# Note on Advanced Statistical Inference

Zepeng CHEN
The HK PolyU

*Date: February 12, 2023*

## 1 Common Families of Distributions

### 1.1 Exponential Family

**Definition 1.1 (Exponential Family)**

*Given a feature map $\phi : \mathcal{X} \to \mathbb{R}^m$ and an $m$-dimensional canonical parameter vector $\boldsymbol{\theta} \in \mathbb{R}^m$, an exponential family is defined as the set $\mathcal{P} = \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^m\}$ where the density function $p_{\boldsymbol{\theta}}$ satisfies the following for a log-partition function $A : \mathbb{R}^m \to \mathbb{R}$:*

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = exp\left(\boldsymbol{\theta}^\top \phi(\mathbf{x}) - A(\boldsymbol{\theta})\right).$$

**Lemma 1.1**

*The log-partition function $A : \mathbb{R}^m \to \mathbb{R}$ can be determined as:*

$$A(\boldsymbol{\theta}) = log\left(\sum_{\mathbf{x}\in\mathcal{X}} exp\left(\boldsymbol{\theta}^\top \phi(\mathbf{x})\right)\right).$$

**Proof** Because

$$\sum_{\mathbf{x}\in\mathcal{X}} p_{\boldsymbol{\theta}}(\mathbf{x}) = 1.$$

■

**Lemma 1.2**

*(i) The gradient of the log-partition function $A$ is the mean of random vector $\phi(\mathbf{x})$:*

$$\nabla A(\boldsymbol{\theta}) = \boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbb{E}_{X\sim p_{\boldsymbol{\theta}}}[\phi(\mathbf{x})].$$

*(ii) The Hessian of the log-partition function $A$ is the covariance matrix of random vector $\phi(\mathbf{x})$:*

$$H_A(\boldsymbol{\theta}) = \mathrm{Cov}_{X\sim p_{\boldsymbol{\theta}}}(\phi(\mathbf{x})).$$

**Proof**

(i) Because

$$\nabla A(\boldsymbol{\theta}) = \frac{\sum_{\mathbf{x}\in\mathcal{X}} e^{\boldsymbol{\theta}^\top\phi(\mathbf{x})}\phi(\mathbf{x})}{\sum_{\mathbf{x}\in\mathcal{X}} e^{\boldsymbol{\theta}^\top\phi(\mathbf{x})}} = \sum_{\mathbf{x}\in\mathcal{X}} \frac{e^{\boldsymbol{\theta}^\top\phi(\mathbf{x})}}{\sum_{\mathbf{x}'\in\mathcal{X}} e^{\boldsymbol{\theta}^\top\phi(\mathbf{x}')}}\phi(\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{X}} p_{\boldsymbol{\theta}}(\mathbf{x})\phi(\mathbf{x}).$$

(ii)

■

**Lemma 1.3**

*The log-partition function $A$ of an exponential family is a convex function.*

**Proof** From probability we know that a covaraince matrix is always positive semi-definte (PSD). Thus, the Hessian of $A$ is a PSD matrix, implying it is a convex function. ■

**Note on** *In other words, $\nabla A(\boldsymbol{\theta})$ is a monotone function of the canonical parameters $\boldsymbol{\theta}$, i.e.,*

$$\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d : \quad (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)^\top \left(\boldsymbol{\mu}_{\boldsymbol{\theta}_2} - \boldsymbol{\mu}_{\boldsymbol{\theta}_1}\right) \geq 0.$$

*Moreover, under the assumption of invertible map, we have*

$$\boldsymbol{\theta} = (\nabla A)^{-1}(\boldsymbol{\mu}).$$

### 1.2 Location-scale Family

## 2 Transformation

## 3 Point Estimation

### 3.1 Maximum Likelihood Method

**Definition 3.1 (Maximum Likelihood Estimator)**

*Given a parameterized family of distributions $\{p_{\boldsymbol{\theta}} : \theta \in \mathbb{R}^d\}$, the maximum liklihood estimator (MLE) of the model parameters from observed samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$ will be*

$$\begin{aligned}
\boldsymbol{\theta}^{MLE} &:= \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \prod_{i=1}^n p_{\boldsymbol{\theta}}(\mathbf{x}_i) \\
&= \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^n log p_{\boldsymbol{\theta}}(\mathbf{x}_i) \\
&= \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} (\frac{1}{n}\sum_{i=1}^n \phi(\mathbf{x}_i))^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \hat{\boldsymbol{\mu}}^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) && \text{(Let } \hat{\boldsymbol{\mu}} \text{ denote the empirical mean)}
\end{aligned}$$

**Lemma 3.1**

*The maximum likelihood problem for fitting canonical parameters of an exponential family is a convex optimization problem.*

**Proof** Obviously the objective function regarding $\boldsymbol{\theta}$ is concave. ■

**Corollary 3.1**

*Since the maximum likelihood problem is a convex optimization problem, by the FOC, we have*

$$\boldsymbol{\theta}^{MLE} = (\nabla A)^{-1}(\hat{\boldsymbol{\mu}}).$$

*In addition, the mean parameter $\boldsymbol{\mu}_{\boldsymbol{\theta}^{MLE}}$ under the maximum likelihood estimator match the empirical mean $\hat{\boldsymbol{\mu}}$:*

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{\theta}^{MLE}} =& \nabla A(\boldsymbol{\theta}^{MLE}) \\ =& \hat{\boldsymbol{\mu}}\end{aligned}$$

**Theorem 3.1 (Central Limit Theorem for Canonical parameter)**

*Consider a sequence of independent random vectors $(\mathbf{x}_i)_{i=1}^{\infty}$ distributed as $p_{\boldsymbol{\theta}}$. Then, for the Maximum Liklikelihood canonical parameter $\boldsymbol{\theta}_n^{MLE}$ from $n$ samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$, the following holds*

$$\sqrt{n}\left(\boldsymbol{\theta}_n^{MLE} - \boldsymbol{\theta}^*\right) \xrightarrow{dist} \mathcal{N}\left(\mathbf{0}, \mathrm{Cov}_{\theta^*}^{-1}(\phi(\mathbf{x}))\right).$$

## 3.2 Method of Moments

**Definition 3.2 (Method of Moments Estimator)**

*Given a parameterized family of distributions $\{p_{\boldsymbol{\theta}} : \theta \in \mathbb{R}^d\}$, the method of moments estimator $\hat{\boldsymbol{\theta}}$ of the model parameters from observed samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$ matches the empirical mean vector, i.e., $\hat{\boldsymbol{\theta}}$ satisfies*

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}}[\phi(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i).$$

## 3.3 Maximum Entropy Principle

**Definition 3.3**

*Given a probability vector $\boldsymbol{q} = [q_1, \cdots, q_k]$ for a discrete random variable $X$, the (Shannon) entropy of $X$ is defined as*

$$H_{\boldsymbol{q}}(X) = \sum_{i=1}^{k} q_i log\frac{1}{q_i}.$$

**Note on** *The entropy value is always non-negative. Moreover, the entropy is upper-bounded by $\log k$ (Jensen's Inequality). Particularly, the upper-bound is achieved by the discrete uniform distribution, i.e., $q_1 = \cdots = q_k = \frac{1}{k}$. This can be proved by solving the entropy maximization* *problem:*

$$\max_{\boldsymbol{q}\in\mathbb{R}^k} \quad \sum_{i=1}^{k} q_i log\frac{1}{q_i}$$
$$s.t. \quad \sum_{i=1}^{k} q_i = 1,$$
$$q_i \geq 0, i = 1, \cdots, k.$$

**Definition 3.4 (Maximum Entropy Principle)**

*Given a set of probability distributions*

$$M_\phi := \left\{ q \in \mathcal{P}_{\mathcal{X}} : \mathbb{E}_{\hat{\boldsymbol{\theta}}}[\phi(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i) \right\},$$

*conduct the inference and base the decision on the distribution maximizing the Entropy function:*

$$\underset{q\in M_\phi}{\text{argmax}} H_q(\mathbf{X}) := \sum_{\mathbf{x}\in\mathcal{X}} q(\mathbf{x}) \log\frac{1}{q(\mathbf{x})}.$$

**Note on** *Entropy measures the uncertainty of a distribution, thus, this principle chooses the most uncertain model based on the given set $M$.*

**Theorem 3.2**

*The distribution that maximizes the entropy is an exponential family model with feature function $\phi$.*

**Proof** Consider the maximum entropy problem

$$\max_{\boldsymbol{q}\in\mathbb{R}^{|\mathcal{X}|}} \quad \sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} log\frac{1}{q_{\mathbf{x}}} = -\sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} log q_{\mathbf{x}}$$
$$\text{s.t.} \quad \sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}}\phi(\mathbf{x}) = \hat{\boldsymbol{\mu}},$$
$$\sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} = 1,$$
$$q_{\mathbf{x}} \geq 0, \mathbf{x} \in \mathcal{X},$$

as a problem without inequality constraints, i.e.,

$$\max_{\boldsymbol{q}\in\mathbb{R}^{|\mathcal{X}|}} \quad -\sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} log q_{\mathbf{x}}$$
$$\text{s.t.} \quad \sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} \begin{bmatrix}\phi(\mathbf{x}) \\ 1\end{bmatrix} = \begin{bmatrix}\hat{\boldsymbol{\mu}} \\ 1\end{bmatrix}.$$

Next we consider its Lagrangian problem

$$\mathcal{L}(\mathbf{q},\gamma) = \sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} \left( -log q_{\mathbf{x}} - \phi(\mathbf{x})^\top \gamma_{1:k} - \gamma_{k+1} \right) + \hat{\boldsymbol{\mu}}^\top \gamma_{1:k} + \gamma_{k+1},$$

the stationary KKT condition

$$\nabla_{q_{\mathbf{x}}} \mathcal{L}(\mathbf{q},\gamma) = -log q^*_{\mathbf{x}} - \phi(\mathbf{x})^\top \gamma_{1:k} - \gamma_{k+1} + 1 = 0$$

leads to

$$q^*_{\mathbf{x}} = exp\left(-\phi(\mathbf{x})^\top \gamma_{1:k} - \gamma_{k+1} + 1\right) \geq 0.$$

Thus, $q^*_{\mathbf{x}}$ is also the optimal solution to the original problem. Moreover,

$$q^*_{\mathbf{x}} \propto exp\left(-\phi(\mathbf{x})^\top \gamma_{1:k}\right)$$

leads to

$$q^*_{\mathbf{x}} = \frac{exp\left(-\phi(\mathbf{x})^\top \gamma_{1:k}\right)}{exp\left(-\phi(\mathbf{x})^\top \gamma_{1:k} - \gamma_{k+1} + 1\right)}$$

due to the constraint that probability $q_{\mathbf{x}}$'s add up to $1$. ■

## 3.4 Connections

**Proposition 3.1 (Equivalence of Method of Moments and MLE)**

*Given a parameterized family of distributions $\{p_{\boldsymbol{\theta}} : \theta \in \mathbb{R}^d\}$ with feature function $\phi$, the method of moments estimator with $\phi$-based moments results in the same estimator as maximum likelihood estimator.*

**Proof** Note that $\boldsymbol{\mu}_{\boldsymbol{\theta}^{\text{MLE}}} = \hat{\boldsymbol{\mu}}$ by Corollary 3.1, and this coincides with the definition of the method of moments estimator. ■