CS242: Probabilistic Graphical Models

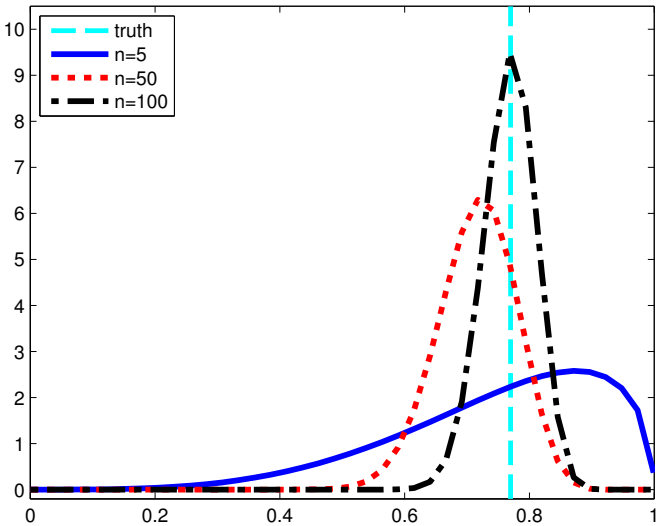
Lecture 8A: Conjugate Priors & Bayesian Learning

Professor Erik Sudderth
Brown University Computer Science
November 1, 2016

Some figures and materials courtesy Iain Murray,
Markov Chain Monte Carlo, MLSS 2009
http://homepages.inf.ed.ac.uk/imurray2/teaching/09mlss/

CS242: Lecture 8A Outline

- Beta priors for Bernoulli distributions
- Dirichlet priors for categorical distributions
- Conjugate priors for exponential family distributions



Bayesian Parameter Estimation

- Suppose I have *L* independent observations sampled from some unknown probability distribution: $x = \{x^{(1)}, \ldots, x^{(L)}\}$
- We have a *likelihood model* with unknown parameters:

$$p(x \mid \theta) = \prod_{\ell=1}^{L} p(x^{(\ell)} \mid \theta)$$

- We have a *prior distribution* on parameters (possible models):

$$p(\theta)$$

- *Posterior distribution* on parameters, given data:

$$p(\theta \mid x) = \frac{1}{p(x)} p(\theta) \prod_{\ell=1}^{L} p(x^{(\ell)} \mid \theta)$$

Bayesian Parameter Estimation

- **Maximum a Posteriori (MAP)** parameter estimate:
  Choose the parameters with largest posterior probability.

$$\hat{\theta} = \arg\max_{\theta} \; p(\theta \mid x) = \arg\max_{\theta} \; p(\theta) \prod_{\ell=1}^{L} p(x^{(\ell)} \mid \theta)$$

- **Conditional Expectation** parameter estimate:
  Set the parameters to the mean of the posterior distribution.

$$\hat{\theta} = E[\theta \mid x] = \int \theta p(\theta \mid x) \; d\theta$$

- *Posterior distribution* on parameters, given data:

$$p(\theta \mid x) = \frac{1}{p(x)} p(\theta) \prod_{\ell=1}^{L} p(x^{(\ell)} \mid \theta)$$

Bayesian Parameter Estimation

- **Maximum a Posteriori (MAP)** parameter estimate:
  Choose the parameters with largest posterior probability.

$$\hat{\theta} = \arg\max_{\theta} \; p(\theta \mid x) = \arg\max_{\theta} \; p(\theta) \prod_{\ell=1}^{L} p(x^{(\ell)} \mid \theta)$$

- **Conditional Expectation** parameter estimate:
  Set the parameters to the mean of the posterior distribution.

$$\hat{\theta} = E[\theta \mid x] = \int \theta p(\theta \mid x) \; d\theta$$

- Both estimators pick parameters with high posterior probability
- Choice of estimator can be formalized via *decision theory*
  (conditional expectation minimizes expected squared error)

Bayesian Learning of Binary Distributions

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

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

$$p(x^{(1)}, \ldots, x^{(L)} \mid \theta) = \theta^{N_1} (1-\theta)^{N_0}$$

$$N_1 = \sum_{\ell=1}^{L} x^{(\ell)} \qquad N_0 = \sum_{\ell=1}^{L} (1 - x^{(\ell)}) = L - N_1$$

Uniform Prior Distribution:

$$p(\theta) = 1 \quad \text{for} \quad 0 \le \theta \le 1.$$

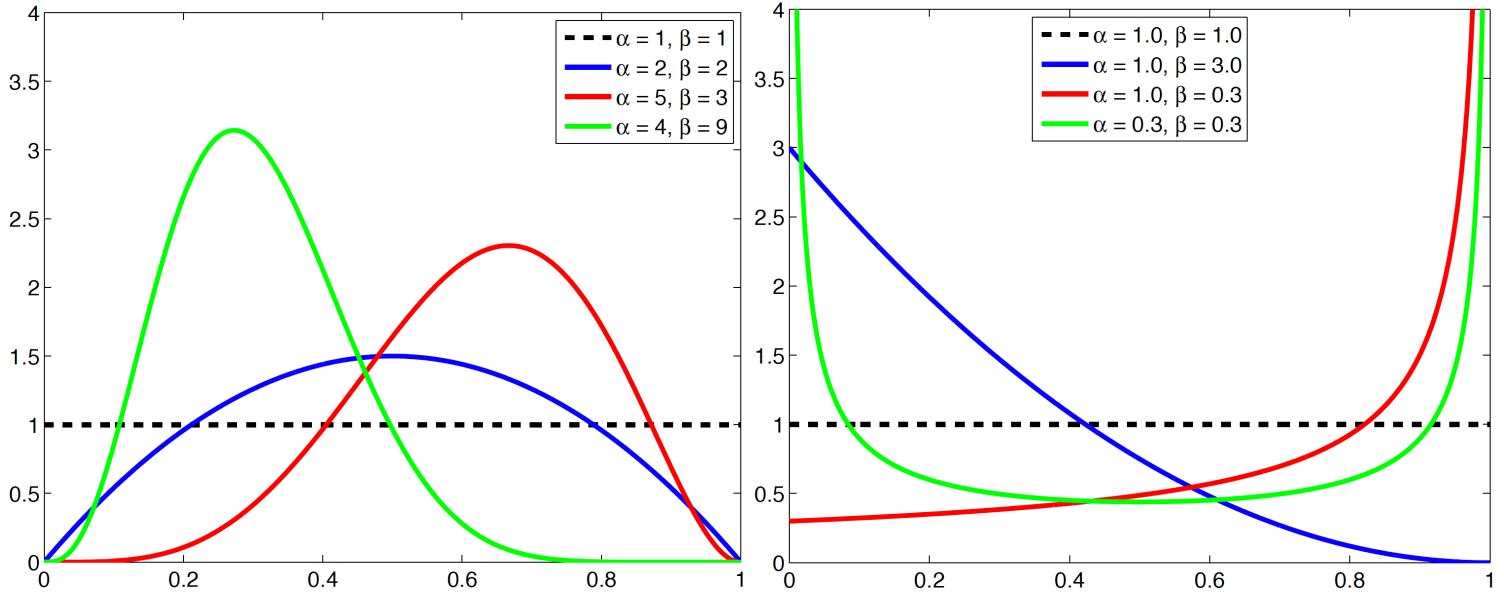
Posterior Distribution:

$$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{p(x)} = \frac{1}{p(x)} \theta^{N_1} (1-\theta)^{N_0} \quad \text{for} \quad 0 \le \theta \le 1.$$

$$p(x) = \int_0^1 p(x \mid \theta) p(\theta) \, d\theta.$$

What is this distribution?

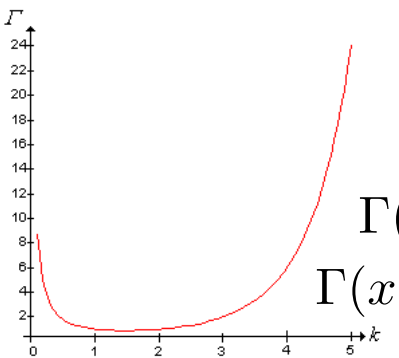
Beta Distributions



Beta probability density function: $\theta \in [0, 1]$

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

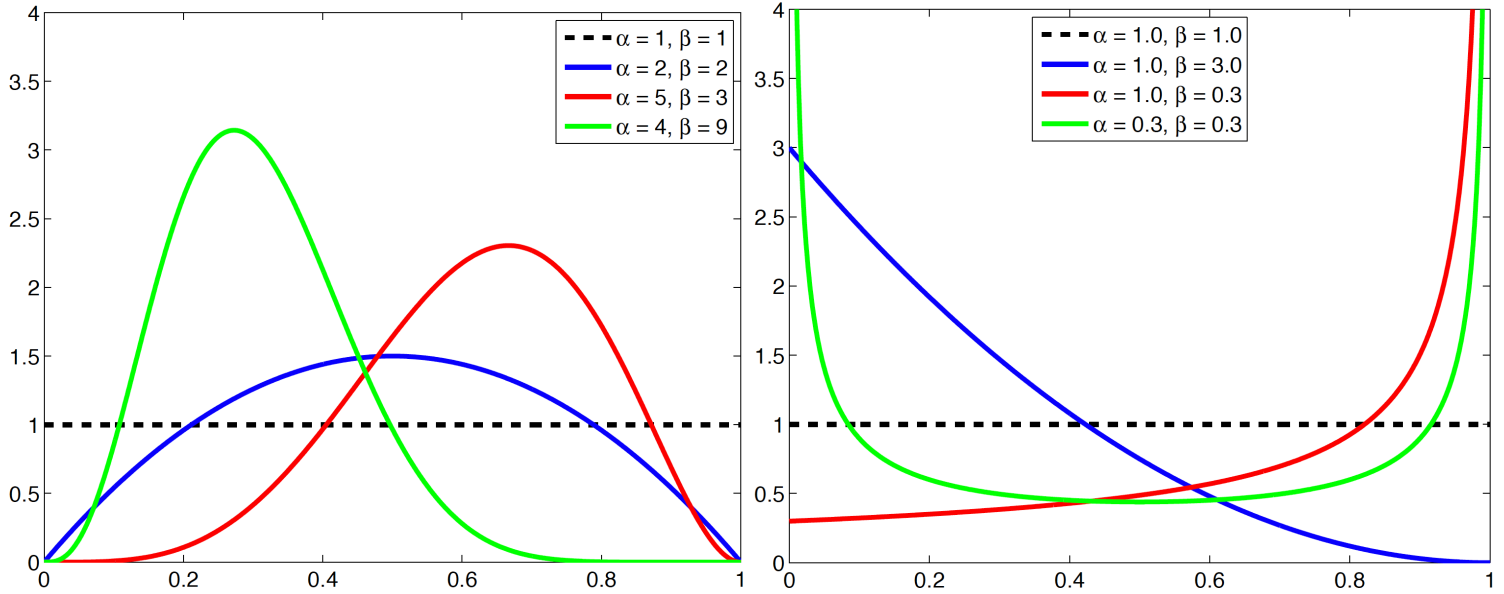
$$B(\alpha, \beta) \triangleq \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad \alpha, \beta > 0$$



$$\Gamma(k) = (k-1)!$$

$$\Gamma(x+1) = x\Gamma(x)$$

Beta Distributions

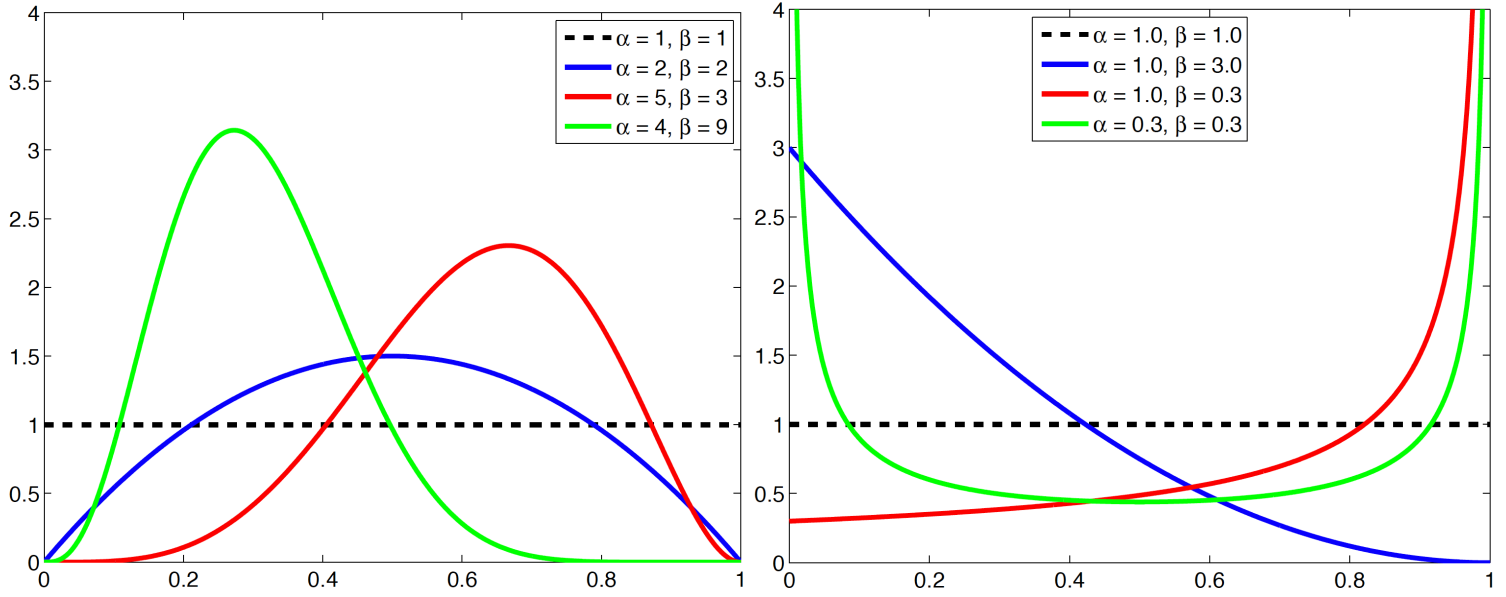


Beta probability density function: $\theta \in [0, 1]$

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

$$\mathbb{E}[\theta] = \frac{\alpha}{\alpha + \beta} \qquad \mathbb{V}[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Beta Distributions



There is a *unique mode* assuming $\alpha > 1, \beta > 1$

$$\text{Mode}[\theta] = \arg\max_{\theta \in [0,1]} \text{Beta}(\theta \mid \alpha, \beta) = \frac{\alpha - 1}{(\alpha - 1) + (\beta - 1)}$$

Otherwise the mode may be degenerate $(\theta = 0 \text{ or } 1)$ or be undefined.

Bayesian Learning of Binary Distributions

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

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

$$p(x^{(1)}, \ldots, x^{(L)} \mid \theta) = \theta^{N_1} (1-\theta)^{N_0}$$

$$N_1 = \sum_{\ell=1}^{L} x^{(\ell)} \qquad N_0 = \sum_{\ell=1}^{L} (1 - x^{(\ell)}) = L - N_1$$

Beta Prior Distribution: $p(\theta) = \mathrm{Beta}(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

Beta Posterior Distribution:

$$p(\theta \mid x) \propto \theta^{N_1+\alpha-1}(1-\theta)^{N_0+\beta-1} \propto \mathrm{Beta}(\theta \mid N_1+\alpha, N_0+\beta)$$

Prior is conjugate to likelihood because posterior distribution in same family.

Bayesian Learning of Binary Distributions

Recommended Estimator: Posterior mean

$$\hat{\theta} = \mathbb{E}[\theta \mid x] = \frac{N_1 + \alpha}{N_1 + \alpha + N_0 + \beta}$$

With uniform prior:

$$\hat{\theta} = \mathbb{E}[\theta \mid x] = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

"add one" to observed counts

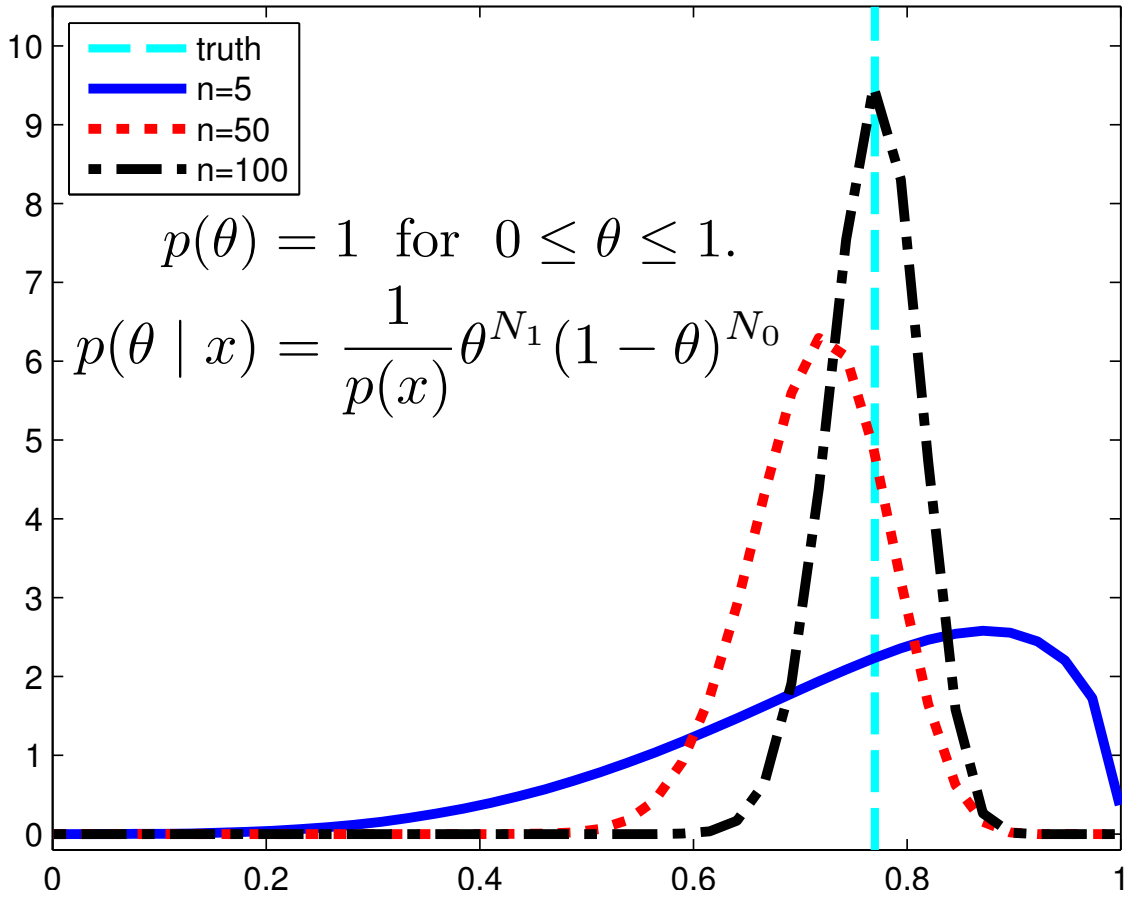
Beta Prior Distribution: $p(\theta) = \text{Beta}(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

Beta Posterior Distribution:

$$p(\theta \mid x) \propto \theta^{N_1+\alpha-1}(1-\theta)^{N_0+\beta-1} \propto \text{Beta}(\theta \mid N_1 + \alpha, N_0 + \beta)$$

Prior is conjugate to likelihood because posterior distribution in same family.

A Sequence of Beta Posteriors



Estimators for Beta Posteriors

Prior: $p(\theta) = \text{Beta}(\theta \mid \alpha, \beta)$

$p(\theta) = \text{Beta}(\theta \mid 1, 1) = 1$

MMSE:

$$\hat{\theta} = \mathbb{E}[\theta \mid x] = \frac{N_1 + \alpha}{N + \alpha + \beta}$$

$$\hat{\theta} = \mathbb{E}[\theta \mid x] = \frac{N_1 + 1}{N + 2}$$

MAP:

$$\hat{\theta} = \arg\max_\theta p(\theta \mid x) = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2}$$

assuming $N_1 + \alpha > 1, N_0 + \beta > 1$

$$\hat{\theta} = \arg\max_\theta \; p(\theta \mid x) = \frac{N_1}{N}$$

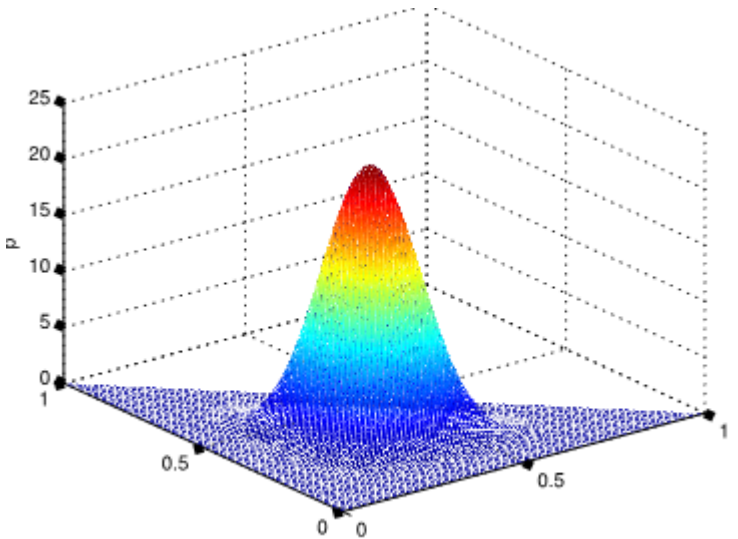
equivalent to maximum likelihood (ML)

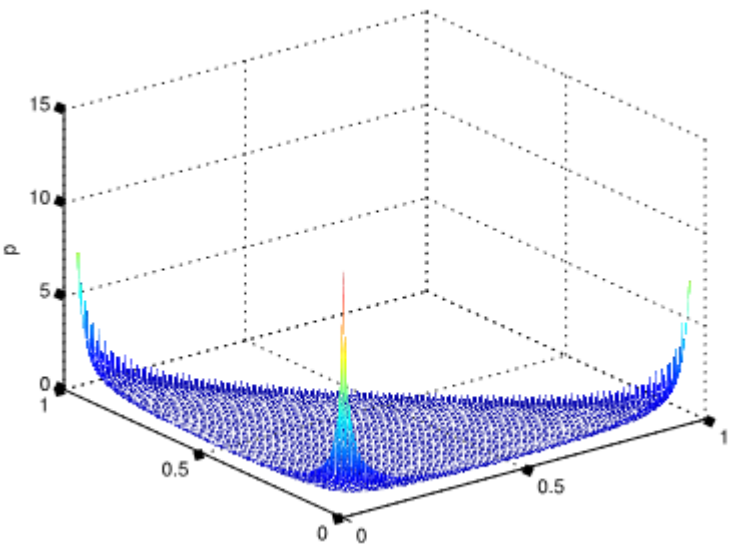
$$p(\theta \mid x) = \text{Beta}(\theta \mid N_1 + \alpha, N_0 + \beta)$$

$$N_1 = \sum_{i=1}^{N} x_i$$
$$N_0 = N - N_1$$

CS242: Lecture 8A Outline

- Beta priors for Bernoulli distributions
- Dirichlet priors for categorical distributions
- Conjugate priors for exponential family distributions





Dirichlet Probability Distributions

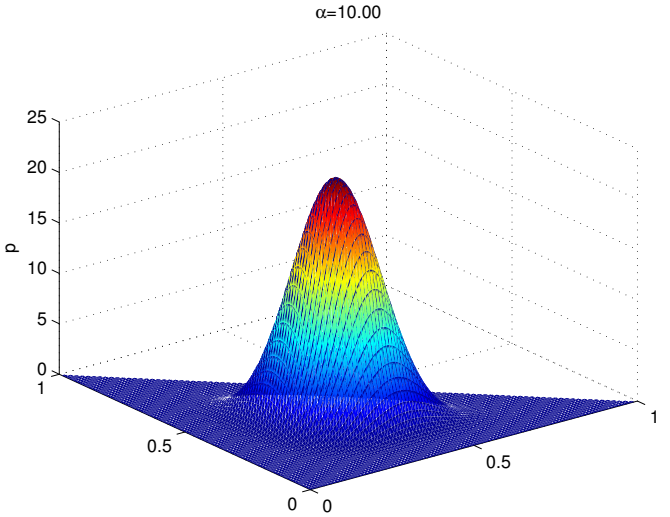
Simplex: Set of possible categorical distributions

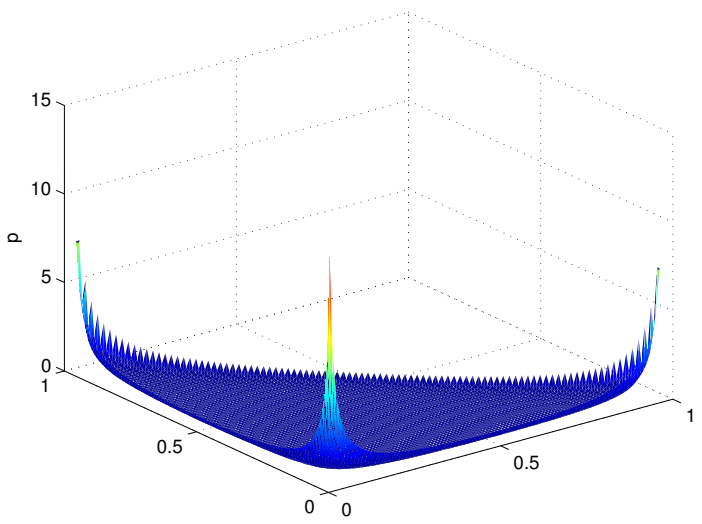
$$\Theta = \{\theta : 0 \leq \theta_k \leq 1, \sum_{k=1}^{K} \theta_K = 1\}$$

Dirichlet: Probability distribution on simplex

$$\mathrm{Dir}(\theta \mid \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \alpha_k > 0$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$



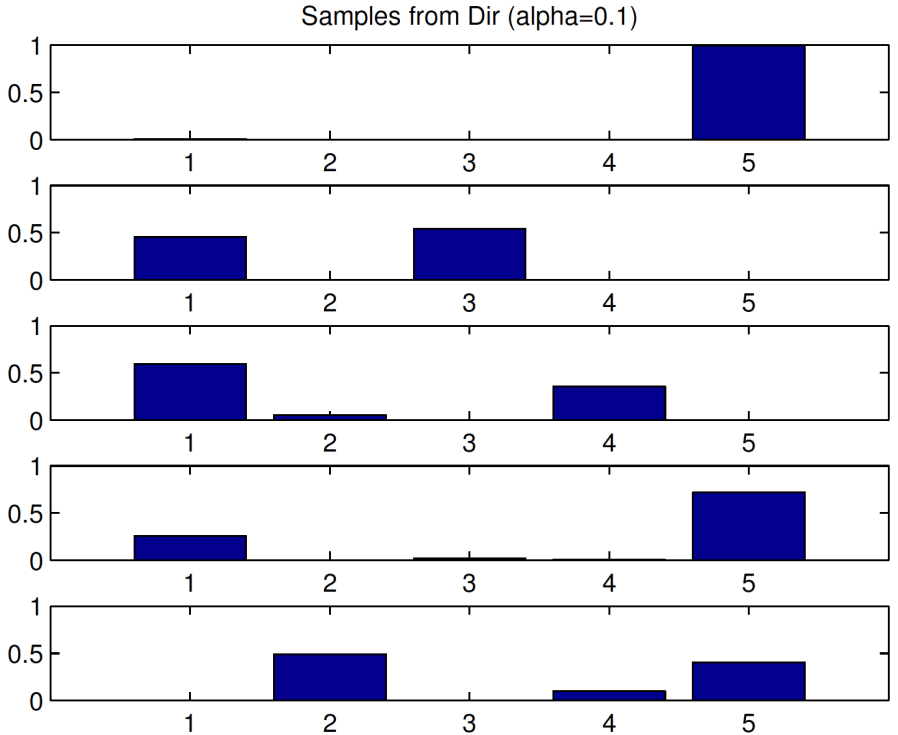


Normalizer: $B(\alpha) \triangleq \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)}$

Mean: $\mathbb{E}[\theta_k] = \frac{\alpha_k}{\alpha_0}$

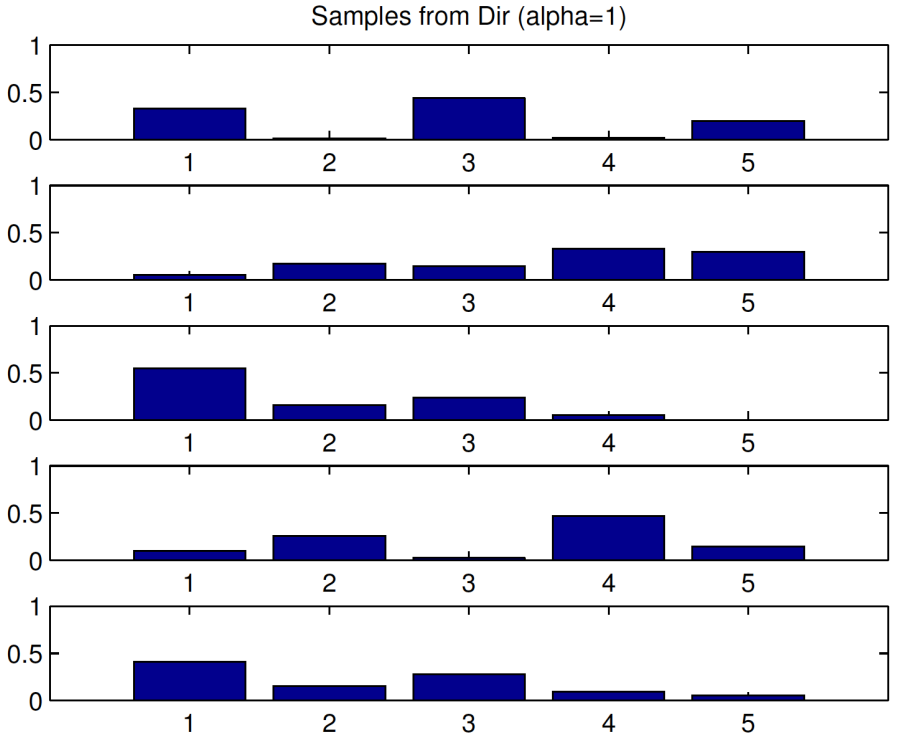
Variance: Inversely proportional to α_0

Samples from Dirichlet Prior Distributions



$$\mathrm{Dir}(\theta \mid 0.1, 0.1, 0.1, 0.1, 0.1)$$

Mean weights classes equally, but biased towards sparse distributions



$$\mathrm{Dir}(\theta \mid 1.0, 1.0, 1.0, 1.0, 1.0)$$

Uniform distribution on probability simplex

Bayes Learning of Categorical Distributions

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

$$\text{Cat}(x \mid \theta) = \prod_{k=1}^{K} \theta_k^{x_k} \qquad x_k \in \{0, 1\}, \ \sum_{k=1}^{K} x_k = 1.$$

$$p(x^{(1)}, \ldots, x^{(L)} \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k} \qquad N_k = \sum_{\ell=1}^{L} x_k^{(\ell)}$$

Dirichlet Prior Distribution:

$$p(\theta) = \text{Dir}(\theta \mid \alpha) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Dirichlet Posterior Distribution:

$$p(\theta \mid x) \propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} \propto \text{Dir}(\theta \mid N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

Prior is conjugate to likelihood because posterior distribution in same family.

Bayes Learning of Categorical Distributions

Recommended Estimator: Posterior mean

$$\hat{\theta}_k = \mathbb{E}[\theta_k \mid x] = \frac{N_k + \alpha_k}{L + \alpha_0}$$

With uniform prior:

$$\hat{\theta}_k = \mathbb{E}[\theta_k \mid x] = \frac{N_k + 1}{L + K}$$

Dirichlet Prior Distribution:

$$p(\theta) = \mathrm{Dir}(\theta \mid \alpha) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Dirichlet Posterior Distribution:

$$p(\theta \mid x) \propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} \propto \mathrm{Dir}(\theta \mid N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

Prior is conjugate to likelihood because posterior distribution in same family.

Possible Dirichlet Priors & Posteriors

Prior:

$$p(\theta) = \mathrm{Dir}(\theta \mid \alpha) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

- *K* parameters (positive numbers)
- *K-1* degrees of freedom define mean:

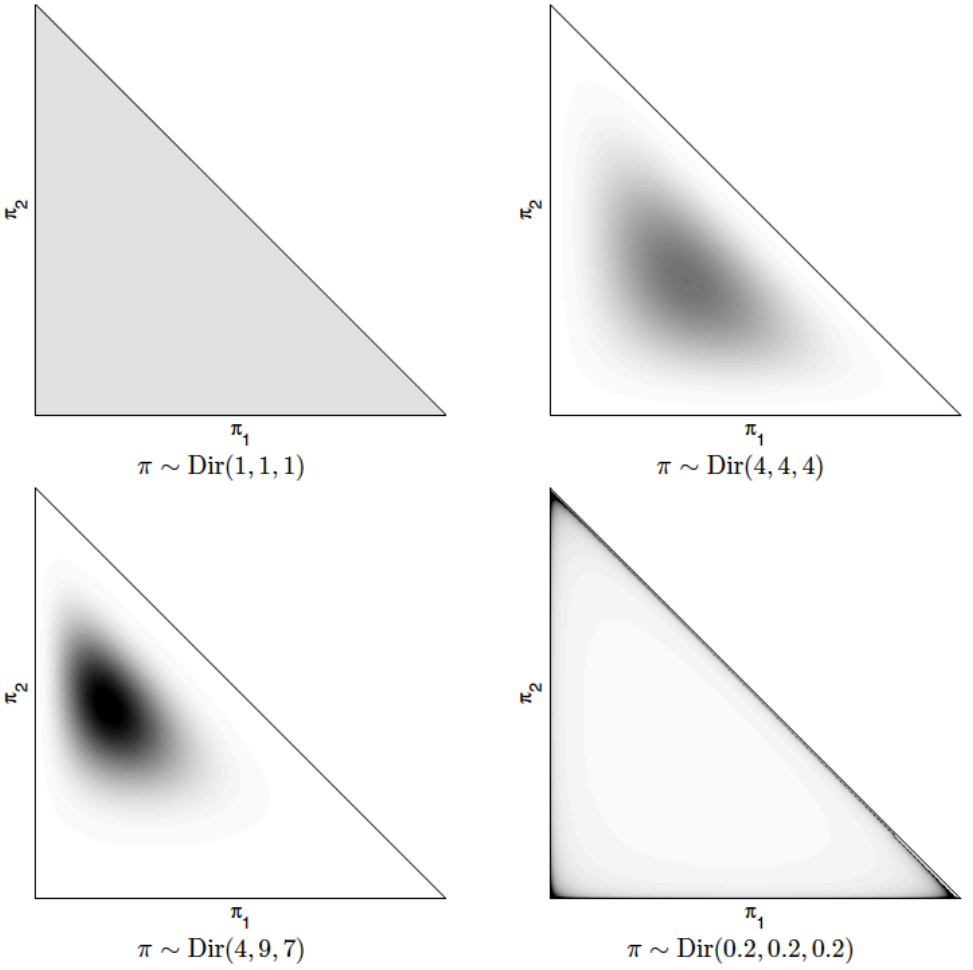
$$\mathbb{E}[\theta_k] = \frac{\alpha_k}{\alpha_0} \qquad \alpha_0 = \sum_{k=1}^{K} \alpha_k$$

- Variance proportional to $1/\alpha_0$
- Favors *sparsity* as $\alpha_0 \to 0$

Posterior:

$$p(\theta \mid x) \propto \mathrm{Dir}(\theta \mid N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

- Posterior mean is weighted average of prior mean and observed counts
- As *N* grows, posterior variance shrinks



Estimators for Dirichlet Posteriors

Prior: $p(\theta) = \mathrm{Dir}(\theta \mid \alpha_1, \ldots, \alpha_K)$

$\alpha_0 = \sum_{k=1}^{K} \alpha_k$

$p(\theta) = \mathrm{Dir}(\theta \mid 1, \ldots, 1) = 1$

MMSE:

$$\hat{\theta}_k = \mathbb{E}[\theta_k \mid x] = \frac{N_k + \alpha_k}{N + \alpha_0}$$

$$\hat{\theta}_k = \mathbb{E}[\theta_k \mid x] = \frac{N_k + 1}{N + K}$$

MAP:

$$\hat{\theta}_k = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}$$

$$\hat{\theta} = \frac{N_k}{N}$$

assuming $N_k + \alpha_k > 1$ for all k

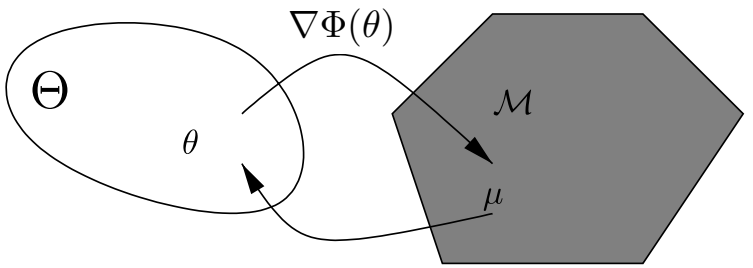
equivalent to maximum likelihood (ML)

$$p(\theta \mid x) = \mathrm{Dir}(\theta \mid N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

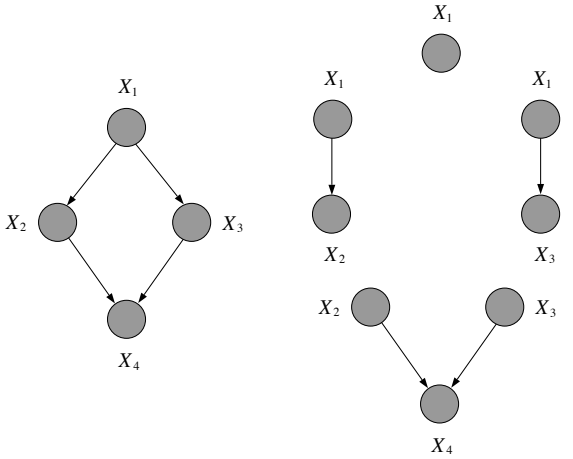
$$N_k = \sum_{n=1}^{N} x_{nk}$$

CS242: Lecture 8A Outline

- Beta priors for Bernoulli distributions
- Dirichlet priors for categorical distributions
- Conjugate priors for exponential family distributions



Learning Directed Graphical Models



$$p(x) = \prod_{i \in \mathcal{V}} p(x_i \mid x_{\Gamma(i)}, \theta_i)$$

Intuition: *Must learn a good predictive model of each node, given its parent nodes*

- Directed factorization allows likelihood to locally decompose:

$$p(x \mid \theta) = p(x_1 \mid \theta_1)p(x_2 \mid x_1, \theta_2)p(x_3 \mid x_1, \theta_3)p(x_4 \mid x_2, x_3, \theta_4)$$

$$\log p(x \mid \theta) = \log p(x_1 \mid \theta_1) + \log p(x_2 \mid x_1, \theta_2) + \log p(x_3 \mid x_1, \theta_3) + \log p(x_4 \mid x_2, x_3, \theta_4)$$

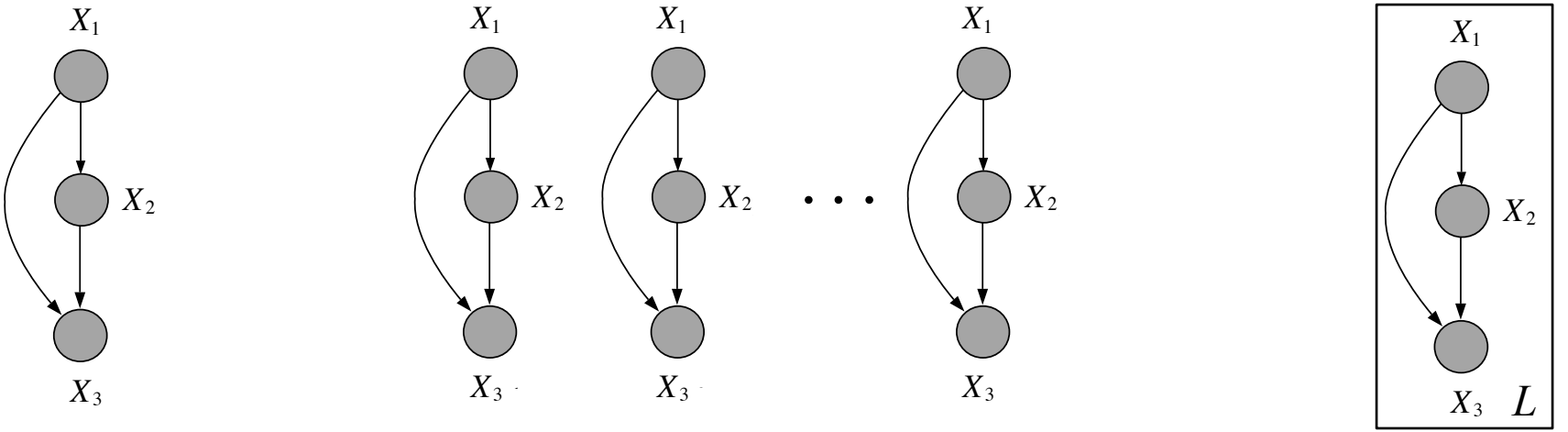
- We often assume a similarly factorized (meta-independent) prior:

$$p(\theta) = p(\theta_1)p(\theta_2)p(\theta_3)p(\theta_4)$$

$$\log p(\theta) = \log p(\theta_1) + \log p(\theta_2) + \log p(\theta_3) + \log p(\theta_4)$$

- We thus have *independent* Bayesian learning problems at each node

Bayesian Learning with Complete Data



- Directed graph encodes statistical structure of single training examples:

$$p(x \mid \theta) = \prod_{\ell=1}^{L} \prod_{i=1}^{N} p(x_i^{(\ell)} \mid x_{\Gamma(i)}^{(\ell)}, \theta_i)$$

- Given completely observed training data, nodes have independent posteriors:

$$p(\theta \mid x) \propto p(\theta) p(x \mid \theta) \propto \prod_{i=1}^{N} \left[p(\theta_i) \prod_{\ell=1}^{L} p(x_i^{(\ell)} \mid x_{\Gamma(i)}^{(\ell)}, \theta_i) \right]$$

Bayesian Learning with Complete Data

- For discrete variables with no parents, parameters define some *Bernoulli/categorical distribution with a beta/Dirichlet conjugate prior*
- More generally, there are multiple categorical distributions per node, one for every *combination* of parent variables
  - ➢ Learning objective decomposes into multiple terms, one for subset of training data with each parent configuration
  - ➢ Apply independent Bayesian learning to each
- How can we generalize to continuous variables? *Exponential families.*

- Given completely observed training data, nodes have independent posteriors:

$$p(\theta \mid x) \propto p(\theta)p(x \mid \theta) \propto \prod_{i=1}^{N} \left[p(\theta_i) \prod_{\ell=1}^{L} p(x_i^{(\ell)} \mid x_{\Gamma(i)}^{(\ell)}, \theta_i) \right]$$

Exponential Families of Distributions

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \nu(x) \exp\{\theta^T \phi(x)\} \qquad Z(\theta) = \int_{\mathcal{X}} \nu(x) \exp\{\theta^T \phi(x)\} \, dx$$

$$= \nu(x) \exp\{\theta^T \phi(x) - \Phi(\theta)\} \qquad \Phi(\theta) = \log Z(\theta)$$

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

$\theta \in \Theta \subseteq \mathbb{R}^d$ ⟶ unknown vector of *natural parameters*, determine particular distribution in this family

$Z(\theta) > 0$ ⟶ normalization constant or *partition function* (for discrete variables, integral becomes sum)

$\nu(x) > 0$ ⟶ *reference measure* independent of parameters (for many models, we simply have $\nu(x) = 1$)

$\Theta = \{\theta \in \mathbb{R}^d \mid Z(\theta) < \infty\}$ *ensures we construct a valid distribution*

ML Estimation for Exponential Families

$$\log p(x^{(\ell)} \mid \theta) = \log \nu(x^{(\ell)}) + \theta^T \phi(x^{(\ell)}) - \Phi(\theta)$$

- Given *L* observations, the *log-likelihood function* equals:

$$L(\theta) = C + \left[\sum_{\ell=1}^{L} \theta^T \phi(x^{(\ell)})\right] - L\Phi(\theta)$$

$$C = \sum_{\ell=1}^{L} \log \nu(x^{(\ell)})$$

- Note that the *negative log-likelihood function is convex!*
- Gradients of the log-likelihood function have a simple form:

$$\nabla L(\theta) = \left[\sum_{\ell=1}^{L} \phi(x^{(\ell)})\right] - L E_\theta[\phi(x)]$$

- At unique global optimum, gradient is 0: $E_\theta[\phi(x)] = \frac{1}{L} \sum_{\ell=1}^{L} \phi(x^{(\ell)})$

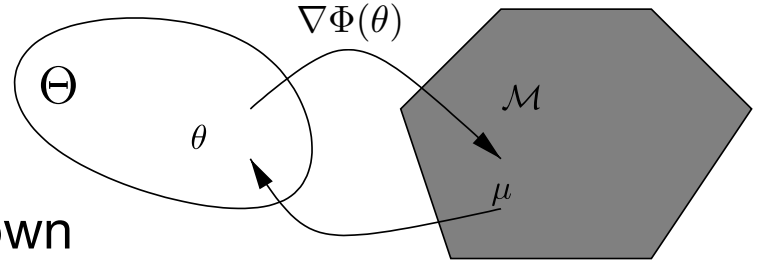
Exponential Families: Inference & Learning

$$\log p(x^{(\ell)} \mid \theta) = \log \nu(x^{(\ell)}) + \theta^T \phi(x^{(\ell)}) - \Phi(\theta)$$

- Canonical parameters & moments:

$$\Theta \triangleq \{\theta \in \mathbb{R}^d \mid \Phi(\theta) < +\infty\}$$

$$\mathcal{M} \triangleq \{\mu \in \mathbb{R}^d \mid \exists\, p \text{ such that } \mathbb{E}_p[\phi(x)] = \mu\}$$



- **Inference:** Find moments of model with known parameters (joint distribution). Computable from marginals!

$$\mu = \nabla_\theta \Phi(\theta) = \mathbb{E}_\theta[\phi(x)] = \sum_{\mathcal{X}} \phi(x) p(x \mid \theta)$$

- **Learning:** Find model parameters matching data moments

$$\mathbb{E}_{\hat{\theta}}[\phi(x)] = \hat{\mu}$$

This is the inverse of the mapping defining inference.

Maximum Likelihood (ML):

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

Parametric & Predictive Sufficiency

Posterior distribution:

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

Predictive likelihood:

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

Theorem 2.1.2. *Let $p(x \mid \theta)$ denote an exponential family with canonical parameters 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:*

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

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*

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

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

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

Sample moments & sample size are:

- *Sufficient to find posterior under **any** prior distribution.*
- *Sufficient to optimally **predict future** data.*

Learning with Conjugate Priors

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

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

Conjugate priors have matched functional forms.

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

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:*

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

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

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

Learning with Conjugate Priors

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

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

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:*

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

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

- Closed form for **posterior distribution**.

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

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

- Closed form for **marginal likelihood**.

Bayes Learning of Categorical Distributions

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

$$\text{Cat}(x \mid \theta) = \prod_{k=1}^{K} \theta_k^{x_k} \qquad x_k \in \{0, 1\}, \ \sum_{k=1}^{K} x_k = 1.$$

$$p(x^{(1)}, \ldots, x^{(L)} \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k} \qquad N_k = \sum_{\ell=1}^{L} x_k^{(\ell)}$$

Dirichlet Prior Distribution:

$$p(\theta) = \text{Dir}(\theta \mid \alpha) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Dirichlet Posterior Distribution:

$$p(\theta \mid x) \propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} \propto \text{Dir}(\theta \mid N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

Prior is conjugate to likelihood because posterior distribution in same family.

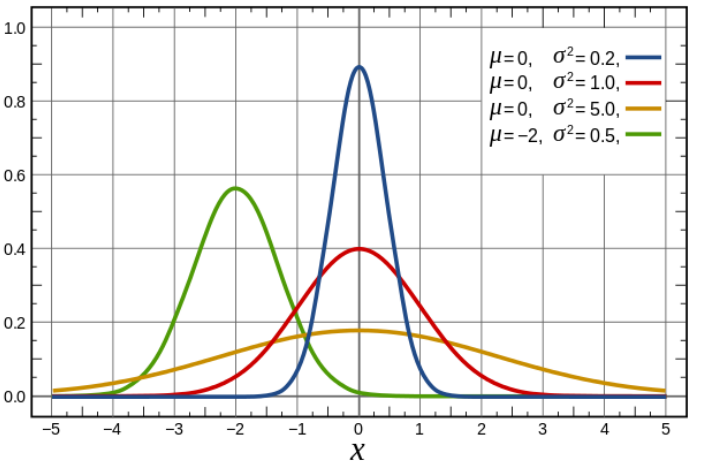
Normal (Gaussian) Random Variables

$$p(x) = \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

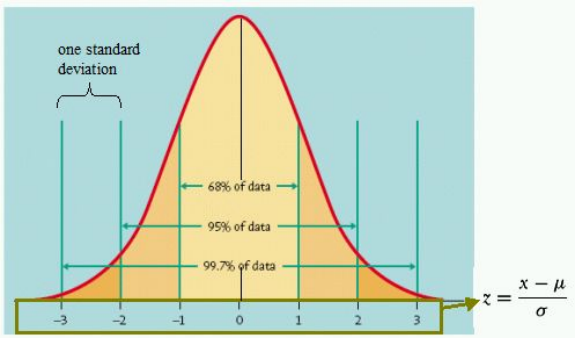
$$E[X] = \mu$$

$$\mathrm{Var}[X] = E[(X-\mu)^2] = \sigma^2$$

$$\sqrt{\mathrm{Var}[X]} = \sigma \text{ is the standard deviation}$$



- Standard deviations provide *confidence intervals*:



$$\int_{\mu-\sigma}^{\mu+\sigma} \mathcal{N}(x \mid \mu, \sigma^2)\, dx \approx 0.68$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \mathcal{N}(x \mid \mu, \sigma^2)\, dx \approx 0.95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} \mathcal{N}(x \mid \mu, \sigma^2)\, dx \approx 0.997$$

Bayesian Learning of Gaussians

Scalar Gaussian Likelihood Function:

$$p(x \mid \mu) = \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Assume variance σ^2 is known and fixed.

Gaussian Prior Distribution:

$$p(\mu) = \mathcal{N}(\mu \mid \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

Gaussian Posterior Distribution:

Prior is conjugate to likelihood.

$$p(\mu \mid x^{(1)}, \ldots, x^{(L)}) = \mathcal{N}(\mu \mid \mu_L, \sigma_L^2) \qquad \frac{1}{\sigma_L^2} = \frac{1}{\sigma_0^2} + \frac{L}{\sigma^2}$$

$$\mu_L = \frac{\sigma^2}{L\sigma_0^2 + \sigma^2}\mu_0 + \frac{L\sigma_0^2}{L\sigma_0^2 + \sigma^2}\bar{x}_L \qquad \bar{x}_L = \frac{1}{L}\sum_{\ell=1}^{L} x^{(\ell)}$$

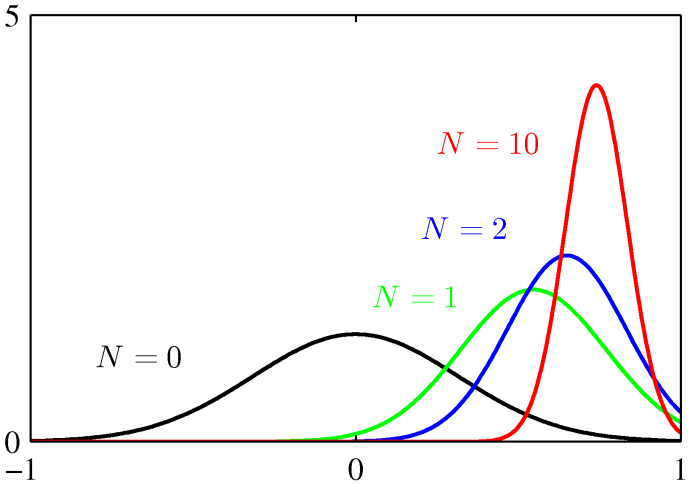
Posterior Mean versus Empirical Mean

Optimal Estimator:

Posterior mean,
Posterior mode, &
Posterior median

$$\hat{\mu} = \mu_L = E[\mu \mid x]$$

$$\mu_L \to \bar{x}_L \text{ as } L \to \infty$$



Example:

Posterior given varying amounts of data N=L

$$\mu = 0.8$$
$$\sigma^2 = 0.1$$

Gaussian Posterior Distribution:

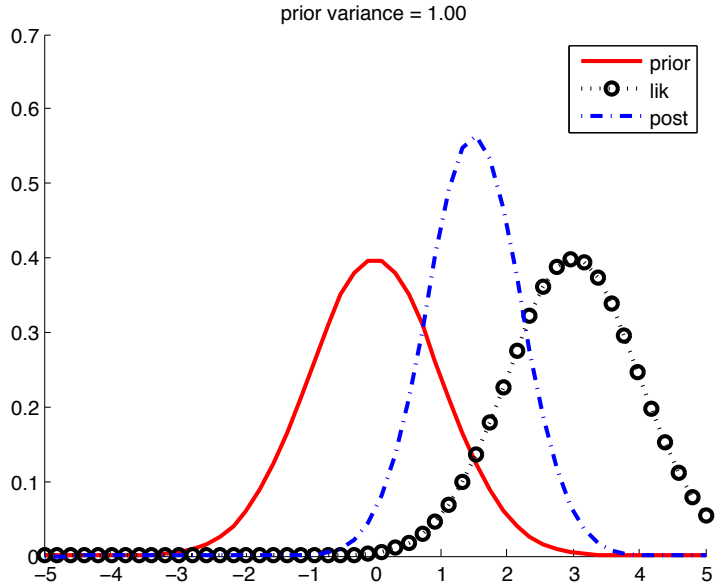
$$p(\mu \mid x^{(1)}, \ldots, x^{(L)}) = \mathcal{N}(\mu \mid \mu_L, \sigma_L^2)$$

$$\mu_L = \frac{\sigma^2}{L\sigma_0^2 + \sigma^2}\mu_0 + \frac{L\sigma_0^2}{L\sigma_0^2 + \sigma^2}\bar{x}_L$$

$$\frac{1}{\sigma_L^2} = \frac{1}{\sigma_0^2} + \frac{L}{\sigma^2}$$

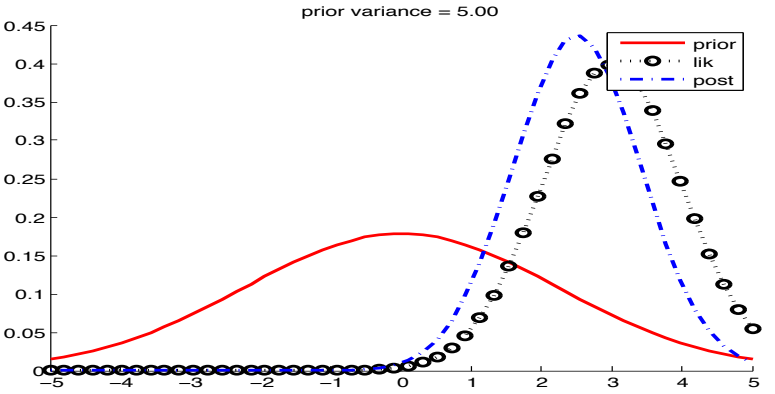
$$\bar{x}_L = \frac{1}{L}\sum_{\ell=1}^{L} x^{(\ell)}$$

Impact of Prior Variance



Example: *Posteriors given same single observation, for two different priors.*

$$\mu_L \to \bar{x}_L \text{ as } \sigma_0^2 \to \infty$$



Gaussian Posterior Distribution:

$$p(\mu \mid x^{(1)}, \ldots, x^{(L)}) = \mathcal{N}(\mu \mid \mu_L, \sigma_L^2)$$

$$\mu_L = \frac{\sigma^2}{L\sigma_0^2 + \sigma^2}\mu_0 + \frac{L\sigma_0^2}{L\sigma_0^2 + \sigma^2}\bar{x}_L$$

$$\frac{1}{\sigma_L^2} = \frac{1}{\sigma_0^2} + \frac{L}{\sigma^2}$$

$$\bar{x}_L = \frac{1}{L}\sum_{\ell=1}^{L} x^{(\ell)}$$

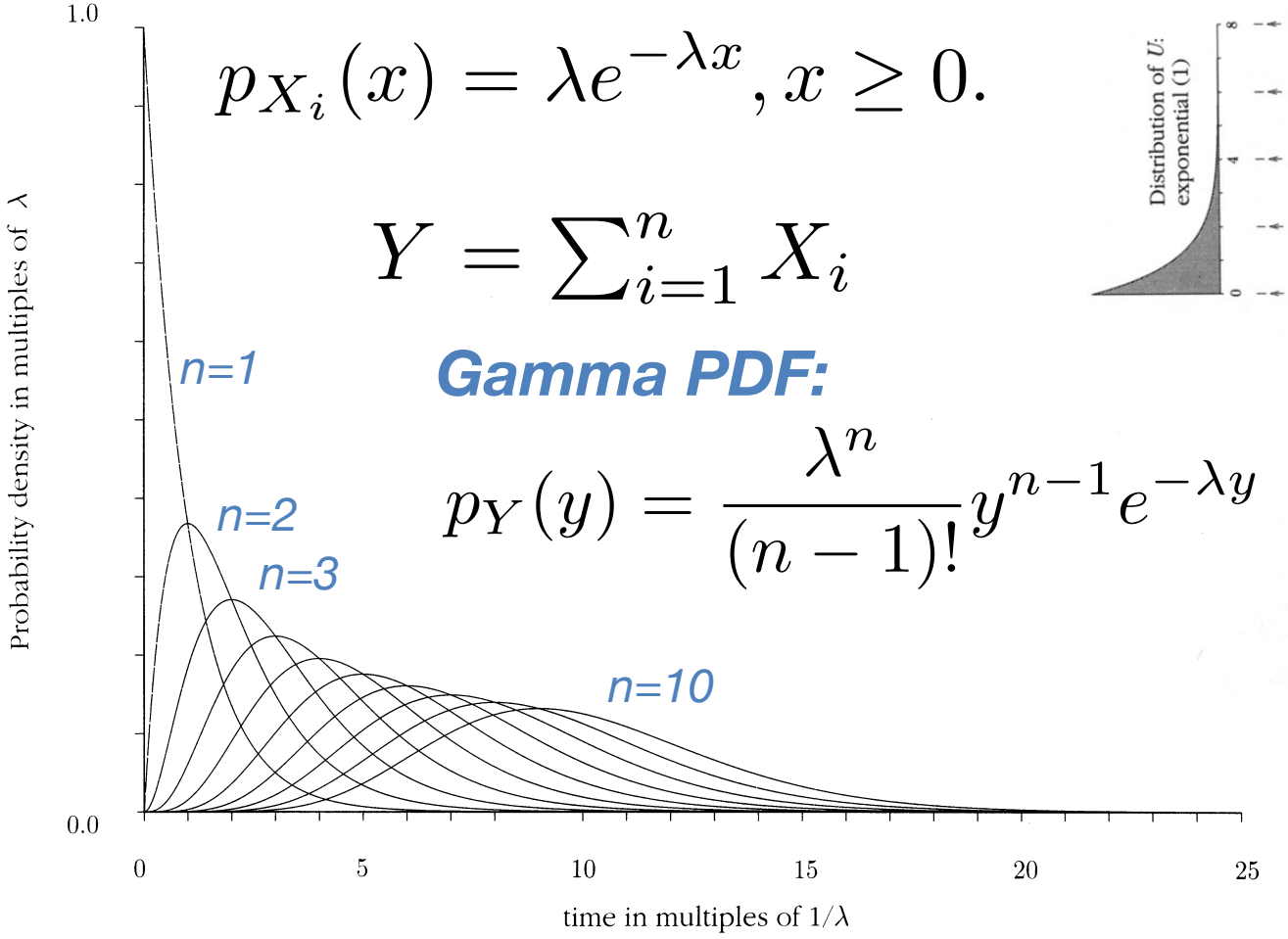
Aside: Sums of Exponential Variables

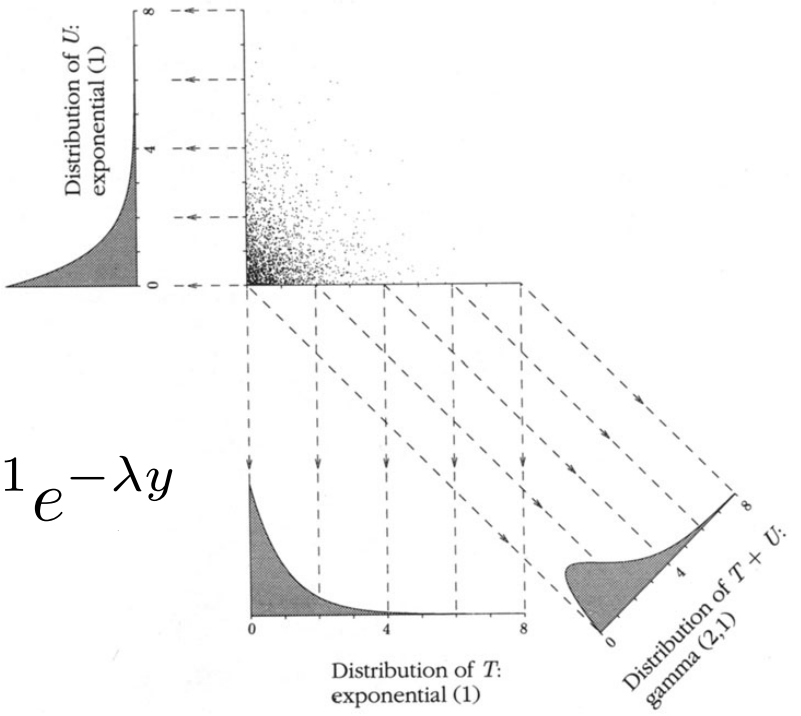
$$p_{X_i}(x) = \lambda e^{-\lambda x}, x \geq 0.$$

$$Y = \sum_{i=1}^{n} X_i$$

Gamma PDF:

$$p_Y(y) = \frac{\lambda^n}{(n-1)!} y^{n-1} e^{-\lambda y}$$





Approximately Gaussian for large n!

Bayesian Learning of Variances

Scalar Gaussian Likelihood Function:

$$p(x \mid \lambda) = \mathcal{N}(x \mid \mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left\{-\frac{\lambda}{2}(x-\mu)^2\right\}$$

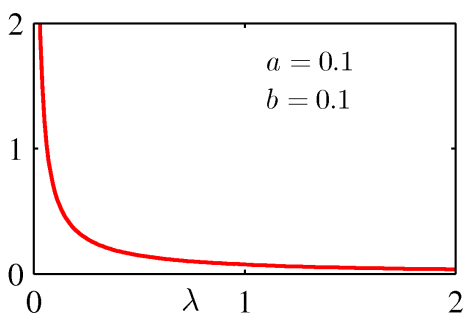
Assume mean μ is known and fixed.

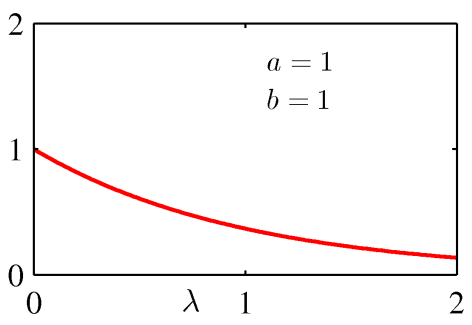
Gamma Prior Distribution on Inverse Variance (precision):

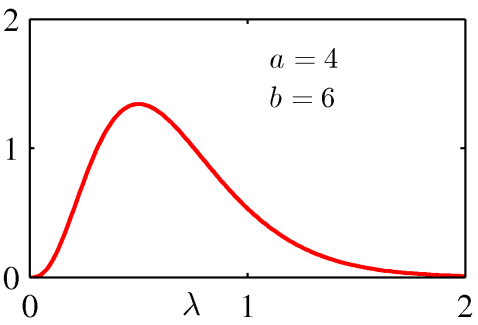
$$p(\lambda) = \text{Gamma}(\lambda \mid a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp\{-b_0 \lambda\}$$

$$E[\lambda] = \frac{a}{b}$$

$$\text{Var}[\lambda] = \frac{a}{b^2}$$







Bayesian Learning of Variances

Scalar Gaussian Likelihood Function:

$$p(x \mid \lambda) = \mathcal{N}(x \mid \mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left\{-\frac{\lambda}{2}(x-\mu)^2\right\}$$

Assume mean μ is known and fixed.

Gamma Prior Distribution on Inverse Variance (precision):

$$p(\lambda) = \text{Gamma}(\lambda \mid a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} \exp\{-b_0 \lambda\}$$

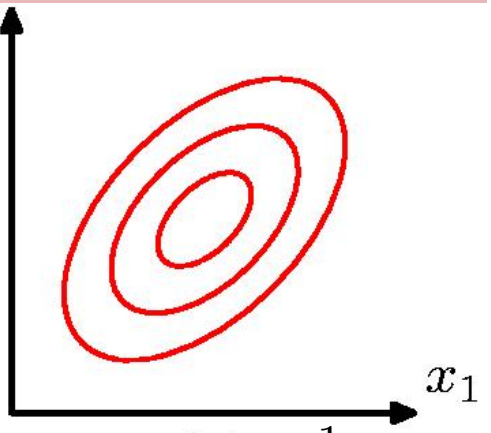
Gamma Posterior Distribution: *Prior is conjugate to likelihood.*

$$p(\lambda \mid x^{(1)}, \ldots, x^{(L)}) = \text{Gamma}(\lambda \mid a_L, b_L)$$

$$a_L = a_0 + \frac{L}{2} \qquad b_L = b_0 + \frac{L}{2}\bar{\sigma}^2 \qquad \bar{\sigma}^2 = \frac{1}{L}\sum_{\ell=1}^{L}(x^{(\ell)} - \mu)^2$$

Multivariate Gaussian Distribution

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{N/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$



Moment Parameterization: Mean & Covariance

$$\mu = \mathbb{E}[x] \in \mathbb{R}^{N \times 1}$$

$$\Sigma = \mathbb{E}[(x-\mu)(x-\mu)^T] = \mathbb{E}[xx^T] - \mu\mu^T \in \mathbb{R}^{N \times N}$$

For simplicity, assume covariance positive definite & invertible.

$$\Lambda = \Sigma^{-1}$$
$$\vartheta = \Sigma^{-1}\mu$$

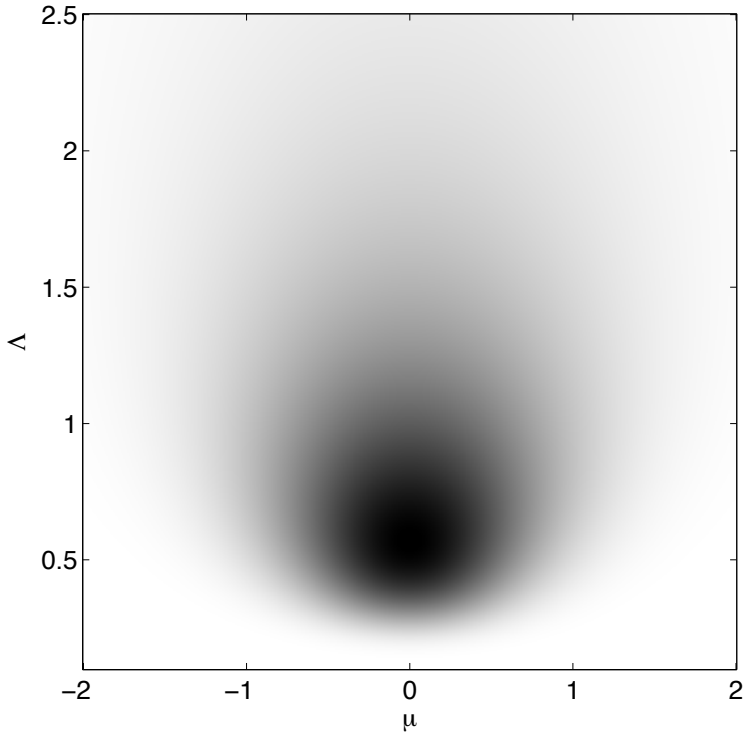
Information Parameterization: Canonical Parameters

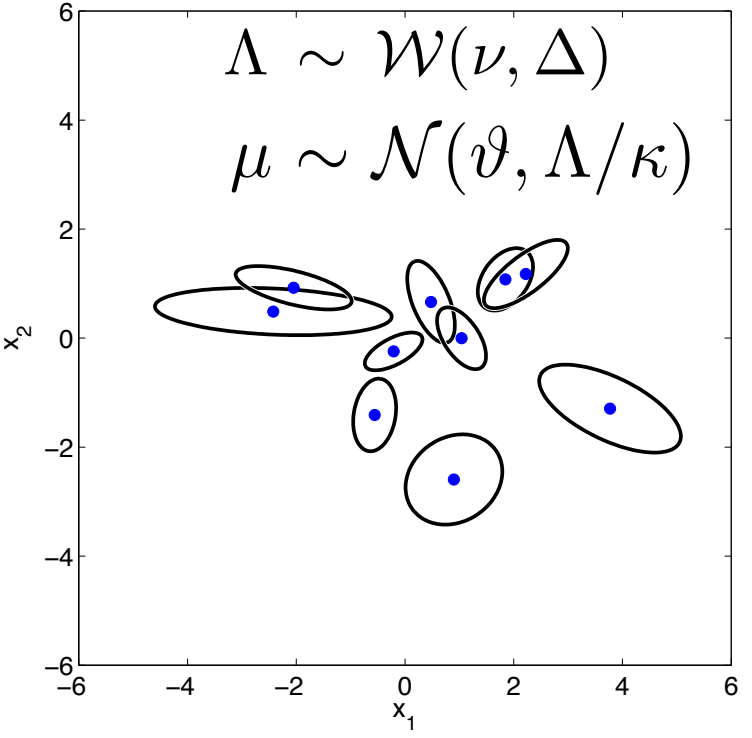
$$\mathcal{N}(x \mid \mu, \Sigma) \propto \exp\left\{-\frac{1}{2}x^T\Sigma^{-1}x + \mu^T\Sigma^{-1}x - \frac{1}{2}\mu^T\Sigma^{-1}\mu\right\}$$

$$\mathcal{N}^{-1}(x \mid \vartheta, \Lambda) \propto \exp\left\{-\frac{1}{2}x^T\Lambda x + \vartheta^T x\right\} \propto \exp\left\{\left[\sum_{s=1}^{N} \vartheta_s x_s - \frac{1}{2}\Lambda_{ss}x_s^2\right] - \left[\sum_{s \neq t} \Lambda_{st} x_s x_t\right]\right\}$$

Recall general exponential family form: $p(x \mid \theta) = \exp\{\theta^T \phi(x) - \Phi(\theta)\}$

Normal-Inverse-Wishart Prior Distributions





- The *Wishart distribution* generalizes gamma to positive definite matrices
- For multivariate normal, conjugate prior is Wishart on inverse covariance, and multivariate Gaussian (with dependent covariance) on mean

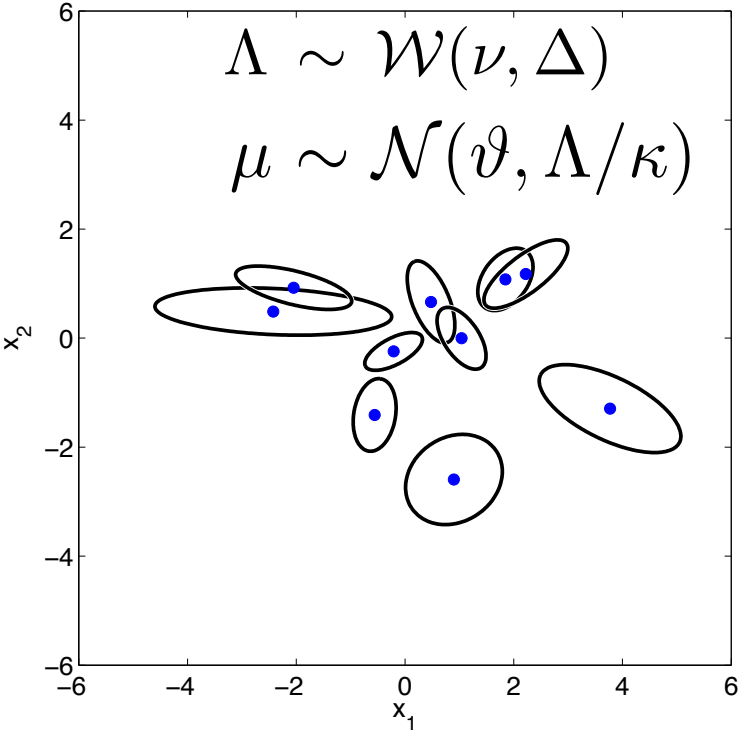
Normal-Inverse-Wishart Prior Distributions

$$\bar{\kappa} = \kappa + L$$

$$\bar{\nu} = \nu + L$$

$$\bar{\kappa}\bar{\vartheta} = \kappa\vartheta + \sum_{\ell=1}^{L} x^{(\ell)}$$

$$\bar{\nu}\bar{\Delta} = \nu\Delta + \sum_{\ell=1}^{L} x^{(\ell)}x^{(\ell)T} + \kappa\vartheta\vartheta^T - \bar{\kappa}\bar{\vartheta}\bar{\vartheta}^T$$



$$p(\mu, \Lambda \mid \kappa, \vartheta, \nu, \Delta) \propto |\Lambda|^{-\left(\frac{\nu+d}{2}+1\right)} \exp\left\{-\frac{1}{2}\operatorname{tr}(\nu\Delta\Lambda^{-1}) - \frac{\kappa}{2}(\mu-\vartheta)^T\Lambda^{-1}(\mu-\vartheta)\right\}$$