Exponential family &
Generalized Linear Models (GLIMs)

40-956 Advanced Topics in AI:
Probabilistic Graphical Models
Sharif University of Technology

Soleymani
Spring 2015
Outline

- Exponential family
  - Many standard distributions are in this family
  - Similarities among learning algorithms for different models in this family:
    - ML estimation has a simple form for exponential families
      - *moment matching* of sufficient statistics
    - Bayesian learning is simplest for exponential families
  - They have a maximum entropy interpretation

- GLIMs as to parameterize conditional distributions that have an exponential distribution on a variable for each value of parent
Exponential family: canonical parameterization

\[ P(x|\eta) = \frac{1}{Z(\eta)} h(x) \exp\{\eta^T T(x)\} \]

\[ Z(\eta) = \int h(x) \exp\{\eta^T T(x)\} \, dx \]

\[ P(x|\eta) = h(x) \exp\{\eta^T T(x) - \ln Z(\eta)\} \]

- \( T: \mathcal{X} \rightarrow \mathbb{R}^K \): sufficient statistics function
- \( \eta \): natural or canonical parameters
- \( h: \mathcal{X} \rightarrow \mathbb{R}^+ \): reference measure independent of parameters
- \( Z \): Normalization factor or partition function \((0 < Z(\eta) < \infty)\)
Example: Bernoulli

\[ P(x|\theta) = \theta^x (1 - \theta)^{1-x} \]

\[ = \exp \left\{ \ln \left( \frac{\theta}{1-\theta} \right)^x + \ln(1 - \theta) \right\} \]

- \( \eta = \ln \left( \frac{\theta}{1-\theta} \right) \)
- \( \eta = \ln \left( \frac{\theta}{1-\theta} \right) \Rightarrow \theta = \frac{\eta}{\eta + 1} = \frac{1}{1+e^{-\eta}} \)
- \( T(x) = x \)
- \( A(\eta) = -\ln(1 - \theta) = \ln(1 + e^\eta) \)
- \( h(x) = 1 \)
Example: Gaussian

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

1. \[ \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} \]

2. \[ \Rightarrow \mu = -\frac{\eta_1}{2\eta_2}, \sigma^2 = -\frac{1}{2\eta_2} \]

3. \[ T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \]

4. \[ A(\eta) = -\ln \left( \sqrt{2\pi\sigma} \exp \left\{ \frac{\mu^2}{2\sigma^2} \right\} \right) = -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln(-2\eta_2) - \frac{\eta_1^2}{4\eta_2} \]

5. \[ h(x) = 1 \]
Example: Multinomial

\[ P(x|\theta) = \prod_{k=1}^{K} \theta_{k}^{x_{k}} \quad \sum_{k=1}^{K} \theta_{k} = 1 \]

\[ P(x|\theta) = \exp \left\{ \sum_{k=1}^{K} x_{k} \ln \theta_{k} \right\} \]

\[ = \exp \left\{ \sum_{k=1}^{K-1} x_{k} \ln \theta_{k} + \left( 1 - \sum_{k=1}^{K-1} x_{k} \right) \ln \left( 1 - \sum_{k=1}^{K-1} \theta_{k} \right) \right\} \]

- \[ \eta = [\eta_{1}, ..., \eta_{K}]^{T} = \left[ \ln \left( \frac{\theta_{1}}{1 - \sum_{k=1}^{K-1} \theta_{k}} \right), ..., \ln \left( \frac{\theta_{K-1}}{1 - \sum_{k=1}^{K-1} \theta_{k}} \right) \right]^{T} \]
- \[ \eta = \left[ \ln \left( \frac{\theta_{1}}{\theta_{K}} \right), ..., \ln \left( \frac{\theta_{K-1}}{\theta_{K}} \right) \right]^{T} \Rightarrow \theta_{k} = \frac{e^{\eta_{k}}}{\sum_{j=1}^{K} e^{\eta_{j}}} \]
- \[ T(x) = [x_{1}, ..., x_{K-1}]^{T} \]
- \[ A(\eta) = - \ln \theta_{K} = - \ln \left( 1 - \sum_{k=1}^{K-1} \theta_{k} \right) = \ln \left( \sum_{k=1}^{K} e^{\eta_{j}} \right) \]
Multiple exponential families may encode the same set of distributions

We want the parameter space \( \{\eta | 0 < Z(\eta) < \infty\} \) to be:

- **Convex set**
- **Non-redundant:** \( \eta \neq \eta' \Rightarrow P(x|\eta) \neq P(x|\eta') \)
  - The function from \( \theta \) to \( \eta \) is invertible
  - Example: invertible function from \( \theta \) to \( \eta \) in the Bernoulli example
    \[
    \theta = \frac{1}{1+e^{-\eta}}
    \]
Examples of non-exponential distributions

- Uniform
- Laplace
- Student t-distribution
Moments

\[ A(\eta) = \ln Z(\eta) \]

\[ Z(\eta) = \int h(x) \exp\{\eta^T T(x)\} \, dx \]

\[ \nabla_\eta A(\eta) = \frac{\nabla_\eta Z(\eta)}{Z(\eta)} = \frac{\int h(x) T(x) \exp\{\eta^T T(x)\} \, dx}{Z(\eta)} \]

\[ = \int T(x) \frac{h(x) \exp\{\eta^T T(x)\}}{Z(\eta)} \, dx = E_{P(x|\eta)}[T(x)] \]

\[ \Rightarrow \nabla_\eta A(\eta) = E_\eta [T(x)] \quad \text{The first derivative of } A(\eta) \text{ is the mean of sufficient statistics} \]

\[ \nabla_\eta^2 A(\eta) = E_\eta [T(x)T(x)^T] - E_\eta [T(x)] E_\eta [T(x)]^T = Cov_\eta [T(x)] \]

The i-th derivative gives the i-th centered moment of sufficient statistics.
The moment parameters $\theta$ can be derived as a function of the natural or canonical parameters:

$$\nabla_\eta A(\eta) = E_\eta[T(x)]$$

For many distributions, we have $\theta \equiv E_\eta[T(x)]$ \Rightarrow $\nabla_\eta A(\eta) = \theta$

- $A(\eta)$ is convex since $\nabla_\eta^2 A(\eta) = \text{Cov}_\eta[T(x)] \succeq 0$
- Covariance matrix is always positive semi-definite \Rightarrow Hessian $\nabla_\eta^2 A(\eta)$ is positive semi-definite, and hence that $A(\eta) = \ln Z(\eta)$ is a convex function of $\eta$. 
Exponential family: moment parameterization

- A distribution in the exponential family can also be parameterized by the **moment parameterization**:

  \[
P(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp\{\psi(\theta)^T T(x)\} \]

  \[
Z(\theta) = \int h(x) \exp\{\psi(\theta)^T T(x)\} \, dx
\]

- If \( \nabla_\eta^2 A(\eta) > 0 \Rightarrow \nabla_\eta A(\eta) \) is ascending \( \Rightarrow \psi^{-1}(\eta) = \theta = \nabla_\eta A(\eta) \) is ascending and thus is 1-to-1

- The mapping from the moments to the canonical parameters is invertible (1-to-1 relationship): \( \eta = \psi(\theta) \)
Sufficiency

- Bayesian:
  \[ \theta \perp X|T(X) \]

- Frequentist:
  \[ P(x|T(x), \theta) = P(x|T(x)) \]

Sufficiency in both frequentist and Bayesian frameworks implies a factorization of \( P(x|\theta) \) (Neyman factorization theorem):

\[
\begin{align*}
P(x, T(x), \theta) &= f(T(x), \theta)g(x, T(x)) \\
P(x, \theta) &= f(T(x), \theta)g(x, T(x)) \\
P(x|\theta) &= f'(T(x), \theta)g(x, T(x))
\end{align*}
\]
MLE for exponential family

\[
\ell(\eta; \mathcal{D}) = \ln P(\mathcal{D} | \eta) = \ln \prod_{n=1}^{N} h(x^{(n)}) \exp\{\eta^T T(x^{(n)}) - A(\eta)\}
\]

\[
= \ln \prod_{n=1}^{N} h(x^{(n)}) + \eta^T \left( \sum_{n=1}^{N} T(x^{(n)}) \right) - NA(\eta)
\]

Concave function

\[
\nabla_\eta \ell(\eta; \mathcal{D}) = 0 \Rightarrow \sum_{n=1}^{N} T(x^{(n)}) - NA(\eta) = 0
\]

\[
\Rightarrow \nabla_\eta A(\eta) = \frac{\sum_{n=1}^{N} T(x^{(n)})}{N}
\]

\[
\Rightarrow \nabla_\eta A(\hat{\eta}) = E_\hat{\eta}[T(x)] = \frac{\sum_{n=1}^{N} T(x^{(n)})}{N}
\]

moment matching
Maximum entropy models

- Among all distributions with certain moments of interest, the exponential family is the most random (makes fewest assumptions or structure)
  - Out of all distributions which reproduce the observed sufficient statistics, the exponential family distribution (roughly) makes the fewest additional assumptions.

- The unique distribution maximizing the entropy, subject to the constraint that these moments are exactly matched, is then an exponential family distribution.
Maximum entropy

- **Constraints:**

\[
E[f_k] = \sum_x f_k(x)P(x) = F_k
\]

- **Maximum entropy (maxent):** pick the distribution with maximum entropy subject to the constraints \( L(P, \lambda) \)

\[
= -\sum_x P(x) \log P(x) + \lambda_0 \left(1 - \sum_x P(x)\right) + \sum_k \lambda_k \left(F_k - \sum_x f_k(x)P(x)\right)
\]

\[
\nabla L = 0 \Rightarrow P(x) = \frac{1}{Z} \exp \left\{ -\sum_k \lambda_k f_k(x) \right\}
\]

\[
Z = \sum_x \exp \left\{ -\sum_k \lambda_k f_k(x) \right\}
\]

\( f_k(x) \): an arbitrary function

\( F_k \): constant
Maximum entropy: constraints

- Constants in the constraints:
  - $F_k$ measure the empirical counts on the training data
  - $F_k = \frac{\sum_{n=1}^{N} f_k(x^{(n)})}{N}$

- These constraints also ensure consistency automatically.
Exponential family: summary

- Many famous distribution are in the exponential family

- Important properties for learning with exponential families:
  - Gradients of log partition function gives expected sufficient statistics, or moments, of some current model
    - Moments of any distribution in exponential family can be easily computed by taking the derivatives of the log normalizer
  - The Hessian of the log partition function is positive semi-definite and so the log partition function is convex
  - Among all distributions with certain moments of interest, the exponential family has the highest entropy
Generalized linear models (GLIMs)

- Conditional relationship between $Y$ and $X$

Examples:

- Linear regression: $P(y|x, w, \sigma^2) = \mathcal{N}(y|w^T x, \sigma^2)$

- Discriminative linear classifier (two class)
  - Logistic regression: $P(y|x, w) = Ber(y|\sigma(w^T x))$
  - Probit regression: $P(y|x, w) = Ber(y|\Phi(w^T x))$ where $\Phi$ is the cdf of $\mathcal{N}(0,1)$
Generalized linear models (GLIMs)

- $P(y|x)$ is a generalized linear model if:
  - $x$ enters into the model via a linear combination $w^T x$
  - The conditional mean of $P(y|x)$ is expressed as $f(w^T x)$:
    - $f$ is called the response function
    - $\mu = E[y|x] = f(w^T x)$
  - The distribution of $y$ is characterized by an exponential family distribution (with conditional mean $f(w^T x)$)

- We have two choices in the specification of a GLIM:
  - The choice of the exponential family distribution
    - Usually constrained by the nature of $Y$
  - The choice of the response function $f$
    - the principal degree of freedom in the specification of a GLIM
    - However, we need to impose constraints on this function (e.g., $f$ must be in $[0,1]$ for Bernoulli distribution on $y$)
The relation between vars. in a GLIMs

\[ \theta \rightarrow \xi \leftarrow x \quad \xi \xrightarrow{f} \mu \xrightarrow{\psi} \eta \]
Canonical response function

- Canonical response function: $f(.) = \psi^{-1}(.)$ or $\xi = \eta$
  - In this case, the choice of the exponential family density completely determines the GLIM

<table>
<thead>
<tr>
<th>Model</th>
<th>Canonical response function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>$\mu = \eta$</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>$\mu = 1/(1 + e^{-\eta})$</td>
</tr>
<tr>
<td>multinomial</td>
<td>$\mu_i = \eta_i / \sum_j e^{\eta_j}$</td>
</tr>
<tr>
<td>Poisson</td>
<td>$\mu = e^{\eta}$</td>
</tr>
<tr>
<td>gamma</td>
<td>$\mu = -\eta^{-1}$</td>
</tr>
</tbody>
</table>

- The constraints on the range of $f$ are automatically satisfied.
  - $\mu = f(\eta)$ are guaranteed to be possible values of the conditional expectation (i.e., $f(\eta) = \psi^{-1}(\eta) = \frac{dA(\eta)}{d\eta} = E[Y|\eta]$)
Log likelihood for GLIMs

\[
\ell(\eta; D) = \ln P(D|\eta) \\
= \ln \prod_{n=1}^{N} h(y^{(n)}) \exp \left\{ \eta^{(n)} y^{(n)} - A(\eta^{(n)}) \right\} \\
= \sum_{n=1}^{N} \ln h(y^{(n)}) + \sum_{n=1}^{N} \left( \eta^{(n)} y^{(n)} - A(\eta^{(n)}) \right)
\]

- \( \eta^{(n)} = \psi(\mu^{(n)}) \) and \( \mu^{(n)} = f(\theta^T x^{(n)}) \)
- In the case of canonical response function \( \eta^{(n)} = \theta^T x^{(n)} \)

\[
\ell(\theta; D) = \sum_{n=1}^{N} \ln h(y^{(n)}) + \theta^T \sum_{n=1}^{N} x^{(n)} y^{(n)} - \sum_{n=1}^{N} A(\theta^T x^{(n)})
\]

Sufficient statistics for \( \theta \)
Gradient of log likelihood

\[
\nabla_\theta l(\eta; D) = \sum_{n=1}^{N} \frac{dl}{d\eta^{(n)}} \nabla_\theta \eta^{(n)} = \sum_{n=1}^{N} \left( y^{(n)} - \frac{dA(\eta^{(n)})}{d\eta^{(n)}} \right) \nabla_\theta \eta^{(n)}
\]

\[
= \sum_{n=1}^{N} \left( y^{(n)} - \mu^{(n)} \right) \frac{d\eta^{(n)}}{d\mu^{(n)}} \frac{d\mu^{(n)}}{d\xi^{(n)}} x^{(n)}
\]

- In the case of canonical response function \( \eta^{(n)} = \xi^{(n)} \):

\[
\nabla_\theta l(\theta; D) = \sum_{n=1}^{N} \left( y^{(n)} - \mu^{(n)} \right) x^{(n)}
\]

\[
\mu^{(n)} = f(\theta^T x^{(n)})
\]
Online learning for GLIMs

- An LMS like algorithm as a generic stochastic gradient descent for GLIMs:

$$\theta^{t+1} = \theta^t + \rho \left( y^{(n)} - \mu^{(n)t} \right) x^{(n)}$$

$$\mu^{(n)t} = f \left( \theta^T x^{(n)} \right)$$

Similar to Least Mean Squares (LMS) algorithm

- If we do not use the canonical response function only scaling coefficients due to the derivatives of $f(.)$ and $\psi(.)$ will also incorporated into the step size
Batch learning for GLIMs

- For the canonical response functions:

\[ \nabla_{\theta} l(\theta; D) = \sum_{n=1}^{N} (y^{(n)} - \mu^{(n)}) x^{(n)} = X^T (y - \mu) \]

\[ H = \frac{d^2 l}{d\theta d\theta^T} = \frac{dl}{d\theta} \sum_{n=1}^{N} (y^{(n)} - \mu^{(n)}) x^{(n)} = - \sum_{n=1}^{N} x^{(n)} \frac{d\mu^{(n)}}{d\theta^T} \]

\[ = - \sum_{n=1}^{N} x^{(n)} \frac{d\mu^{(n)}}{d\eta^{(n)}} \frac{d\eta^{(n)}}{d\theta^T} \]

- Since \( \eta^{(n)} = \theta^T x^{(n)} \)

\[ H = - \sum_{n=1}^{N} x^{(n)} \frac{d\mu^{(n)}}{d\eta^{(n)}} x^{(n)^T} = -X^T WX \]

\[ W = \text{diag} \left( \frac{d\mu^{(1)}}{d\eta^{(1)}}, \ldots, \frac{d\mu^{(N)}}{d\eta^{(N)}} \right) \]

\[ \frac{d\mu^{(n)}}{d\eta^{(n)}} = \frac{d^2 A}{d\eta^{(n)}} \]
Batch learning for GLIMs: Newton-Rafson

\[
\theta^{t+1} = \theta^t + (X^T W^t X)^{-1} X^T (y - \mu^t)
\]

\[
= (X^T W^t X)^{-1} [X^T W^t X \theta^t + X^T (y - \mu^t)]
\]

\[
\Rightarrow \theta^{t+1} = (X^T W^t X)^{-1} X^T W^t z^t
\]

\[
z^t = \eta^t + W^{-1} (y - \mu^t)
\]

Iterative Reweighted Least Squares (IRLS)
Linear regression

- Cost function (according to MLE where $P(y|x) = \mathcal{N}(y|w^T x, \sigma^2)$):
  \[
  J(\theta) = \frac{1}{2} \sum_{n=1}^{N} \left( \theta^T x^{(n)} - y^{(n)} \right)^2
  \]
  \[
  \nabla_{\theta} J(\theta) = 0 \Rightarrow \theta = (X^T X)^{-1} X^T y
  \]

- Online learning (LMS):
  \[
  \theta^{t+1} = \theta^t + \rho \left( y^{(n)} - \theta^T x^{(n)} \right) x^{(n)}
  \]

- IRLS:
  \[
  \theta^{t+1} = (X^T W^t X)^{-1} X^T W^t z^t
  = (X^T X)^{-1} X^T (X \theta^t + y - \mu^t)
  = \theta^t + (X^T X)^{-1} X^T (y - \mu^t)
  \]

Canonical response function:
\[
\mu(x) = \theta^T x = \eta(x)
\]
\[
\frac{d\mu}{d\eta} = 1 \Rightarrow W = I
\]
Logistic regression

\[ \mu(x) = \frac{1}{1 + e^{-\eta(x)}} \]

- Canonical response function \( \eta = \xi = \theta^T x \)

- IRLS:

\[ \frac{d\mu}{d\eta} = \mu(1 - \mu) \]

\[ W = \begin{bmatrix} \mu^{(1)}(1 - \mu^{(1)}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mu^{(N)}(1 - \mu^{(N)}) \end{bmatrix} \]

\[ \theta^{t+1} = (X^T W^t X)^{-1} X^T W^t z^t \]

\[ z^t = X\theta^t + W^{t-1}(y - \mu^t) \]
References

- Jordan, Chapter 8.
- Koller & Friedman, Chapter 8.1-8.3.