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

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \qquad Z(\boldsymbol{\theta}) = \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x}$$

$$= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] \qquad A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$$

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

$\theta \in \Theta \longrightarrow$ unknown vector of *natural parameters*, determine particular distribution in this family

$Z(\theta) > 0 \longrightarrow$ normalization constant or *partition function*, ensuring this is a valid probability distribution

$h(x) > 0 \longrightarrow$ *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?

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \qquad Z(\boldsymbol{\theta}) = \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x}$$

$$= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] \qquad A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$$

- 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

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \qquad Z(\boldsymbol{\theta}) = \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x}$$

$$= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] \qquad A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$$

- Bernoulli and binomial (2 classes) $\phi(x) = \mathbb{I}(x=1) = x$
- Categorical and multinomial (K classes)

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

- Scalar Gaussian $\phi(x) = [x, x^2]$
- Multivariate Gaussian $\phi(x) = [x, xx^T]$

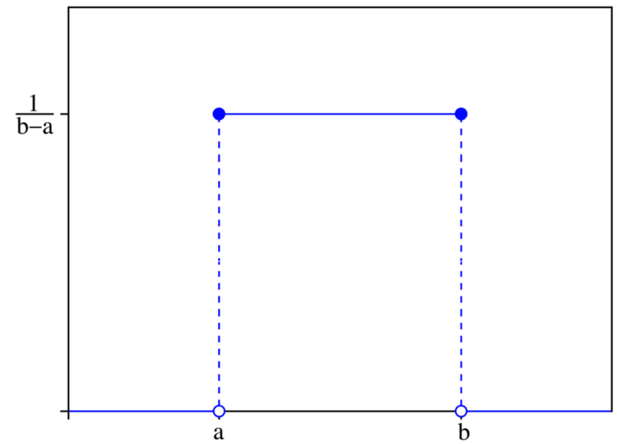
- Poisson $h(x) = \frac{1}{x!}, \phi(x) = x$

- Dirichlet and beta
- Gamma and exponential
- …

Non-Exponential Families

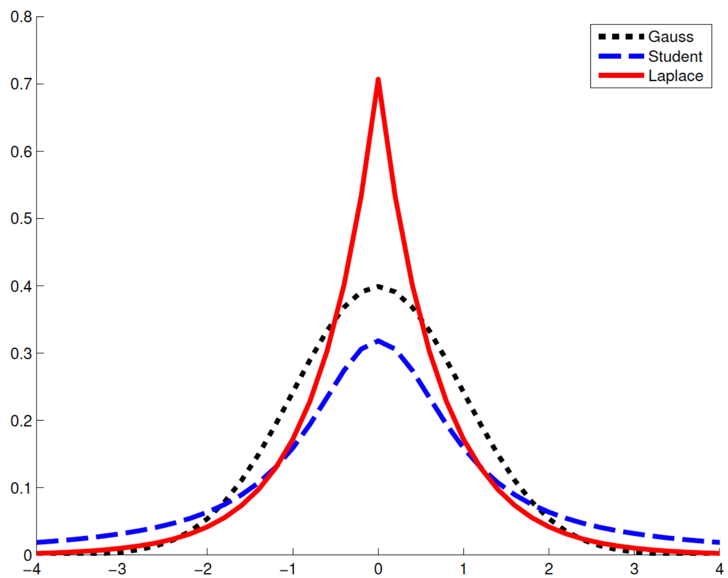
- Uniform distribution

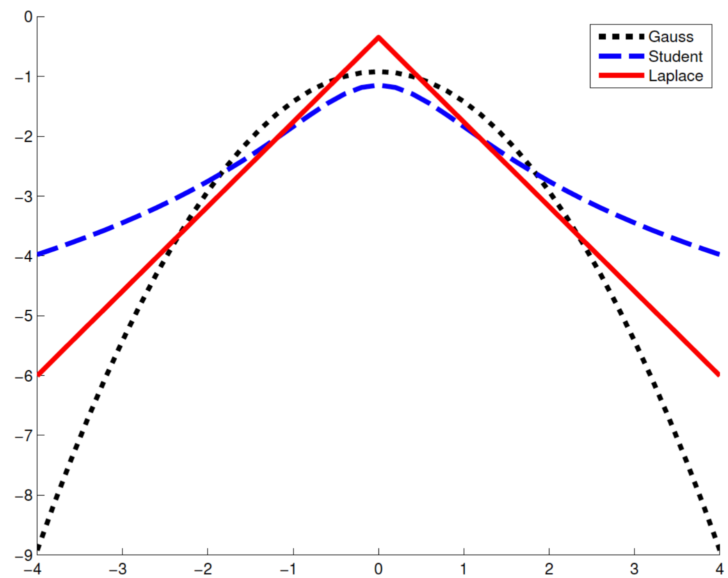
$$\mathrm{Unif}(x \mid a, b) = \frac{1}{b-a}\mathbb{I}(a \leq x \leq b)$$



- Laplace and Student-t distributions

$$\mathrm{Lap}(x \mid \mu, \lambda) = \frac{\lambda}{2}\exp(-\lambda|x-\mu|)$$

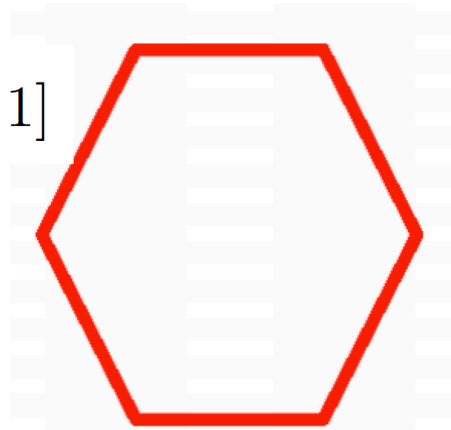


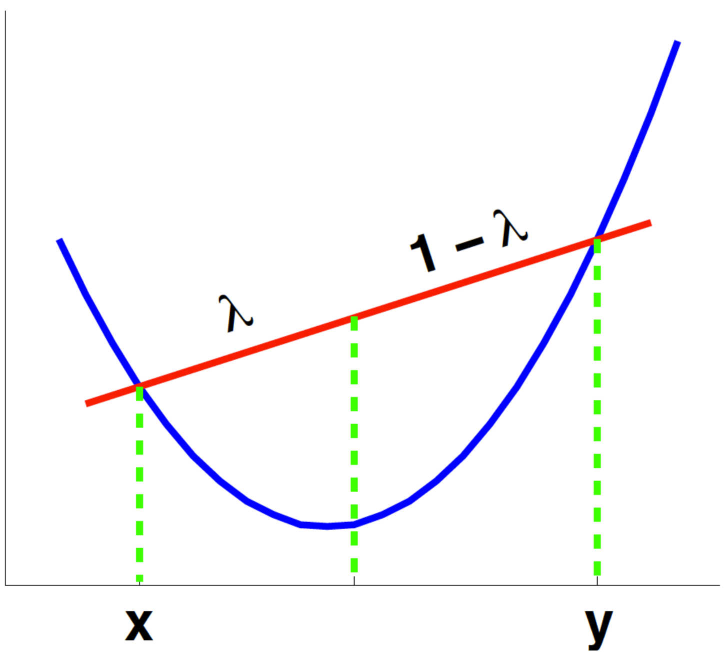


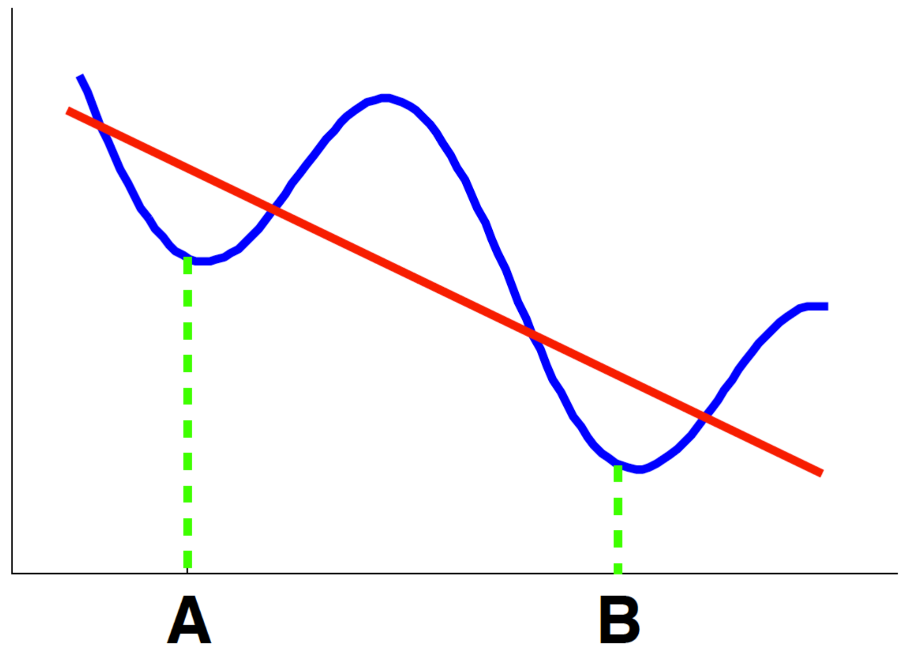
Convexity

$$\lambda\boldsymbol{\theta} + (1-\lambda)\boldsymbol{\theta}' \in \mathcal{S}, \quad \forall\, \lambda \in [0,1]$$

$$\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{S}$$

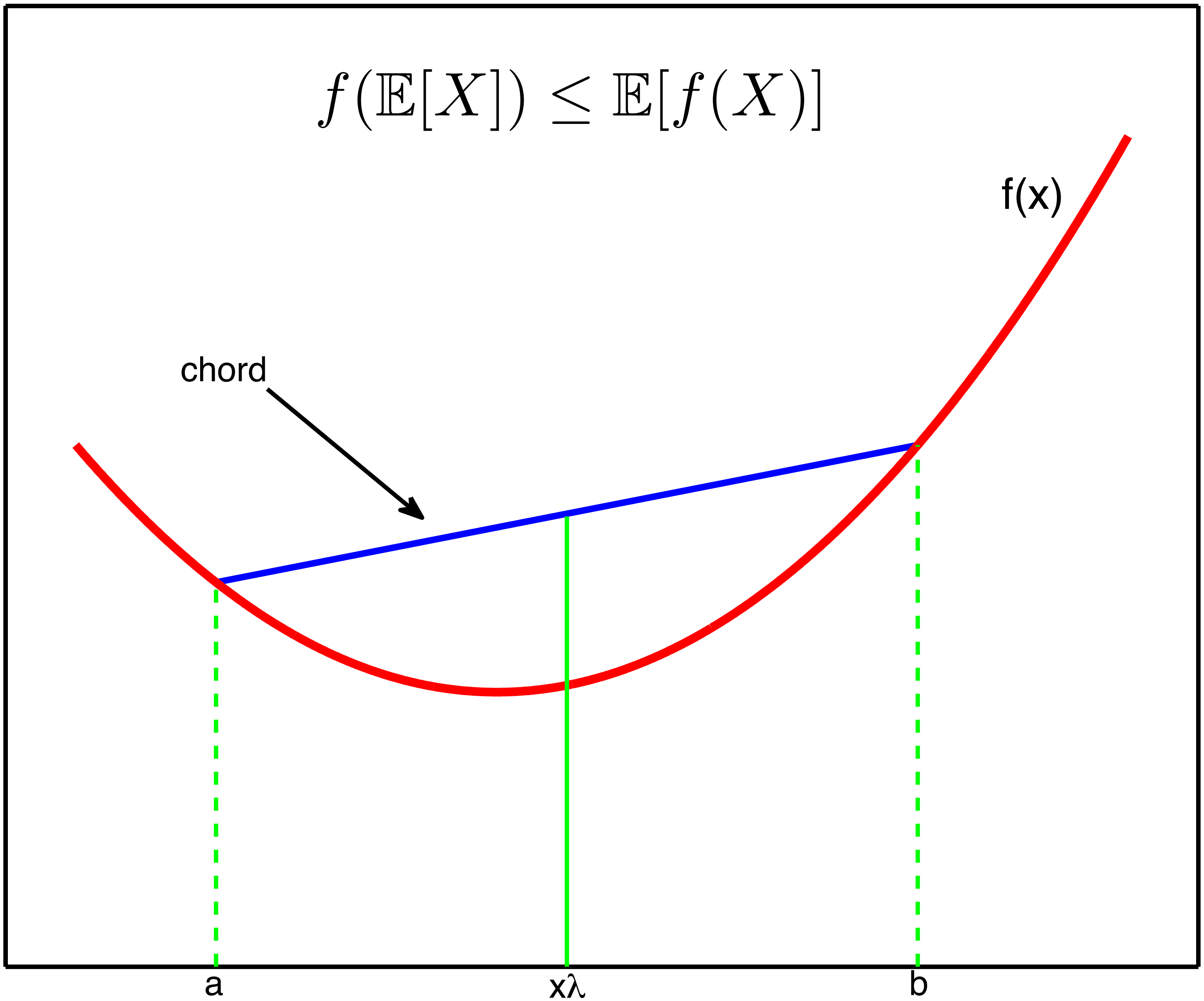






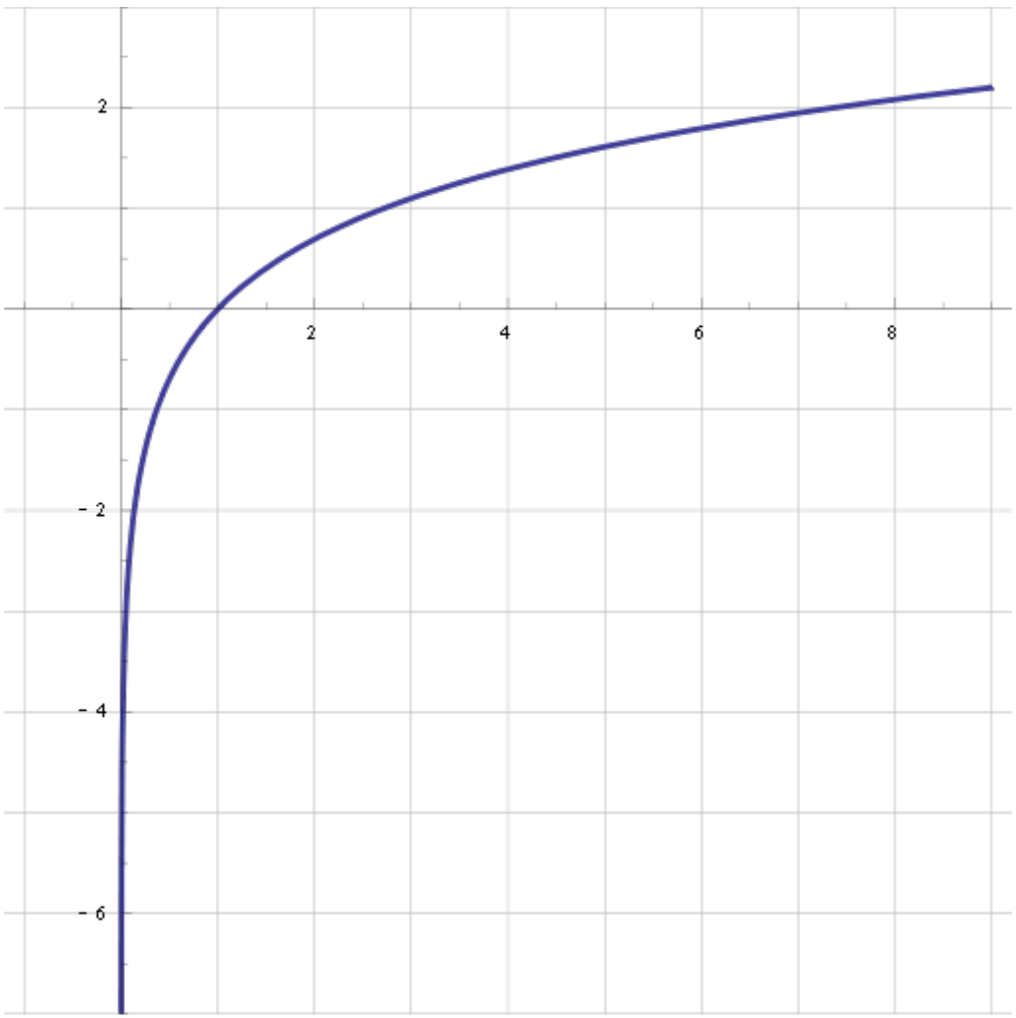
$$f(\lambda\boldsymbol{\theta} + (1-\lambda)\boldsymbol{\theta}') \le \lambda f(\boldsymbol{\theta}) + (1-\lambda) f(\boldsymbol{\theta}')$$

Convexity & Jensen's Inequality



Concavity & Jensen's Inequality

$$\ln(\mathbb{E}[X]) \geq \mathbb{E}[\ln(X)]$$



Log Partition Function

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \qquad Z(\boldsymbol{\theta}) = \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x}$$

$$= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] \qquad A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$$

- Derivatives of log partition function have an intuitive form:

$$\nabla_\theta A(\theta) = \mathbb{E}_\theta[\phi(x)]$$

$$\nabla_\theta^2 A(\theta) = \mathrm{Cov}_\theta[\phi(x)] = \mathbb{E}_\theta[\phi(x)\phi(x)^T] - \mathbb{E}_\theta[\phi(x)]\mathbb{E}_\theta[\phi(x)]^T$$

- 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 Θ 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:

$$H(p) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

discrete entropy
(concave, non-negative)

$$H(p) = -\int_{\mathcal{X}} p(x) \log p(x) \, dx$$

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:

$$D(p \,||\, q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, dx \qquad\qquad D(p \,||\, q) \;\geq\; 0$$

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

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

$$I(p_{xy}) \triangleq D(p_{xy} \,||\, p_x p_y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_{xy}(x,y) \log \frac{p_{xy}(x,y)}{p_x(x) p_y(y)} \, dy \, dx$$
$$= H(p_x) + H(p_y) - H(p_{xy})$$

Learning in Exponential Families

$$
\begin{aligned}
p(\mathbf{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] & Z(\boldsymbol{\theta}) &= \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x} \\
&= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] & A(\boldsymbol{\theta}) &= \log Z(\boldsymbol{\theta})
\end{aligned}
$$

- Given any *target* probability distribution $\tilde{p}(x)$, the closest exponential family distribution *matches moments*:

$$\hat{\theta} = \arg\min_{\theta} D(\tilde{p} \,||\, p_\theta) \quad \Longleftrightarrow \quad \mathbb{E}_{\hat{\theta}}[\phi_a(x)] = \int_{\mathcal{X}} \phi_a(x)\, \tilde{p}(x)\, dx$$

- 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:

$$\hat{\theta} = \arg\max_{\theta} \sum_{\ell=1}^{L} \log p(x^{(\ell)} \mid \theta) = \arg\min_{\theta} D(\tilde{p} \,||\, p_\theta) \quad \Longleftrightarrow \quad \mathbb{E}_{\hat{\theta}}[\phi_a(x)] = \frac{1}{L} \sum_{\ell=1}^{L} \phi_a(x^{(\ell)})$$

Maximum Entropy Models

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \qquad Z(\boldsymbol{\theta}) = \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x}$$

$$= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] \qquad A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$$

- 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

$$\mathbb{E}_{\hat{\theta}}[\phi_a(x)] = \mu_a \qquad h(x) = 1$$

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:

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

$$p(\bar{x} \mid x^{(1)}, \dots, x^{(L)}, \lambda) = \int_\Theta p(\bar{x} \mid \theta)\, p(\theta \mid x^{(1)}, \dots, x^{(L)}, \lambda)\ d\theta$$

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

$$\boldsymbol{\phi}(x^{(1)}, \dots, x^{(L)}) \triangleq \left\{ \frac{1}{L} \sum_{\ell=1}^{L} \phi_a(x^{(\ell)}) \;\middle|\; a \in \mathcal{A} \right\} \tag{2.24}$$

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*

$$p(\theta \mid x^{(1)}, \dots, x^{(L)}, \lambda) = p(\theta \mid \boldsymbol{\phi}(x^{(1)}, \dots, x^{(L)})\,, L, \lambda) \tag{2.25}$$

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

$$p(\bar{x} \mid x^{(1)}, \dots, x^{(L)}, \lambda) = p(\bar{x} \mid \boldsymbol{\phi}(x^{(1)}, \dots, x^{(L)})\,, L, \lambda) \tag{2.26}$$

Learning with Conjugate Priors

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

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

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

Proposition 2.1.4. *Let $p(x \mid \theta)$ denote an exponential family with canonical parameters θ, 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:*

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

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

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

$$\log p(x^{(1)}, \ldots, x^{(L)} \mid \lambda) = \Omega\big(\bar{\lambda}\big) - \Omega(\lambda) + \sum_{\ell=1}^{L} \log \nu(x^{(\ell)}) \tag{2.33}$$

Learning with Conjugate Priors

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

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

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

Proposition 2.1.4. *Let $p(x \mid \theta)$ denote an exponential family with canonical parameters θ, 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:*

$$p(\theta \mid x^{(1)}, \ldots, x^{(L)}, \lambda) = p\big(\theta \mid \bar{\lambda}\big) \qquad (2.31)$$

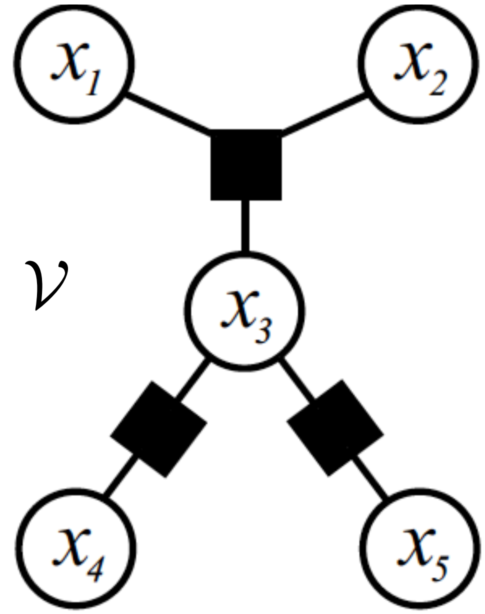
$$\bar{\lambda}_0 = \lambda_0 + L \qquad \bar{\lambda}_a = \frac{\lambda_0 \lambda_a + \sum_{\ell=1}^{L} \phi_a(x^{(\ell)})}{\lambda_0 + L} \qquad a \in \mathcal{A} \qquad (2.32)$$

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

- Prior expected values λ_a of the *d* sufficient statistics
- A measure of confidence in those prior expectations, expressed as a positive number of *pseudo-observations* λ_0

Factor Graphs & Exponential Families

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



$\mathcal{F} \longrightarrow$ set of hyperedges linking subsets of nodes $f \subseteq \mathcal{V}$

$\mathcal{V} \longrightarrow$ set of N nodes or vertices, $\{1, 2, \ldots, N\}$

$Z \longrightarrow$ 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:

$$\theta_f \triangleq \{\theta_{fa} \mid a \in \mathcal{A}_f\}$$

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