum of product of entries of Θ and $\Phi(\mathbf{x})$.

Minimal Exponential Family

• An exponential family is minimal if there does not exist a nonzero tensor $\mathbf{U} \in \mathbb{R}^{k_1\times k_2\times k_3}$ such that $\langle\langle \mathbf{U}, \Phi(\mathbf{x})\rangle\rangle$ is equal to a constant for all $\mathbf{x} \in \mathcal{X}$.

Truncated Exponential Family

• Truncated exponential family is a set of parametric probability distributions resulting from truncating the support of an exponential family. They share the same parametric form with their non-truncated counterparts up to a normalizing constant.

Learning Exponential Family

• If Φ and $\mathcal{X}$ are known, then learning an exponential family distribution is equivalent to learning Θ.
• There is no known method (without any abstract condition) that is both computationally and statistically efficient for learning Θ of a minimal truncated exponential family distribution.

Maximum Likelihood Estimator

• The MLE of the parametric family $f_{\mathbf{x}}(\cdot;\Theta)$ minimizes

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

• The MLE is
1. Consistent
2. Asymptotically normal
3. Asymptotically efficient
4. Computationally hard

Takeaway

We provide a computationally efficient proxy for the maximum likelihood estimator for learning exponential family distributions.

Algorithm

Loss Function

• Given n samples $\mathbf{x}^{(1)} \cdots, \mathbf{x}^{(n)}$ of $\mathbf{x}$, we propose the following computationally tractable loss function

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

where $\mathit{\Phi}(\cdot) \coloneqq \Phi(\cdot) - \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\Phi(\mathbf{x})]$ with $\mathcal{U}_{\mathcal{X}}$ being the uniform distribution over $\mathcal{X}$.

• The loss function $\mathcal{L}_n(\Theta)$ is an empirical average of the inverse of the function of $\mathbf{x}$ that the probability density $f_{\mathbf{x}}(\mathbf{x};\Theta)$ is proportional to.

Estimator

• The estimator $\hat{\Theta}_n$ is obtained by minimizing $\mathcal{L}_n(\Theta)$ over all Θ in the constraint set Λ, i.e.,

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

• We implement a projected gradient descent with $O(\mathrm{poly}(k_1k_2/\epsilon))$ iterations to solve the above convex minimization problem.

Main Results

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

• $\arg\min \mathcal{L}(\Theta) = \arg\min D(\mathcal{U}_{\mathcal{X}} \,\|\, f_{\mathbf{x}}(\cdot;\Theta^* - \Theta))$
where $\mathcal{L}(\Theta) = \mathbb{E}\Big[\exp\left(-\langle\langle \Theta, \mathit{\Phi}(\mathbf{x})\rangle\rangle\right)\Big]$ is the population version of $\mathcal{L}_n(\Theta)$ and $D(\cdot \,\|\, \cdot)$ is the Kullback-Leibler (KL) divergence.

• $\mathcal{L}(\Theta)$ is minimized if and only if $\Theta = \Theta^*$.

2 - $\hat{\Theta}_n$ is asymptotically consistent and normal

• The traditional MLE is intractable.
• Our M-estimation is tractable (but not asymptotically efficient).

3 - Parameter recovery with an ℓ_2 error of α with:
- $O(\mathrm{poly}(k_1k_2/\alpha))$ samples and
- $O(\mathrm{poly}(k_1k_2/\alpha))$ computations.

• Our work does not require any stringent conditions common in the literature, e.g., incoherence, dependency, sparse eigenvalue or restricted strong convexity.
• Learning graphical models focuses on local assumptions on the parameters such as node-wise-sparsity while our work focuses on global structures on the parameters (e.g., a low-rank constraint).

Examples

Our framework can capture various constraints on the natural parameters including:
1. Decomposition of Θ as a sparse matrix
2. Decomposition of Θ as a low-rank matrix
3. Decomposition of Θ as a sparse matrix and a low-rank matrix

Open Question

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