Exponential Family and Generalized Linear Models

Piyush Rai

Probabilistic Machine Learning (CS772A)

Aug 22, 2017

Exponential Family (Pitman, Darmois, Koopman, Late 1930s)

- Defines a class of distributions. An Exponential Family distribution is of the form

$$p(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})]$$

Exponential Family (Pitman, Darmois, Koopman, Late 1930s)

- Defines a class of distributions. An Exponential Family distribution is of the form

$$\boxed{p(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] \quad = \quad h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]}$$

Exponential Family (Pitman, Darmois, Koopman, Late 1930s)

- Defines a class of distributions. An Exponential Family distribution is of the form

$$\boxed{p(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] \quad = \quad h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]}$$

- $\boldsymbol{x} \in \mathcal{X}^m$ is the random variable being modeled (where $\mathcal{X}$ denotes some space, e.g., $\mathbb{R}$ or $\{0, 1\}$)

Exponential Family (Pitman, Darmois, Koopman, Late 1930s)

- Defines a class of distributions. An Exponential Family distribution is of the form

$$\boxed{p(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] \quad = \quad h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]}$$

- $\boldsymbol{x} \in \mathcal{X}^m$ is the random variable being modeled (where $\mathcal{X}$ denotes some space, e.g., $\mathbb{R}$ or $\{0, 1\}$)
- $\theta \in \mathbb{R}^d$: Natural parameters or canonical parameters defining the distribution

Exponential Family (Pitman, Darmois, Koopman, Late 1930s)

- Defines a class of distributions. An Exponential Family distribution is of the form

$$\boxed{p(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] \quad = \quad h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]}$$

- $\boldsymbol{x} \in \mathcal{X}^m$ is the random variable being modeled (where $\mathcal{X}$ denotes some space, e.g., $\mathbb{R}$ or $\{0, 1\}$)
- $\theta \in \mathbb{R}^d$: Natural parameters or canonical parameters defining the distribution
- $\phi(\boldsymbol{x}) \in \mathbb{R}^d$: Sufficient statistics (another random variable)

Exponential Family (Pitman, Darmois, Koopman, Late 1930s)

- Defines a class of distributions. An Exponential Family distribution is of the form

$$\boxed{p(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] \quad = \quad h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]}$$

- $\boldsymbol{x} \in \mathcal{X}^m$ is the random variable being modeled (where $\mathcal{X}$ denotes some space, e.g., $\mathbb{R}$ or $\{0, 1\}$)
- $\theta \in \mathbb{R}^d$: Natural parameters or canonical parameters defining the distribution
- $\phi(\boldsymbol{x}) \in \mathbb{R}^d$: Sufficient statistics (another random variable)
  - Why "sufficient": $p(\boldsymbol{x}|\theta)$ as a function of θ depends on $\boldsymbol{x}$ only via $\phi(\boldsymbol{x})$

Exponential Family (Pitman, Darmois, Koopman, Late 1930s)

- Defines a class of distributions. An Exponential Family distribution is of the form

$$\boxed{p(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] \quad = \quad h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]}$$

- $\boldsymbol{x} \in \mathcal{X}^m$ is the random variable being modeled (where $\mathcal{X}$ denotes some space, e.g., $\mathbb{R}$ or $\{0,1\}$)
- $\theta \in \mathbb{R}^d$: Natural parameters or canonical parameters defining the distribution
- $\phi(\boldsymbol{x}) \in \mathbb{R}^d$: Sufficient statistics (another random variable)
  - Why "sufficient": $p(\boldsymbol{x}|\theta)$ as a function of θ depends on $\boldsymbol{x}$ only via $\phi(\boldsymbol{x})$
- $Z(\theta) = \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$: Partition function

Exponential Family (Pitman, Darmois, Koopman, Late 1930s)

- Defines a class of distributions. An Exponential Family distribution is of the form

$$\boxed{p(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] \quad = \quad h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]}$$

- $\boldsymbol{x} \in \mathcal{X}^m$ is the random variable being modeled (where $\mathcal{X}$ denotes some space, e.g., $\mathbb{R}$ or $\{0,1\}$)
- $\theta \in \mathbb{R}^d$: Natural parameters or canonical parameters defining the distribution
- $\phi(\boldsymbol{x}) \in \mathbb{R}^d$: Sufficient statistics (another random variable)
  - Why "sufficient": $p(\boldsymbol{x}|\theta)$ as a function of θ depends on $\boldsymbol{x}$ only via $\phi(\boldsymbol{x})$
- $Z(\theta) = \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$: Partition function
- $A(\theta) = \log Z(\theta)$: Log-partition function (also called the cumulant function)

Exponential Family (Pitman, Darmois, Koopman, Late 1930s)

- Defines a class of distributions. An Exponential Family distribution is of the form

$$\boxed{p(\boldsymbol{x}|\theta) = \frac{1}{Z(\theta)} h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] \quad = \quad h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]}$$

- $\boldsymbol{x} \in \mathcal{X}^m$ is the random variable being modeled (where $\mathcal{X}$ denotes some space, e.g., $\mathbb{R}$ or $\{0, 1\}$)
- $\theta \in \mathbb{R}^d$: Natural parameters or canonical parameters defining the distribution
- $\phi(\boldsymbol{x}) \in \mathbb{R}^d$: Sufficient statistics (another random variable)
  - Why "sufficient": $p(\boldsymbol{x}|\theta)$ as a function of θ depends on $\boldsymbol{x}$ only via $\phi(\boldsymbol{x})$
- $Z(\theta) = \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$: Partition function
- $A(\theta) = \log Z(\theta)$: Log-partition function (also called the cumulant function)
- $h(\boldsymbol{x})$: A constant (doesn't depend on θ)

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of the Binomial distribution

$$\text{Binomial}(x|N, p) = \binom{N}{x} p^x (1-p)^{N-x}$$

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of the Binomial distribution

$$\text{Binomial}(x|N,p) = \binom{N}{x} p^x (1-p)^{N-x}$$

where N: number of trials

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of the Binomial distribution

$$\text{Binomial}(x|N,p) = \binom{N}{x} p^x (1-p)^{N-x}$$

where N: number of trials, x (a scalar): number of successes

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of the Binomial distribution

$$\text{Binomial}(x|N, p) = \binom{N}{x} p^x (1-p)^{N-x}$$

where N: number of trials, x (a scalar): number of successes, p: probability is success in each trial

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of the Binomial distribution

$$\text{Binomial}(x|N,p) = \binom{N}{x} p^x (1-p)^{N-x}$$

where N: number of trials, x (a scalar): number of successes, p: probability is success in each trial

- Can re-express the above representation of the Binomial as

$$\binom{N}{x} \exp\left[x \log\left(\frac{p}{1-p}\right) + N \log(1-p)\right]$$

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of the Binomial distribution

$$\text{Binomial}(x|N,p) = \binom{N}{x} p^x (1-p)^{N-x}$$

where N: number of trials, x (a scalar): number of successes, p: probability is success in each trial

- Can re-express the above representation of the Binomial as

$$\binom{N}{x} \exp\left[x \log\left(\frac{p}{1-p}\right) + N\log(1-p)\right]$$

- $h(x) = \binom{N}{x}$

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of the Binomial distribution

$$\text{Binomial}(x|N,p) = \binom{N}{x} p^x (1-p)^{N-x}$$

where N: number of trials, x (a scalar): number of successes, p: probability is success in each trial

- Can re-express the above representation of the Binomial as

$$\binom{N}{x} \exp\left[x \log\left(\frac{p}{1-p}\right) + N\log(1-p)\right]$$

- $h(x) = \binom{N}{x}$
- $\theta = \log\left(\frac{p}{1-p}\right)$

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of the Binomial distribution

$$\text{Binomial}(x|N,p) = \binom{N}{x} p^x (1-p)^{N-x}$$

where N: number of trials, x (a scalar): number of successes, p: probability is success in each trial

- Can re-express the above representation of the Binomial as

$$\binom{N}{x} \exp\left[x \log\left(\frac{p}{1-p}\right) + N\log(1-p)\right]$$

- $h(x) = \binom{N}{x}$
- $\theta = \log\left(\frac{p}{1-p}\right)$, and $p = \frac{1}{1+\exp(-\theta)}$

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of the Binomial distribution

$$\text{Binomial}(x|N,p) = \binom{N}{x} p^x (1-p)^{N-x}$$

where N: number of trials, x (a scalar): number of successes, p: probability is success in each trial

- Can re-express the above representation of the Binomial as

$$\binom{N}{x}\exp\left[x\log\left(\frac{p}{1-p}\right) + N\log(1-p)\right]$$

- $h(x) = \binom{N}{x}$
- $\theta = \log\left(\frac{p}{1-p}\right)$, and $p = \frac{1}{1+\exp(-\theta)}$
- $\phi(x) = x$

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of the Binomial distribution

$$\text{Binomial}(x|N,p) = \binom{N}{x} p^x (1-p)^{N-x}$$

where N: number of trials, x (a scalar): number of successes, p: probability is success in each trial

- Can re-express the above representation of the Binomial as

$$\binom{N}{x} \exp\left[x \log\left(\frac{p}{1-p}\right) + N\log(1-p)\right]$$

- $h(x) = \binom{N}{x}$
- $\theta = \log\left(\frac{p}{1-p}\right)$, and $p = \frac{1}{1+\exp(-\theta)}$
- $\phi(x) = x$
- $A(\theta) = -N\log(1-p)$

Binomial as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of the Binomial distribution

$$\text{Binomial}(x|N,p) = \binom{N}{x} p^x (1-p)^{N-x}$$

where N: number of trials, x (a scalar): number of successes, p: probability is success in each trial

- Can re-express the above representation of the Binomial as

$$\binom{N}{x} \exp\left[x \log\left(\frac{p}{1-p}\right) + N\log(1-p)\right]$$

- $h(x) = \binom{N}{x}$
- $\theta = \log\left(\frac{p}{1-p}\right)$, and $p = \frac{1}{1+\exp(-\theta)}$
- $\phi(x) = x$
- $A(\theta) = -N\log(1-p) = N\log(1+\exp(\theta))$

(Univariate) Gaussian as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

(Univariate) Gaussian as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of a univariate Gaussian

$$\begin{aligned}
\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] &= \frac{1}{\sqrt{2\pi}}\exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right] \\
&= \frac{1}{\sqrt{2\pi}}\exp\left[\begin{bmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix}^\top \begin{bmatrix} x \\ x^2\end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} - \log\sigma\right)\right]
\end{aligned}$$

(Univariate) Gaussian as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of a univariate Gaussian

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[\begin{bmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} - \log\sigma\right)\right]$$

- $h(x) = \frac{1}{\sqrt{2\pi}}$

(Univariate) Gaussian as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of a univariate Gaussian

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi}}\exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right]$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left[\begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix}^\top \begin{bmatrix}x\\ x^2\end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} - \log\sigma\right)\right]$$

- $h(x) = \frac{1}{\sqrt{2\pi}}$

- $\theta = \begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix} = \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix}$

(Univariate) Gaussian as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of a univariate Gaussian

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[\begin{bmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} - \log\sigma\right)\right]$$

- $h(x) = \frac{1}{\sqrt{2\pi}}$
- $\theta = \begin{bmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix} = \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix}$, and $\begin{bmatrix}\mu \\ \sigma^2\end{bmatrix} = \begin{bmatrix}-\frac{\theta_1}{2\theta_2} \\ -\frac{1}{2\theta_2}\end{bmatrix}$

(Univariate) Gaussian as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp[\theta^\top\phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of a univariate Gaussian

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi}}\exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right]$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left[\begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix}^\top\begin{bmatrix}x\\ x^2\end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} - \log\sigma\right)\right]$$

- $h(x) = \frac{1}{\sqrt{2\pi}}$
- $\theta = \begin{bmatrix}\frac{\mu}{\sigma^2}\\ -\frac{1}{2\sigma^2}\end{bmatrix} = \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix}$, and $\begin{bmatrix}\mu\\ \sigma^2\end{bmatrix} = \begin{bmatrix}-\frac{\theta_1}{2\theta_2}\\ -\frac{1}{2\theta_2}\end{bmatrix}$
- $\phi(x) = \begin{bmatrix}x\\ x^2\end{bmatrix}$

(Univariate) Gaussian as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of a univariate Gaussian

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[\begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} - \log\sigma\right)\right]$$

- $h(x) = \frac{1}{\sqrt{2\pi}}$
- $\theta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$, and $\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} -\frac{\theta_1}{2\theta_2} \\ -\frac{1}{2\theta_2} \end{bmatrix}$
- $\phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$
- $A(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma$

(Univariate) Gaussian as Exponential Family

- Let's try to write the Binomial distribution in the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp[\theta^\top \phi(\boldsymbol{x}) - A(\theta)]$$

- Recall the standard definition of a univariate Gaussian

$$\begin{aligned}\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] &= \frac{1}{\sqrt{2\pi}}\exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right] \\ &= \frac{1}{\sqrt{2\pi}}\exp\left[\begin{bmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix}^\top \begin{bmatrix}x \\ x^2\end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} - \log\sigma\right)\right]\end{aligned}$$

- $h(x) = \frac{1}{\sqrt{2\pi}}$
- $\theta = \begin{bmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix} = \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix}$, and $\begin{bmatrix}\mu \\ \sigma^2\end{bmatrix} = \begin{bmatrix}-\frac{\theta_1}{2\theta_2} \\ -\frac{1}{2\theta_2}\end{bmatrix}$
- $\phi(x) = \begin{bmatrix}x \\ x^2\end{bmatrix}$
- $A(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma = \frac{-\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi)$

A General Trick

A general trick to represent any distribution (assuming it is exp-family dist.) in exp-family form

- Write the given distribution as $\exp(\log p())$ and simplify, e.g., for the Binomial

$$\begin{aligned}
\exp\left(\log \text{Binomial}(x|N,p)\right) &= \exp\left(\log \binom{N}{x} p^x (1-p)^{N-x}\right) \\
&= \exp\left(\log \binom{N}{x} + x\log p + (N-x)\log(1-p)\right) \\
&= \binom{N}{x} \exp\left(x \log \frac{p}{1-p} - N\log(1-p)\right)
\end{aligned}$$

- Now compare the resulting expression with the exponential family form

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x})\exp(\theta^\top \phi(\boldsymbol{x}) - A(\theta))$$

.. to identify the natural parameters, sufficient statistics, log-partition function, etc.

Other Examples

- Many other distribution belong to the exponential family
  - Bernoulli
  - Beta
  - Gamma
  - Multinoulli/Multinomial
  - Dirichlet
  - Multivariate Gaussian
  - .. and many more (`https://en.wikipedia.org/wiki/Exponential_family`)

Other Examples

- Many other distribution belong to the exponential family
  - Bernoulli
  - Beta
  - Gamma
  - Multinoulli/Multinomial
  - Dirichlet
  - Multivariate Gaussian
  - .. and many more (`https://en.wikipedia.org/wiki/Exponential_family`)
- Note: Not all distributions belong to the exponential family, e.g.,
  - Uniform distribution ($x \sim \text{Unif}(a, b)$)
  - Student-t distribution
  - Mixture distributions (e.g., mixture of Gaussians)

Other Examples

- Many other distribution belong to the exponential family
  - Bernoulli
  - Beta
  - Gamma
  - Multinoulli/Multinomial
  - Dirichlet
  - Multivariate Gaussian
  - .. and many more (`https://en.wikipedia.org/wiki/Exponential_family`)
- Note: Not all distributions belong to the exponential family, e.g.,
  - Uniform distribution ($x \sim \text{Unif}(a, b)$)
  - Student-t distribution
  - Mixture distributions (e.g., mixture of Gaussians)
- If the support of the distribution depends on its parameters, then it is not an exp. family dist.

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$
- Exercise: Assume θ to be a scalar (thus $\phi(\boldsymbol{x})$ is also scalar).

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$
- Exercise: Assume θ to be a scalar (thus $\phi(\boldsymbol{x})$ is also scalar). Show that the first and the second derivatives of $A(\theta)$ are

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$
- Exercise: Assume θ to be a scalar (thus $\phi(\boldsymbol{x})$ is also scalar). Show that the first and the second derivatives of $A(\theta)$ are

$$\frac{dA}{d\theta} = \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]$$

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$
- Exercise: Assume θ to be a scalar (thus $\phi(\boldsymbol{x})$ is also scalar). Show that the first and the second derivatives of $A(\theta)$ are

$$\frac{dA}{d\theta} = \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \text{mean}[\phi(\boldsymbol{x})]$$

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$
- Exercise: Assume θ to be a scalar (thus $\phi(\boldsymbol{x})$ is also scalar). Show that the first and the second derivatives of $A(\theta)$ are

$$
\begin{aligned}
\frac{dA}{d\theta} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \text{mean}[\phi(\boldsymbol{x})] \\
\frac{d^2A}{d\theta^2} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi^2(\boldsymbol{x})] - \left[\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]\right]^2
\end{aligned}
$$

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$
- Exercise: Assume θ to be a scalar (thus $\phi(\boldsymbol{x})$ is also scalar). Show that the first and the second derivatives of $A(\theta)$ are

$$
\begin{aligned}
\frac{dA}{d\theta} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \text{mean}[\phi(\boldsymbol{x})] \\
\frac{d^2A}{d\theta^2} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi^2(\boldsymbol{x})] - \left[\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]\right]^2 = \text{var}[\phi(\boldsymbol{x})]
\end{aligned}
$$

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$
- Exercise: Assume θ to be a scalar (thus $\phi(\boldsymbol{x})$ is also scalar). Show that the first and the second derivatives of $A(\theta)$ are

$$\begin{aligned} \frac{dA}{d\theta} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \text{mean}[\phi(\boldsymbol{x})] \\ \frac{d^2A}{d\theta^2} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi^2(\boldsymbol{x})] - \left[\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]\right]^2 = \text{var}[\phi(\boldsymbol{x})] \end{aligned}$$

- Note: The above result also holds when θ and $\phi(\boldsymbol{x})$ are vector-valued (the "var" will be "covar")

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$
- Exercise: Assume θ to be a scalar (thus $\phi(\boldsymbol{x})$ is also scalar). Show that the first and the second derivatives of $A(\theta)$ are

$$
\begin{aligned}
\frac{dA}{d\theta} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \text{mean}[\phi(\boldsymbol{x})] \\
\frac{d^2A}{d\theta^2} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi^2(\boldsymbol{x})] - \left[\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]\right]^2 = \text{var}[\phi(\boldsymbol{x})]
\end{aligned}
$$

- Note: The above result also holds when θ and $\phi(\boldsymbol{x})$ are vector-valued (the "var" will be "covar")
- Important: $A(\theta)$ is a **convex function** of θ.

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$
- Exercise: Assume θ to be a scalar (thus $\phi(\boldsymbol{x})$ is also scalar). Show that the first and the second derivatives of $A(\theta)$ are

$$\begin{aligned}\frac{dA}{d\theta} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \text{mean}[\phi(\boldsymbol{x})] \\ \frac{d^2A}{d\theta^2} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi^2(\boldsymbol{x})] - \left[\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]\right]^2 = \text{var}[\phi(\boldsymbol{x})]\end{aligned}$$

- Note: The above result also holds when θ and $\phi(\boldsymbol{x})$ are vector-valued (the "var" will be "covar")
- Important: $A(\theta)$ is a **convex function** of θ. Why?

Log-Partition Function

- $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$ is the log-partition function
- $A(\theta)$ is also called the cumulant function
- Derivatives of $A(\theta)$ can be used to generate the cumulants of the sufficient statistics $\phi(\boldsymbol{x})$
- Exercise: Assume θ to be a scalar (thus $\phi(\boldsymbol{x})$ is also scalar). Show that the first and the second derivatives of $A(\theta)$ are

$$\begin{aligned}
\frac{dA}{d\theta} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \text{mean}[\phi(\boldsymbol{x})] \\
\frac{d^2A}{d\theta^2} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi^2(\boldsymbol{x})] - \left[\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]\right]^2 = \text{var}[\phi(\boldsymbol{x})]
\end{aligned}$$

- Note: The above result also holds when θ and $\phi(\boldsymbol{x})$ are vector-valued (the "var" will be "covar")
- Important: $A(\theta)$ is a **convex function** of θ. Why?
- Exercise: For Binomial, using its expression of $A(\theta)$, derive the first and second cumulants of $\phi(x)$

MLE for Exponential Family Distributions

- Suppose we have data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn i.i.d. from an exponential family distribution

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp\left[\theta^\top \phi(\boldsymbol{x}) - A(\theta)\right]$$

MLE for Exponential Family Distributions

- Suppose we have data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn i.i.d. from an exponential family distribution

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp\left[\theta^\top \phi(\boldsymbol{x}) - A(\theta)\right]$$

- To do MLE, we need the overall likelihood. This is simply a product of the individual likelihoods

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta)$$

MLE for Exponential Family Distributions

- Suppose we have data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn i.i.d. from an exponential family distribution

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp\left[\theta^\top \phi(\boldsymbol{x}) - A(\theta)\right]$$

- To do MLE, we need the overall likelihood. This is simply a product of the individual likelihoods

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) - NA(\theta)\right]$$

MLE for Exponential Family Distributions

- Suppose we have data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn i.i.d. from an exponential family distribution

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp\left[\theta^\top \phi(\boldsymbol{x}) - A(\theta)\right]$$

- To do MLE, we need the overall likelihood. This is simply a product of the individual likelihoods

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) - NA(\theta)\right] = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

MLE for Exponential Family Distributions

- Suppose we have data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn i.i.d. from an exponential family distribution

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp\left[\theta^\top \phi(\boldsymbol{x}) - A(\theta)\right]$$

- To do MLE, we need the overall likelihood. This is simply a product of the individual likelihoods

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) - NA(\theta)\right] = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- To estimate θ (as we'll see shortly), we only need $\phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$ and N

MLE for Exponential Family Distributions

- Suppose we have data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn i.i.d. from an exponential family distribution

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp\left[\theta^\top \phi(\boldsymbol{x}) - A(\theta)\right]$$

- To do MLE, we need the overall likelihood. This is simply a product of the individual likelihoods

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) - NA(\theta)\right] = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- To estimate θ (as we'll see shortly), we only need $\phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$ and N
- Size of $\phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$ does not grow with N (same as the size of each $\phi(\boldsymbol{x}_i)$)

MLE for Exponential Family Distributions

- Suppose we have data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn i.i.d. from an exponential family distribution

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp\left[\theta^\top \phi(\boldsymbol{x}) - A(\theta)\right]$$

- To do MLE, we need the overall likelihood. This is simply a product of the individual likelihoods

$$p(\mathcal{D}|\theta) = \prod_{i=1}^N p(\boldsymbol{x}_i|\theta) = \left[\prod_{i=1}^N h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \sum_{i=1}^N \phi(\boldsymbol{x}_i) - NA(\theta)\right] = \left[\prod_{i=1}^N h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- To estimate θ (as we'll see shortly), we only need $\phi(\mathcal{D}) = \sum_{i=1}^N \phi(\boldsymbol{x}_i)$ and N
- Size of $\phi(\mathcal{D}) = \sum_{i=1}^N \phi(\boldsymbol{x}_i)$ does not grow with N (same as the size of each $\phi(\boldsymbol{x}_i)$)
- Only exponential family distributions have finite-sized sufficient statistics

MLE for Exponential Family Distributions

- Suppose we have data $\mathcal{D} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ drawn i.i.d. from an exponential family distribution

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp\left[\theta^\top \phi(\boldsymbol{x}) - A(\theta)\right]$$

- To do MLE, we need the overall likelihood. This is simply a product of the individual likelihoods

$$p(\mathcal{D}|\theta) = \prod_{i=1}^N p(\boldsymbol{x}_i|\theta) = \left[\prod_{i=1}^N h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \sum_{i=1}^N \phi(\boldsymbol{x}_i) - NA(\theta)\right] = \left[\prod_{i=1}^N h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- To estimate θ (as we'll see shortly), we only need $\phi(\mathcal{D}) = \sum_{i=1}^N \phi(\boldsymbol{x}_i)$ and N
- Size of $\phi(\mathcal{D}) = \sum_{i=1}^N \phi(\boldsymbol{x}_i)$ does not grow with N (same as the size of each $\phi(\boldsymbol{x}_i)$)
- Only exponential family distributions have finite-sized sufficient statistics
  - No need to store all the data; can simply store and recursively update the sufficient statistics with more and more data

MLE for Exponential Family Distributions

- Suppose we have data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn i.i.d. from an exponential family distribution

$$p(\boldsymbol{x}|\theta) = h(\boldsymbol{x}) \exp\left[\theta^\top \phi(\boldsymbol{x}) - A(\theta)\right]$$

- To do MLE, we need the overall likelihood. This is simply a product of the individual likelihoods

$$p(\mathcal{D}|\theta) = \prod_{i=1}^N p(\boldsymbol{x}_i|\theta) = \left[\prod_{i=1}^N h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \sum_{i=1}^N \phi(\boldsymbol{x}_i) - NA(\theta)\right] = \left[\prod_{i=1}^N h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- To estimate θ (as we'll see shortly), we only need $\phi(\mathcal{D}) = \sum_{i=1}^N \phi(\boldsymbol{x}_i)$ and N
- Size of $\phi(\mathcal{D}) = \sum_{i=1}^N \phi(\boldsymbol{x}_i)$ does not grow with N (same as the size of each $\phi(\boldsymbol{x}_i)$)
- Only exponential family distributions have finite-sized sufficient statistics
  - No need to store all the data; can simply store and recursively update the sufficient statistics with more and more data
  - Very useful when doing probabilistic/Bayesian inference with large-scale data sets. Also useful in online parameter estimation problems.

MLE for Exponential Family Distributions

- The likelihood is of the form $p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$
- The log-likelihood is (ignoring constant w.r.t. θ): $\log p(\mathcal{D}|\theta) = \theta^\top \phi(\mathcal{D}) - NA(\theta)$

MLE for Exponential Family Distributions

- The likelihood is of the form $p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$
- The log-likelihood is (ignoring constant w.r.t. θ): $\log p(\mathcal{D}|\theta) = \theta^\top \phi(\mathcal{D}) - NA(\theta)$
- Note: This is concave in θ (since $-A(\theta)$ is concave). Maximization will yield a global maxima of θ

MLE for Exponential Family Distributions

- The likelihood is of the form $p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$
- The log-likelihood is (ignoring constant w.r.t. θ): $\log p(\mathcal{D}|\theta) = \theta^\top \phi(\mathcal{D}) - NA(\theta)$
- Note: This is concave in θ (since $-A(\theta)$ is concave). Maximization will yield a global maxima of θ
- MLE for exp-fam distributions can also be seen as doing moment-matching. To see this, note that

$$\nabla_\theta \left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

MLE for Exponential Family Distributions

- The likelihood is of the form $p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$
- The log-likelihood is (ignoring constant w.r.t. θ): $\log p(\mathcal{D}|\theta) = \theta^\top \phi(\mathcal{D}) - NA(\theta)$
- Note: This is concave in θ (since $-A(\theta)$ is concave). Maximization will yield a global maxima of θ
- MLE for exp-fam distributions can also be seen as doing moment-matching. To see this, note that

$$\nabla_\theta \left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] = \phi(\mathcal{D}) - N\nabla_\theta[A(\theta)]$$

MLE for Exponential Family Distributions

- The likelihood is of the form $p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$
- The log-likelihood is (ignoring constant w.r.t. θ): $\log p(\mathcal{D}|\theta) = \theta^\top \phi(\mathcal{D}) - NA(\theta)$
- Note: This is concave in θ (since $-A(\theta)$ is concave). Maximization will yield a global maxima of θ
- MLE for exp-fam distributions can also be seen as doing moment-matching. To see this, note that

$$\nabla_\theta \left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \quad = \quad \phi(\mathcal{D}) - N\nabla_\theta[A(\theta)] \quad = \quad \phi(\mathcal{D}) - N\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]$$

MLE for Exponential Family Distributions

- The likelihood is of the form $p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$
- The log-likelihood is (ignoring constant w.r.t. θ): $\log p(\mathcal{D}|\theta) = \theta^\top \phi(\mathcal{D}) - NA(\theta)$
- Note: This is concave in θ (since $-A(\theta)$ is concave). Maximization will yield a global maxima of θ
- MLE for exp-fam distributions can also be seen as doing moment-matching. To see this, note that

$$\nabla_\theta \left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] = \phi(\mathcal{D}) - N\nabla_\theta[A(\theta)] = \phi(\mathcal{D}) - N\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) - N\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]$$

- Therefore, at the "optimal" (i.e., MLE) $\hat{\theta}$, where the derivative is 0, the following must hold

$$\boxed{\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \frac{1}{N}\sum_{i=1}^{N} \phi(\boldsymbol{x}_i)}$$

MLE for Exponential Family Distributions

- The likelihood is of the form $p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$
- The log-likelihood is (ignoring constant w.r.t. θ): $\log p(\mathcal{D}|\theta) = \theta^\top \phi(\mathcal{D}) - NA(\theta)$
- Note: This is concave in θ (since $-A(\theta)$ is concave). Maximization will yield a global maxima of θ
- MLE for exp-fam distributions can also be seen as doing moment-matching. To see this, note that

$$\nabla_\theta \left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] = \phi(\mathcal{D}) - N\nabla_\theta[A(\theta)] = \phi(\mathcal{D}) - N\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) - N\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]$$

- Therefore, at the "optimal" (i.e., MLE) $\hat{\theta}$, where the derivative is 0, the following must hold

$$\boxed{\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \frac{1}{N}\sum_{i=1}^{N} \phi(\boldsymbol{x}_i)}$$

- This is basically matching the expected moments of the distribution with empirical moments ("empirical" here means what we compute using the observed data)

Moment Matching: An Example

- Given N observations $x_1, \ldots, x_N$ from a univariate Gaussian $N(x|\mu, \sigma^2)$, doing moment-matching

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

Moment Matching: An Example

- Given N observations $x_1, \ldots, x_N$ from a univariate Gaussian $N(x|\mu, \sigma^2)$, doing moment-matching

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

- The "true", i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$. Therefore

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

Moment Matching: An Example

- Given N observations $x_1, \ldots, x_N$ from a univariate Gaussian $N(x|\mu, \sigma^2)$, doing moment-matching

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

- The "true", i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$. Therefore

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- For a univariate Gaussian, note that $\mathbb{E}[x] = \mu$

Moment Matching: An Example

- Given N observations $x_1, \ldots, x_N$ from a univariate Gaussian $N(x|\mu, \sigma^2)$, doing moment-matching

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

- The "true", i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$. Therefore

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- For a univariate Gaussian, note that $\mathbb{E}[x] = \mu$ and $\mathbb{E}[x^2] = \text{var}[x] + \mathbb{E}[x]^2$

Moment Matching: An Example

- Given N observations $x_1, \ldots, x_N$ from a univariate Gaussian $N(x|\mu, \sigma^2)$, doing moment-matching

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

- The "true", i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$. Therefore

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- For a univariate Gaussian, note that $\mathbb{E}[x] = \mu$ and $\mathbb{E}[x^2] = \text{var}[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$

Moment Matching: An Example

- Given N observations $x_1, \ldots, x_N$ from a univariate Gaussian $N(x|\mu, \sigma^2)$, doing moment-matching

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

- The "true", i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$. Therefore

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- For a univariate Gaussian, note that $\mathbb{E}[x] = \mu$ and $\mathbb{E}[x^2] = \text{var}[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$
- Thus we have two equations and two unknowns

Moment Matching: An Example

- Given N observations $x_1, \ldots, x_N$ from a univariate Gaussian $N(x|\mu, \sigma^2)$, doing moment-matching

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

- The "true", i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$. Therefore

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- For a univariate Gaussian, note that $\mathbb{E}[x] = \mu$ and $\mathbb{E}[x^2] = \text{var}[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$
- Thus we have two equations and two unknowns
- From the first equation, we immediately get $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

Moment Matching: An Example

- Given N observations $x_1, \ldots, x_N$ from a univariate Gaussian $N(x|\mu, \sigma^2)$, doing moment-matching

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

- The "true", i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$. Therefore

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- For a univariate Gaussian, note that $\mathbb{E}[x] = \mu$ and $\mathbb{E}[x^2] = \text{var}[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$
- Thus we have two equations and two unknowns
- From the first equation, we immediately get $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- From the second equation, we get $\sigma^2 = \mathbb{E}[x^2] - \mu^2$

Moment Matching: An Example

- Given N observations $x_1, \ldots, x_N$ from a univariate Gaussian $N(x|\mu, \sigma^2)$, doing moment-matching

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

- The "true", i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$. Therefore

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- For a univariate Gaussian, note that $\mathbb{E}[x] = \mu$ and $\mathbb{E}[x^2] = \text{var}[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$
- Thus we have two equations and two unknowns
- From the first equation, we immediately get $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- From the second equation, we get $\sigma^2 = \mathbb{E}[x^2] - \mu^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$

Moment Matching: An Example

- Given N observations $x_1, \ldots, x_N$ from a univariate Gaussian $N(x|\mu, \sigma^2)$, doing moment-matching

$$\mathbb{E}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$$

- The "true", i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$. Therefore

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- For a univariate Gaussian, note that $\mathbb{E}[x] = \mu$ and $\mathbb{E}[x^2] = \text{var}[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$
- Thus we have two equations and two unknowns
- From the first equation, we immediately get $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- From the second equation, we get $\sigma^2 = \mathbb{E}[x^2] - \mu^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Bayesian Inference for Exponential Family Distributions

- We saw that the total likelihood given N i.i.d. observations $\mathcal{D}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

Bayesian Inference for Exponential Family Distributions

- We saw that the total likelihood given N i.i.d. observations $\mathcal{D}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- Let's choose the following prior (note: it looks similar in terms of θ within the exponent)

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]}$$

Bayesian Inference for Exponential Family Distributions

- We saw that the total likelihood given N i.i.d. observations $\mathcal{D}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- Let's choose the following prior (note: it looks similar in terms of θ within the exponent)

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]}$$

- Ignoring the prior's log-partition function $A_c(\nu_0, \boldsymbol{\tau}_0) = \log \int_\theta h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right] d\theta$

Bayesian Inference for Exponential Family Distributions

- We saw that the total likelihood given N i.i.d. observations $\mathcal{D}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- Let's choose the following prior (note: it looks similar in terms of θ within the exponent)

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]}$$

- Ignoring the prior's log-partition function $A_c(\nu_0, \boldsymbol{\tau}_0) = \log \int_\theta h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right] d\theta$

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]}$$

Bayesian Inference for Exponential Family Distributions

- We saw that the total likelihood given N i.i.d. observations $\mathcal{D}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- Let's choose the following prior (note: it looks similar in terms of θ within the exponent)

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]}$$

- Ignoring the prior's log-partition function $A_c(\nu_0, \boldsymbol{\tau}_0) = \log \int_\theta h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right] d\theta$

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]}$$

- Comparing the prior's form with the likelihood, we notice that

Bayesian Inference for Exponential Family Distributions

- We saw that the total likelihood given N i.i.d. observations $\mathcal{D}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- Let's choose the following prior (note: it looks similar in terms of θ within the exponent)

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]}$$

- Ignoring the prior's log-partition function $A_c(\nu_0, \boldsymbol{\tau}_0) = \log \int_\theta h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right] d\theta$

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]}$$

- Comparing the prior's form with the likelihood, we notice that
  - ν_0 is like the number of "pseudo-observations" coming from the prior

Bayesian Inference for Exponential Family Distributions

- We saw that the total likelihood given N i.i.d. observations $\mathcal{D}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- Let's choose the following prior (note: it looks similar in terms of θ within the exponent)

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]}$$

- Ignoring the prior's log-partition function $A_c(\nu_0, \boldsymbol{\tau}_0) = \log \int_\theta h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right] d\theta$

$$\boxed{p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]}$$

- Comparing the prior's form with the likelihood, we notice that
  - ν_0 is like the number of "pseudo-observations" coming from the prior
  - τ_0 is the total sufficient statistics of these ν_0 pseudo-observations

The Posterior Distribution

- As we saw, the likelihood is

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- And the prior we chose is

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

The Posterior Distribution

- As we saw, the likelihood is

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \quad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- And the prior we chose is

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta)\exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

- For this form of the prior, the posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ will be

$$\boxed{p(\theta|\mathcal{D}) \propto h(\theta)\exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]}$$

The Posterior Distribution

- As we saw, the likelihood is

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- And the prior we chose is

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

- For this form of the prior, the posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ will be

$$\boxed{p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]}$$

- Note that the posterior has the same form as the prior

The Posterior Distribution

- As we saw, the likelihood is

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- And the prior we chose is

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

- For this form of the prior, the posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ will be

$$\boxed{p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]}$$

- Note that the posterior has the same form as the prior; such a prior is called a **conjugate prior**

The Posterior Distribution

- As we saw, the likelihood is

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- And the prior we chose is

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

- For this form of the prior, the posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ will be

$$\boxed{p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]}$$

- Note that the posterior has the same form as the prior; such a prior is called a **conjugate prior** (note: all exponential family distributions have a conjugate prior having a form shown as above)

The Posterior Distribution

- As we saw, the likelihood is

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- And the prior we chose is

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

- For this form of the prior, the posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ will be

$$\boxed{p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]}$$

- Note that the posterior has the same form as the prior; such a prior is called a **conjugate prior** (note: all exponential family distributions have a conjugate prior having a form shown as above)
- Thus posterior hyperparams ${\nu_0}', {\tau_0}'$ are obtained by simply adding "stuff" to prior's hyperparams

The Posterior Distribution

- As we saw, the likelihood is

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- And the prior we chose is

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

- For this form of the prior, the posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ will be

$$\boxed{p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]}$$

- Note that the posterior has the same form as the prior; such a prior is called a **conjugate prior** (note: all exponential family distributions have a conjugate prior having a form shown as above)

- Thus posterior hyperparams ${\nu_0}', {\tau_0}'$ are obtained by simply adding "stuff" to prior's hyperparams

$${\nu_0}' \quad \leftarrow \quad \nu_0 + N \qquad\qquad \text{(no. of pseudo-obs + no. of actual obs)}$$

The Posterior Distribution

- As we saw, the likelihood is

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- And the prior we chose is

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta)\exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

- For this form of the prior, the posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ will be

$$\boxed{p(\theta|\mathcal{D}) \propto h(\theta)\exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]}$$

- Note that the posterior has the same form as the prior; such a prior is called a **conjugate prior** (note: all exponential family distributions have a conjugate prior having a form shown as above)

- Thus posterior hyperparams ${\nu_0}', {\tau_0}'$ are obtained by simply adding "stuff" to prior's hyperparams

$${\nu_0}' \leftarrow \nu_0 + N \qquad \text{(no. of pseudo-obs + no. of actual obs)}$$

$${\tau_0}' \leftarrow \tau_0 + \phi(\mathcal{D}) \qquad \text{(total suff-stats from pseudo-obs + total suff-stats from actual obs)}$$

The Posterior Distribution

- As we saw, the likelihood is

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- And the prior we chose is

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

- For this form of the prior, the posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ will be

$$\boxed{p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]}$$

- Note that the posterior has the same form as the prior; such a prior is called a **conjugate prior** (note: all exponential family distributions have a conjugate prior having a form shown as above)
- Thus posterior hyperparams ${\nu_0}', {\tau_0}'$ are obtained by simply adding "stuff" to prior's hyperparams

$$\begin{aligned} {\nu_0}' &\leftarrow \nu_0 + N && \text{(no. of pseudo-obs + no. of actual obs)} \\ {\tau_0}' &\leftarrow \tau_0 + \phi(\mathcal{D}) && \text{(total suff-stats from pseudo-obs + total suff-stats from actual obs)} \end{aligned}$$

- Note: Prior's log-partition function $A_c(\nu_0, \boldsymbol{\tau}_0)$ updates to posterior's: $A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))$

The Posterior Distribution

- Assuming the prior $p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$, the posterior was

$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]$$

The Posterior Distribution

- Assuming the prior $p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$, the posterior was

$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]$$

- Assuming $\tau_0 = \nu_0 \bar{\tau}_0$, we can also write the prior as $p(\theta|\nu_0, \bar{\tau}_0) \propto \exp\left[\theta^\top \nu_0 \bar{\tau}_0 - \nu_0 A(\theta)\right]$

The Posterior Distribution

- Assuming the prior $p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$, the posterior was

$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N) A(\theta)\right]$$

- Assuming $\tau_0 = \nu_0 \bar{\tau}_0$, we can also write the prior as $p(\theta|\nu_0, \bar{\tau}_0) \propto \exp\left[\theta^\top \nu_0 \bar{\tau}_0 - \nu_0 A(\theta)\right]$
- Can think of $\bar{\tau}_0 = \tau_0/\nu_0$ as the average sufficient statistics per pseudo-observation

The Posterior Distribution

- Assuming the prior $p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$, the posterior was

$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]$$

- Assuming $\tau_0 = \nu_0 \bar{\tau}_0$, we can also write the prior as $p(\theta|\nu_0, \bar{\tau}_0) \propto \exp\left[\theta^\top \nu_0 \bar{\tau}_0 - \nu_0 A(\theta)\right]$
- Can think of $\bar{\tau}_0 = \tau_0/\nu_0$ as the average sufficient statistics per pseudo-observation
- The posterior can be written as

$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\nu_0 + N) \frac{\nu_0 \bar{\tau}_0 + \phi(\mathcal{D})}{\nu_0 + N} - (\nu_0 + N)A(\theta)\right]$$

The Posterior Distribution

- Assuming the prior $p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$, the posterior was

$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]$$

- Assuming $\tau_0 = \nu_0 \bar{\tau}_0$, we can also write the prior as $p(\theta|\nu_0, \bar{\tau}_0) \propto \exp\left[\theta^\top \nu_0 \bar{\tau}_0 - \nu_0 A(\theta)\right]$
- Can think of $\bar{\tau}_0 = \tau_0/\nu_0$ as the average sufficient statistics per pseudo-observation
- The posterior can be written as

$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\nu_0 + N) \frac{\nu_0 \bar{\tau}_0 + \phi(\mathcal{D})}{\nu_0 + N} - (\nu_0 + N)A(\theta)\right]$$

- Denoting $\bar{\phi} = \frac{\phi(D)}{N}$ as the average suff-stats per real observation, the posterior updates are

The Posterior Distribution

- Assuming the prior $p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$, the posterior was

$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N) A(\theta)\right]$$

- Assuming $\tau_0 = \nu_0 \bar{\tau}_0$, we can also write the prior as $p(\theta|\nu_0, \bar{\tau}_0) \propto \exp\left[\theta^\top \nu_0 \bar{\tau}_0 - \nu_0 A(\theta)\right]$
- Can think of $\bar{\tau}_0 = \tau_0/\nu_0$ as the average sufficient statistics per pseudo-observation
- The posterior can be written as

$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\nu_0 + N) \frac{\nu_0 \bar{\tau}_0 + \phi(\mathcal{D})}{\nu_0 + N} - (\nu_0 + N) A(\theta)\right]$$

- Denoting $\bar{\phi} = \frac{\phi(D)}{N}$ as the average suff-stats per real observation, the posterior updates are

$$\nu_0{}' \leftarrow \nu_0 + N$$

The Posterior Distribution

- Assuming the prior $p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$, the posterior was

$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]$$

- Assuming $\tau_0 = \nu_0 \bar{\tau}_0$, we can also write the prior as $p(\theta|\nu_0, \bar{\tau}_0) \propto \exp\left[\theta^\top \nu_0 \bar{\tau}_0 - \nu_0 A(\theta)\right]$
- Can think of $\bar{\tau}_0 = \tau_0/\nu_0$ as the average sufficient statistics per pseudo-observation
- The posterior can be written as

$$p(\theta|\mathcal{D}) \propto h(\theta) \exp\left[\theta^\top (\nu_0 + N) \frac{\nu_0 \bar{\tau}_0 + \phi(\mathcal{D})}{\nu_0 + N} - (\nu_0 + N)A(\theta)\right]$$

- Denoting $\bar{\phi} = \frac{\phi(D)}{N}$ as the average suff-stats per real observation, the posterior updates are

$$\begin{aligned} {\nu_0}' &\leftarrow \nu_0 + N \\ \bar{\tau}_0' &\leftarrow \frac{\nu_0 \bar{\tau}_0 + N\bar{\phi}}{\nu_0 + N} \end{aligned}$$

The Posterior Distribution

- Assuming the prior $p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta)\exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$, the posterior was

$$p(\theta|\mathcal{D}) \propto h(\theta)\exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]$$

- Assuming $\tau_0 = \nu_0\bar{\tau}_0$, we can also write the prior as $p(\theta|\nu_0, \bar{\tau}_0) \propto \exp\left[\theta^\top \nu_0\bar{\tau}_0 - \nu_0 A(\theta)\right]$
- Can think of $\bar{\tau}_0 = \tau_0/\nu_0$ as the average sufficient statistics per pseudo-observation
- The posterior can be written as

$$p(\theta|\mathcal{D}) \propto h(\theta)\exp\left[\theta^\top(\nu_0 + N)\frac{\nu_0\bar{\tau}_0 + \phi(\mathcal{D})}{\nu_0 + N} - (\nu_0 + N)A(\theta)\right]$$

- Denoting $\bar{\phi} = \frac{\phi(D)}{N}$ as the average suff-stats per real observation, the posterior updates are

$$\begin{aligned} {\nu_0}' &\leftarrow \nu_0 + N \\ \bar{\tau}_0' &\leftarrow \frac{\nu_0\bar{\tau}_0 + N\bar{\phi}}{\nu_0 + N} \end{aligned}$$

- Note that the posterior hyperparam $\bar{\tau}_0'$ is a convex combination of the average suff-stats $\bar{\tau}_0$ of the ν_0 pseudo-observations and the average suff-stats $\bar{\phi}$ of the N actual observations

Posterior Predictive Distribution

- Assume some past (training) data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from an exp. family distribution

Posterior Predictive Distribution

- Assume some past (training) data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from an exp. family distribution
- Assme some test data $\mathcal{D}' = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{N'}\}$ from the same distribution ($N' \geq 1$)

Posterior Predictive Distribution

- Assume some past (training) data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from an exp. family distribution
- Assme some test data $\mathcal{D}' = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{N'}\}$ from the same distribution ($N' \geq 1$)
- The posterior predictive distribution of $\mathcal{D}'$ (probability distribution of new data given old data)

$$p(\mathcal{D}'|\mathcal{D}) = \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta$$

Posterior Predictive Distribution

- Assume some past (training) data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from an exp. family distribution
- Assme some test data $\mathcal{D}' = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{N'}\}$ from the same distribution ($N' \geq 1$)
- The posterior predictive distribution of $\mathcal{D}'$ (probability distribution of new data given old data)

$$p(\mathcal{D}'|\mathcal{D}) = \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta$$

- We've already seen some specific examples of computing the posterior predictive dist., e.g.,

Posterior Predictive Distribution

- Assume some past (training) data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from an exp. family distribution
- Assme some test data $\mathcal{D}' = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{N'}\}$ from the same distribution ($N' \geq 1$)
- The posterior predictive distribution of $\mathcal{D}'$ (probability distribution of new data given old data)

$$p(\mathcal{D}'|\mathcal{D}) = \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta$$

- We've already seen some specific examples of computing the posterior predictive dist., e.g.,
    - Beta-Bernoulli case: Posterior predictive distribution of next coin toss

Posterior Predictive Distribution

- Assume some past (training) data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from an exp. family distribution
- Assme some test data $\mathcal{D}' = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{N'}\}$ from the same distribution ($N' \geq 1$)
- The posterior predictive distribution of $\mathcal{D}'$ (probability distribution of new data given old data)

$$p(\mathcal{D}'|\mathcal{D}) = \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta$$

- We've already seen some specific examples of computing the posterior predictive dist., e.g.,
  - Beta-Bernoulli case: Posterior predictive distribution of next coin toss
  - Bayesian linear regression: Posterior predictive distribution of the response y_* of test input $\boldsymbol{x}_*$

Posterior Predictive Distribution

- Assume some past (training) data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ generated from an exp. family distribution
- Assme some test data $\mathcal{D}' = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{N'}\}$ from the same distribution ($N' \geq 1$)
- The posterior predictive distribution of $\mathcal{D}'$ (probability distribution of new data given old data)

$$p(\mathcal{D}'|\mathcal{D}) = \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta$$

- We've already seen some specific examples of computing the posterior predictive dist., e.g.,
  - Beta-Bernoulli case: Posterior predictive distribution of next coin toss
  - Bayesian linear regression: Posterior predictive distribution of the response y_* of test input $\boldsymbol{x}_*$
- Nice Property: If the likelihood is an exponential family distribution, prior is conjugate (and thus is the posterior), the posterior predictive always has a closed form expression (shown next)

Posterior Predictive Distribution

- Recall the form of the likelihood $p(\mathcal{D}|\theta)$ for exp. family dist.

$$p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- The conjugate prior was

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]$$

Posterior Predictive Distribution

- Recall the form of the likelihood $p(\mathcal{D}|\theta)$ for exp. family dist.

$$p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- The conjugate prior was

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]$$

- For this choice of the conjugate prior, the posterior was shown to be

$$p(\theta|\mathcal{D}) = h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]$$

Posterior Predictive Distribution

- Recall the form of the likelihood $p(\mathcal{D}|\theta)$ for exp. family dist.

$$p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- The conjugate prior was

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]$$

- For this choice of the conjugate prior, the posterior was shown to be

$$p(\theta|\mathcal{D}) = h(\theta) \exp\left[\theta^\top (\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]$$

- For the test data $\mathcal{D}'$, the likelihood will be

$$p(\mathcal{D}'|\theta) = \left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}') - N'A(\theta)\right] \qquad \text{where} \quad \phi(\mathcal{D}') = \sum_{i=1}^{N'} \phi(\tilde{\boldsymbol{x}}_i)$$

Posterior Predictive Distribution

- Therefore the posterior predictive distribution will be

$$p(\mathcal{D}'|\mathcal{D}) = \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta$$

Posterior Predictive Distribution

- Therefore the posterior predictive distribution will be

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta \\
&= \int \underbrace{\left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right]}_{\text{constant w.r.t. } \theta} \exp\left[\theta^\top \phi(\mathcal{D}') - N'A(\theta)\right] h(\theta)\exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta) - \underbrace{A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))}_{\text{constant w.r.t. } \theta}\right] d\theta
\end{aligned}
$$

Posterior Predictive Distribution

- Therefore the posterior predictive distribution will be

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta \\
&= \int \underbrace{\left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right]}_{\text{constant w.r.t. } \theta} \exp\left[\theta^\top \phi(\mathcal{D}') - N'A(\theta)\right] h(\theta) \exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta) - \underbrace{A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))}_{\text{constant w.r.t. } \theta}\right] d\theta
\end{aligned}
$$

- The above gets simplified further into

Posterior Predictive Distribution

- Therefore the posterior predictive distribution will be

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta \\
&= \int \underbrace{\left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right]}_{\text{constant w.r.t. } \theta} \exp\left[\theta^\top \phi(\mathcal{D}') - N' A(\theta)\right] h(\theta) \exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta) - \underbrace{A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))}_{\text{constant w.r.t. } \theta}\right] d\theta
\end{aligned}
$$

- The above gets simplified further into

$$
p(\mathcal{D}'|\mathcal{D}) = \left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right] \frac{\int h(\theta) \exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - (\nu_0 + N + N')A(\theta)\right] d\theta}{\exp\left[A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]}
$$

Posterior Predictive Distribution

- Therefore the posterior predictive distribution will be

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta \\
&= \int \underbrace{\left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right]}_{\text{constant w.r.t. } \theta} \exp\left[\theta^\top \phi(\mathcal{D}') - N'A(\theta)\right] h(\theta) \exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta) - \underbrace{A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))}_{\text{constant w.r.t. } \theta}\right] d\theta
\end{aligned}
$$

- The above gets simplified further into

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right] \frac{\int h(\theta) \exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - (\nu_0 + N + N')A(\theta)\right] d\theta}{\exp\left[A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]} \\
&= \left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{\exp\left[A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]}
\end{aligned}
$$

Posterior Predictive Distribution

- Therefore the posterior predictive distribution will be

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta \\
&= \int \underbrace{\left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right]}_{\text{constant w.r.t. } \theta} \exp\left[\theta^\top \phi(\mathcal{D}') - N'A(\theta)\right] h(\theta) \exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta) - \underbrace{A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))}_{\text{constant w.r.t. } \theta}\right] d\theta
\end{aligned}
$$

- The above gets simplified further into

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{\int h(\theta)\exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - (\nu_0 + N + N')A(\theta)\right] d\theta}{\exp\left[A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]} \\
&= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{\exp\left[A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]}
\end{aligned}
$$

where $Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) = \int h(\theta)\exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - (\nu_0 + N + N')A(\theta)\right] d\theta$

Posterior Predictive Distribution

- Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the posterior predictive distribution as

Posterior Predictive Distribution

- Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the posterior predictive distribution as

$$p(\mathcal{D}'|\mathcal{D}) = \left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{Z_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))}$$

Posterior Predictive Distribution

- Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the posterior predictive distribution as

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{Z_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))} \\
&= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \exp\left[A_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]
\end{aligned}
$$

Posterior Predictive Distribution

- Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the posterior predictive distribution as

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{Z_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))} \\
&= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \exp\left[A_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]
\end{aligned}
$$

- Therefore the posterior predictive is proportional to ..

Posterior Predictive Distribution

- Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the posterior predictive distribution as

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{Z_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))} \\
&= \left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right] \exp\left[A_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]
\end{aligned}
$$

- Therefore the posterior predictive is proportional to ..
    - .. the ratio of two partition functions of two "posterior distributions" (one with $N + N'$ examples and the other with N examples)

Posterior Predictive Distribution

- Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the posterior predictive distribution as

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{Z_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))} \\
&= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \exp\left[A_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]
\end{aligned}
$$

- Therefore the posterior predictive is proportional to ..
  - .. the ratio of two partition functions of two "posterior distributions" (one with $N + N'$ examples and the other with N examples)
  - .. or exponential of the difference of the corresponding log-partition functions

Posterior Predictive Distribution

- Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the posterior predictive distribution as

$$\begin{aligned} p(\mathcal{D}'|\mathcal{D}) &= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{Z_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))} \\ &= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \exp\left[A_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right] \end{aligned}$$

- Therefore the posterior predictive is proportional to ..
  - .. the ratio of two partition functions of two "posterior distributions" (one with $N + N'$ examples and the other with N examples)
  - .. or exponential of the difference of the corresponding log-partition functions
- Note that the form of Z_c (and A_c) will simply depend on the chosen conjugate prior

Posterior Predictive Distribution

- Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the posterior predictive distribution as

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{Z_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))} \\
&= \left[\prod_{i=1}^{N'} h(\tilde{\boldsymbol{x}}_i)\right] \exp\left[A_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]
\end{aligned}
$$

- Therefore the posterior predictive is proportional to ..
  - .. the ratio of two partition functions of two "posterior distributions" (one with $N + N'$ examples and the other with N examples)
  - .. or exponential of the difference of the corresponding log-partition functions
- Note that the form of Z_c (and A_c) will simply depend on the chosen conjugate prior
- Very useful result. Also holds for $N = 0$

Posterior Predictive Distribution

- Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the posterior predictive distribution as

$$
\begin{aligned}
p(\mathcal{D}'|\mathcal{D}) &= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{Z_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))} \\
&= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \exp\left[A_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]
\end{aligned}
$$

- Therefore the posterior predictive is proportional to ..
  - .. the ratio of two partition functions of two "posterior distributions" (one with $N + N'$ examples and the other with N examples)
  - .. or exponential of the difference of the corresponding log-partition functions
- Note that the form of Z_c (and A_c) will simply depend on the chosen conjugate prior
- Very useful result. Also holds for $N = 0$
  - In the $N = 0$ case, $p(\mathcal{D}') = \int p(\mathcal{D}'|\theta)p(\theta)d\theta$ is simply the marginal likelihood of $\mathcal{D}'$

Exponential Family and GLM

Generalized Linear Models (GLM)

- (Probabilistic) Linear regression: when response y is real-valued

$$p(y|\boldsymbol{x},\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$$

Generalized Linear Models (GLM)

- (Probabilistic) Linear regression: when response y is real-valued

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$$

- Logistic regression: when response y is binary (0/1)

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \text{Bernoulli}(\sigma(\boldsymbol{w}^\top \boldsymbol{x})) = [\sigma(\boldsymbol{w}^\top \boldsymbol{x})]^y [1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x})]^{1-y}$$

where $\sigma(\boldsymbol{w}^\top \boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{w}^\top \boldsymbol{x})} = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})}$

Generalized Linear Models (GLM)

- (Probabilistic) Linear regression: when response y is real-valued

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$$

- Logistic regression: when response y is binary (0/1)

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \text{Bernoulli}(\sigma(\boldsymbol{w}^\top \boldsymbol{x})) = [\sigma(\boldsymbol{w}^\top \boldsymbol{x})]^y [1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x})]^{1-y}$$

  where $\sigma(\boldsymbol{w}^\top \boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{w}^\top \boldsymbol{x})} = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})}$

- In both, the model depends on the inputs $\boldsymbol{x}$ as $\boldsymbol{w}^\top \boldsymbol{x}$

Generalized Linear Models (GLM)

- (Probabilistic) Linear regression: when response y is real-valued

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$$

- Logistic regression: when response y is binary (0/1)

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \text{Bernoulli}(\sigma(\boldsymbol{w}^\top \boldsymbol{x})) = [\sigma(\boldsymbol{w}^\top \boldsymbol{x})]^y [1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x})]^{1-y}$$

where $\sigma(\boldsymbol{w}^\top \boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{w}^\top \boldsymbol{x})} = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})}$

- In both, the model depends on the inputs $\boldsymbol{x}$ as $\boldsymbol{w}^\top \boldsymbol{x}$
- Can we extend it to other type of outputs?

Generalized Linear Models (GLM)

- (Probabilistic) Linear regression: when response y is real-valued

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$$

- Logistic regression: when response y is binary (0/1)

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \text{Bernoulli}(\sigma(\boldsymbol{w}^\top \boldsymbol{x})) = [\sigma(\boldsymbol{w}^\top \boldsymbol{x})]^y[1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x})]^{1-y}$$

  where $\sigma(\boldsymbol{w}^\top \boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{w}^\top \boldsymbol{x})} = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})}$

- In both, the model depends on the inputs $\boldsymbol{x}$ as $\boldsymbol{w}^\top \boldsymbol{x}$
- Can we extend it to other type of outputs?
- Solution: Model the output using an exp-fam distribution (Gaussian and Bernoulli already are!)

$$\boxed{p(y|\eta) = h(y)\exp(\eta y - A(\eta))} \qquad \text{(Generalized Linear Model (GLM))}$$

Generalized Linear Models (GLM)

- (Probabilistic) Linear regression: when response y is real-valued

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1})$$

- Logistic regression: when response y is binary (0/1)

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \text{Bernoulli}(\sigma(\boldsymbol{w}^\top \boldsymbol{x})) = [\sigma(\boldsymbol{w}^\top \boldsymbol{x})]^y [1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x})]^{1-y}$$

where $\sigma(\boldsymbol{w}^\top \boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{w}^\top \boldsymbol{x})} = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})}$

- In both, the model depends on the inputs $\boldsymbol{x}$ as $\boldsymbol{w}^\top \boldsymbol{x}$
- Can we extend it to other type of outputs?
- Solution: Model the output using an exp-fam distribution (Gaussian and Bernoulli already are!)

$$\boxed{p(y|\eta) = h(y)\exp(\eta y - A(\eta))} \qquad \text{(Generalized Linear Model (GLM))}$$

.. where η is a scalar-valued natural parameter (depends on $\boldsymbol{x}$) and sufficient statistics $\phi(y) = y$

Generalized Linear Models: Formally

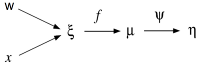
- The GLM is of the form $p(y|\eta) = h(y) \exp(\eta y - A(\eta))$ where η depends on $\boldsymbol{x}$

Generalized Linear Models: Formally

- The GLM is of the form $p(y|\eta) = h(y)\exp(\eta y - A(\eta))$ where η depends on $\boldsymbol{x}$
- The inputs $\boldsymbol{x}$ appear in the model only as a linear combination $\xi = \boldsymbol{w}^\top \boldsymbol{x}$

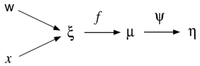
Generalized Linear Models: Formally

- The GLM is of the form $p(y|\eta) = h(y)\exp(\eta y - A(\eta))$ where η depends on $\boldsymbol{x}$
- The inputs $\boldsymbol{x}$ appear in the model only as a linear combination $\xi = \boldsymbol{w}^\top \boldsymbol{x}$



Generalized Linear Models: Formally

- The GLM is of the form $p(y|\eta) = h(y)\exp(\eta y - A(\eta))$ where η depends on $\boldsymbol{x}$
- The inputs $\boldsymbol{x}$ appear in the model only as a linear combination $\xi = \boldsymbol{w}^\top \boldsymbol{x}$

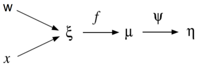


- Conditional mean μ of a response y's distribution is modeled via a response function f

$$\mu = \mathbb{E}[y] = f(\xi) = f(\boldsymbol{w}^\top \boldsymbol{x})$$

Generalized Linear Models: Formally

- The GLM is of the form $p(y|\eta) = h(y)\exp(\eta y - A(\eta))$ where η depends on $\boldsymbol{x}$
- The inputs $\boldsymbol{x}$ appear in the model only as a linear combination $\xi = \boldsymbol{w}^\top \boldsymbol{x}$



- Conditional mean μ of a response y's distribution is modeled via a response function f

$$\mu = \mathbb{E}[y] = f(\xi) = f(\boldsymbol{w}^\top \boldsymbol{x})$$

- for (probabilistic) linear regression, f is identity, i.e., $\mu = f(\boldsymbol{w}^\top \boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$,

Generalized Linear Models: Formally

- The GLM is of the form $p(y|\eta) = h(y)\exp(\eta y - A(\eta))$ where η depends on $\boldsymbol{x}$
- The inputs $\boldsymbol{x}$ appear in the model only as a linear combination $\xi = \boldsymbol{w}^\top \boldsymbol{x}$

$$\begin{tikzcd} \mathrm{w} \arrow[dr] & & & \\ & \xi \arrow[r, "f"] & \mu \arrow[r, "\Psi"] & \eta \\ x \arrow[ur] & & & \end{tikzcd}$$

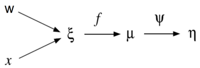
- Conditional mean μ of a response y's distribution is modeled via a response function f

$$\mu = \mathbb{E}[y] = f(\xi) = f(\boldsymbol{w}^\top \boldsymbol{x})$$

- for (probabilistic) linear regression, f is identity, i.e., $\mu = f(\boldsymbol{w}^\top \boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$,
- for logistic regression, f is sigmoid, i.e., $\mu = f(\boldsymbol{w}^\top \boldsymbol{x}) = \exp(\boldsymbol{w}^\top \boldsymbol{x})/(1 + \exp(\boldsymbol{w}^\top \boldsymbol{x}))$

Generalized Linear Models: Formally

- The GLM is of the form $p(y|\eta) = h(y)\exp(\eta y - A(\eta))$ where η depends on $\boldsymbol{x}$
- The inputs $\boldsymbol{x}$ appear in the model only as a linear combination $\xi = \boldsymbol{w}^\top \boldsymbol{x}$



- Conditional mean μ of a response y's distribution is modeled via a response function f

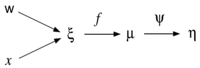
$$\mu = \mathbb{E}[y] = f(\xi) = f(\boldsymbol{w}^\top \boldsymbol{x})$$

- for (probabilistic) linear regression, f is identity, i.e., $\mu = f(\boldsymbol{w}^\top \boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$,
- for logistic regression, f is sigmoid, i.e., $\mu = f(\boldsymbol{w}^\top \boldsymbol{x}) = \exp(\boldsymbol{w}^\top \boldsymbol{x})/(1 + \exp(\boldsymbol{w}^\top \boldsymbol{x}))$

- The natural parameter $\eta = \psi(\mu)$ where ψ is the link function

Generalized Linear Models: Formally

- The GLM is of the form $p(y|\eta) = h(y)\exp(\eta y - A(\eta))$ where η depends on $\boldsymbol{x}$
- The inputs $\boldsymbol{x}$ appear in the model only as a linear combination $\xi = \boldsymbol{w}^\top \boldsymbol{x}$



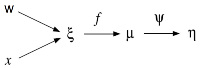
- Conditional mean μ of a response y's distribution is modeled via a response function f

$$\mu = \mathbb{E}[y] = f(\xi) = f(\boldsymbol{w}^\top \boldsymbol{x})$$

  - for (probabilistic) linear regression, f is identity, i.e., $\mu = f(\boldsymbol{w}^\top \boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$,
  - for logistic regression, f is sigmoid, i.e., $\mu = f(\boldsymbol{w}^\top \boldsymbol{x}) = \exp(\boldsymbol{w}^\top \boldsymbol{x})/(1 + \exp(\boldsymbol{w}^\top \boldsymbol{x}))$
- The natural parameter $\eta = \psi(\mu)$ where ψ is the link function
- Note: Some GLM can be represented as $p(y|\eta,\phi) = h(y,\phi)\exp(\frac{\eta y - A(\eta)}{\phi})$ where ϕ is a dispersion parameter (Gaussian/gamma GLMs use this representation)

GLM with Canonical Response Function

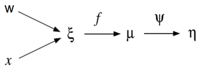
- A GLM has a canonical response function f if $f = \psi^{-1}$



- For such a GLM, $\eta = \psi(\mu) = \psi(f(\boldsymbol{w}^\top \boldsymbol{x})) = \boldsymbol{w}^\top \boldsymbol{x}$

GLM with Canonical Response Function

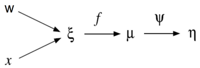
- A GLM has a canonical response function f if $f = \psi^{-1}$



- For such a GLM, $\eta = \psi(\mu) = \psi(f(\boldsymbol{w}^\top \boldsymbol{x})) = \boldsymbol{w}^\top \boldsymbol{x}$
- E.g., for logistic regression $\eta = \log \frac{\mu}{1-\mu} = \boldsymbol{w}^\top \boldsymbol{x}_n$

GLM with Canonical Response Function

- A GLM has a canonical response function f if $f = \psi^{-1}$

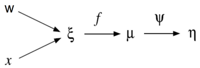


- For such a GLM, $\eta = \psi(\mu) = \psi(f(\boldsymbol{w}^\top \boldsymbol{x})) = \boldsymbol{w}^\top \boldsymbol{x}$
- E.g., for logistic regression $\eta = \log \frac{\mu}{1-\mu} = \boldsymbol{w}^\top \boldsymbol{x}_n$
- Thus, for Canonical GLMs

$$\begin{aligned} p(y|\eta) &= h(y)\exp(\eta y - A(\eta)) \\ &= h(y)\exp(y\boldsymbol{w}^\top \boldsymbol{x} - A(\eta)) \end{aligned}$$

GLM with Canonical Response Function

- A GLM has a canonical response function f if $f = \psi^{-1}$



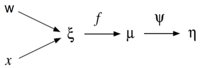
- For such a GLM, $\eta = \psi(\mu) = \psi(f(\boldsymbol{w}^\top \boldsymbol{x})) = \boldsymbol{w}^\top \boldsymbol{x}$
- E.g., for logistic regression $\eta = \log \frac{\mu}{1-\mu} = \boldsymbol{w}^\top \boldsymbol{x}_n$
- Thus, for Canonical GLMs

$$\begin{aligned} p(y|\eta) &= h(y)\exp(\eta y - A(\eta)) \\ &= h(y)\exp(y\boldsymbol{w}^\top \boldsymbol{x} - A(\eta)) \end{aligned}$$

- This form makes parameter estimation in canonical GLM easy (e.g., gradients easy to compute)

GLM with Canonical Response Function

- A GLM has a canonical response function f if $f = \psi^{-1}$



- For such a GLM, $\eta = \psi(\mu) = \psi(f(\boldsymbol{w}^\top \boldsymbol{x})) = \boldsymbol{w}^\top \boldsymbol{x}$
- E.g., for logistic regression $\eta = \log \frac{\mu}{1-\mu} = \boldsymbol{w}^\top \boldsymbol{x}_n$
- Thus, for Canonical GLMs

$$\begin{aligned} p(y|\eta) &= h(y)\exp(\eta y - A(\eta)) \\ &= h(y)\exp(y\boldsymbol{w}^\top \boldsymbol{x} - A(\eta)) \end{aligned}$$

- This form makes parameter estimation in canonical GLM easy (e.g., gradients easy to compute)
- We will focus on canonical GLMs only (these are the most common)

MLE for GLM

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n) \exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - A(\eta_n))$$

MLE for GLM

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n) \exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{w}^\top \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

MLE for GLM

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n) \exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{w}^\top \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

- Convexity of $A(\eta)$ guarantees a global optima. Taking derivative w.r.t. $\boldsymbol{w}$

$$\sum_{n=1}^{N} \left(y_n \boldsymbol{x}_n - A'(\eta_n) \frac{d\eta_n}{d\boldsymbol{w}} \right)$$

MLE for GLM

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n) \exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{w}^\top \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

- Convexity of $A(\eta)$ guarantees a global optima. Taking derivative w.r.t. $\boldsymbol{w}$

$$\sum_{n=1}^{N} \left(y_n \boldsymbol{x}_n - A'(\eta_n) \frac{d\eta_n}{d\boldsymbol{w}} \right) = \sum_{n=1}^{N} (y_n \boldsymbol{x}_n - \mu_n \boldsymbol{x}_n)$$

MLE for GLM

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n) \exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{w}^\top \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

- Convexity of $A(\eta)$ guarantees a global optima. Taking derivative w.r.t. $\boldsymbol{w}$

$$\sum_{n=1}^{N} \left(y_n \boldsymbol{x}_n - A'(\eta_n) \frac{d\eta_n}{d\boldsymbol{w}} \right) = \sum_{n=1}^{N} (y_n \boldsymbol{x}_n - \mu_n \boldsymbol{x}_n) = \sum_{n=1}^{N} (y_n - \mu_n) \boldsymbol{x}_n$$

MLE for GLM

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n) \exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{w}^\top \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

- Convexity of $A(\eta)$ guarantees a global optima. Taking derivative w.r.t. $\boldsymbol{w}$

$$\sum_{n=1}^{N} \left(y_n \boldsymbol{x}_n - A'(\eta_n) \frac{d\eta_n}{d\boldsymbol{w}} \right) = \sum_{n=1}^{N} (y_n \boldsymbol{x}_n - \mu_n \boldsymbol{x}_n) = \sum_{n=1}^{N} (y_n - \mu_n) \boldsymbol{x}_n$$

where $\mu_n = f(\boldsymbol{w}^\top \boldsymbol{x}_n)$ and 'f' $(= \psi^{-1})$ depends on type of response y, e.g.,

MLE for GLM

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n)\exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{w}^\top \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

- Convexity of $A(\eta)$ guarantees a global optima. Taking derivative w.r.t. $\boldsymbol{w}$

$$\sum_{n=1}^{N}\left(y_n \boldsymbol{x}_n - A'(\eta_n)\frac{d\eta_n}{d\boldsymbol{w}}\right) = \sum_{n=1}^{N}(y_n \boldsymbol{x}_n - \mu_n \boldsymbol{x}_n) = \sum_{n=1}^{N}(y_n - \mu_n)\boldsymbol{x}_n$$

where $\mu_n = f(\boldsymbol{w}^\top \boldsymbol{x}_n)$ and 'f' $(= \psi^{-1})$ depends on type of response y, e.g.,

  - Real-valued y (linear regression): f is identity, i.e., $\mu_n = \boldsymbol{w}^\top \boldsymbol{x}_n$

MLE for GLM

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n) \exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{w}^\top \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

- Convexity of $A(\eta)$ guarantees a global optima. Taking derivative w.r.t. $\boldsymbol{w}$

$$\sum_{n=1}^{N} \left(y_n \boldsymbol{x}_n - A'(\eta_n) \frac{d\eta_n}{d\boldsymbol{w}} \right) = \sum_{n=1}^{N} (y_n \boldsymbol{x}_n - \mu_n \boldsymbol{x}_n) = \sum_{n=1}^{N} (y_n - \mu_n) \boldsymbol{x}_n$$

where $\mu_n = f(\boldsymbol{w}^\top \boldsymbol{x}_n)$ and 'f' $(= \psi^{-1})$ depends on type of response y, e.g.,

  - Real-valued y (linear regression): f is identity, i.e., $\mu_n = \boldsymbol{w}^\top \boldsymbol{x}_n$
  - Binary y (logistic regression): f is logistic function, i.e., $\mu_n = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x}_n)}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x}_n)}$

MLE for GLM

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n) \exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{w}^\top \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

- Convexity of $A(\eta)$ guarantees a global optima. Taking derivative w.r.t. $\boldsymbol{w}$

$$\sum_{n=1}^{N} \left(y_n \boldsymbol{x}_n - A'(\eta_n) \frac{d\eta_n}{d\boldsymbol{w}} \right) = \sum_{n=1}^{N} (y_n \boldsymbol{x}_n - \mu_n \boldsymbol{x}_n) = \sum_{n=1}^{N} (y_n - \mu_n) \boldsymbol{x}_n$$

where $\mu_n = f(\boldsymbol{w}^\top \boldsymbol{x}_n)$ and 'f' $(= \psi^{-1})$ depends on type of response y, e.g.,

  - Real-valued y (linear regression): f is identity, i.e., $\mu_n = \boldsymbol{w}^\top \boldsymbol{x}_n$
  - Binary y (logistic regression): f is logistic function, i.e., $\mu_n = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x}_n)}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x}_n)}$
  - Count-valued y (Poisson regression): $\mu_n = \exp(\boldsymbol{w}^\top \boldsymbol{x}_n)$

MLE for GLM

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n) \exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{w}^\top \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

- Convexity of $A(\eta)$ guarantees a global optima. Taking derivative w.r.t. $\boldsymbol{w}$

$$\sum_{n=1}^{N} \left(y_n \boldsymbol{x}_n - A'(\eta_n) \frac{d\eta_n}{d\boldsymbol{w}} \right) = \sum_{n=1}^{N} (y_n \boldsymbol{x}_n - \mu_n \boldsymbol{x}_n) = \sum_{n=1}^{N} (y_n - \mu_n) \boldsymbol{x}_n$$

where $\mu_n = f(\boldsymbol{w}^\top \boldsymbol{x}_n)$ and 'f' $(= \psi^{-1})$ depends on type of response y, e.g.,

  - Real-valued y (linear regression): f is identity, i.e., $\mu_n = \boldsymbol{w}^\top \boldsymbol{x}_n$
  - Binary y (logistic regression): f is logistic function, i.e., $\mu_n = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x}_n)}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x}_n)}$
  - Count-valued y (Poisson regression): $\mu_n = \exp(\boldsymbol{w}^\top \boldsymbol{x}_n)$
  - Positive reals y (gamma regression): $\mu_n = -(\boldsymbol{w}^\top \boldsymbol{x}_n)^{-1}$

MLE for GLM

- Log likelihood

$$L(\eta) = \log p(Y|\eta) = \log \prod_{n=1}^{N} h(y_n) \exp(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - A(\eta_n)) = \sum_{n=1}^{N} \log h(y_n) + \boldsymbol{w}^\top \sum_{n=1}^{N} y_n \boldsymbol{x}_n - \sum_{n=1}^{N} A(\eta_n)$$

- Convexity of $A(\eta)$ guarantees a global optima. Taking derivative w.r.t. $\boldsymbol{w}$

$$\sum_{n=1}^{N} \left(y_n \boldsymbol{x}_n - A'(\eta_n) \frac{d\eta_n}{d\boldsymbol{w}} \right) = \sum_{n=1}^{N} (y_n \boldsymbol{x}_n - \mu_n \boldsymbol{x}_n) = \sum_{n=1}^{N} (y_n - \mu_n) \boldsymbol{x}_n$$

where $\mu_n = f(\boldsymbol{w}^\top \boldsymbol{x}_n)$ and 'f' $(= \psi^{-1})$ depends on type of response y, e.g.,

  - Real-valued y (linear regression): f is identity, i.e., $\mu_n = \boldsymbol{w}^\top \boldsymbol{x}_n$
  - Binary y (logistic regression): f is logistic function, i.e., $\mu_n = \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x}_n)}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x}_n)}$
  - Count-valued y (Poisson regression): $\mu_n = \exp(\boldsymbol{w}^\top \boldsymbol{x}_n)$
  - Positive reals y (gamma regression): $\mu_n = -(\boldsymbol{w}^\top \boldsymbol{x}_n)^{-1}$

- To estimate $\boldsymbol{w}$ via MLE (or MAP), either set the derivative to zero or use iterative methods (e.g., gradient descent, iteratively reweighted least squares, etc.)

Bayesian Inference for GLM

- If likelihood is conjugate to the prior on $\boldsymbol{w}$, Bayesian inference can be done in closed form

Bayesian Inference for GLM

- If likelihood is conjugate to the prior on $\boldsymbol{w}$, Bayesian inference can be done in closed form
  - Example: Bayesian linear regression with Gaussian likelihood and Gaussian prior on $\boldsymbol{w}$

Bayesian Inference for GLM

- If likelihood is conjugate to the prior on $\boldsymbol{w}$, Bayesian inference can be done in closed form
  - Example: Bayesian linear regression with Gaussian likelihood and Gaussian prior on $\boldsymbol{w}$
- Otherwise, approximate Bayesian inference is needed (e.g., Laplace, MCMC, variational inf, etc.)

Bayesian Inference for GLM

- If likelihood is conjugate to the prior on $\boldsymbol{w}$, Bayesian inference can be done in closed form
  - Example: Bayesian linear regression with Gaussian likelihood and Gaussian prior on $\boldsymbol{w}$
- Otherwise, approximate Bayesian inference is needed (e.g., Laplace, MCMC, variational inf, etc.)
  - Example: Bayesian logistic regression with sigmoid-Bernoulli likelihood and Gaussian prior on $\boldsymbol{w}$

Bayesian Inference for GLM

- If likelihood is conjugate to the prior on $\boldsymbol{w}$, Bayesian inference can be done in closed form
  - Example: Bayesian linear regression with Gaussian likelihood and Gaussian prior on $\boldsymbol{w}$
- Otherwise, approximate Bayesian inference is needed (e.g., Laplace, MCMC, variational inf, etc.)
  - Example: Bayesian logistic regression with sigmoid-Bernoulli likelihood and Gaussian prior on $\boldsymbol{w}$
- Interesting class project idea: Design simple inference algorithms for non-conjugate GLM

Summary

- Exp. family distributions are very useful for modeling diverse types of data/parameters

Summary

- Exp. family distributions are very useful for modeling diverse types of data/parameters
- Conjugate priors to exp. family distributions make parameter updates very simple

Summary

- Exp. family distributions are very useful for modeling diverse types of data/parameters
- Conjugate priors to exp. family distributions make parameter updates very simple
- Other quantities such as posterior predictive can be computed in closed form

Summary

- Exp. family distributions are very useful for modeling diverse types of data/parameters
- Conjugate priors to exp. family distributions make parameter updates very simple
- Other quantities such as posterior predictive can be computed in closed form
- Useful in designing generative classification models.

Summary

- Exp. family distributions are very useful for modeling diverse types of data/parameters
- Conjugate priors to exp. family distributions make parameter updates very simple
- Other quantities such as posterior predictive can be computed in closed form
- Useful in designing generative classification models. Choosing class-conditional from exponential family with conjugate priors helps in parameter estimation

Summary

- Exp. family distributions are very useful for modeling diverse types of data/parameters
- Conjugate priors to exp. family distributions make parameter updates very simple
- Other quantities such as posterior predictive can be computed in closed form
- Useful in designing generative classification models. Choosing class-conditional from exponential family with conjugate priors helps in parameter estimation
- Useful in designing generative models for unsupervised learning

Summary

- Exp. family distributions are very useful for modeling diverse types of data/parameters
- Conjugate priors to exp. family distributions make parameter updates very simple
- Other quantities such as posterior predictive can be computed in closed form
- Useful in designing generative classification models. Choosing class-conditional from exponential family with conjugate priors helps in parameter estimation
- Useful in designing generative models for unsupervised learning
- Also used in designing GLMs

Summary

- Exp. family distributions are very useful for modeling diverse types of data/parameters
- Conjugate priors to exp. family distributions make parameter updates very simple
- Other quantities such as posterior predictive can be computed in closed form
- Useful in designing generative classification models. Choosing class-conditional from exponential family with conjugate priors helps in parameter estimation
- Useful in designing generative models for unsupervised learning
- Also used in designing GLMs
- We will see several use cases when we discuss approximate inference algoritms (e.g., Gibbs sampling, and especially variational inference)