Lecture 4

Conjugate Prior and Exponential Family

Instructor: Shibo Li

shiboli@cs.fsu.edu

Department of Computer Science
Florida State University

Outline

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

Bayes' Rule Revisited

- 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 Revisited

- 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

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

Bayes' Rule Revisited

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

Bayes' Rule

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

Conjugate prior

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

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

Conjugate prior - definition

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

- 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

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

Commonly Used Examples

- Beta prior is conjugate to Bernoulli likelihoods

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

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

Commonly Used Examples

- Beta prior is conjugate to Bernoulli likelihoods

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

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

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

Commonly Used Examples

- Beta prior is conjugate to Bernoulli likelihoods

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

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

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

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

- Beta prior is conjugate to Bernoulli likelihoods

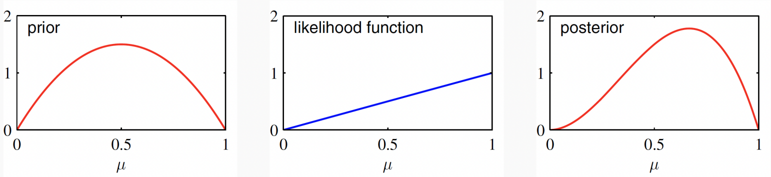
$$p(\mu|a,b) = \mathrm{Beta}(\mu|a,b) \propto \mu^{a-1}(1-\mu)^{b-1}$$

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

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

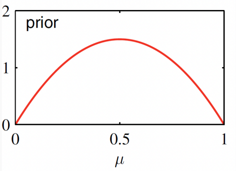
$$p(\mu|x) = \mathrm{Beta}(a+x, b+1-x)$$ Note x is either 0 or 1

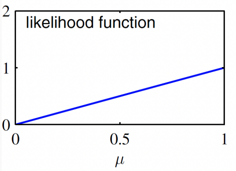
Prior vs. Posterior

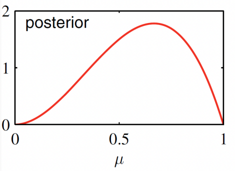


$\mathrm{Beta}(\mu|2,2)$ $\qquad$ $p(x=1|\mu) = \mu^1(1-\mu)^{1-1} = \mu$

Prior vs. Posterior







$\text{Beta}(\mu|2,2)$ $\quad p(x=1|\mu) = \mu^1(1-\mu)^{1-1} = \mu$ $\quad \text{Beta}(\mu|3,2)$

Commonly Used Examples

- Dirichlet prior is conjugate to categorical likelihoods

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

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

Commonly Used Examples

- Dirichlet prior is conjugate to categorical likelihoods

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

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

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

Commonly Used Examples

- Dirichlet prior is conjugate to categorical likelihoods

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

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

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

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

Commonly Used Examples

- Gamma prior is conjugate to Gaussian likelihood

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

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

Commonly Used Examples

- Gamma prior is conjugate to Gaussian likelihood

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

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

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

- Gamma prior is conjugate to Gaussian likelihood

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

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

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

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

Commonly Used Examples

- Wishart prior is conjugate to multivariate Gaussian likelihood

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

- Wishart prior is conjugate to multivariate Gaussian likelihood

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

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

- Wishart prior is conjugate to multivariate Gaussian likelihood

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

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

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

- Wishart prior is conjugate to multivariate Gaussian likelihood

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

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

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

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

Criticisms about conjugate priors

- 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

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

Exponential Family

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

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

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

Exponential Family

- 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 ***x***

Exponential Family

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

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

Natural parameters: the parameters that determine the distribution

Exponential Family

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

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

Normalizer/partition function

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

Exponential Family

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

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

Normalizer/partition function

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

This is a function of η !!!

Exponential Family

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

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

$$= h(\mathbf{x}) \exp\left(\mathbf{u}(\mathbf{x})^\top \boldsymbol{\eta} - \underbrace{\log Z(\boldsymbol{\eta})}\right)$$

Log normalizer/partition function

Exponential Family - Example

- Many distributions belong to the exponential family

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

Exponential Family - Example

- Many distributions belong to the exponential family

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$$
$$= \exp\left(x \log \mu + (1-x) \log(1-\mu)\right)$$

Exponential Family - Example

- Many distributions belong to the exponential family

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

Exponential Family - Example

- Many distributions belong to the exponential family

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

Exponential Family - Example

- Many distributions belong to the exponential family

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

Exponential Family - Example

- Many distributions belong to the exponential family

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

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

Exponential Family - Example

- Many distributions belong to the exponential family

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

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

↓

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

Exponential Family - Example

- Many distributions belong to the exponential family

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

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

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \Longrightarrow Z(\eta) = \frac{1}{1-\mu} = 1 + \exp(\eta)$$

Exponential Family - Example

- Multivariate Gaussian distribution

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

Exponential Family - Example

- Multivariate Gaussian distribution

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

$$= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \exp\big(\mathrm{tr}(-\frac{1}{2}\mathbf{x}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}) + \mathbf{x}^\top(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\big)$$

- Multivariate Gaussian distribution

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

$$= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \exp\big(\mathrm{tr}(\boxed{-\frac{1}{2}\mathbf{x}\mathbf{x}^\top}\boldsymbol{\Sigma}^{-1}) + \boxed{\mathbf{x}^\top}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\big)$$

- Multivariate Gaussian distribution

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

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

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

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

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

44

- Multivariate Gaussian distribution

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

$$= |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \exp\left(\mathrm{tr}(\boxed{-\frac{1}{2}\mathbf{x}\mathbf{x}^\top}\boldsymbol{\Sigma}^{-1}) + \boxed{\mathbf{x}^\top}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right)$$

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

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

- Multivariate Gaussian distribution

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

$$= \underbrace{|2\pi\boldsymbol{\Sigma}|^{-1/2} \exp(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}\exp\Big(\mathrm{tr}(-\frac{1}{2}\mathbf{x}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}) + \mathbf{x}^\top(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\Big)$$

$$Z(\boldsymbol{\eta}) = |2\pi\boldsymbol{\Sigma}|^{1/2} \exp(\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$$

- Multivariate Gaussian distribution

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

$$= \underbrace{|2\pi\boldsymbol{\Sigma}|^{-1/2} \exp(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}\exp\big(\mathrm{tr}(-\frac{1}{2}\mathbf{x}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}) + \mathbf{x}^\top(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\big)$$

$$Z(\boldsymbol{\eta}) = |2\pi\boldsymbol{\Sigma}|^{1/2} \exp(\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$$

$$= |2\pi\boldsymbol{\Sigma}|^{1/2} \exp(\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$$

Exponential Family - Example

- Multivariate Gaussian distribution

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

$$= \underbrace{|2\pi\boldsymbol{\Sigma}|^{-1/2}\exp(-\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}\exp\Big(\mathrm{tr}(-\frac{1}{2}\mathbf{x}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}) + \mathbf{x}^\top(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\Big)$$

$$Z(\boldsymbol{\eta}) = |2\pi\boldsymbol{\Sigma}|^{1/2}\exp(\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$$

$$= |2\pi\boldsymbol{\Sigma}|^{1/2}\exp(\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$$

$$= |2\pi\boldsymbol{\eta}_1^{-1}|^{1/2}\exp(\frac{1}{2}\boldsymbol{\eta}_2^\top\boldsymbol{\eta}_1^{-1}\boldsymbol{\eta}_2)$$

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

48

Exponential Family - Spoiler

- We can see the mapping between the (commonly used) expectation parameters and the natural parameters. $\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \end{pmatrix}$ $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\Sigma} \end{pmatrix}$
- 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.

Exponential Family - Important Property

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

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

Exponential Family - Important Property

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

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

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

Exponential Family - Important Property

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

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

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

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

Exponential Family - Important Property

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

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

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

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

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

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

Exponential Family - Important Property

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

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

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

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

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

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

We will review this property when we study the approximate inference

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

Leave it as your exercise

Exponential Family - MLE

- Suppose we have N iid observations $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

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 $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

How to estimate the natural parameters $\boldsymbol{\eta}$?

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

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

Exponential Family - MLE

- Suppose we have N iid observations $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

How to estimate the natural parameters $\boldsymbol{\eta}$?

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

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

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

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

Exponential Family - MLE

- Suppose we have N iid observations $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

How to estimate the natural parameters $\boldsymbol{\eta}$?

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

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

$$\nabla \log\left(p(\mathcal{D}|\boldsymbol{\eta})\right) = \mathbf{0} \quad \Rightarrow \quad \nabla \log Z(\boldsymbol{\eta}) = \underbrace{\frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)}_{\text{Empirical Mean}}$$

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

Exponential Family - Conjugate Prior

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

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

Exponential Family - Conjugate Prior

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

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

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

Exponential Family - Conjugate Prior

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

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

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

$$\Downarrow$$

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

Exponential Family - Conjugate Prior

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

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

ν : pseudo count

weighted sum of the sufficient statistics from prior and observations

Exponential Family - Conjugate Prior

- 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