Probabilistic Graphical Models

Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

Lecture 7:
Exponential Families, Conjugate Priors, and Factor Graphs

Some figures courtesy Michael Jordan’s draft textbook,
An Introduction to Probabilistic Graphical Models
Exponential Families of Distributions

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

\[ Z(\theta) = \int_{x^m} h(x) \exp[\theta^T \phi(x)] dx \]

\[ = h(x) \exp[\theta^T \phi(x) - A(\theta)] \]

\[ A(\theta) = \log Z(\theta) \]

\[ \phi(x) \in \mathbb{R}^d \quad \text{fixed vector of sufficient statistics (features), specifying the family of distributions} \]

\[ \theta \in \Theta \quad \text{unknown vector of natural parameters, determine particular distribution in this family} \]

\[ Z(\theta) > 0 \quad \text{normalization constant or partition function, ensuring this is a valid probability distribution} \]

\[ h(x) > 0 \quad \text{reference measure independent of parameters (for many models, we simply have } h(x) = 1) \]

To ensure this construction is valid, we take

\[ \Theta = \{ \theta \in \mathbb{R}^d \mid Z(\theta) < \infty \} \]
Why the Exponential Family?

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

\[ Z(\theta) = \int_{\mathcal{X}} h(x) \exp[\theta^T \phi(x)] dx \]

\[ = h(x) \exp[\theta^T \phi(x) - A(\theta)] \]

\[ A(\theta) = \log Z(\theta) \]

- Many standard distributions are in this family, and by studying exponential families, we study them all simultaneously
- Explains similarities among learning algorithms for different models, and makes it easier to derive new algorithms:
  - ML estimation takes a simple form for exponential families: *moment matching* of sufficient statistics
  - Bayesian learning is simplest for exponential families: they are the only distributions with *conjugate priors*
- They have a *maximum entropy* interpretation: Among all distributions with certain moments of interest, the exponential family is the most random (makes fewest assumptions)
Examples of Exponential Families

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

\[ Z(\theta) = \int_{x^m} h(x) \exp[\theta^T \phi(x)] dx \]

\[ = h(x) \exp[\theta^T \phi(x) - A(\theta)] \]

\[ A(\theta) = \log Z(\theta) \]

- Bernoulli and binomial (2 classes)

- Categorical and multinomial (K classes)

\[ \phi(x) = \mathbb{I}(x = 1) = x \]

\[ \phi(x) = [\mathbb{I}(x = 1), \ldots, \mathbb{I}(x = K - 1)] \]

- Scalar Gaussian

\[ \phi(x) = [x, x^2] \]

- Multivariate Gaussian

\[ \phi(x) = [x, xx^T] \]

- Poisson

\[ h(x) = \frac{1}{x!}, \phi(x) = x \]

- Dirichlet and beta

- Gamma and exponential

- …
Non-Exponential Families

• Uniform distribution
  \[
  \text{Unif}(x \mid a, b) = \frac{1}{b-a} \mathbb{I}(a \leq x \leq b)
  \]

• Laplace and Student-t distributions
  \[
  \text{Lap}(x \mid \mu, \lambda) = \frac{\lambda}{2} \exp\left(-\lambda|x - \mu|\right)
  \]
Convexity

\[
\lambda \theta + (1 - \lambda) \theta' \in S, \quad \forall \lambda \in [0, 1]
\]
\[
\theta, \theta' \in S
\]

\[
f(\lambda \theta + (1 - \lambda) \theta') \leq \lambda f(\theta) + (1 - \lambda) f(\theta')
\]
Convexity & Jensen’s Inequality

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$
Concavity & Jensen’s Inequality

\[ \ln(\mathbb{E}[X]) \geq \mathbb{E}[\ln(X)] \]
Log Partition Function

\[
p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp[\theta^T \phi(x)] \\
= h(x) \exp[\theta^T \phi(x) - A(\theta)] \\
A(\theta) = \log Z(\theta)
\]

- Derivatives of log partition function have an intuitive form:
  \[
  \nabla_\theta A(\theta) = \mathbb{E}_\theta[\phi(x)] \\
  \nabla_\theta^2 A(\theta) = \text{Cov}_\theta[\phi(x)] = \mathbb{E}_\theta[\phi(x)\phi(x)^T] - \mathbb{E}_\theta[\phi(x)]\mathbb{E}_\theta[\phi(x)]^T
  \]

- Important consequences for learning with exponential families:
  - Finding gradients is equivalent to finding expected sufficient statistics, or *moments*, of some current model
  - The Hessian is positive definite so \( A(\theta) \) is convex
  - This in turn implies that the parameter space \( \Theta \) is convex
  - Learning is a convex problem: No local optima!
  *At least when we have complete observations...*
A Little Information Theory

- The **entropy** is a natural measure of the inherent uncertainty (difficulty of compression) of some random variable:

\[
H(p) = - \sum_{x \in \mathcal{X}} p(x) \log p(x) \quad \text{discrete entropy (concave, non-negative)}
\]

\[
H(p) = - \int_{\mathcal{X}} p(x) \log p(x) \, dx \quad \text{differential entropy (concave, real-valued)}
\]

- The **relative entropy** or **Kullback-Leibler (KL) divergence** is then a non-negative, but asymmetric, “distance” between a given pair of probability distributions:

\[
D(p \mid \mid q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, dx \quad D(p \mid \mid q) \geq 0
\]

The KL divergence equals zero iff \( p(x) = q(x) \) almost everywhere.

- The **mutual information** measures dependence between a pair of random variables:

\[
I(p_{xy}) \triangleq D(p_{xy} \mid p_x p_y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_{xy}(x, y) \log \frac{p_{xy}(x, y)}{p_x(x)p_y(y)} \, dy \, dx
\]

\[
= H(p_x) + H(p_y) - H(p_{xy})
\]
Learning in Exponential Families

\[ p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp[\theta^T \phi(x)] \quad Z(\theta) = \int_{x^m} h(x) \exp[\theta^T \phi(x)] dx \]
\[ = h(x) \exp[\theta^T \phi(x) - A(\theta)] \quad A(\theta) = \log Z(\theta) \]

- Given any target probability distribution \( \tilde{p}(x) \), the closest exponential family distribution matches moments:
  \[ \hat{\theta} = \arg \min_{\theta} D(\tilde{p} \mid \mid p_\theta) \quad \iff \quad \mathbb{E}_{\hat{\theta}}[\phi_a(x)] = \int_X \phi_a(x) \tilde{p}(x) dx \]

- Given \( L \) samples, their empirical distribution equals
  \[ \tilde{p}(x) = \frac{1}{L} \sum_{\ell=1}^L \delta_{x^{(\ell)}}(x) \]

- For exponential families, maximum likelihood estimation always minimizes KL divergence from empirical distribution:
  \[ \hat{\theta} = \arg \max_{\theta} \sum_{\ell=1}^L \log p(x^{(\ell)} \mid \theta) = \arg \min_{\theta} D(\tilde{p} \mid \mid p_\theta) \quad \iff \quad \mathbb{E}_{\hat{\theta}}[\phi_a(x)] = \frac{1}{L} \sum_{\ell=1}^L \phi_a(x^{(\ell)}) \]
Maximum Entropy Models

\[
p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp[\theta^T \phi(x)] \\
Z(\theta) = \int_{\mathcal{X}} h(x) \exp[\theta^T \phi(x)] dx \\
= h(x) \exp[\theta^T \phi(x) - A(\theta)] \\
A(\theta) = \log Z(\theta)
\]

• Consider a collection of d target statistics \( \phi_a(x) \), whose expectations with respect to some distribution \( \tilde{p}(x) \) are

\[
\int_{\mathcal{X}} \phi_a(x) \tilde{p}(x) dx = \mu_a
\]

• The unique distribution \( \hat{p}(x) \) maximizing the entropy \( H(\hat{p}) \), subject to the constraint that these moments are exactly matched, is then an exponential family distribution with

\[
\mathbb{E}_{\hat{p}}[\phi_a(x)] = \mu_a, \quad h(x) = 1
\]

Out of all distributions which reproduce the observed sufficient statistics, the exponential family distribution (roughly) makes the fewest additional assumptions.
Parametric & Predictive Sufficiency

Posterior distributions and predictive likelihoods:

\[
p(\theta \mid x^{(1)}, \ldots, x^{(L)}, \lambda) = \frac{p(x^{(1)}, \ldots, x^{(L)} \mid \theta, \lambda) p(\theta \mid \lambda)}{\int_{\Theta} p(x^{(1)}, \ldots, x^{(L)} \mid \theta, \lambda) p(\theta \mid \lambda) d\theta} \propto p(\theta \mid \lambda) \prod_{\ell=1}^{L} p(x^{(\ell)} \mid \theta)
\]

\[
p(\bar{x} \mid x^{(1)}, \ldots, x^{(L)}, \lambda) = \int_{\Theta} p(\bar{x} \mid \theta) p(\theta \mid x^{(1)}, \ldots, x^{(L)}, \lambda) d\theta
\]

**Theorem 2.1.2.** Let \( p(x \mid \theta) \) denote an exponential family with canonical parameters \( \theta \), and \( p(\theta \mid \lambda) \) a corresponding prior density. Given \( L \) independent, identically distributed samples \( \{x^{(\ell)}\}_{\ell=1}^{L} \), consider the following statistics:

\[
\phi(x^{(1)}, \ldots, x^{(L)}) \triangleq \left\{ \frac{1}{L} \sum_{\ell=1}^{L} \phi_a(x^{(\ell)}) \mid a \in A \right\}
\] (2.24)

These empirical moments, along with the sample size \( L \), are then said to be parametric sufficient for the posterior distribution over canonical parameters, so that

\[
p(\theta \mid x^{(1)}, \ldots, x^{(L)}, \lambda) = p(\theta \mid \phi(x^{(1)}, \ldots, x^{(L)}), L, \lambda)
\] (2.25)

Equivalently, they are predictive sufficient for the likelihood of new data \( \bar{x} \):

\[
p(\bar{x} \mid x^{(1)}, \ldots, x^{(L)}, \lambda) = p(\bar{x} \mid \phi(x^{(1)}, \ldots, x^{(L)}), L, \lambda)
\] (2.26)
Learning with Conjugate Priors

\[ p(x \mid \theta) = \nu(x) \exp \left\{ \sum_{a \in A} \theta_a \phi_a(x) - \Phi(\theta) \right\} \]
\[ \Phi(\theta) = \log \int_x \nu(x) \exp \left\{ \sum_{a \in A} \theta_a \phi_a(x) \right\} \, dx \]
\[ p(\theta \mid \lambda) = \exp \left\{ \sum_{a \in A} \theta_a \lambda_0 \lambda_a - \lambda_0 \Phi(\theta) - \Omega(\lambda) \right\} \]
\[ \Omega(\lambda) = \log \int_\Theta \exp \left\{ \sum_{a \in A} \theta_a \lambda_0 \lambda_a - \lambda_0 \Phi(\theta) \right\} \, d\theta \]
\[ \Lambda \triangleq \left\{ \lambda \in \mathbb{R}^{\lvert A \rvert + 1} \mid \Omega(\lambda) < \infty \right\} \]

**Proposition 2.1.4.** Let \( p(x \mid \theta) \) denote an exponential family with canonical parameters \( \theta \), and \( p(\theta \mid \lambda) \) a family of conjugate priors defined as in eq. (2.28). Given \( L \) independent samples \( \{x^{(\ell)}\}_{\ell=1}^L \), the posterior distribution remains in the same family:

\[ p(\theta \mid x^{(1)}, \ldots, x^{(L)}, \lambda) = p(\theta \mid \bar{\lambda}) \quad (2.31) \]
\[ \bar{\lambda}_0 = \lambda_0 + L \quad \bar{\lambda}_a = \frac{\lambda_0 \lambda_a + \sum_{\ell=1}^L \phi_a(x^{(\ell)})}{\lambda_0 + L} \quad a \in A \quad (2.32) \]

Integrating over \( \Theta \), the log–likelihood of the observations can then be compactly written using the normalization constant of eq. (2.29):

\[ \log p(x^{(1)}, \ldots, x^{(L)} \mid \lambda) = \Omega(\bar{\lambda}) - \Omega(\lambda) + \sum_{\ell=1}^L \log \nu(x^{(\ell)}) \quad (2.33) \]
Learning with Conjugate Priors

\[
p(x \mid \theta) = \nu(x) \exp \left\{ \sum_{a \in A} \theta_a \phi_a(x) - \Phi(\theta) \right\} \\
\Phi(\theta) = \log \int_{\mathcal{X}} \nu(x) \exp \left\{ \sum_{a \in A} \theta_a \phi_a(x) \right\} \, dx
\]

\[
p(\theta \mid \lambda) = \exp \left\{ \sum_{a \in A} \theta_a \lambda_0 \lambda_a - \lambda_0 \Phi(\theta) - \Omega(\lambda) \right\} \\
\Omega(\lambda) = \log \int_{\Theta} \exp \left\{ \sum_{a \in A} \theta_a \lambda_0 \lambda_a - \lambda_0 \Phi(\theta) \right\} \, d\theta
\]

\[
\Lambda \triangleq \left\{ \lambda \in \mathbb{R}^{\left|A\right|+1} \mid \Omega(\lambda) < \infty \right\}
\]

**Proposition 2.1.4.** Let \( p(x \mid \theta) \) denote an exponential family with canonical parameters \( \theta \), and \( p(\theta \mid \lambda) \) a family of conjugate priors defined as in eq. (2.28). Given \( L \) independent samples \( \{x^{(\ell)}\}_{\ell=1}^{L} \), the posterior distribution remains in the same family:

\[
p(\theta \mid x^{(1)}, \ldots, x^{(L)}, \lambda) = p(\theta \mid \bar{\lambda}) \tag{2.31}
\]

\[
\bar{\lambda}_0 = \lambda_0 + L \\
\bar{\lambda}_a = \frac{\lambda_0 \lambda_a + \sum_{\ell=1}^{L} \phi_a(x^{(\ell)})}{\lambda_0 + L} \quad a \in A \tag{2.32}
\]

For an exponential family, the conjugate prior is defined by:

- Prior expected values \( \lambda_a \) of the \( d \) sufficient statistics
- A measure of confidence in those prior expectations, expressed as a positive number of *pseudo-observations* \( \lambda_0 \)
Factor Graphs & Exponential Families

\[ p(x) = \frac{1}{Z(\theta)} \prod_{f \in \mathcal{F}} \psi_f(x_f \mid \theta_f) \]

- \( \mathcal{F} \rightarrow \) set of hyperedges linking subsets of nodes \( f \subseteq \mathcal{V} \)
- \( \mathcal{V} \rightarrow \) set of \( N \) nodes or vertices, \( \{1, 2, \ldots, N\} \)
- \( Z \rightarrow \) normalization constant (partition function)

- A **factor graph** is created from non-negative potential functions
- To guarantee non-negativity, we typically define potentials as

\[
\psi_f(x_f \mid \theta_f) = \nu_f(x_f) \exp \left\{ \sum_{a \in \mathcal{A}_f} \theta_{fa} \phi_{fa}(x_f) \right\}
\]

Local exponential family:
\[
\theta_f \triangleq \{ \theta_{fa} \mid a \in \mathcal{A}_f \}
\]

\[
p(x \mid \theta) = \left( \prod_{f \in \mathcal{F}} \nu_f(x_f) \right) \exp \left\{ \sum_{f \in \mathcal{F}} \sum_{a \in \mathcal{A}_f} \theta_{fa} \phi_{fa}(x_f) - \Phi(\theta) \right\}
\]

\[ \Phi(\theta) = \log Z(\theta) \]