Lecture 3

Exponential Families

Prof. Dahua Lin
dhlin@ie.cuhk.edu.hk

Roadmap

1. Basic formulation
2. Minimal and overcomplete representations
3. Mean parameters and gradient map
4. Conjugate Prior

Definition

An **exponential family** $\mathcal{P}$ over a measure space $\mathcal{X}$:

$$p_\theta(\mathbf{x}) = \frac{h(\mathbf{x})}{Z(\boldsymbol{\theta})} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{\phi}(\mathbf{x})\right) = h(\mathbf{x}) \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})\right)$$

- **sufficient statistics**: $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^d$.
- **canonical parameter function**: $\boldsymbol{\eta} : \Theta \to \mathbb{R}^d$.
- **partition function**: $Z : \Theta \to \mathbb{R}$.
- **base density**: h over $\mathcal{X}$.

Partition Function

- The **partition function** is given by:

$$Z(\boldsymbol{\theta}) = \int_{\mathcal{X}} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{\phi}(\mathbf{x})\right) h(\mathbf{x}) \nu(d\mathbf{x})$$

- The **log-partition function** given by $A(\boldsymbol{\theta}) = \log(Z(\boldsymbol{\theta}))$ is often used instead of $Z(\boldsymbol{\theta})$.

Parameter Space

- An exponential family is essentially determined by the *domain* $\mathcal{X}$ and the *sufficient statistics* ϕ.
- The set of valid parameters is $\Theta = \{\boldsymbol{\theta} : Z(\boldsymbol{\theta}) < \infty\}$.
- An exponential family can be parameterized in many ways. When $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, it is said to be in the **canonical form**.

Examples

- Many important families of distributions are exponential families:
    - Binomial distribution
    - Poisson distribution
    - Normal distribution
    - Exponential distribution
    - Beta distribution
    - And many more ......

Bernoulli Distribution

A **Bernoulli distribution** describes an *event* that may or may not happen.

- domain: $\{0, 1\}$
- parameter: $\theta \in (0, 1)$
- pdf:

$$p_\theta(x) = \begin{cases} 1 - \theta & (x = 0) \\ \theta & (x = 1) \end{cases}$$

- sufficient stats: $\phi(x) = x$
- canonical params:

$$\eta(\theta) = \log\left(\frac{\theta}{1 - \theta}\right)$$

- base: $h(x) = 1$ *w.r.t.* counting
- partition function: $Z(\theta) = \frac{1}{1-\theta}$

Poisson Distribution

A **Poisson distribution** characterizes the number of independent events occurring in a certain rate λ within a unit time.

- domain: $\mathbb{N} = \{0, 1, \ldots\}$
- parameter: $\lambda \in \mathbb{R}_+$
- pdf:

$$p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- sufficient stats: $\phi(x) = x$
- canonical params: $\eta(\lambda) = \log(\lambda)$
- base: $h(x) = 1/x!$
- partition function: $Z(\lambda) = e^\lambda$

Exponential Distribution

An **exponential distribution** characterizes the time interval between independent events occurring at a certain rate λ.

- domain: $\mathbb{N}$
- parameter: $\lambda \in \mathbb{R}_+$
- pdf:
$$p_\lambda(x) = \lambda e^{-\lambda x}$$
- sufficient stats: $\phi(x) = x$
- canonical params: $\eta(\lambda) = -\lambda$
- base: $h(x) = 1$
- partition function: $Z(\lambda) = \lambda^{-1}$

Normal Distribution

Normal distributions are the most widely used distributions in probabilistic analysis to describe real-valued variables.

- domain: $\mathbb{R}$
- parameter: $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$
- pdf:

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

- sufficient stats: $\phi(x) = (x, -x^2/2)$
- canonical params:

$$\eta(\mu, \sigma^2) = (\mu/\sigma^2, 1/\sigma^2)$$

- base: $h(x) = 1$
- partition function: $Z(\theta) = \sqrt{2\pi\sigma^2} \exp\left(\mu^2/(2\sigma^2)\right)$

Normal Distribution in Canonical Form

The normal distributions are often parameterized in the **canonical form** in Bayesian analysis.

- Canonical parameters:
  - **potential** coefficient: $h = \mu/\sigma^2$.
  - **precision** coefficient: $J = 1/\sigma^2 > 0$.

- Probability density function:

$$p_{h,J}(x) = \frac{1}{Z(h,J)} \exp\left(-\frac{J}{2}x^2 + hx\right),$$

  with

$$Z(h,J) = \sqrt{2\pi J^{-1}} \exp(h^2/J).$$

- An exponential family over $\mathbb{R}$ with a quadratic exponent is **normal**.

Regular Family

We will focus on exponential families in the *canonical form*:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \exp\left(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})\right).$$

The set of all valid *canonical parameters* is:

$$\Omega(\mathcal{P}) = \left\{\boldsymbol{\theta} \in \mathbb{R}^d : \int_{\mathcal{X}} \exp\left(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})\right) h(d\mathbf{x}) < +\infty\right\}$$

The exponential family $\mathcal{P}$ is called a **regular family**, if $\Omega(\mathcal{P})$ is an *open* subset of $\mathbb{R}^d$. We restrict our attention to *regular families*.

Bernoulli in Canonical Forms

An exponential family $\mathcal{P}$ can be parameterized in different ways. Consider the *Bernoulli distributions* over $\{0, 1\}$:

- **Form-A**

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp(\theta x)$$

  with $Z(\theta) = 1 + e^{\theta}$.

- **Form-B**

$$p(x|\theta_0, \theta_1) = \frac{1}{Z(\theta_0, \theta_1)} \exp(\theta_0(1 - x) + \theta_1 x)$$

  with $Z(\theta_0, \theta_1) = e_0^{\theta} + e_1^{\theta}$.

Minimal and Overcomplete

Consider an exponential family $\mathcal{P}$ parameterized as:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \exp\left(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})\right).$$

- This parameterized form is called an **overcomplete representation** of $\mathcal{P}$, if there exist $\mathbf{a} \in \mathbb{R}^d - \{0\}$ and $b \in \mathbb{R}$ such that

  $$\mathbf{a}^T \boldsymbol{\phi}(\mathbf{x}) = b$$

  holds almost everywhere.
- Otherwise, it is called a **minimal representation**.

Identifiability

Let $\mathcal{P}[\Omega] = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Omega\}$ be a parameterized family:

- $\mathcal{P}[\Omega]$ is called **identifiable** when each distribution in $P \in \mathcal{P}$ corresponds to a unique parameter $\boldsymbol{\theta} \in \Omega$:

$$P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2} \implies \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2.$$

- **Identifiability** indicates whether different parameters can always be distinguishable purely based on observed samples. In other words, if $\mathcal{P}[\Omega]$ is **not identifiable**, then

$$\exists \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Omega : \quad \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \ \& \ P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}.$$

Minimality and Identifiability

If a parameterized exponential family $\mathcal{P}[\Omega]$ is overcomplete, then $\mathcal{P}[\Omega]$ is not identifiable.

Proof:

- There exist $(\mathbf{a}, b)$, such that $\mathbf{a} \neq \mathbf{0}$ and $\mathbf{a}^T \boldsymbol{\phi}(\mathbf{x}) = b$.
- Given $\boldsymbol{\theta} \in \Omega$, then we can show:

$$P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta} + \lambda \mathbf{a}}, \quad \forall \lambda \in \mathbb{R}.$$

Is the converse also true?

- We will answer this later.

Bernoulli Revisit

- **Form-A** with sufficient stats x
  - It is minimal and identifiable.
- **Form-B** with sufficient stats $(1-x, x)$.
  - It is overcomplete, as

$$1 \cdot (1-x) + 1 \cdot x = 1.$$

and not identifiable:

$$P_{(\theta_1, \theta_2)} = P_{(\theta_1+\lambda, \theta_2+\lambda)}, \quad \forall \lambda \in \mathbb{R}.$$

Another Example

Consider the **categorical distribution** parameterized in a canonical form, with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

$$p(x) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{i=1}^{k} \theta_i \delta_i(x)\right),$$

with $x \in \{1, \ldots, k\}$ and $Z(\boldsymbol{\theta}) = \sum_{i=1}^{k} \exp\left(\theta_i\right)$.

- Questions
  - Is it a minimal representation?
  - Is it identifiable?
  - If it is not minimal, how to make it into a minimal representation?

Mean Parameters

- The expectation of sufficient statistics are called **mean parameters**:

$$\boldsymbol{\mu} = E_p[\boldsymbol{\phi}(x)] = \int_{\mathcal{X}} \boldsymbol{\phi}(\mathbf{x}) p(\mathbf{x}) \nu(d\mathbf{x}).$$

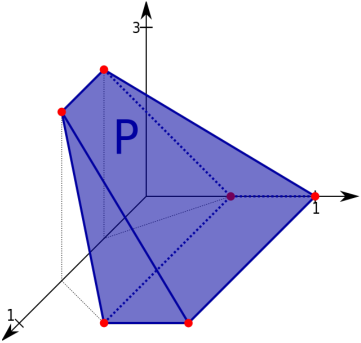
- The *mean parameters* provide an alternative way to parameterize an exponential family.
    - Under *certain* conditions, the distribution in an exponential family is *uniquely* determined by the *mean parameters*.

Realizable Mean Parameters

- Not every vector in $\mathbb{R}^b$ can be a mean parameter.
- Given a sufficient stats ϕ, we say a distribution p **realizes** a *mean parameter* $\boldsymbol{\mu}$ if $E_p[\boldsymbol{\phi}(X)] = \boldsymbol{\mu}$.
- The set of **(realizable) mean parameters** for a given sufficient stats ϕ is:
$$\mathcal{M}_\phi = \left\{ \boldsymbol{\mu} \in \mathbb{R}^d : \exists p \text{ s.t. } E_p[\boldsymbol{\phi}(X)] = \boldsymbol{\mu} \right\}$$
Here, p is **arbitrary**, not restricted to the exponential family.
- $\mathcal{M}_\phi$ is a convex set. Why?

Convex Polytopes

- Given a set $C \subset \mathbb{R}^d$, the **convex hull** of C, denoted by $\mathrm{conv}(C)$, is the set of all *convex combinations* of elements in C.
- $\mathrm{conv}(C)$ is the minimum convex set that contains C.
- A convex hull of some finite set is called a **convex polytope**.
- Convex polytopes are compact.



Probability Simplex

- Given a finite space $\mathcal{X}$, the **probability simplex** over $\mathcal{X}$ is defined as:

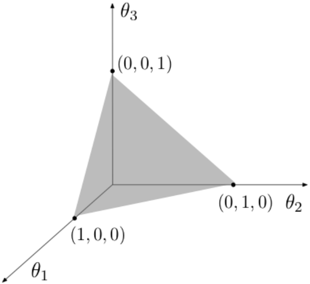
$$\mathcal{S}(\mathcal{X}) \triangleq \left\{ f \in \mathbb{R}_+^{\mathcal{X}} : \sum_{x \in \mathcal{X}} f(x) = 1 \right\}.$$

- When $\mathcal{X} = \{1, \dots, n\}$, $\mathcal{S}(\mathcal{X})$ reduces to:

$$\mathcal{S}_{n-1} \triangleq \mathcal{S}(\mathbb{R}^n) = \left\{ \mathbf{x} \in \mathbb{R}_+^n : \mathbf{1}^T \mathbf{x} = 1 \right\}$$

- $\mathcal{S}_{n-1}$ is an $(n-1)$-dimensional convex polytope:

$$\mathcal{S}_{n-1} = \operatorname{conv}(\mathbf{e}_1, \dots, \mathbf{e}_n)$$



Polytope of Mean Parameters

- When the sample space $\mathcal{X}$ is finite, given any $\phi : \mathcal{X} \to \mathbb{R}^d$, the set $\mathcal{M}_\phi$ is a convex polytope:

$$\mathcal{M}_\phi = \mathrm{conv}\left\{\boldsymbol{\phi}(x) : x \in \mathcal{X}\right\}$$

- Each $\boldsymbol{\mu} \in \mathcal{M}_\phi$ can be written as

$$\mu = \sum_{x \in \mathcal{X}} \alpha(x)\boldsymbol{\phi}(x) \quad \text{with } \alpha \in \mathcal{S}(\mathcal{X})$$

Log-partition Function

- The **log-partition function** given by

$$A(\boldsymbol{\theta}) = \log \int_{\mathcal{X}} \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})) h(d\mathbf{x})$$

  has the following properties:
  - First-order
  $$\nabla A(\boldsymbol{\theta}) = E_{p_\theta}[\phi(X)]$$
  - Second-order
  $$\nabla^2 A(\boldsymbol{\theta}) = \mathrm{Cov}_{p_\theta}[\phi(X)]$$

- $A(\boldsymbol{\theta})$ is a convex function and thus the parameter set $\Omega = \{\boldsymbol{\theta} : A(\boldsymbol{\theta}) < \infty\}$ is a convex set.

Log-partition Function (cont'd)

- For an overcomplete representation, A **is not** strictly convex.
  - **Proof:** We have $\mathbf{a}^T \boldsymbol{\phi}(x) = b$ for some $(\mathbf{a}, b)$, thus

$$\mathrm{Var}_{p_\theta}[\mathbf{a}^T \boldsymbol{\phi}(X)] = \mathbf{a}^T \mathrm{Cov}_{p_\theta}[\boldsymbol{\phi}(X)]\mathbf{a} = 0.$$

  Therefore:

$$\mathbf{a}^T \nabla^2 A(\boldsymbol{\theta})\mathbf{a} = 0.$$

- For a minimal representation, A **is** strictly convex.
  - **Proof:** Given arbitrary $\mathbf{a}$, we have $\mathrm{Var}[\mathbf{a}^T \boldsymbol{\phi}(X)] > 0$, and thus $\mathbf{a}^T \nabla^2 A(\theta)\mathbf{a} > 0$.

Gradient Map

- The **gradient map** defined as

$$\nabla A : \theta \mapsto E_{p_\theta}[\phi(X)]$$

  is a *mapping* from the canonical parameters Ω to the mean parameters $\mathcal{M}$.

- Two questions:
  - When is ∇A injective *(i.e. one-to-one)*?
  - When is ∇A surjective *onto* $\mathcal{M}$?

Gradient Map (cont'd)

- The *gradient map* is injective if and only if the exponential representation is minimal.
- **Proof:**
  - If it is minimal, then A is *strictly convex*, and thus

$$\langle \nabla A(\boldsymbol{\theta}) - \nabla A(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle > 0$$

  - If it is overcomplete, there exists an affine subset of canonical parameters that corresponds to a single distribution, thus the same mean parameter.
- **We now answer a question left earlier:**
  - An exponential family with a minimal representation is identifiable.

Gradient Map (cont'd)

- With a *minimal representation*, ∇A is *onto* $\mathcal{M}^\circ$, the *interior* of $\mathcal{M}$.
  - Each mean parameter $\boldsymbol{\mu} \in \mathcal{M}^\circ$ is uniquely realized by a canonical parameter $\boldsymbol{\theta} \in \Omega$.
- Given $\boldsymbol{\mu} \in \mathcal{M}^\circ$, there can be many distributions that realize $\boldsymbol{\mu}$, among which there is one that maximizes the entropy, which is in the exponential family associated with ϕ (we will see this).

Maximum Entropy Problem

- Given a distribution over $\mathcal{X}$, with density function p *w.r.t.* the base measure μ its **entropy** is defined to be:

$$H(p) \triangleq -\int_{\mathcal{X}} p(\mathbf{x}) \log p(\mathbf{x}) \mu(d\mathbf{x}).$$

- Given a statistic function ϕ and $\boldsymbol{\mu} \in \mathcal{M}_\phi$, the **maximum entropy** problem is defined as:

$$\text{maximize } H(p) \quad \text{s.t.} \quad p \in \mathcal{P}(\mathcal{X}) \text{ and } E_p[\boldsymbol{\phi}(X)] = \boldsymbol{\mu}$$

  Here, $\mathcal{P}(\mathcal{X})$ is the space of all distributions over $\mathcal{X}$.

- Solution?

Optimal Solution to Maximum Entropy

- The optimal solution $\hat{p}$ to the maximum entropy problem is given by

$$\hat{p}(\mathbf{x}) = \frac{1}{Z} \exp\left(\hat{\boldsymbol{\theta}}^T \boldsymbol{\phi}(\mathbf{x})\right) \quad \text{with } E_{\boldsymbol{\theta}}[\boldsymbol{\phi}(X)] = \boldsymbol{\mu}.$$

- When $\mathcal{X}$ is finite, this can be shown using the method of Lagrange multipliers.

- For general $\mathcal{X}$, the proof can be generalized using the tools in functional analysis.

Convex Conjugate

Consider a real-valued function $f : \Omega \to \mathbb{R}$: $\Omega \subset \mathbb{R}^d$:

- The **convex conjugate** of f is defined to be

$$f^*(\mathbf{y}) \triangleq \sup_{\mathbf{x} \in \Omega} \left(\mathbf{y}^T \mathbf{x} - f(\mathbf{x}) \right)$$

- f^* is always *convex* no matter whether f is convex, and thus $\mathrm{dom}(f^*) = \{\mathbf{y} \in \mathbb{R}^d : f^*(\mathbf{y}) < +\infty\}$ is *convex*.

- **Fenchel's inequality**

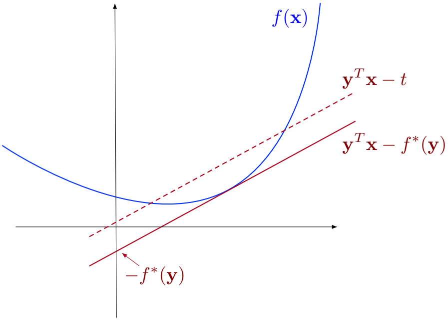
$$f(\mathbf{x}) + f^*(\mathbf{y}) \geq \mathbf{y}^T \mathbf{x}, \quad \forall \mathbf{x} \in \mathrm{dom}(f), \mathbf{y} \in \mathrm{dom}(f^*)$$

Convex Conjugate (cont'd)

- $\forall \mathbf{y} \in \mathrm{dom}(f^*)$, $\mathbf{y}^T\mathbf{x} - f^*(\mathbf{y})$ is a *supporting plane* of $f(\mathbf{x})$.
- For the **biconjugate** f^{**}, $\mathrm{epi}(f^{**}) = \mathrm{conv}(\mathrm{epi}(f))$.
- **(Fenchel-Moreau theorem)** $f^{**} = f$ iff f is convex and lower semi-continuous. Under such conditions:

$$f^*(\mathbf{y}) = \sup_{\mathbf{x}\in\mathrm{dom}(f)} \left(\mathbf{x}^T\mathbf{y} - f(\mathbf{x})\right)$$

$$f(\mathbf{x}) = \sup_{\mathbf{y}\in\mathrm{dom}(f^*)} \left(\mathbf{x}^T\mathbf{y} - f^*(\mathbf{y})\right)$$



Dual Coupling

Given a convex and lower semi-continuous function f and its convex conjugate f^*:

- For each $\mathbf{x} \in \operatorname{dom}(f)$, define

$$\hat{\mathbf{y}}(\mathbf{x}) \triangleq \underset{\mathbf{y}}{\operatorname{argmax}} \left\{ \mathbf{y}^T \mathbf{x} - f^*(\mathbf{y}) \right\}$$

- For each $\mathbf{y} \in \operatorname{dom}(f^*)$, define

$$\hat{\mathbf{x}}(\mathbf{y}) \triangleq \underset{\mathbf{x}}{\operatorname{argmax}} \left\{ \mathbf{y}^T \mathbf{x} - f(\mathbf{x}) \right\}$$

- We have $\hat{\mathbf{x}} \left(\hat{\mathbf{y}}(\mathbf{x}) \right) = \mathbf{x}$. Thus, we call $(\mathbf{x}, \hat{\mathbf{y}}(\mathbf{x}))$ **dually coupled**.

Convex Conjugate of Log-partition

- The *convex conjugate* to a log-partition function A is

$$A^*(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta} \in \Omega} \left(\boldsymbol{\theta}^T \boldsymbol{\mu} - A(\boldsymbol{\theta}) \right)$$

- Supreme attained at $\hat{\boldsymbol{\theta}}$ iff $(\hat{\boldsymbol{\theta}}, \boldsymbol{\mu})$ iff

$$E_{\hat{\boldsymbol{\theta}}}[\phi(X)] = \nabla A(\hat{\boldsymbol{\theta}}) = \boldsymbol{\mu}$$

- Under such condition, $(\hat{\boldsymbol{\theta}}, \boldsymbol{\mu})$ is **dually coupled**.
    - In other words, the canonical parameter $\boldsymbol{\theta}$ is *dually coupled* with the corresponding mean parameter $\boldsymbol{\mu} = \nabla A(\boldsymbol{\theta})$.

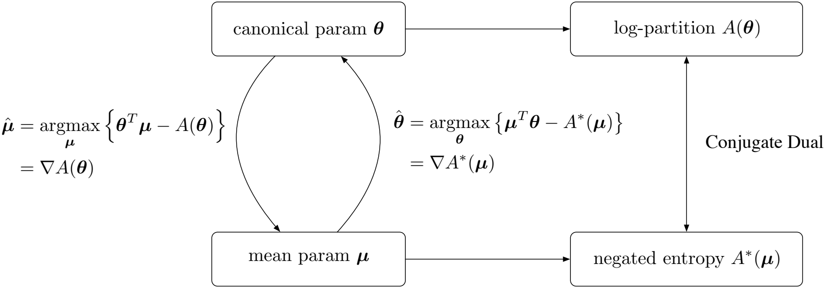
Convex Conjugate of Log-partition (cont'd)

- Then, A^* is actually the negated entropy:

$$A^*(\boldsymbol{\mu}) = \begin{cases} -H\left(p_{\hat{\boldsymbol{\theta}}(\boldsymbol{\mu})}\right) & (\boldsymbol{\mu} \in \mathcal{M}^\circ) \\ +\infty & (\boldsymbol{\mu} \notin \overline{\mathcal{M}}) \end{cases}$$

- With a *minimal representation*, ∇A maps Ω one-to-one onto $\mathcal{M}^\circ$, while ∇A^* is the inverse map.

Summary of the Conjugate Relations



Prior and Posterior

- In *Bayesian analysis*, we usually place a **prior** with density $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$ over the parameter space Ω.
- $\boldsymbol{\theta}$ is linked to observations $\mathcal{D} = \mathbf{x}_{1:n}$ via a **likelihood model**: $f(\mathbf{x}|\boldsymbol{\theta})$.
- The **posterior** conditioned on $\mathcal{D}$ is

$$p(\boldsymbol{\theta}|\mathcal{D};\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha},\mathcal{D})} p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{i=1}^{n} f(\mathbf{x}_i|\boldsymbol{\theta})$$

- Computing the *posterior distribution* is generally very difficult.
  - It requires the integration over the parameter space.
- However, when the prior is *conjugate* to the likelihood model, the computation can be drastically simplified.

Conjugate Prior

- A prior with density $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$ is called a **conjugate prior** to the likelihood model $f(\mathbf{x}|\boldsymbol{\theta})$, if the posterior conditioned on $\mathcal{D} = x_{1:n}$ is in the same parameterized family, *i.e.* in the form

$$p(\boldsymbol{\theta}|\mathcal{D};\boldsymbol{\alpha}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha} \oplus \mathcal{D}).$$

- $\oplus : \Omega \times \mathcal{X} \to \Omega$ is left-associative and satisfies

$$\alpha \oplus \mathbf{x} \oplus \mathbf{y} = \alpha \oplus \mathbf{y} \oplus \mathbf{x}$$

- With $D = \mathbf{x}_{1:n}$,

$$\boldsymbol{\alpha} \oplus \mathcal{D} \triangleq \alpha \oplus \mathbf{x}_1 \oplus \cdots \oplus \mathbf{x}_n$$

  The result is independent of the order of samples.

Conjugate Prior for Exponential Families

- Generally, **conjugate pairs** in *exponential families* are as follows:
  - Prior:

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \beta) = \exp(\boldsymbol{\alpha}^T \boldsymbol{\eta}(\boldsymbol{\theta}) - \beta a(\boldsymbol{\theta}) - A(\boldsymbol{\alpha}, \kappa))$$

  - Likelihood:

$$f(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{\phi}(\mathbf{x}) - \gamma a(\boldsymbol{\theta}))$$

- Given a dataset $\mathcal{D} = \mathbf{x}_{1:n}$, the posterior remains in the same family, with parameters updated to:

$$(\boldsymbol{\alpha}, \beta) \oplus \mathcal{D} = \left(\alpha + \sum_{i=1}^{n} \phi(\mathbf{x}_i),\ \beta + n\gamma\right)$$

CP for Exponential Families (cont'd)

- The family of *conjugate priors* is largely determined by the *likelihood model*, particularly by the form of $\boldsymbol{\eta}(\boldsymbol{\theta})$ and $a(\boldsymbol{\theta})$.
- A family of *prior distributions* can serve as the *conjugate priors* to different *likelihood model*.

Example: Beta-Bernoulli

- Prior: Beta distribution

$$p(\theta|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \text{ with } B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- Likelihood: Bernoulli distribution

$$f(x|\theta) = \theta^x \cdot (1-\theta)^{1-x}, \quad \text{with } x \in \{0, 1\}$$

- Posterior: remains a Beta distribution:

$$\theta|\mathcal{D} \sim \mathrm{Beta}\left(\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n}(1-x_i)\right)$$

Example: Normal-Normal

- Prior: Normal distribution

$$\theta|\mu_0, \sigma_0^2 \sim \mathcal{N}(\mu_0, \sigma_0^2) = \mathcal{N}_c(\sigma_0^{-2}\mu_0,\ \sigma_0^{-2})$$

Here, $\mathcal{N}_c$ denotes the canonical form of normal distribution.

- Likelihood: Normal distribution (fixed variance)

$$x|\theta \sim \mathcal{N}(\theta,\ \sigma^2)$$

- Posterior: remains a Normal distribution

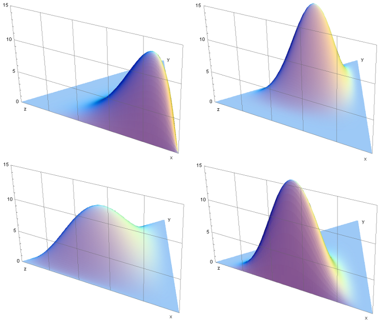
$$\theta|\mathcal{D} \sim \mathcal{N}_c\left(\sigma_0^{-2}\mu_0 + \sigma^{-2}\sum_{i=1}^{n} x_i,\ \sigma_0^{-2} + n\sigma^{-2}\right)$$

Dirichlet Distribution

- **Dirichlet distribution** is a distribution over $\mathcal{S}_{n-1}$.
- It is often used as a *conjugate prior* to *Categorical distributions* or *Multinomial distributions*.
- With $\boldsymbol{\alpha} \in \mathbb{R}^n_{++}$ as the parameter, its density is

$$p_{\boldsymbol{\alpha}}(x) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{n} x_i^{\alpha_i - 1}$$

with $B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)}$



Dirichlet Distribution (cont'd)

- Mean: $E[X_i] = \frac{\alpha_i}{\alpha_0}$ with $\alpha_0 = \alpha_1 + \ldots + \alpha_n$.
- Covariance:

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0+1)} & (i = j) \\ \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)} & (i \neq j) \end{cases}$$

- Mode:

$$\left(\frac{\alpha_i - 1}{\alpha_0 - n} \right)_{1:n}$$

- Marginal:

$$X_i \sim \mathrm{Beta}(\alpha_i, \alpha_0 - \alpha_i)$$

Dirichlet Distribution (cont'd)

- Dirichlet distributions are an *exponential family*:
  - Canonical parameter: $\boldsymbol{\eta}(\boldsymbol{\alpha}) = (\alpha_i - 1)_{1:n}$
  - Sufficient stats: $\boldsymbol{\phi}(\mathbf{x}) = (\log(x_i))_{1:n}$
  - Log-partition function:

$$\log B(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \log \Gamma(\alpha_i) - \log \Gamma(\alpha_0)$$

  - Hence,

$$E_{\boldsymbol{\alpha}}[\log(X_i)] = \frac{\partial \log B(\boldsymbol{\alpha})}{\partial \alpha_i} = \psi(\alpha_i) - \psi(\alpha_0)$$

    Here, ψ is the *digamma function*. **Note:** This equation is very important in deriving the inference algorithm for Latent Dirichlet Allocation (LDA).

Predictive Distribution

- Given $\mathcal{D} = \mathbf{x}_{1:n}$, the **predictive distribution** of a new sample $\mathbf{x}$:

$$p(\mathbf{x}|\mathcal{D}) = \int_{\Omega} f(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \beta) \nu(d\boldsymbol{\theta})$$

- With *exponential family* and *conjugate prior*, we have

$$p(\mathbf{x}|\mathcal{D}) = h(\mathbf{x}) \exp\left(A\left(\boldsymbol{\alpha} + \boldsymbol{\phi}(\mathbf{x}), \beta + \gamma\right) - A(\boldsymbol{\alpha}, \beta)\right)$$

Prove this as an exercise.

Common Conjugate Priors

Prior	**Likelihood parameter**
Beta	the *probability parameter* of *Bernoulli*, *Binomial*, *Geometric* or *Negative Binomial*
Normal	the *mean parameter* of *Normal*
InverseGamma	the *variance parameter* of *Normal*
Gamma	the *rate parameter* of *Exponential* or *Poisson*, or the *precision parameter* of *Normal*

Common Conjugate Priors (cont'd)

Prior	Likelihood parameter
Beta Dirichlet	the *probability vector* of *Categorical* or *Multinomial*
Multivariate Normal	the *mean vector* of *Multivariate Normal*
InverseWishart	the *covariance matrix* of *Multivariate Normal*
Wishart	the *precision matrix* of *Multivariate Normal*