Computer Science

CSC535: Probabilistic Graphical Models

The Exponential Family

Prof. Jason Pacheco

Some material from: Prof. Erik Sudderth

Outline

- Definition & Examples
- Conjugate Prior
- Parameters & Properties

Outline

- Definition & Examples
- Conjugate Prior
- Parameters & Properties

The Exponential Family

- Class of parametric distributions with PMF/PDF characterized by:
  - *Parameters*
  - *Sufficient statistics* of the random variable (RV)
  - Other functions of the RV and parameters for normalization

- Includes many well-known discrete and continuous distributions:
  - Gaussian
  - Bernoulli
  - Binomial
  - Multinomial
  - Beta
  - Gamma
  - Poisson
  - many many more…

The Exponential Family

Definition Let X be a RV with *sufficient statistics* $\phi(x) \in \mathbb{R}^d$. An exponential family distribution with *natural parameters* $\eta \in \mathbb{R}^d$ has PMF/PDF,

$$p(x) = h(x) \exp\left\{\eta^T \phi(x) - A(\eta)\right\}$$

With *base measure* $h(x)$ and *log-partition function:*

$$A(\eta) = \log \int \exp\left\{\eta^T \phi(x)\right\} h(x) dx$$

Why the Exponential Family?

$$p(x) = h(x) \exp\left\{\eta^T \phi(x) - A(\eta)\right\}$$

$$A(\eta) = \log \int \exp\left\{\eta^T \phi(x)\right\} h(x) dx$$

$\phi(x) \in \mathbb{R}^d$ $\longrightarrow$ vector of *sufficient statistics* (features) defining the family

$\eta \subseteq \mathbb{R}^d$ $\longrightarrow$ vector of *natural parameters* indexing particular distributions

- Includes many popular probability distributions: *Bernoulli (binary), Categorical, Poisson (counts), Exponential (positive), Gaussian (real), …*
- Maximum likelihood (ML) learning is simple: *moment matching of sufficient statistics*
- Bayesian learning is simple: *conjugate priors are available*
- The *maximum entropy* interpretation: Among all distributions with certain moments of interest, the exponential family is the most random (makes fewest assumptions)

Example: Gaussian

$$p(x) = \mathcal{N}(x \mid m, \sigma^2)$$

Example: Gaussian

$$p(x) = \mathcal{N}(x \mid m, \sigma^2)$$

Normal PDF

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

Example: Gaussian

$$p(x) = \mathcal{N}(x \mid m, \sigma^2)$$

Normal PDF $$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2}(x-m)^2\sigma^{-2} \right\}$$

Move σ^{-1} inside $$= \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(x-m)^2\sigma^{-2} - \frac{1}{2}\log\sigma^2 \right\}$$

Example: Gaussian

$$p(x) = \mathcal{N}(x \mid m, \sigma^2)$$

Normal PDF $$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}(x-m)^2\sigma^{-2}\right\}$$

Move σ^{-1} inside $$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-m)^2\sigma^{-2} - \frac{1}{2}\log\sigma^2\right\}$$

Expand quadratic $$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\sigma^{-2} + x\sigma^{-2}m - \frac{1}{2}m^2\sigma^{-2} - \frac{1}{2}\log\sigma^2\right\}$$

Example: Gaussian

$$p(x) = \mathcal{N}(x \mid m, \sigma^2)$$

Normal PDF $$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}(x-m)^2\sigma^{-2}\right\}$$

Move σ^{-1} inside $$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-m)^2\sigma^{-2} - \frac{1}{2}\log\sigma^2\right\}$$

Expand quadratic $$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\sigma^{-2} + x\sigma^{-2}m - \frac{1}{2}m^2\sigma^{-2} - \frac{1}{2}\log\sigma^2\right\}$$

Vectorize $$= \frac{1}{\sqrt{2\pi}} \exp\left\{\begin{pmatrix} \sigma^{-2}m \\ -\frac{1}{2}\sigma^{-2} \end{pmatrix}^T \begin{pmatrix} x \\ x^2 \end{pmatrix} - \frac{1}{2}m^2\sigma^{-2} - \frac{1}{2}\log\sigma^2\right\}$$

Example: Gaussian

$$p(x) = \mathcal{N}(x \mid m, \sigma^2)$$

Normal PDF $$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}(x-m)^2\sigma^{-2}\right\}$$

Move σ^{-1} inside $$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-m)^2\sigma^{-2} - \frac{1}{2}\log\sigma^2\right\}$$

Expand quadratic $$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\sigma^{-2} + x\sigma^{-2}m - \frac{1}{2}m^2\sigma^{-2} - \frac{1}{2}\log\sigma^2\right\}$$

Vectorize $$= \frac{1}{\sqrt{2\pi}} \exp\left\{\begin{pmatrix} \sigma^{-2}m \\ -\frac{1}{2}\sigma^{-2} \end{pmatrix}^T \begin{pmatrix} x \\ x^2 \end{pmatrix} - \frac{1}{2}m^2\sigma^{-2} - \frac{1}{2}\log\sigma^2\right\}$$

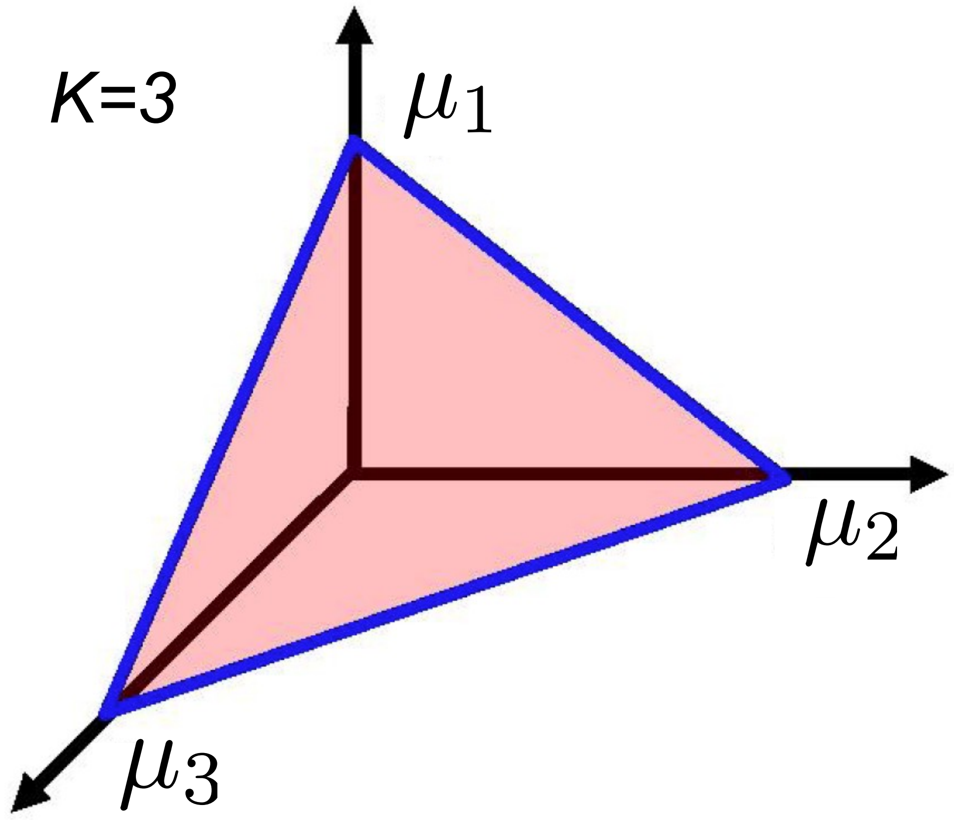
$$h(x) = \frac{1}{\sqrt{2\pi}}, \quad \eta = \begin{pmatrix} \sigma^{-2}m \\ -\frac{1}{2}\sigma^{-2} \end{pmatrix}, \quad \phi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}, \quad A(\eta) = \frac{1}{2}m^2\sigma^{-2} + \frac{1}{2}\log\sigma^2$$

Example: Categorical Distribution

Categorical Distribution: Single roll of a (possibly biased) die

$$\mathrm{Cat}(x \mid \mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$\mathcal{X} = \{0,1\}^K, \sum_{k=1}^{K} x_k = 1$$



$$0 \leq \mu_k \leq 1 \qquad \sum_{k=1}^{K} \mu_k = 1$$

$$\mu_k = \mathbb{E}[x_k]$$

Example: Categorical Distribution

Categorical Distribution: Single roll of a (possibly biased) die

$$\mathrm{Cat}(x \mid \mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

Mapping for normalized parameters:

$$\eta_k = \log \mu_k$$

Exponential Family Form:

$$\mathrm{Cat}(x \mid \eta) = \exp\left\{\sum_{k=1}^{K} \eta_k x_k - A(\eta)\right\}$$

$$A(\eta) = \log\left(\sum_{\ell=1}^{K} \exp(\eta_\ell)\right)$$

Exponential family form is not unique

- A linear subspace of exponential family parameters gives the same probabilities, because the features are linearly dependent: $\sum_k x_k = 1$

$$\mathrm{Cat}(x \mid \eta) = \mathrm{Cat}(x \mid \eta + c)$$ For any scalar constant c

Example: Bernoulli Distribution

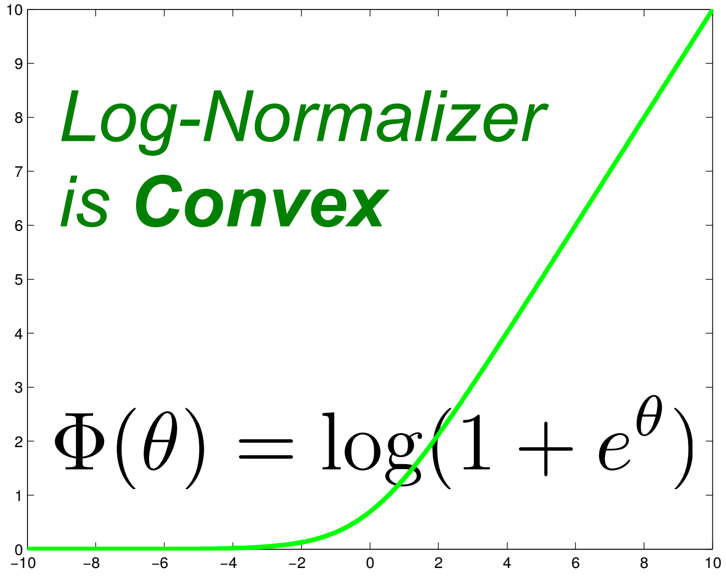
Bernoulli Distribution: Single toss of a (possibly biased) coin

$$\mathrm{Ber}(x \mid \mu) = \mu^x (1-\mu)^{1-x} \qquad x \in \{0,1\}$$

$$\mathbb{E}[x \mid \mu] = \mathbb{P}[x=1] = \mu \qquad 0 \le \mu \le 1$$

Exponential Family Form: Derivation on board

$$\mathrm{Ber}(x \mid \theta) = \exp\{\theta x - \Phi(\theta)\} \qquad \theta \in \Theta = \mathbb{R}$$

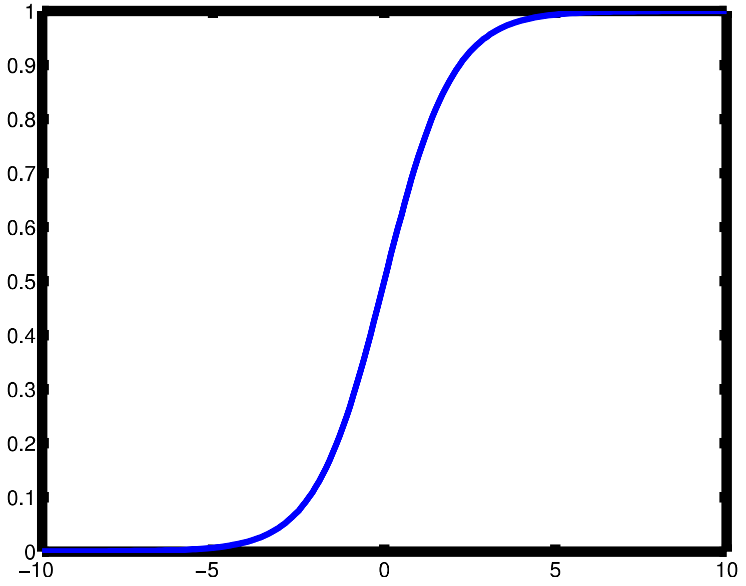


Logistic Function:

$$\mu = (1 + \exp(-\theta))^{-1}$$

Logit Function:

$$\theta = \log(\mu) - \log(1-\mu)$$



Minimal Exponential Families

In a *minimal* exponential family representation, the features must be linearly independent. **Example:**

$$\mathrm{Ber}(x \mid \theta) = \exp\{\theta x - \Phi(\theta)\}$$

In *overcomplete* exponential family representation, features and/or sufficient statistics are linearly dependent and multiple parameters give same distribution. **Example:**

$$\mathrm{Ber}(x \mid \theta) = \exp\{\theta_1 x + \theta_2(1 - x) - \Phi(\theta_1, \theta_2)\}$$

Outline

- Definition & Examples
- Conjugate Prior
- Parameters & Properties

Conjugate Prior

- Given latent variable θ and data x we are often interested in the posterior distribution $p(\theta \mid x)$

- The property of conjugacy ensures that our posterior distribution takes a closed-form

Definition We say that prior $p(\theta)$ is conjugate to likelihood $q(x \mid \theta)$ if and only if the posterior $p(\theta \mid x)$ belongs to the *same functional family* as the prior distribution.

Remark If the above holds, then we also refer to $p(\theta)$ and $q(x \mid \theta)$ as a *conjugate pair*.

Exponential Family Conjugacy

Theorem All likelihoods $q(x \mid \theta)$ in the exponential family have a conjugate prior $p(\theta)$, which is an exponential family (possibly different)

Proof Let $\{x_i\}_{i=1}^N$ be iid from an expfam likelihood,

$$q(x_i \mid \theta) = h(x_i) \exp\left\{\theta^T \phi(x_i) - A(\theta)\right\}$$

Let θ have expfam prior with parameters $\eta = (\eta_1^T, \eta_2 \in \mathbb{R})^T$ and,

$$p(\theta \mid \eta) = g(\theta) \exp\left\{\eta_1^T \theta - \eta_2 A(\theta) - B(\eta)\right\}$$

with log-partition $B(\eta)$ and sufficient statistics vector $\phi(\theta) = (\theta^T, A(\theta))^T$

Then…

Exponential Family Conjugacy

$$p(\theta \mid x_{1:N}, \eta) \propto p(\theta \mid \eta) \prod_{i=1}^{N} q(x_i \mid \theta)$$

Exponential Family Conjugacy

$$p(\theta \mid x_{1:N}, \eta) \propto p(\theta \mid \eta) \prod_{i=1}^{N} q(x_i \mid \theta)$$

Def'n of p & q

$$= g(\theta) \exp\left\{\eta_1^T \theta - \eta_2 A(\theta) - B(\eta)\right\} \prod_{i=1}^{N} h(x_i) \exp\left\{\theta^T \phi(x_i) - A(\theta)\right\}$$

Exponential Family Conjugacy

$$p(\theta \mid x_{1:N}, \eta) \propto p(\theta \mid \eta) \prod_{i=1}^{N} q(x_i \mid \theta)$$

Def'n of p & q

$$= g(\theta) \exp\left\{\eta_1^T \theta - \eta_2 A(\theta) - B(\eta)\right\} \prod_{i=1}^{N} h(x_i) \exp\left\{\theta^T \phi(x_i) - A(\theta)\right\}$$

Collect terms

$$\propto g(\theta) \exp\left\{\theta^T \left(\eta_1 + \sum_{i=1}^{N} \phi(x_i)\right) - (\eta_2 + N) A(\theta)\right\}$$

Exponential Family Conjugacy

$$p(\theta \mid x_{1:N}, \eta) \propto p(\theta \mid \eta) \prod_{i=1}^{N} q(x_i \mid \theta)$$

Def'n of p & q

$$= g(\theta) \exp\left\{\eta_1^T \theta - \eta_2 A(\theta) - B(\eta)\right\} \prod_{i=1}^{N} h(x_i) \exp\left\{\theta^T \phi(x_i) - A(\theta)\right\}$$

Collect terms

$$\propto g(\theta) \exp\left\{\theta^T \left(\eta_1 + \sum_{i=1}^{N} \phi(x_i)\right) - (\eta_2 + N) A(\theta)\right\}$$

Def'n of p

$$\propto p(\theta \mid \widetilde{\eta})$$

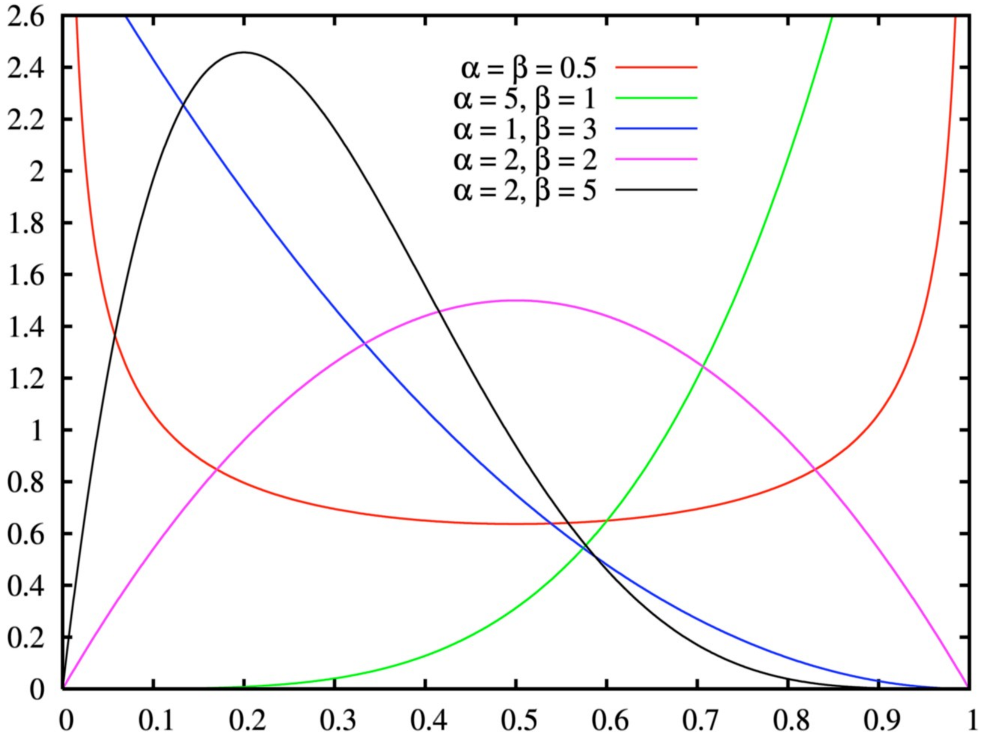
Where posterior parameters are: $\widetilde{\eta} = (\eta_1^T + \sum_{i=1}^{N} \phi(x_i)^T, \eta_2 + N)^T$

Example: Beta-Bernoulli

Bernoulli *A.k.a. the **coinflip** distribution on <u>binary</u> RVs* $X \in \{0, 1\}$

$$\text{Bernoulli}(X \mid \theta) = \theta^X (1-\theta)^{(1-X)}$$

Beta *distribution on* $\theta \in (0, 1)$ *with* $\alpha, \beta > 0$ *has PDF,*



$$\text{Beta}(\theta \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

For N coinflips $x_1, \ldots, x_N$ the posterior is,

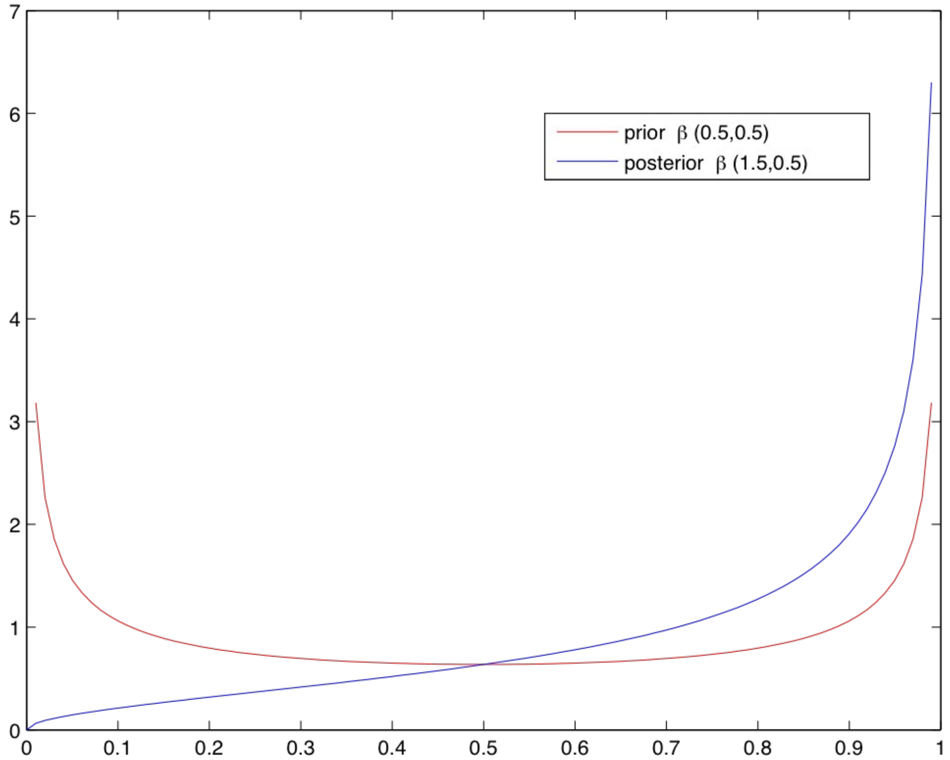
$$\text{Beta}(\theta \mid \alpha + \sum_i x_i, \beta + N - \sum_i x_i)$$

Example: Beta-Bernoulli

After a single coinflip of heads (x=1) the posterior is…

$$p(\theta \mid X = 1, \alpha, \beta) = \mathrm{Beta}(\theta \mid \widetilde{\alpha}, \widetilde{\beta})$$

$$\widetilde{\alpha} = \alpha + x \qquad \widetilde{\beta} = \beta + 1 - x$$



The prior (red) is a fair coin,

$$\mathrm{Beta}(\theta \mid \alpha = 0.5, \beta = 0.5)$$

After observing one flip, the posterior (blue) concentrates on heads,

$$\mathrm{Beta}(\theta \mid \widetilde{\alpha} = 1.5, \widetilde{\beta} = 0.5)$$

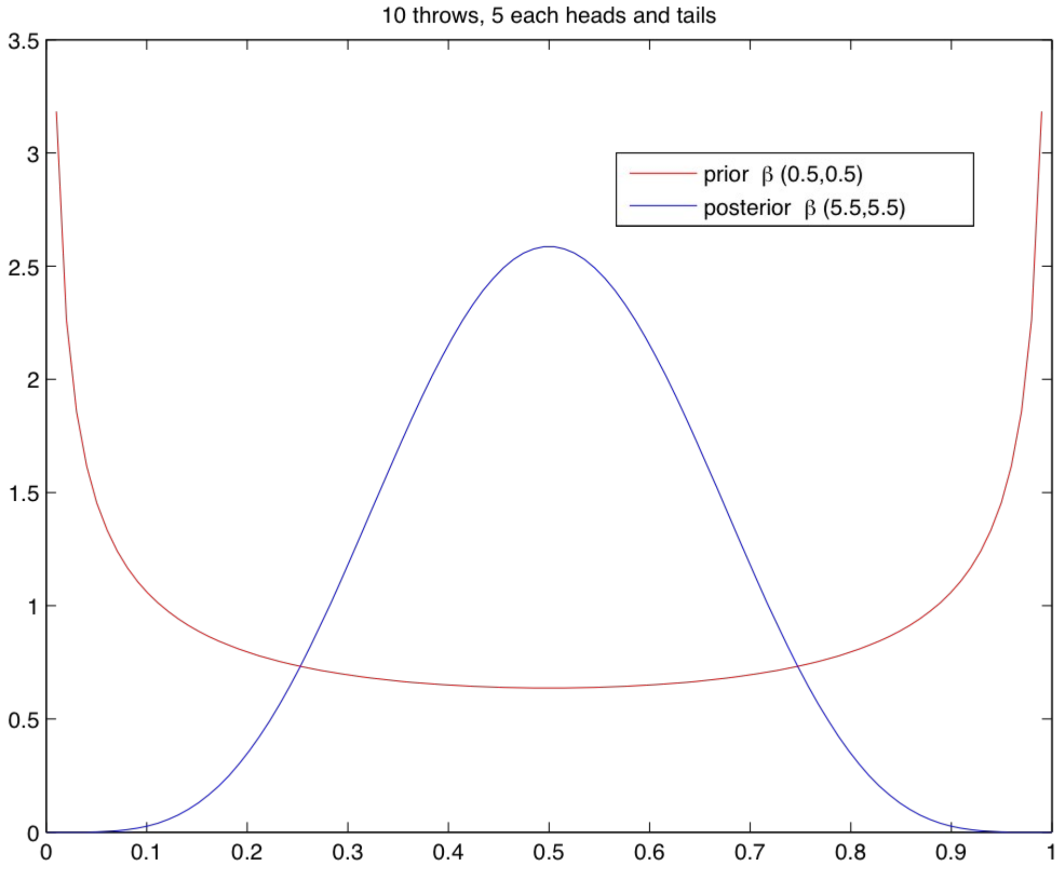
What do you expect if we flip N=10 times with 5 heads and 5 tails?

Example: Beta-Bernoulli

After a N=10 flips (5 heads, 5 tails) we have…

$$p(\theta \mid X = 1, \alpha = 0.5, \beta = 0.5) = \mathrm{Beta}(\theta \mid \widetilde{\alpha}, \widetilde{\beta})$$

$$\widetilde{\alpha} = 0.5 + 5 = 5.5 \qquad \widetilde{\beta} = 0.5 + 10 - 5 = 5.5$$



Posterior concentrates on fair coin $\theta = 0.5$

Other Conjugate Pairs

Likelihood	Model Parameters	Conjugate Prior
Normal	Mean	Normal
Normal	Mean / Variance	Normal-Inv-Gamma
Multivariate Normal	Mean / Variance	Normal-Inv-Wishart
Multinomial	Probability vector	Dirichlet
Gamma	Rate	Gamma
Poisson	Rate	Gamma
Exponential	Rate	Gamma

Wikipedia has a nice list of standard conjugate forms…

https://en.wikipedia.org/wiki/Conjugate_prior

Outline

- Definition & Examples
- Conjugate Prior
- Parameters & Properties

Mean Parameters

We use *natural parameters* η in the exponential family canonical form,

$$p_\eta(x) = h(x) \exp\left\{\eta^T \phi(x) - A(\eta)\right\}$$

Alternate set of *mean parameters* given by expected sufficient stats,

$$\mu_i = \mathbb{E}_{p_\eta}[\phi_i(x)]$$

If family is <u>minimal</u> then there is an invertible mapping between mean/natural parameters

Example Gaussian $\mathcal{N}(x \mid m, \sigma^2)$ with sufficient stats $\phi(x) = (x, x^2)^T$,

$$\mu = \begin{pmatrix} \mathbb{E}[x] \\ \mathbb{E}[x^2] \end{pmatrix} = \begin{pmatrix} m \\ \sigma^2 + m^2 \end{pmatrix} \Leftrightarrow \eta(\mu) = \begin{pmatrix} \sigma^{-2} m \\ -\frac{1}{2}\sigma^{-2} \end{pmatrix}$$

Log-Partition Function

Derivatives of the log-partition (w.r.t. η) yield moments of sufficient stats

$$\mu_i = \mathbb{E}_{p_\eta}[\phi_i(x)] = \frac{\partial}{\partial \eta_i} A(\eta) \qquad \mathrm{Var}_{p_\eta}[\phi_i(x)] = \frac{\partial^2}{\partial^2 \eta_i^2} A(\eta)$$

Example Gaussian $\mathcal{N}(x \mid m, \sigma^2)$ with sufficient stats $\phi(x) = (x, x^2)^T$,

$$A(\eta) = -\tfrac{\eta_1^2}{4\eta_2} - \tfrac{1}{2}\log(-2\eta_2) \qquad \eta = \begin{pmatrix} m\sigma^{-2} \\ -\tfrac{1}{2}\sigma^{-2} \end{pmatrix}$$

$$\frac{\partial}{\partial \eta_1} A(\eta) = -\frac{1}{2}\frac{\eta_1}{\eta_2}$$

Log-Partition Function

Derivatives of the log-partition (w.r.t. η) yield moments of sufficient stats

$$\mu_i = \mathbb{E}_{p_\eta}[\phi_i(x)] = \frac{\partial}{\partial \eta_i} A(\eta) \qquad \mathrm{Var}_{p_\eta}[\phi_i(x)] = \frac{\partial^2}{\partial^2 \eta_i^2} A(\eta)$$

Example Gaussian $\mathcal{N}(x \mid m, \sigma^2)$ with sufficient stats $\phi(x) = (x, x^2)^T$,

$$A(\eta) = -\tfrac{\eta_1^2}{4\eta_2} - \tfrac{1}{2}\log(-2\eta_2) \qquad \eta = \begin{pmatrix} m\sigma^{-2} \\ -\tfrac{1}{2}\sigma^{-2} \end{pmatrix}$$

$$\frac{\partial}{\partial \eta_1} A(\eta) = -\frac{1}{2}\frac{\eta_1}{\eta_2} = m$$

Log-Partition Function

Derivatives of the log-partition (w.r.t. η) yield moments of sufficient stats

$$\mu_i = \mathbb{E}_{p_\eta}[\phi_i(x)] = \frac{\partial}{\partial \eta_i} A(\eta) \qquad \mathrm{Var}_{p_\eta}[\phi_i(x)] = \frac{\partial^2}{\partial^2 \eta_i^2} A(\eta)$$

Example Gaussian $\mathcal{N}(x \mid m, \sigma^2)$ with sufficient stats $\phi(x) = (x, x^2)^T$,

$$A(\eta) = -\tfrac{\eta_1^2}{4\eta_2} - \tfrac{1}{2}\log(-2\eta_2) \qquad \eta = \begin{pmatrix} m\sigma^{-2} \\ -\frac{1}{2}\sigma^{-2} \end{pmatrix}$$

$$\frac{\partial}{\partial \eta_1} A(\eta) = -\frac{1}{2}\frac{\eta_1}{\eta_2} = m = \mathbb{E}[\phi_1(x) = x]$$

Maximum Likelihood Estimation for Exponential Families

Theorem $A(\eta)$ is a **convex** function of the natural parameters η

Proof The second derivative is a positive semidefinite covariance matrix

$$\nabla^2_\eta A(\eta) = \mathrm{Cov}(\phi(x)) \succeq 0$$

Important consequences for learning with exponential families:

- Finding *gradients* is equivalent to finding expected sufficient statistics, or moments, of some current model. This *is an inference problem!*
- Convexity of log-partition implies *parameter space is convex*
- Learning is a convex problem: *No local optima!*
  At least when we have complete observations…

Maximum Likelihood Estimation for Exponential Families

Log-likelihood of observation x_i is given by,

$$\log p(x_i \mid \eta) = \log h(x_i) + \eta^T \phi(x_i) - A(\eta)$$

Given N iid observations, the *log-likelihood function* equals:

$$\mathcal{L}(\eta) = \left[\sum_{i=1}^{N} \eta^T \phi(x_i)\right] - NA(\eta) + \text{const.}$$

At unique global optimum, the zero-gradient gives:

$$\nabla_\eta \mathcal{L}(\eta) = \nabla_\eta \left[\sum_{i=1}^{N} \eta^T \phi(x_i)\right] - N\nabla A(\eta)$$

Maximum Likelihood Estimation for Exponential Families

Log-likelihood of observation x_i is given by,

$$\log p(x_i \mid \eta) = \log h(x_i) + \eta^T \phi(x_i) - A(\eta)$$

Given N iid observations, the *log-likelihood function* equals:

$$\mathcal{L}(\eta) = \left[\sum_{i=1}^{N} \eta^T \phi(x_i)\right] - NA(\eta) + \text{const.}$$

At unique global optimum, the zero-gradient gives:

$$\nabla_\eta \mathcal{L}(\eta) = \nabla_\eta \left[\sum_{i=1}^{N} \eta^T \phi(x_i)\right] - N\nabla A(\eta) = \left[\sum_{i=1}^{N} \phi(x_i)\right] - N\mathbb{E}_{p_\eta}[\phi(x)]$$

$$\mathbf{E}_{p_\eta}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N} \phi(x_i)$$

Moment matching conditions

Example: Bernoulli Distribution

Bernoulli Distribution: Single toss of a (possibly biased) coin

$$\mathrm{Ber}(x \mid \mu) = \mu^x (1-\mu)^{1-x} \qquad x \in \{0,1\}$$

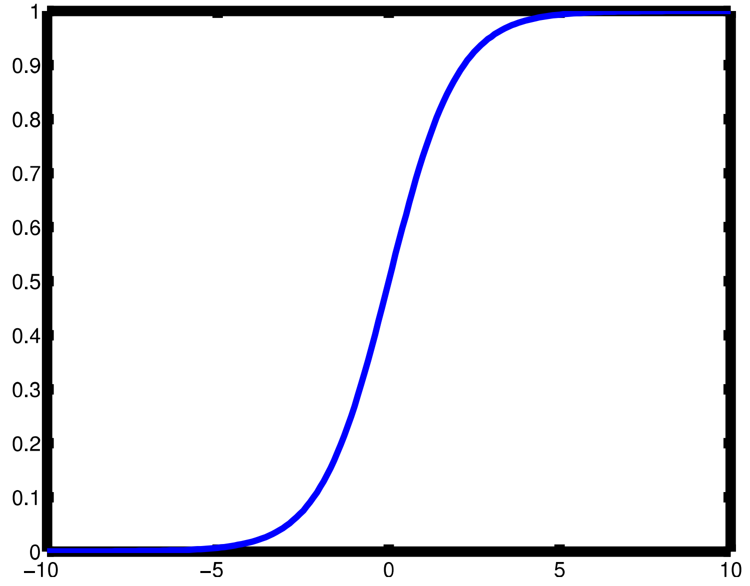
$$\mathbb{E}[x \mid \mu] = \mathbb{P}[x=1] = \mu \qquad 0 \le \mu \le 1$$

Exponential Family Form:

$$\mathrm{Ber}(x \mid \theta) = \exp\{\theta x - \Phi(\theta)\}$$

$$\mu = (1+\exp(-\theta))^{-1}$$

$$\theta = \log(\mu) - \log(1-\mu)$$



Maximum Likelihood from *L* data:

$$\hat{\mu} = \frac{1}{L}\sum_{\ell=1}^{L} x^{(\ell)}$$

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

Other Useful Properties

- Often closed under multiplication / division:

$$p(x \mid \eta_1)p(x \mid \eta_2) \propto p(x \mid \eta_1 + \eta_2) \qquad p(x \mid \eta_1) \div p(x \mid \eta_2) \propto p(x \mid \eta_1 - \eta_2)$$

If $\eta_1 + \eta_2$ valid parameters

If $\eta_1 - \eta_2$ valid parameters

- Posterior predictive of conjugate pair typically closed-form
- The *maximum entropy distribution* of data is in exponential family
- *Kullback-Leibler* (KL) divergence between two expfams closed-form
- Minimum KL(p||q) with q in expfam given by moment matching,

$$\mathbb{E}_p[\phi(x)] = \mathbb{E}_q[\phi(x)]$$

True for any distribution p

Summary

- Family of distributions with PMF/PDF of the form:

$$p(x) = h(x) \exp\left\{\eta^T \phi(x) - A(\eta)\right\}$$

Base Measure

Natural Parameters

Sufficient Statistics

- Log-Partition: $$A(\eta) = \log \int \exp\left\{\eta^T \phi(x)\right\} h(x) dx$$

- Alternate *mean parameters* as expected sufficient statistics or derivatives of log-partition:

$$\mu_i = \mathbb{E}_{p_\eta}[\phi_i(x)] = \frac{\partial}{\partial \eta_i} A(\eta)$$

Summary

- **Lots of useful properties**
  - Allows simultaneous study of many popular probability distributions:
    Bernoulli (binary), Categorical, Poisson (counts), Exponential (positive), Gaussian (real), …
  - Maximum likelihood (ML) learning is simple: *moment matching of sufficient statistics*
  - Bayesian learning is simple: *conjugate priors are available*
    Beta, Dirichlet, Gamma, Gaussian, Wishart, …
  - The *maximum entropy* interpretation: Among all distributions with certain moments of interest, the exponential family is the most random (makes fewest assumptions)
  - Parametric and predictive *sufficiency*: For arbitrarily large datasets, optimal learning is possible from a finite-dimensional set of statistics (streaming, big data)

- **All exponential family likelihoods have conjugate priors**
  - Means posterior is same distribution as prior
  - Inference reduces to computing posterior parameters