Probabilistic Graphical Models

Lecture X: Exponential Families

Matthew Walter

TTI-Chicago

May 18, 2020

Exponential Families

- We have considered different representations for complex distributions
- Representations of global structures (Bayesian networks, Markov networks)
- Representations of local structure (CPDs, potentials)
- Now, we will consider *families of distributions*
- Multinomial distribution over K outcomes vs. set of all multinomials over K outcomes
- Gaussian distribution vs. the set of all Gaussian distributions
- Graphical model vs. the set of all graphical models with the same structure and CPD parametrization

Exponential Families

An exponential family is a class of distributions that share the same functional form

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

where

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

is the *finite* partition function, and

- $\boldsymbol{\tau} : \boldsymbol{X} \rightarrow \mathbb{R}^k$ is a **sufficient statistics** function
- $\Theta \subseteq \mathbb{R}^M$ is a convex **parameter space**
- $\boldsymbol{t} : \mathbb{R}^M \rightarrow \mathbb{R}^K$ is a **natural parameter function** (typically, $t(\theta) = \theta$, i.e., the *canonical form*)

Exponential Families

An exponential family is a class of distributions that share the same functional form

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^{\top} \boldsymbol{\tau}(\boldsymbol{x})\right)$$

where

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^{\top} \boldsymbol{\tau}(\boldsymbol{x})\right)$$

is the *finite* partition function, and

- $\boldsymbol{\tau} : \boldsymbol{X} \rightarrow \mathbb{R}^k$ is a **sufficient statistics** function
- $\Theta \subseteq \mathbb{R}^M$ is a convex **parameter space**
- $\boldsymbol{t} : \mathbb{R}^M \rightarrow \mathbb{R}^K$ is a **natural parameter function** (typically, $t(\theta) = \theta$, i.e., the *canonical form*)

Gives rise to the **parametric family** $\mathcal{P} = \{P_\theta : \boldsymbol{\theta} \in \Theta\}$ (defined by *canonical parameter* θ)

Exponential Families: Sufficient Statistics

- Why do we refer to $\tau(\boldsymbol{X})$ as a *sufficient statistic*?
- Sufficiency characterizes what is essential in a dataset
- A *statistic* is any function on the sample space that isn't a function of the parameter
- We say that $\tau(\boldsymbol{X})$ is *sufficient* if there is no information in $\boldsymbol{X}$ about θ that isn't available in $\tau(\boldsymbol{X})$
- Consider θ to be a random variable
- In the Bayesian sense, $\tau(\boldsymbol{X})$ is sufficient if

$$\theta \perp \boldsymbol{X} \mid \tau(\boldsymbol{X})$$

Example: Bernoulli Distribution

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$
$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

- $\boldsymbol{\tau}(X) = \begin{bmatrix} \mathbb{1}(X=1) & \mathbb{1}(X=0) \end{bmatrix}$ (not particularly interesting)
- $\boldsymbol{t}(\theta) = \begin{bmatrix} \ln\theta & \ln(1-\theta) \end{bmatrix}$
- $\theta \in [0, 1]$
- $Z(\theta) = 1$

Example: Bernoulli Distribution

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$
$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

- $\boldsymbol{\tau}(X) = \begin{bmatrix}\mathbb{1}(X=1) & \mathbb{1}(X=0)\end{bmatrix}$ (not particularly interesting)
- $\boldsymbol{t}(\theta) = \begin{bmatrix}\ln\theta & \ln(1-\theta)\end{bmatrix}$
- $\theta \in [0,1]$
- $Z(\theta) = 1$

$$X = 0 : \exp\left(1 \cdot \ln\theta + 0 \cdot \ln(1-\theta)\right) = \theta$$
$$X = 1 : \exp\left(0 \cdot \ln\theta + 1 \cdot \ln(1-\theta)\right) = 1-\theta$$

Example: Gaussian Distribution

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

$$\mathcal{P} = \{P_\theta : \boldsymbol{\theta} \in \Theta\}$$

- $\boldsymbol{\tau}(x) = \begin{bmatrix} x & x^2 \end{bmatrix}$
- $\boldsymbol{\theta} = \begin{bmatrix} \mu & \sigma^2 \end{bmatrix} \in \mathbb{R} \times \mathbb{R}^+$
- $\boldsymbol{t}(\mu, \sigma^2) = \begin{bmatrix} \frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2} \end{bmatrix}$
- $Z(\mu, \sigma^2) = \sqrt{2\pi}\sigma \exp\left(\frac{\mu^2}{2\sigma^2}\right)$

Example: Gaussian Distribution

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

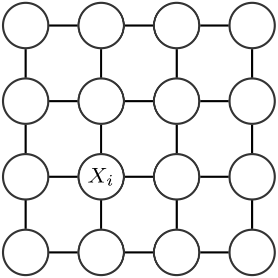
$$\mathcal{P} = \{P_\theta : \boldsymbol{\theta} \in \Theta\}$$

- $\boldsymbol{\tau}(x) = \begin{bmatrix} x & x^2 \end{bmatrix}$
- $\boldsymbol{\theta} = \begin{bmatrix} \mu & \sigma^2 \end{bmatrix} \in \mathbb{R} \times \mathbb{R}^+$
- $\boldsymbol{t}(\mu, \sigma^2) = \begin{bmatrix} \frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2} \end{bmatrix}$
- $Z(\mu, \sigma^2) = \sqrt{2\pi}\sigma \exp\left(\frac{\mu^2}{2\sigma^2}\right)$

$$p(x) = \frac{1}{Z(\mu, \sigma^2)} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Example: Ising Model

- $\boldsymbol{\tau}(x) = \{x_i \ \forall i \in \mathcal{V}\} \cup \{x_i x_j \ \forall (i,j) \in \mathcal{E}\}$
- $\boldsymbol{\theta} = \{x_i \ \forall i \in \mathcal{V}\} \cup \{x_i x_j \ \forall (i,j) \in \mathcal{E}\}$
- $\boldsymbol{t}(\theta) = \theta$ (canonical form)



$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_i \theta_i x_i + \sum_{(i,j)\in\mathcal{E}} \theta_{ij} x_i x_j \right)$$

More generally, many graphical models can be represented as exponential families (e.g., all graphical models over discrete random variables)

Why Exponential Families?

- Consider the **expected sufficient statistics**

$$\mu = \mathbb{E}_P[\tau(\boldsymbol{X})]$$

Natural Parameters

- A special case is when $\boldsymbol{t}$ is the identity function
- $\boldsymbol{\theta}$ are the **natural parameters** for the sufficient statistic $\boldsymbol{\tau}$
- We define the **natural parameter space** as the set of allowable natural parameters

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^K : \int \exp\bigl(\boldsymbol{\theta}^\top \boldsymbol{\tau}(\boldsymbol{x})\bigr) d\boldsymbol{x} < \infty \right\}$$

- Gives rise to the **linear exponential family**

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\bigl(\boldsymbol{\theta}^\top \boldsymbol{\tau}(\boldsymbol{x})\bigr)$$

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

Example: Gaussian (Revisited)

- Let $\boldsymbol{\theta} = \begin{bmatrix} \mu/\sigma^2 & -1/2\sigma^2 \end{bmatrix}$
- Z is only defined (finite) with $\theta_2 < 0$
- Thus, natural parameter space is $\mathbb{R} \times \mathbb{R}^-$

$$
\begin{aligned}
P_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \frac{1}{Z(\boldsymbol{\theta})} \exp\bigl(\boldsymbol{\theta}^\top \boldsymbol{\tau}(\boldsymbol{x})\bigr) \\
Z(\boldsymbol{\theta}) &= \sum_{\boldsymbol{x}} \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\tau}(\boldsymbol{x})\right) \\
&= \int_{-\infty}^{\infty} \exp\left(\theta_1 x + \theta_2 x^2\right) dx
\end{aligned}
$$

Log-Linear Models

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right) \quad \text{where} \quad Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

- Recall, that log-linear models are expressed as

$$P(X_1, X_2, \ldots, X_n) \propto \exp\left(\sum_{i=1}^{k} \theta_i f_i(\boldsymbol{D}_i)\right)$$

where f_i is a feature function with scope $\boldsymbol{D}_i$

Log-Linear Models

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right) \quad \text{where} \quad Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

- Recall, that log-linear models are expressed as

$$P(X_1, X_2, \ldots, X_n) \propto \exp\left(\sum_{i=1}^{k} \theta_i f_i(\boldsymbol{D}_i)\right)$$

  where f_i is a feature function with scope $\boldsymbol{D}_i$
- This is a linear exponential family where the sufficient statistics are

$$\boldsymbol{\tau}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{d}_1) & f_2(\boldsymbol{d}_2) & \ldots & f_k(\boldsymbol{d}_k) \end{bmatrix}$$

Log-Linear Models

$$P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right) \quad \text{where} \quad Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

- Recall, that log-linear models are expressed as

$$P(X_1, X_2, \dots, X_n) \propto \exp\left(\sum_{i=1}^{k} \theta_i f_i(\boldsymbol{D}_i)\right)$$

  where f_i is a feature function with scope $\boldsymbol{D}_i$
- This is a linear exponential family where the sufficient statistics are

$$\boldsymbol{\tau}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{d}_1) & f_2(\boldsymbol{d}_2) & \dots & f_k(\boldsymbol{d}_k) \end{bmatrix}$$

- Thus, **discrete Markov networks are linear exponential families**

Product Distributions

- An **exponential factor family** Φ is defined by $\boldsymbol{\tau}, \boldsymbol{t}, \boldsymbol{A}, \Theta$ with a factor

$$\phi_{\boldsymbol{\theta}} = \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$$

- Can readily show that the product of exponential factors is an exponential family
- Similarly, the product of linear exponential factors is a linear exponential family

Bayesian Networks as Exponential Families

- Bayesian network with exponential CPDs define an exponential family
- Consider a tabular CPD $P(X \mid \boldsymbol{U})$:

Bayesian Networks as Exponential Families

- Bayesian network with exponential CPDs define an exponential family
- Consider a tabular CPD $P(X \mid \boldsymbol{U})$:
  - Let $\boldsymbol{\tau}(X) = \left[\mathbb{1}(X = x, \boldsymbol{U} = \boldsymbol{u}) : x \in Val(X), \boldsymbol{u} \in Val(\boldsymbol{U})\right]$

Bayesian Networks as Exponential Families

- Bayesian network with exponential CPDs define an exponential family
- Consider a tabular CPD $P(X \mid \boldsymbol{U})$:
  - Let $\boldsymbol{\tau}(X) = \left[\mathbb{1}(X = x, \boldsymbol{U} = \boldsymbol{u}) : x \in Val(X), \boldsymbol{u} \in Val(\boldsymbol{U})\right]$
  - $\boldsymbol{t}(\boldsymbol{\theta}) = \left[\ln P(x \mid \boldsymbol{u}) : x \in Val(X), \boldsymbol{u} \in Val(\boldsymbol{U})\right]$

Bayesian Networks as Exponential Families

- Bayesian network with exponential CPDs define an exponential family
- Consider a tabular CPD $P(X \mid \boldsymbol{U})$:
  - Let $\boldsymbol{\tau}(X) = \left[\mathbb{1}(X = x, \boldsymbol{U} = \boldsymbol{u}) : x \in Val(X), \boldsymbol{u} \in Val(\boldsymbol{U})\right]$
  - $\boldsymbol{t}(\boldsymbol{\theta}) = \left[\ln P(x \mid \boldsymbol{u}) : x \in Val(X), \boldsymbol{u} \in Val(\boldsymbol{U})\right]$
  - $P(x \mid \boldsymbol{u}) = \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$

Bayesian Networks as Exponential Families

- Bayesian network with exponential CPDs define an exponential family
- Consider a tabular CPD $P(X \mid \boldsymbol{U})$:
  - Let $\boldsymbol{\tau}(X) = \left[\mathbb{1}(X = x, \boldsymbol{U} = \boldsymbol{u}) : x \in Val(X), \boldsymbol{u} \in Val(\boldsymbol{U})\right]$
  - $\boldsymbol{t}(\boldsymbol{\theta}) = \left[\ln P(x \mid \boldsymbol{u}) : x \in Val(X), \boldsymbol{u} \in Val(\boldsymbol{U})\right]$
  - $P(x \mid \boldsymbol{u}) = \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$
  - Not a linear exponential family
- BN that is the product of CPDs that have exponential form defines an exponential family

Bayesian Networks as Exponential Families

- Bayesian network with exponential CPDs define an exponential family
- Consider a tabular CPD $P(X \mid \boldsymbol{U})$:
  - Let $\boldsymbol{\tau}(X) = \left[\mathbb{1}(X = x, \boldsymbol{U} = \boldsymbol{u}) : x \in Val(X), \boldsymbol{u} \in Val(\boldsymbol{U})\right]$
  - $\boldsymbol{t}(\boldsymbol{\theta}) = \left[\ln P(x \mid \boldsymbol{u}) : x \in Val(X), \boldsymbol{u} \in Val(\boldsymbol{U})\right]$
  - $P(x \mid \boldsymbol{u}) = \exp\left(\boldsymbol{t}(\boldsymbol{\theta})^\top \boldsymbol{\tau}(\boldsymbol{x})\right)$
  - Not a linear exponential family
- BN that is the product of CPDs that have exponential form defines an exponential family
- Importantly, the CPDs have to be normalized