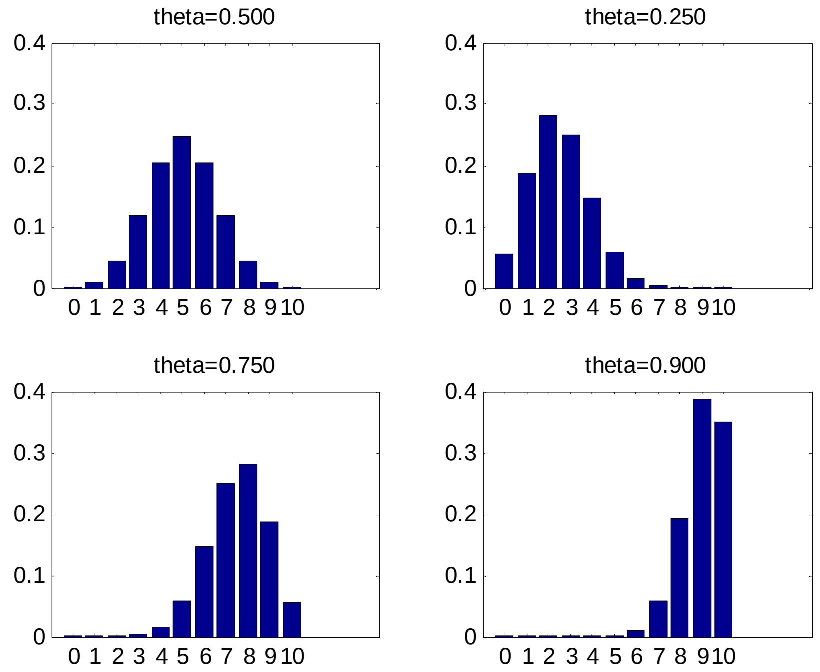
CS340 Machine learning
Bayesian statistics 2

Binomial distribution (count data)

- X ~ Binom(θ, N), X ∈ {0,1,…,N}

$$P(X = x|\theta, N) = \binom{N}{x} \theta^x (1-\theta)^{N-x}$$



Bernoulli distribution (binary data)

- Binomial distribution when N=1 is called the Bernoulli distribution.
- We write X ~ Ber(θ), $X \in \{0,1\}$

$$p(X) = \theta^X (1-\theta)^{1-X}$$

- So p(X=1) = θ, p(X=0) = 1-θ

Bernoulli likelihood function

- The likelihood is

$$
\begin{aligned}
L(\theta) &= p(D|\theta) = \prod_{n=1}^{N} p(x_n|\theta) \\
&= \prod_n \theta^{I(x_n=1)}(1-\theta)^{I(x_n=0)} \\
&= \theta^{\sum_n I(x_n=1)}(1-\theta)^{\sum_n I(x_n=0)} \\
&= \theta^{N_1}(1-\theta)^{N_0}
\end{aligned}
$$

We say that N_0 and N_1 are sufficient statistics of D for θ

This is the same as the Binomial likelihood function, up to constant factors.

Summary of beta-Bernoulli model

- Prior $p(\theta) = \mathrm{Beta}(\theta|\alpha_1, \alpha_0) = \frac{1}{B(\alpha_1, \alpha_0)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$
- Likelihood $p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$
- Posterior $p(\theta|D) = \mathrm{Beta}(\theta|\alpha_1 + N_1, \alpha_0 + N_0)$
- Posterior predictive $p(X=1|D) = \frac{\alpha_1 + N_1}{\alpha_1 + \alpha_0 + N}$

Marginal likelihood

- When performing Bayesian model selection and empirical Bayes estimation, we will need

$$p(D) = \int p(D|\theta)p(\theta)d\theta$$

- This is given by a ratio of the posterior and prior normalizing constants

$$\begin{aligned} p(\theta|D) &= \frac{p(\theta)p(D|\theta)}{p(D)} \\ &= \frac{1}{p(D)}\left[\frac{1}{B(\alpha_1,\alpha_0)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}\right]\left[\theta^{N_1}(1-\theta)^{N_0}\right] \\ &= \frac{\theta^{\alpha_1'-1}(1-\theta)^{\alpha_0'-1}}{B(\alpha_1',\alpha_0')} \\ p(D) &= \frac{B(\alpha_1',\alpha_0')}{B(\alpha_1,\alpha_0)} \end{aligned}$$

$$\alpha_1' = \alpha_1 + N_1, \quad \alpha_0' = \alpha_0 + N_0$$

Summary of beta-Bernoulli model

- Prior $p(\theta) = \mathrm{Beta}(\theta|\alpha_1, \alpha_0) = \frac{1}{B(\alpha_1, \alpha_0)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$
- Likelihood $p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$
- Posterior $p(\theta|D) = \mathrm{Beta}(\theta|\alpha_1 + N_1, \alpha_0 + N_0)$
- Posterior predictive $p(X=1|D) = \frac{\alpha_1 + N_1}{\alpha_1 + \alpha_0 + N}$
- Marginal likelihood

$$p(D) = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)} = \frac{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_1 + N_1 + \alpha_0 + N_0)} \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)}$$

From coins to dice

- Let $(X_1,\ldots X_d) \mid N \sim \text{Multinomial}(\theta, N)$

$$P(x_1,\ldots,x_d|\theta,N) = \binom{N}{x_1 \;\ldots\; x_d} \prod_{i=1}^{d} \theta_i^{x_i}$$

X_i's no longer conditionally independent since $\sum_i x_i = N$

$$= \frac{N!}{x_1!x_2!\ldots x_d!} \prod_{i=1}^{d} \theta_i^{x_i}$$

We also require $\sum_i \theta_i = 1$.

$$= (\sum_i x_i)! \prod_i \frac{\theta_i^{x_i}}{x_i!}$$

$X_i \in \{0,\ldots,N\}$ = number of times face i occurs

Multinomial(θ, 1)

- Let $(X_1,\ldots X_d) \sim$ Multinomial(θ, 1)

$$P(x_1,\ldots,x_d|\theta) \quad = \quad \prod_{i=1}^{d} \theta_i^{x_i}$$

- Since $\sum_i X_i = 1$, only one "bit" can be on, eg (0,1,0) means face 2 occurred.
- Let $X \in \{1,\ldots,d\}$ represent the event that occurred.

$$P(X=k|\theta) \quad = \quad \prod_{i=1}^{d} \theta_i^{I(X=k)} = \theta_k$$

Likelihood function for the multinomial

- Let D = (X_1, …, X_N), where $X_i \in \{1,\ldots,K\}$

$$P(D|\theta) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \theta_k^{x_{nk}} = \prod_k \theta_k^{\sum_n x_{nk}} = \prod_k \theta_k^{N_k}$$

- The N_i are the sufficient statistics

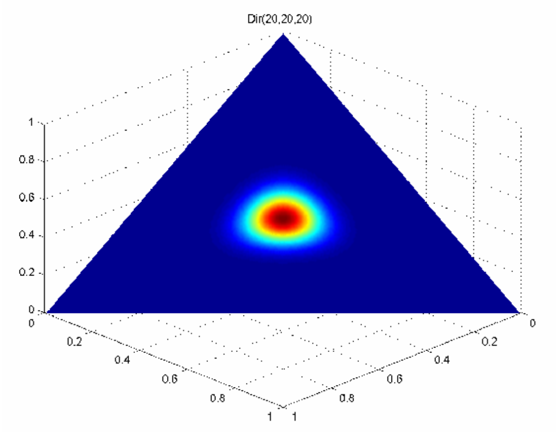
Dirichlet distribution

- Generalization of Beta to K dimensions $E[x_k] = \alpha_k/(\sum_j \alpha_j)$

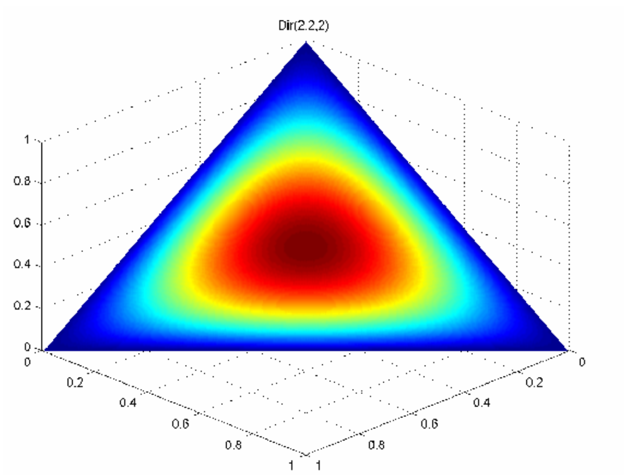
$$p(x|\alpha) = \mathcal{D}(x|\alpha) = \frac{1}{Z(\alpha)} \cdot x_1^{\alpha_1 - 1} \cdot x_2^{\alpha_0 - 1} \cdots x_K^{\alpha_K - 1} I(\sum_{k=1}^{K} x_k - 1)$$

- Normalization constant

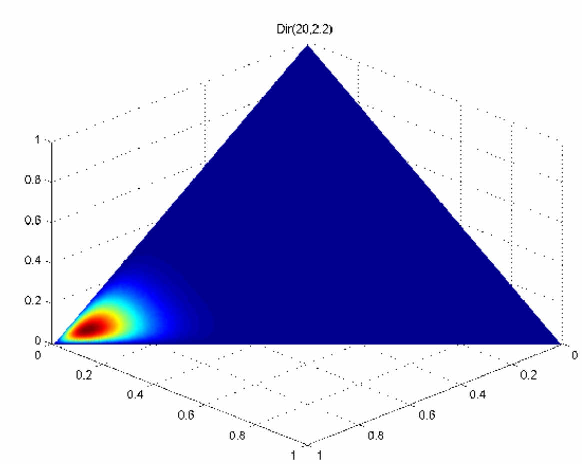
$$Z(\alpha) = \int \cdots \int x_1^{\alpha_1 - 1} \cdots x_K^{\alpha_K - 1} dx_1 \cdots dx_K = \frac{\prod_{j=1}^{K} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{K} \alpha_j)}$$



(20,20,20)



(2,2,2)



(20,2,2)

Summary of Dirichlet-multinomial model

- $X_n \sim$ Mult(θ,1), $p(X_n=k) = \theta_k$
- Prior $p(\theta) = \mathrm{Dir}(\theta|\alpha_1, \ldots, \alpha_K) = \frac{1}{Z(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$
- Likelihood $p(D|\theta) = \prod_{k=1}^{K} \theta_k^{N_k}$
- Posterior $p(\theta|D) = \mathrm{Dir}(\theta|\alpha_1 + N_1, \ldots, \alpha_K + N_K)$
- Posterior predictive $p(X = k|D) = \frac{\alpha_k + N_k}{\sum_{k'} \alpha_{k'} + N_{k'}}$
- Marginal likelihood

$$p(D) = \frac{Z(\vec{N} + \vec{\alpha})}{Z(\vec{\alpha})} = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)}$$

Normal-Normal model

- Consider estimating the mean of a Gaussian whose variance is known.
- The natural conjugate prior is Gaussian.
- So the posterior is also Gaussian ("Gaussian times Gaussian gives Gaussian").
- The algebra is rather messy (see handout), so we will just state and interpret the results.

Normal-normal model

- Likelihood

$$p(D|\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

$$\propto \exp\left(-\frac{N}{2\sigma^2}(\overline{x} - \mu)^2\right) \propto \mathcal{N}(\overline{x}|\mu, \frac{\sigma^2}{N})$$

- Natural conjugate prior

$$p(\mu) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \propto \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

- Posterior

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$p(\mu|D) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\overline{x} = \sigma_N^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{N\overline{x}}{\sigma^2}\right)$$

$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2} = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

$$\lambda_N = N\lambda + \lambda_0$$

Posterior mean

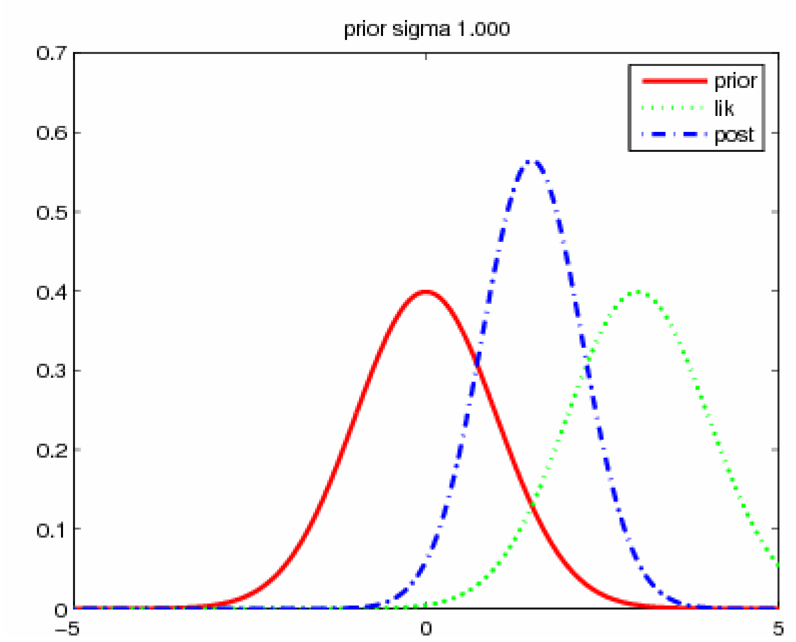
- Consider N=1.

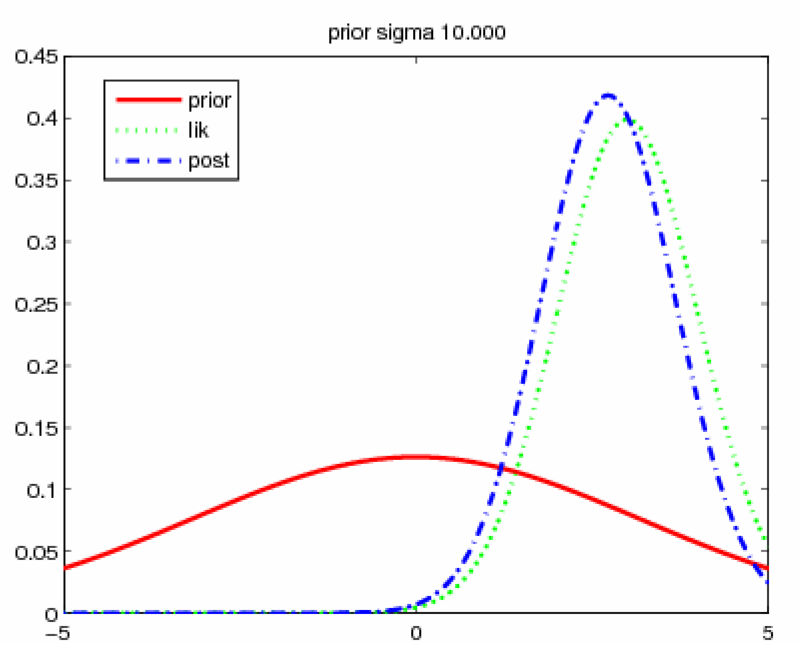
$$
\begin{aligned}
\mu_1 &= \frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0 + \frac{\sigma_0^2}{\sigma^2+\sigma_0^2}x \\
&= \mu_0 + (x-\mu_0)\frac{\sigma_0^2}{\sigma^2+\sigma_0^2} \\
&= x - (x-\mu_0)\frac{\sigma^2}{\sigma^2+\sigma_0^2}
\end{aligned}
$$

Convex comb of prior and MLE

Prior plus data correction term

Data shrunk towards prior



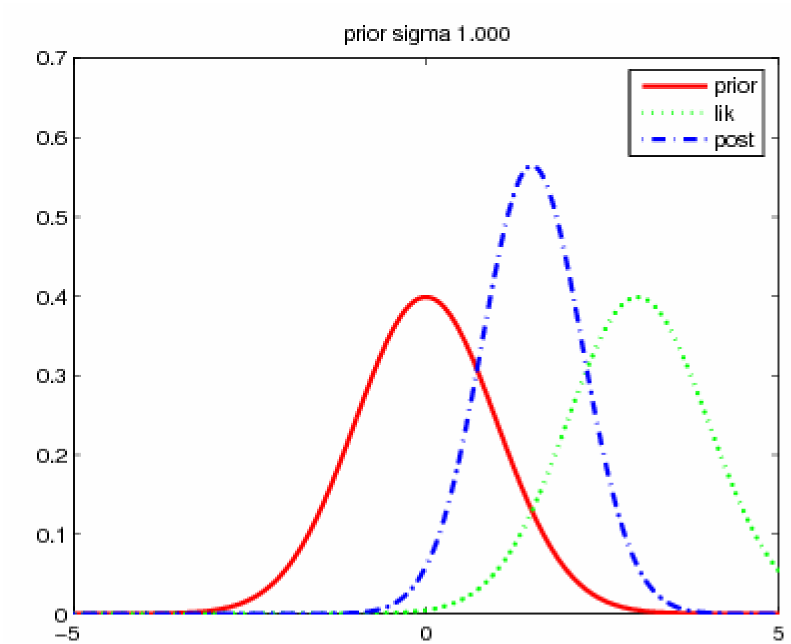


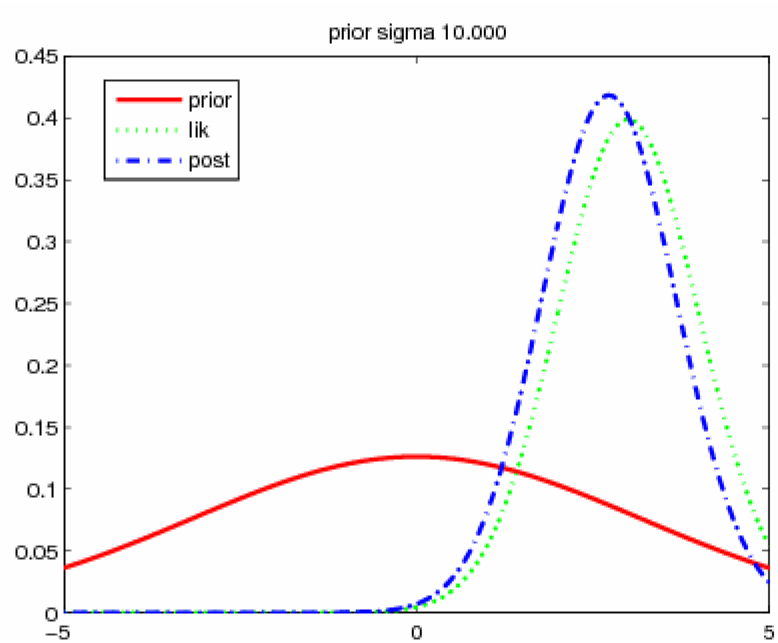
Posterior precision

- Precision = 1/variance, $\lambda = 1/\sigma^2$.
- Precisions add, means are averaged.

$$
\begin{aligned}
p(\mu|D,\lambda) &= \mathcal{N}(\mu|\mu_N, \lambda_N^{-1}) \\
\lambda_N &= \lambda_0 + N\lambda \\
\mu_N &= \frac{\overline{x}N\lambda + \mu_0\lambda_0}{\lambda_N} = w\overline{x} + (1-w)\mu_0
\end{aligned}
$$

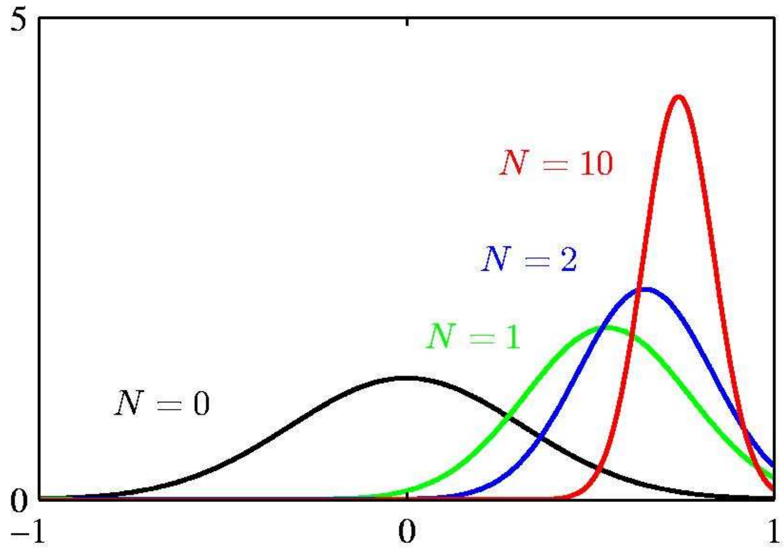
$$w = \frac{N\lambda}{\lambda_N}$$





Sequential updating

- Suppose true mean=0.8, true variance=0.1.
- p(μ|D) rapidly approaches a delta function centered on the true mean.



Bishop 2.12

Posterior predictive distribution

- The predictive variance is the observation noise σ^2 plus the uncertainty about μ, σ^2_N

$$\begin{aligned} p(x|D) &= \int p(x|\mu)p(\mu|D)d\mu \\ &= \int \mathcal{N}(x|\mu,\sigma^2)\mathcal{N}(\mu|\mu_N,\sigma_N^2)d\mu \\ &= \mathcal{N}(x|\mu_N,\sigma_N^2+\sigma^2) \end{aligned}$$

- Or, future X = prior mean μ + noise ε

$$\begin{aligned} X &= \mu + \epsilon \\ \mu &\sim \mathcal{N}(\mu_n,\sigma_n^2) \\ \epsilon &\sim \mathcal{N}(0,\sigma^2) \\ E[X] &= E[\mu]+E[\epsilon] = \mu_n + 0 \\ \text{Var}\,[X] &= \text{Var}\,[\mu] + \text{Var}\,[\epsilon] = \sigma_n^2+\sigma^2 \end{aligned}$$

Summary of Normal-Normal model

- Prior $$p(\mu) = \mathcal{N}(\mu|\mu_0, (\lambda_0)^{-1})$$
- Likelihood $$p(D|\mu) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1})$$
- Posterior

$$\begin{aligned} p(\mu|D) &= \mathcal{N}(\mu|\mu_N, (\lambda_N)^{-1}) \\ \lambda_N &= \lambda_0 + N\lambda \\ \mu_N &= \frac{\overline{x}N\lambda + \mu_0\lambda_0}{\lambda_N} \end{aligned}$$

- Posterior predictive

$$p(x|D) = \mathcal{N}(x|\mu_N, \sigma_N^2 + \sigma^2)$$

- Marginal likelihood

Too messy to print here (see handout)