

# Lecture 4

## Conjugate Prior and Exponential Family

Instructor: Shibo Li

[shiboli@cs.fsu.edu](mailto:shiboli@cs.fsu.edu)



Department of Computer Science  
Florida State University

- Bayes' rule review
- Conjugate prior definition
- Commonly used examples
- Exponential family
- Conjugate prior for exponential family

- We have reviewed MLE and MAP, which give point estimations of the parameters
- However, ideally, we want a posterior ***distribution*** of the parameters given the data (rather than points)

- We have reviewed MLE and MAP, which give point estimations of the parameters
- However, ideally, we want a posterior **distribution** of the parameters given the data (rather than points)

Bayes' Rule 
$$p(\boldsymbol{\theta}|\mathbf{D}) = \frac{p(\boldsymbol{\theta}, \mathbf{D})}{p(\mathbf{D})} = \frac{p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- However, in most cases, the posterior distribution cannot be computed following Bayes' Rule (due to the intractable integration)

$$\text{Bayes' Rule} \quad p(\boldsymbol{\theta}|\mathbf{D}) = \frac{p(\boldsymbol{\theta}, \mathbf{D})}{p(\mathbf{D})} = \frac{p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- We now introduce a special case, where we can compute the posterior distribution analytically

$$p(\boldsymbol{\theta}|\mathbf{D}) = \frac{p(\boldsymbol{\theta}, \mathbf{D})}{p(\mathbf{D})} = \frac{p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta})d\boldsymbol{\theta}}$$

$$p(\boldsymbol{\theta}|\mathbf{D}) = \frac{p(\boldsymbol{\theta}, \mathbf{D})}{p(\mathbf{D})} = \frac{p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- Definition: Given a prior distribution and a likelihood function, if the corresponding posterior distribution has the **same functional form** as the prior distribution, we call the prior distribution is **conjugate to** the likelihood.

- Beta prior is conjugate to Bernoulli likelihoods

$$p(\mu|a, b) = \text{Beta}(\mu|a, b) \propto \mu^{a-1}(1-\mu)^{b-1}$$

- Beta prior is conjugate to Bernoulli likelihoods

$$p(\mu|a, b) = \text{Beta}(\mu|a, b) \propto \mu^{a-1}(1-\mu)^{b-1}$$

$$p(x|\mu) = \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \quad \text{Bernoulli}$$

- Beta prior is conjugate to Bernoulli likelihoods

$$p(\mu|a, b) = \text{Beta}(\mu|a, b) \propto \mu^{a-1}(1-\mu)^{b-1}$$

$$p(x|\mu) = \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \quad \text{Bernoulli}$$

$$p(\mu|x) \propto p(\mu|a, b)p(x|\mu) \propto \boxed{\mu^{a+x-1}(1-\mu)^{b+1-x-1}}$$

- Beta prior is conjugate to Bernoulli likelihoods

$$p(\mu|a, b) = \text{Beta}(\mu|a, b) \propto \mu^{a-1}(1-\mu)^{b-1}$$

$$p(x|\mu) = \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \quad \text{Bernoulli}$$

$$p(\mu|x) \propto p(\mu|a, b)p(x|\mu) \propto \boxed{\mu^{a+x-1}(1-\mu)^{b+1-x-1}}$$

$$p(\mu|x) = \text{Beta}(a+x, b+1-x)$$

- Beta prior is conjugate to Bernoulli likelihoods

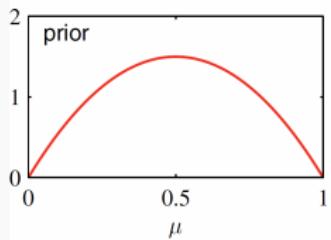
$$p(\mu|a, b) = \text{Beta}(\mu|a, b) \propto \mu^{a-1}(1-\mu)^{b-1}$$

$$p(x|\mu) = \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \quad \text{Bernoulli}$$

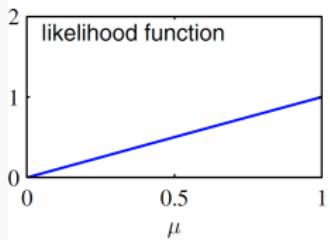
$$p(\mu|x) \propto p(\mu|a, b)p(x|\mu) \propto \boxed{\mu^{a+x-1}(1-\mu)^{b+1-x-1}}$$

$$p(\mu|x) = \text{Beta}(a+x, b+1-x) \quad \text{Note } x \text{ is either 0 or 1}$$

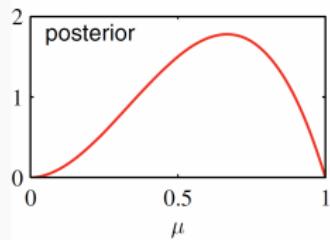
# Prior vs. Posterior



$\text{Beta}(\mu|2, 2)$

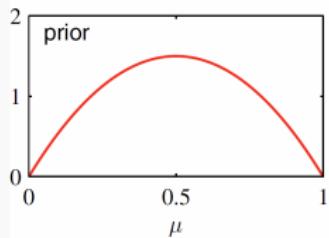


$$p(x = 1|\mu) = \mu^1(1 - \mu)^{1-1} = \mu$$

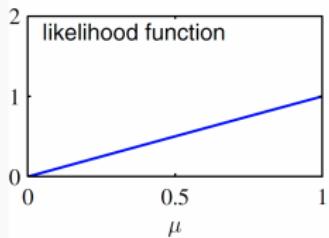


posterior

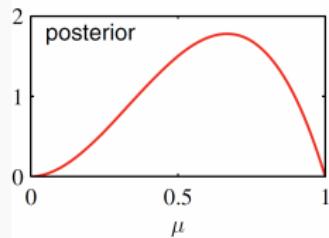
# Prior vs. Posterior



$\text{Beta}(\mu|2, 2)$



$$p(x=1|\mu) = \mu^1(1-\mu)^{1-1} = \mu$$



$\text{Beta}(\mu|3, 2)$

- Dirichlet prior is conjugate to categorical likelihoods

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

- Dirichlet prior is conjugate to categorical likelihoods

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

$$p(\boldsymbol{\mu}|\mathbf{x}) \propto p(\boldsymbol{\mu}|\boldsymbol{\alpha})p(\mathbf{x}|\boldsymbol{\mu}) \propto \prod_{k=1}^K \boxed{\mu_k^{\alpha_k + x_k - 1}}$$

- Dirichlet prior is conjugate to categorical likelihoods

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

$$p(\boldsymbol{\mu}|\mathbf{x}) \propto p(\boldsymbol{\mu}|\boldsymbol{\alpha})p(\mathbf{x}|\boldsymbol{\mu}) \propto \prod_{k=1}^K \boxed{\mu_k^{\alpha_k + x_k - 1}}$$

$$p(\boldsymbol{\mu}|\mathbf{x}) = \text{Dir}(\boldsymbol{\mu}|\mathbf{x} + \boldsymbol{\alpha})$$

- Gamma prior is conjugate to Gaussian likelihood

$$p(\lambda|a, b) = \text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$p(x|\mu, \lambda) = \mathcal{N}(x|\mu, \lambda^{-1}) = \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left(-\frac{\lambda}{2}(x - \mu)^2\right)$$

- Gamma prior is conjugate to Gaussian likelihood

$$p(\lambda|a, b) = \text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$p(x|\mu, \lambda) = \mathcal{N}(x|\mu, \lambda^{-1}) = \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left(-\frac{\lambda}{2}(x - \mu)^2\right)$$

$$p(\lambda|x) \propto p(\lambda|a, b)p(x|\mu, \lambda) \propto \lambda^{1/2+a-1} \exp\left(-\lambda\left(b + \frac{1}{2}(x - \mu)^2\right)\right)$$

- Gamma prior is conjugate to Gaussian likelihood

$$p(\lambda|a, b) = \text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$p(x|\mu, \lambda) = \mathcal{N}(x|\mu, \lambda^{-1}) = \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left(-\frac{\lambda}{2}(x - \mu)^2\right)$$

$$p(\lambda|x) \propto p(\lambda|a, b)p(x|\mu, \lambda) \propto \lambda^{1/2+a-1} \boxed{\exp\left(-\lambda\left(b + \frac{1}{2}(x - \mu)^2\right)\right)}$$

$$p(\lambda|x) = \text{Gam}\left(a + \frac{1}{2}, b + \frac{1}{2}(x - \mu)^2\right)$$

- Wishart prior is conjugate to multivariate Gaussian likelihood

$$p(\boldsymbol{\Lambda}) = \mathcal{W}(\boldsymbol{\Lambda} | \mathbf{W}, \nu) \propto |\boldsymbol{\Lambda}|^{(\nu - d - 1)/2} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)$$

- Wishart prior is conjugate to multivariate Gaussian likelihood

$$p(\boldsymbol{\Lambda}) = \mathcal{W}(\boldsymbol{\Lambda} | \mathbf{W}, \nu) \propto |\boldsymbol{\Lambda}|^{(\nu - d - 1)/2} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)$$

$$p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \propto |\boldsymbol{\Lambda}|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu})\right)$$

- Wishart prior is conjugate to multivariate Gaussian likelihood

$$p(\boldsymbol{\Lambda}) = \mathcal{W}(\boldsymbol{\Lambda} | \mathbf{W}, \nu) \propto |\boldsymbol{\Lambda}|^{(\nu-d-1)/2} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)$$

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \propto |\boldsymbol{\Lambda}|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x}-\boldsymbol{\mu})\right)$$

$$p(\boldsymbol{\Lambda}|\mathbf{x}) \propto |\boldsymbol{\Lambda}|^{(\nu+1-d-1)/2} \exp\left(-\frac{1}{2}\text{tr}\left([\mathbf{W}^{-1} + (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]\boldsymbol{\Lambda}\right)\right)$$

- Wishart prior is conjugate to multivariate Gaussian likelihood

$$p(\boldsymbol{\Lambda}) = \mathcal{W}(\boldsymbol{\Lambda} | \mathbf{W}, \nu) \propto |\boldsymbol{\Lambda}|^{(\nu - d - 1)/2} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)$$

$$p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \propto |\boldsymbol{\Lambda}|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu})\right)$$

$$p(\boldsymbol{\Lambda} | \mathbf{x}) \propto |\boldsymbol{\Lambda}|^{(\nu + 1 - d - 1)/2} \exp\left(-\frac{1}{2}\text{tr}([\mathbf{W}^{-1} + (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] \boldsymbol{\Lambda})\right)$$

$$p(\boldsymbol{\Lambda} | \mathbf{x}) = \mathcal{W}([\mathbf{W}^{-1} + (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top]^{-1}, \nu + 1)$$

- There are criticisms about the conjugate priors. People question that they are constructed just for computational convenience. Are they really appropriate? Are there any more appropriate priors, which, however, are difficult to compute the posterior?

- A family of distributions that play the central role in approximate Bayesian inference

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\right)$$

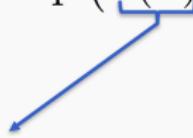
- A family of distributions that play the central role in approximate Bayesian inference

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\right)$$

Any non-negative function that ensures a finite integral over the support;  
Usually simply takes constant 1

- A family of distributions that play the central role in approximate Bayesian inference

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp \left( \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta} \right)$$



**Sufficient statistics:** a (vector) function of  $\mathbf{x}$

- A family of distributions that play the central role in approximate Bayesian inference

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp \left( \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta} \right)$$

Natural parameters: the parameters that determine the distribution

- A family of distributions that play the central role in approximate Bayesian inference

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta})$$



Normalizer/partition function

$$Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}) d\mathbf{x}$$

- A family of distributions that play the central role in approximate Bayesian inference

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta})$$


Normalizer/partition function

$$Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}) d\mathbf{x}$$

This is a function of  $\boldsymbol{\eta}$  !!!

- A family of distributions that play the central role in approximate Bayesian inference

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\eta}) &= \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp \left( \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta} \right) \\ &= h(\mathbf{x}) \exp \left( \mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta} - \log Z(\boldsymbol{\eta}) \right) \end{aligned}$$

Log normalizer/partition function

- Many distributions belong to the exponential family

$$\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$$

- Many distributions belong to the exponential family

$$\begin{aligned}\text{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ &= \exp(x \log \mu + (1-x) \log(1-\mu))\end{aligned}$$

- Many distributions belong to the exponential family

$$\begin{aligned}\text{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ &= \exp(x \log \mu + (1-x) \log(1-\mu)) \\ &= \exp\left(x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right)\end{aligned}$$

- Many distributions belong to the exponential family

$$\begin{aligned}\text{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ &= \exp(x \log \mu + (1-x) \log(1-\mu)) \\ &= \exp\left(x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right) \\ &= (1-\mu) \exp\left(x \log \frac{\mu}{1-\mu}\right)\end{aligned}$$

- Many distributions belong to the exponential family

$$\text{Bern}(x|\mu) = (1 - \mu) \exp\left(x \log \frac{\mu}{1 - \mu}\right)$$

- Many distributions belong to the exponential family

$$\text{Bern}(x|\mu) = (1 - \mu) \exp\left(x \log \frac{\mu}{1 - \mu}\right)$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\right)$$

- Many distributions belong to the exponential family

$$\text{Bern}(x|\mu) = (1 - \mu) \exp\left(x \log \frac{\mu}{1 - \mu}\right)$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\right)$$



$$\begin{aligned} u(x) &= x \\ h(x) &= 1 \end{aligned} \qquad \eta = \ln\left(\frac{\mu}{1 - \mu}\right)$$

- Many distributions belong to the exponential family

$$\text{Bern}(x|\mu) = (1 - \mu) \exp\left(x \log \frac{\mu}{1 - \mu}\right)$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\right)$$

$$\eta = \ln\left(\frac{\mu}{1 - \mu}\right) \quad \xrightarrow{\hspace{1cm}} \quad Z(\eta) = \frac{1}{1 - \mu} = 1 + \exp(\eta)$$

- Multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

- Multivariate Gaussian distribution

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \exp\left(\text{tr}\left(-\frac{1}{2}\mathbf{x}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\right) + \mathbf{x}^\top (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right)\end{aligned}$$

- Multivariate Gaussian distribution

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \exp\left(\text{tr}\left(-\frac{1}{2}\mathbf{x}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\right) + \boxed{\mathbf{x}^\top}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right)\end{aligned}$$

- Multivariate Gaussian distribution

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \exp\left(\text{tr}\left(-\frac{1}{2}\mathbf{x}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\right) + \mathbf{x}^\top (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right)\end{aligned}$$

$$\mathbf{u}(\mathbf{x}) = \begin{pmatrix} -\frac{1}{2}\mathbf{x}\mathbf{x}^\top \\ \mathbf{x} \end{pmatrix} \quad \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \end{pmatrix}$$

Matrix form does not matter, because the trace operator is equivalent to the inner product after vectorization

$$\text{tr}(\mathbf{A}^\top \mathbf{B}) = \text{vec}(\mathbf{A})^\top \text{vec}(\mathbf{B})$$

44

- Multivariate Gaussian distribution

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \exp\left(\text{tr}\left(-\frac{1}{2}\mathbf{x}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\right) + \boxed{\mathbf{x}^\top (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}\right)\end{aligned}$$

$$\mathbf{u}(\mathbf{x}) = \begin{pmatrix} -\frac{1}{2}\mathbf{x}\mathbf{x}^\top \\ \mathbf{x} \end{pmatrix} \quad \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \end{pmatrix}$$

$$h(\mathbf{x}) = 1$$

- Multivariate Gaussian distribution

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \exp\left(\text{tr}\left(-\frac{1}{2}\mathbf{x}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\right) + \mathbf{x}^\top (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right) \\ &\quad \underbrace{\qquad\qquad\qquad}_{Z(\boldsymbol{\eta})} \\ Z(\boldsymbol{\eta}) &= |2\pi\boldsymbol{\Sigma}|^{1/2} \exp\left(\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\end{aligned}$$

- Multivariate Gaussian distribution

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \exp\left(\text{tr}\left(-\frac{1}{2}\mathbf{x}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\right) + \mathbf{x}^\top (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right) \\ Z(\boldsymbol{\eta}) &= |2\pi\boldsymbol{\Sigma}|^{1/2} \exp\left(\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\ &= |2\pi\boldsymbol{\Sigma}|^{1/2} \exp\left(\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\end{aligned}$$

- Multivariate Gaussian distribution

$$\begin{aligned}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \exp\left(\text{tr}\left(-\frac{1}{2}\mathbf{x}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\right) + \mathbf{x}^\top (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right) \\ &\quad \underbrace{\qquad\qquad\qquad}_{Z(\boldsymbol{\eta})} \\ Z(\boldsymbol{\eta}) &= |2\pi\boldsymbol{\Sigma}|^{1/2} \exp\left(\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\ &= |2\pi\boldsymbol{\Sigma}|^{1/2} \exp\left(\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\ &= |2\pi\boldsymbol{\eta}_1^{-1}|^{1/2} \exp\left(\frac{1}{2}\boldsymbol{\eta}_2^\top \boldsymbol{\eta}_1^{-1}\boldsymbol{\eta}_2\right)\end{aligned}$$

$$\boxed{\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \end{pmatrix}} \quad \boldsymbol{\eta}_1 \quad \boldsymbol{\eta}_2$$

AQ

- We can see the mapping between the (commonly used) expectation parameters and the natural parameters.
$$\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \end{pmatrix} \quad \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\Sigma} \end{pmatrix}$$
- There is an insightful connection between the two types of parameters in terms of fisher information matrix
$$\mathbf{I} = \mathbb{E}[-\nabla^2 \log(p(\mathbf{x}|\boldsymbol{\eta}))]$$
- This connection exhibits the essence of mean-field variational inference and stochastic variational inference.

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\right)$$

$$\nabla \log Z(\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} \nabla Z(\boldsymbol{\eta})$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta})$$

$$\nabla \log Z(\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} \nabla Z(\boldsymbol{\eta})$$

$$Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}) d\mathbf{x}$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta})$$

$$\nabla \log Z(\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} \nabla Z(\boldsymbol{\eta})$$

$$Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}) d\mathbf{x}$$

$$\nabla Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}) \mathbf{u}(\mathbf{x}) d\mathbf{x}$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta})$$

$$\nabla \log Z(\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} \nabla Z(\boldsymbol{\eta}) \quad Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}) d\mathbf{x}$$

$$\nabla Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}) \mathbf{u}(\mathbf{x}) d\mathbf{x}$$



$$\nabla \log Z(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

The gradient of the log normalizer  
is the expectation of the sufficient  
statistics!

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta})$$

$$\nabla \log Z(\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} \nabla Z(\boldsymbol{\eta})$$

$$Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}) d\mathbf{x}$$

$$\nabla Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}) \mathbf{u}(\mathbf{x}) d\mathbf{x}$$



$$\nabla \log Z(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

The gradient of the log normalizer  
is the expectation of the sufficient  
statistics!

We will review this property when we study the approximate inference

$$\nabla^2 \log Z(\boldsymbol{\eta}) = \text{cov}(\mathbf{u}(\mathbf{x}))$$

Leave it as your exercise

- Suppose we have  $N$  iid observations  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$   
How to estimate the natural parameters  $\boldsymbol{\eta}$  ?

$$p(\mathcal{D}|\boldsymbol{\eta}) = \prod_{n=1}^N h(\mathbf{x}_n) \frac{1}{Z(\boldsymbol{\eta})^N} \exp\left(\boldsymbol{\eta}^\top \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)\right)$$

- Suppose we have  $N$  iid observations  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$   
How to estimate the natural parameters  $\boldsymbol{\eta}$  ?

$$p(\mathcal{D}|\boldsymbol{\eta}) = \prod_{n=1}^N h(\mathbf{x}_n) \frac{1}{Z(\boldsymbol{\eta})^N} \exp\left(\boldsymbol{\eta}^\top \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)\right)$$

$$\log(p(\mathcal{D}|\boldsymbol{\eta})) = -N \log Z(\boldsymbol{\eta}) + \boldsymbol{\eta}^\top \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)$$

- Suppose we have  $N$  iid observations  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$   
How to estimate the natural parameters  $\boldsymbol{\eta}$  ?

$$p(\mathcal{D}|\boldsymbol{\eta}) = \prod_{n=1}^N h(\mathbf{x}_n) \frac{1}{Z(\boldsymbol{\eta})^N} \exp\left(\boldsymbol{\eta}^\top \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)\right)$$

$$\log(p(\mathcal{D}|\boldsymbol{\eta})) = -N \log Z(\boldsymbol{\eta}) + \boldsymbol{\eta}^\top \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)$$

$$\nabla \log(p(\mathcal{D}|\boldsymbol{\eta})) = \mathbf{0} \quad \xrightarrow{\text{blue arrow}} \quad \begin{aligned} \nabla \log Z(\boldsymbol{\eta}) &= \frac{1}{N} \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) \\ &\parallel \\ \mathbb{E}[\mathbf{u}(\mathbf{x})] \end{aligned}$$

- Suppose we have  $N$  iid observations  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$   
How to estimate the natural parameters  $\boldsymbol{\eta}$  ?

$$p(\mathcal{D}|\boldsymbol{\eta}) = \prod_{n=1}^N h(\mathbf{x}_n) \frac{1}{Z(\boldsymbol{\eta})^N} \exp\left(\boldsymbol{\eta}^\top \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)\right)$$

$$\log(p(\mathcal{D}|\boldsymbol{\eta})) = -N \log Z(\boldsymbol{\eta}) + \boldsymbol{\eta}^\top \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)$$

$$\nabla \log(p(\mathcal{D}|\boldsymbol{\eta})) = \mathbf{0} \quad \longrightarrow \quad \nabla \log Z(\boldsymbol{\eta}) = \frac{1}{N} \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)$$

||  
 $\mathbb{E}[\mathbf{u}(\mathbf{x})]$         
Empirical Mean

$$p(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) \propto \frac{1}{Z(\boldsymbol{\eta})^\nu} \exp(\nu \boldsymbol{\eta}^\top \boldsymbol{\chi})$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta})$$

$$p(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) \propto \frac{1}{Z(\boldsymbol{\eta})^\nu} \exp(\nu \boldsymbol{\eta}^\top \boldsymbol{\chi})$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta}\right)$$

$$p(\boldsymbol{\eta}|\mathbf{x}, \boldsymbol{\chi}, \nu) \propto \frac{1}{Z(\boldsymbol{\eta})^{\nu+1}} \exp\left(\boldsymbol{\eta}^\top (\nu \boldsymbol{\chi} + \mathbf{u}(\mathbf{x}))\right)$$

$$p(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) \propto \frac{1}{Z(\boldsymbol{\eta})^\nu} \exp(\nu \boldsymbol{\eta}^\top \boldsymbol{\chi})$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta})$$

$$p(\boldsymbol{\eta}|\mathbf{x}, \boldsymbol{\chi}, \nu) \propto \frac{1}{Z(\boldsymbol{\eta})^{\nu+1}} \exp(\boldsymbol{\eta}^\top (\nu \boldsymbol{\chi} + \mathbf{u}(\mathbf{x})))$$



$$p(\boldsymbol{\eta} | \frac{\nu}{\nu+1}\boldsymbol{\chi} + \frac{1}{\nu+1}\mathbf{u}(\mathbf{x}), \nu+1)$$

$$p(\boldsymbol{\eta}|\mathbf{x}, \boldsymbol{\chi}, \nu) \propto \frac{1}{Z(\boldsymbol{\eta})^{\nu+1}} \exp(\boldsymbol{\eta}^\top (\nu \boldsymbol{\chi} + \mathbf{u}(\mathbf{x})))$$



$$p(\boldsymbol{\eta} | \underbrace{\frac{\nu}{\nu+1}\boldsymbol{\chi} + \frac{1}{\nu+1}\mathbf{u}(\mathbf{x})}_{\text{weighted sum of the sufficient statistics from prior and observations}}, \nu+1)$$

$\nu$  : pseudo count

weighted sum of the sufficient statistics from prior and observations

- What are conjugate priors?
- What is the motivation of conjugate priors?
- What are the conjugate priors to commonly used likelihoods?
- Definition of the exponential family
- How to turn the existing distribution into the standard form of the exponential family
- Properties of the exponential family
- General conjugate priors to the exponential family