

## Lecture 6: Density Parametrization II

Lecturer: Bo Dai

Scribes: Yilun Zhou, Aditya Chandaliya

**Note:** *LaTeX template courtesy of UC Berkeley EECS Department.*

**Disclaimer:** *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

## 6.1 Recap

- Convex Optimization is a *solver* for Machine Learning
- Density Parametrization is a *modeler* for Machine Learning
- In the previous class we proved convergence of gradient descent and stochastic gradient descent by showing  $f(x_t) - f(x_*) \sim O(\frac{1}{t})$
- The Exponential Family
  - $P(x) = h(x) \exp(\eta^T T(x) - A(\eta))$  given  $P(x) \geq 0, \int p(x) = 1$
  - $A(\eta) = \log \int h(x) \exp(\eta^T T(x)) d(x)$  given  $P(x) \geq 0, \int p(x) = 1$ 
    - \*  $\eta$ : natural parameter
    - \*  $T(x)$ : sufficient statistic of the data
    - \*  $h(x)$ : carrier function
    - \*  $A(\eta)$ : log-partition function (cumulant function)

## 6.2 Recap

- The Exponential Family:
  - *Canonical* form:
 
$$P(x) = h(x) \exp(\eta^T T(x) - A(\eta))$$
 given that  $P(x) \geq 0$  and  $\int p(x) dx = 1$ .
  - Log-partition function:
 
$$A(\eta) = \log \int h(x) \exp(\eta^T T(x)) dx$$
    - \*  $\eta$ : Natural parameter.
    - \*  $T(x)$ : Sufficient statistic of the data.
    - \*  $h(x)$ : Carrier function.
    - \*  $A(\eta)$ : Log-partition function (or cumulant function).

## 6.3 New Content

### 6.3.1 Motivation

In machine learning, we can categorize problems into two main types:

- **Supervised learning:**
  - Regression (e.g., modeling with Gaussian distributions).
  - Classification (e.g., modeling with Bernoulli distributions).
- **Unsupervised learning:**
  - Generative modeling (e.g., modeling the distribution of data).

### 6.3.2 Examples of Machine Learning Distributions

#### 6.3.2.1 Gaussian Distribution (for Regression)

In the case of regression, we model the conditional distribution of  $y$  given  $x$  as a Gaussian:

$$y|x, \omega \sim N(\omega^T x, \sigma^2)$$

The probability density function (pdf) is given by:

$$P(y|x, \omega) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \omega^T x)^2}{2\sigma^2}\right)$$

To find the Maximum Likelihood Estimate (MLE), we minimize the sum of squared errors:

$$\text{MLE} = \min_{\omega} \sum_{i=1}^n \|y_i - \omega^T x_i\|^2$$

#### 6.3.2.2 Bernoulli Distribution (for Classification)

In binary classification, we model the outcome  $y \in \{0, 1\}$  using the Bernoulli distribution:

$$y|x, \omega \sim \text{Bernoulli}(p)$$

where  $p$  is modeled by the sigmoid function:

$$p = \text{sigmoid}(\omega^T x) = \frac{1}{1 + \exp(-\omega^T x)}$$

The probability mass function (pmf) is:

$$P(y|p) = p^y (1 - p)^{1-y}$$

### 6.3.3 Exponential Family Examples

The following examples will illustrate how to verify if a function is of the exponential family. The main method to do this is rewriting the function in the canonical form described above.

#### 6.3.3.1 Showing Bernoulli is in Exponential Family

We can express the Bernoulli distribution in exponential family form. Starting with the Bernoulli pmf:

$$p(x) = \pi^x(1 - \pi)^{1-x} \quad \text{where } \pi \in [0, 1]$$

We can rewrite it as:

$$p(x) = \exp(x \log \pi + (1 - x) \log(1 - \pi))$$

Simplifying further:

$$p(x) = \exp\left(x \log\left(\frac{\pi}{1 - \pi}\right) + \log(1 - \pi)\right)$$

In this case:

$$T(x) = x, \quad \eta = \log\left(\frac{\pi}{1 - \pi}\right), \quad h(x) = 1$$

Thus, the natural parameter  $\eta$  is related to the probability  $\pi$  via:

$$\exp(\eta) = \frac{\pi}{1 - \pi}, \quad \pi = \frac{1}{1 + \exp(-\eta)}$$

#### 6.3.3.2 Showing Gaussian is in Exponential Family

Similarly, the Gaussian distribution can also be expressed in exponential family form. Starting with the Gaussian pdf:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

We rewrite this as:

$$\begin{aligned} p(x) &= \exp\left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)\right)\right) \\ &= \exp\left(-\frac{(x - \mu)^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2})\right) \\ &= \exp\left(-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}\right) \end{aligned}$$

In this case:

$$T(x) = [x^2, x], \quad \eta = \left[-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right], \quad h(x) = 1$$

And the log-partition function is:

$$A(\eta) = -\frac{\mu^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}$$

### 6.3.4 Properties of the Log-Partition Function $A(\eta)$

The log-partition function  $A(\eta)$  has several important properties:

1. **Convexity of  $A(\eta)$ :**

$$A(t\eta_1 + (1-t)\eta_2) \leq tA(\eta_1) + (1-t)A(\eta_2)$$

This convexity is a key property for optimization, ensuring that the function has no local minima.

2. **Gradient of  $A(\eta)$ :** The gradient of  $A(\eta)$  with respect to  $\eta$  is the expected value of the sufficient statistics  $T(x)$  under the distribution:

$$\frac{dA(\eta)}{d\eta} = \frac{1}{Q(\eta)} \frac{\partial Q(\eta)}{\partial \eta} \quad (6.1)$$

$$= \int \frac{h(x) \exp(\eta^T T(x))}{\int h(x) \exp(\eta^T T(x)) dx} T(x) dx \quad (6.2)$$

$$= \int p(x) T(x) dx \quad (6.3)$$

$$= \mathbb{E}_{P_\eta(x)}[T(x)] \quad (6.4)$$

### 6.3.5 Log-Likelihood and MLE in the Exponential Family

In the exponential family, the likelihood function  $P_\theta(x)$  is given by:

$$P_\theta(x) = \exp(f_\theta(x) - A(\theta))$$

where  $f_\theta(x)$  is the function of the data and  $\theta$  represents the model parameters.

The log-likelihood for a dataset  $x_1, x_2, \dots, x_n$  is:

$$L(\theta) = \sum_{i=1}^n \log P_\theta(x_i) = \sum_{i=1}^n [f_\theta(x_i) - A(\theta)]$$

The gradient of the log-likelihood with respect to  $\theta$  is:

$$\frac{dL(\theta)}{d\theta} = \sum_{i=1}^n \frac{\partial f_\theta(x_i)}{\partial \theta} - n \mathbb{E}_{P_\theta(x)} \left[ \frac{\partial f_\theta(x)}{\partial \theta} \right]$$

This shows the balance between the observed data and the expected value of the sufficient statistics under the model.

Additionally, the second derivative (Hessian) of  $A(\theta)$  provides information about the variance of the sufficient statistics:

$$\frac{\partial^2 A(\theta)}{\partial^2 \theta} = (\mathbb{E}_{P_\theta(x)}[T^2(x)]) - \mathbb{E}_{P_\theta(x)}[T(x)]^2 = \text{cov}(T(x))$$

This non-negative second derivative ensures the convexity of  $A(\theta)$  and supports the optimization of the log-likelihood.

Another note from this is that the  $k$ th derivative of the partition function is the  $k$ th moment of the partition function.