Machine Learning

CSE 6363 (Fall 2016)

Lecture 4 Bayesian Learning

Heng Huang, Ph.D.

Department of Computer Science and Engineering

Probability Distribution

- Let's start from a question
- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - You say: Please flip it a few times

  - You say: The probability is:
  - **He says: Why???**
  - You say: Because…

Ref: Carlos Guestrin

Thumbtack – Binomial Distribution

- P(Heads) = θ, P(Tails) = 1 - θ

- Flips are:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence D of α_H Heads and α_T Tails

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

Ref: Carlos Guestrin

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Binomial distribution
- Learning θ is an optimization problem
  - What's the objective function?
- MLE: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

Ref: Carlos Guestrin

Your First Learning Algorithm

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \quad \ln P(\mathcal{D} \mid \theta) \\ &= \arg\max_{\theta} \quad \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero:

$$\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0$$

Ref: Carlos Guestrin

What about Prior

$$\hat{\theta} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: $\theta = 3/5$, I can prove it!
- Billionaire says: Wait, I know that the thumbtack is "close" to 50-50. What can you?
- **You say: I can learn it the Bayesian way…**
- Rather than estimating a single θ, we obtain a distribution over possible values of θ

Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

Ref: Carlos Guestrin

Bayesian Learning for Thumbtack

$$P(\theta \mid \mathcal{D}) \;\propto\; P(\mathcal{D} \mid \theta)P(\theta)$$

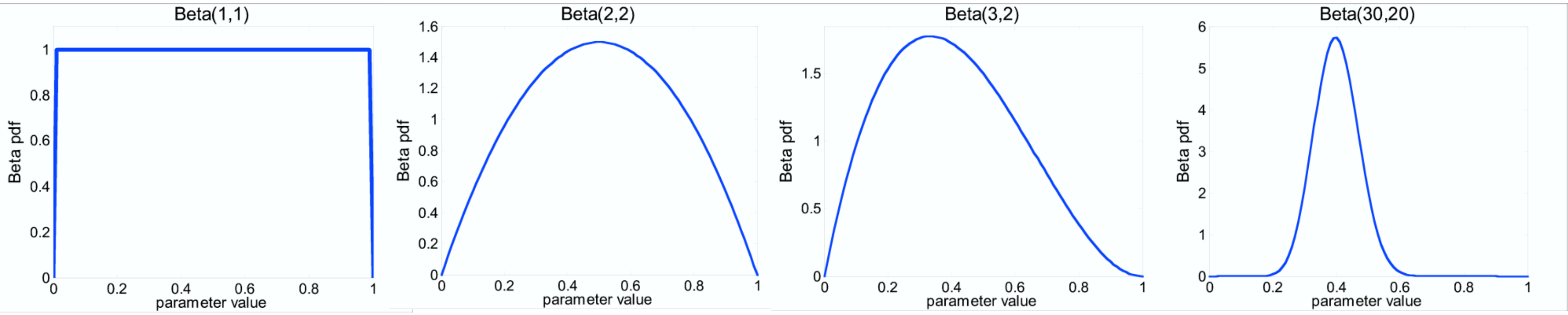
- Likelihood function is simply Binomial:

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Closed-form representation of posterior
  - **For Binomial, conjugate prior is Beta distribution**

Beta Prior Distribution – P(θ)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$



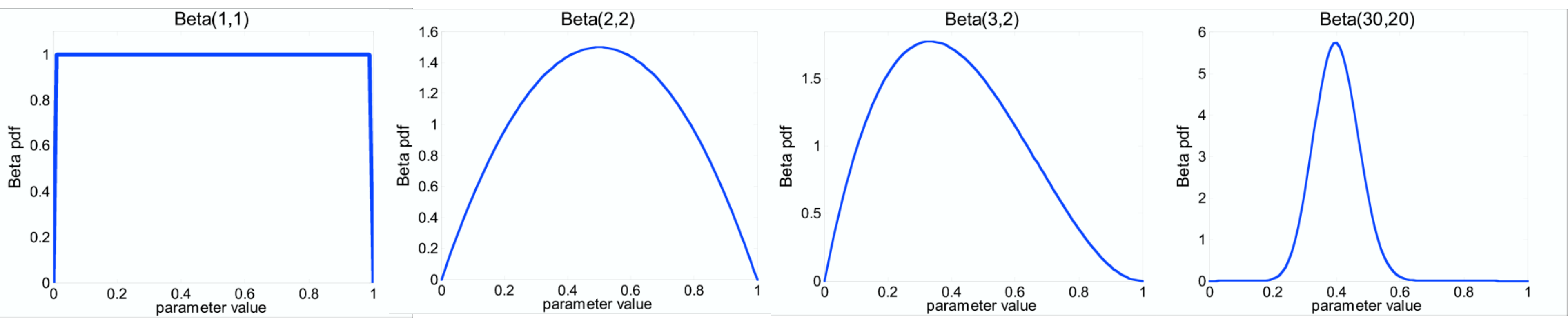
- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$

Ref: Carlos Guestrin

Posterior Distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: α_H Heads and α_T Tails
- Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



Ref: Carlos Guestrin

Conjugate Priors

- A Bayesian estimate of μ requires a prior $p(\mu)$.
- A prior is called conjugate if, when multiplied by the likelihood $p(D|\mu)$, the resulting posterior is in the same parametric family as the prior. (Closed under Bayesian updating.)
- The Beta prior is conjugate to the Bernoulli likelihood

$$\begin{aligned} P(\mu|D) &\propto P(D|\mu)P(\mu) \\ &\propto [\mu^n(1-\mu)^m][\mu^{a-1}\mu^{b-1}] \\ &= \mu^{n+a-1}(1-\mu)^{m+b-1} \end{aligned}$$

  where n is the number of heads and m is the number of tails.
- a, b are hyperparameters (parameters of the prior) and correspond to the number of "virtual" heads/tails (pseudo counts). $N_0 = a + b$ is called the effective sample size (strength) of the prior. $a = b = 1$ is a uniform prior (Laplace smoothing).

Ref: Kevin Murphy

The Beta Distribution

- To ensure the prior is normalized, we define

$$P(\mu|a,b) = \text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

where the gamma function is defined as

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$$

Note that $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1) = 1$. Also, for integers, $\Gamma(x+1) = x!$.

- The normalization constant $1/Z(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ ensures

$$\int_0^1 \text{Beta}(\mu|a,b)d\mu = 1$$

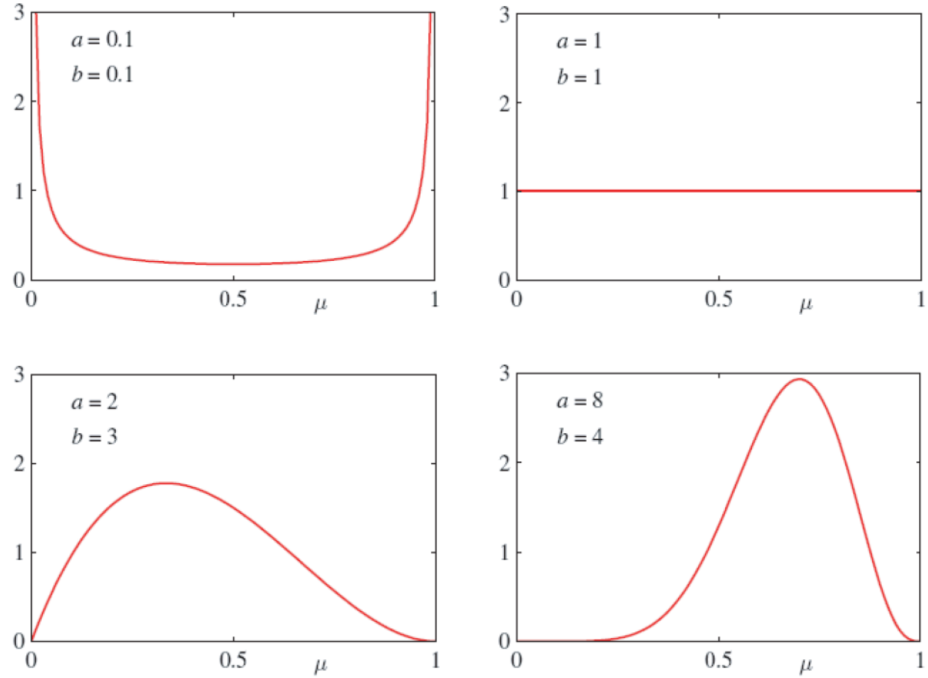
Ref: Kevin Murphy

The Beta Distribution

If $\mu \sim Be(a, b)$, then

$$E\mu = \frac{a}{a+b}$$

$$\mathrm{var}(\mu) = E[(\mu - E[\mu])^2] = \int_0^1 \left(\mu - \frac{a}{a+b}\right)^2 \mathrm{Beta}(\mu|a,b)\, d\mu = \frac{ab}{(a+b)^2(a+b+1)}$$

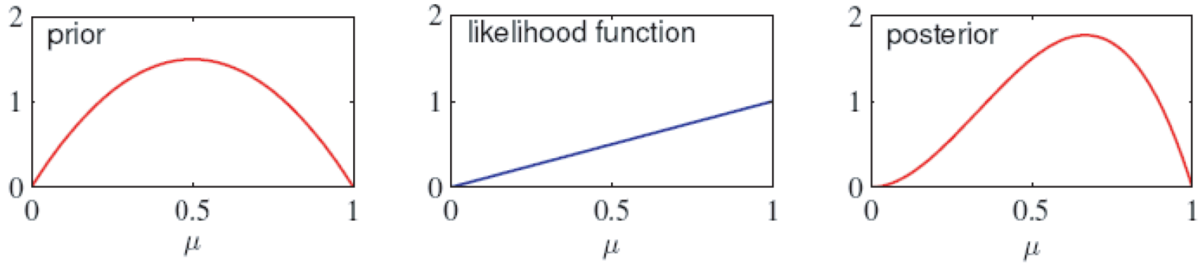


Ref: Kevin Murphy

Bayesian Updating in Pictures

- Start with $Be(\mu|a = 2, b = 2)$ and observe $x = 1$, so the posterior is $Be(\mu|a = 3, b = 2)$.

```
thetas = 0:0.01:1;
alphaH = 2; alphaT = 2; Nh=1; Nt=0; N = Nh+Nt;
prior = betapdf(thetas, alphaH, alphaT);
lik = choose(N,Nh) * thetas.^Nh .* (1-thetas).^Nt;
post = betapdf(thetas, alphaH+Nh, alphaT+Nt);
```



Ref: Kevin Murphy

Posterior Predictive Distribution

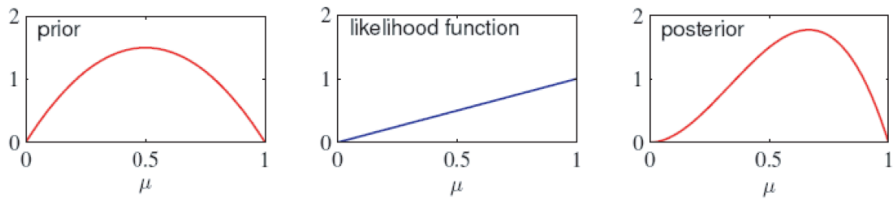
- The posterior predictive distribution is

$$\begin{aligned} p(X=1|D) &= \int_0^1 p(X=1|\mu)p(\mu|D)d\mu \\ &= \int_0^1 \mu p(\mu|D)d\mu = E[\mu|D] = \frac{n+a}{n+m+a+b} \end{aligned}$$

- With a uniform prior $a = b = 1$, we get Laplace's rule of succession

$$p(X=1|N_h, N_t) = \frac{N_h + 1}{N_h + N_t + 2}$$

- Start with $Be(\mu|a=2, b=2)$ and observe $x=1$ to get $Be(\mu|a=3, b=2)$, so the mean shifts from $E[\mu] = 2/4$ to $E[\mu|D] = 3/5$.



Ref: Kevin Murphy

Effect of Prior Strength

- Let $N = N_h + N_t$ be number of samples (observations).
- Let N' be the number of pseudo observations (strength of prior) and define the prior means

$$\alpha_h = N'\alpha'_h, \quad \alpha_t = N'\alpha'_t, \quad \alpha'_h + \alpha'_t = 1$$

- Then posterior mean is a convex combination of the prior mean and the MLE (where $\lambda = N'/(N + N')$):

$$\begin{aligned} P(X = h|\alpha_h, \alpha_t, N_h, N_t) &= \frac{\alpha_h + N_h}{\alpha_h + N_h + \alpha_t + N_t} \\ &= \frac{N'\alpha'_h + N_h}{N + N'} \\ &= \frac{N'}{N + N'}\alpha'_h + \frac{N}{N + N'}\frac{N_h}{N} \\ &= \lambda\alpha'_h + (1 - \lambda)\frac{N_h}{N} \end{aligned}$$

Ref: Kevin Murphy

Effect of Prior Strength

- Suppose we have a uniform prior $\alpha'_h = \alpha'_t = 0.5$, and we observe $N_h = 3$, $N_t = 7$.
- Weak prior $N' = 2$. Posterior prediction:

$$P(X = h|\alpha_h = 1, \alpha_t = 1, N_h = 3, N_t = 7) = \frac{3+1}{3+1+7+1} = \frac{1}{3} \approx 0.33$$

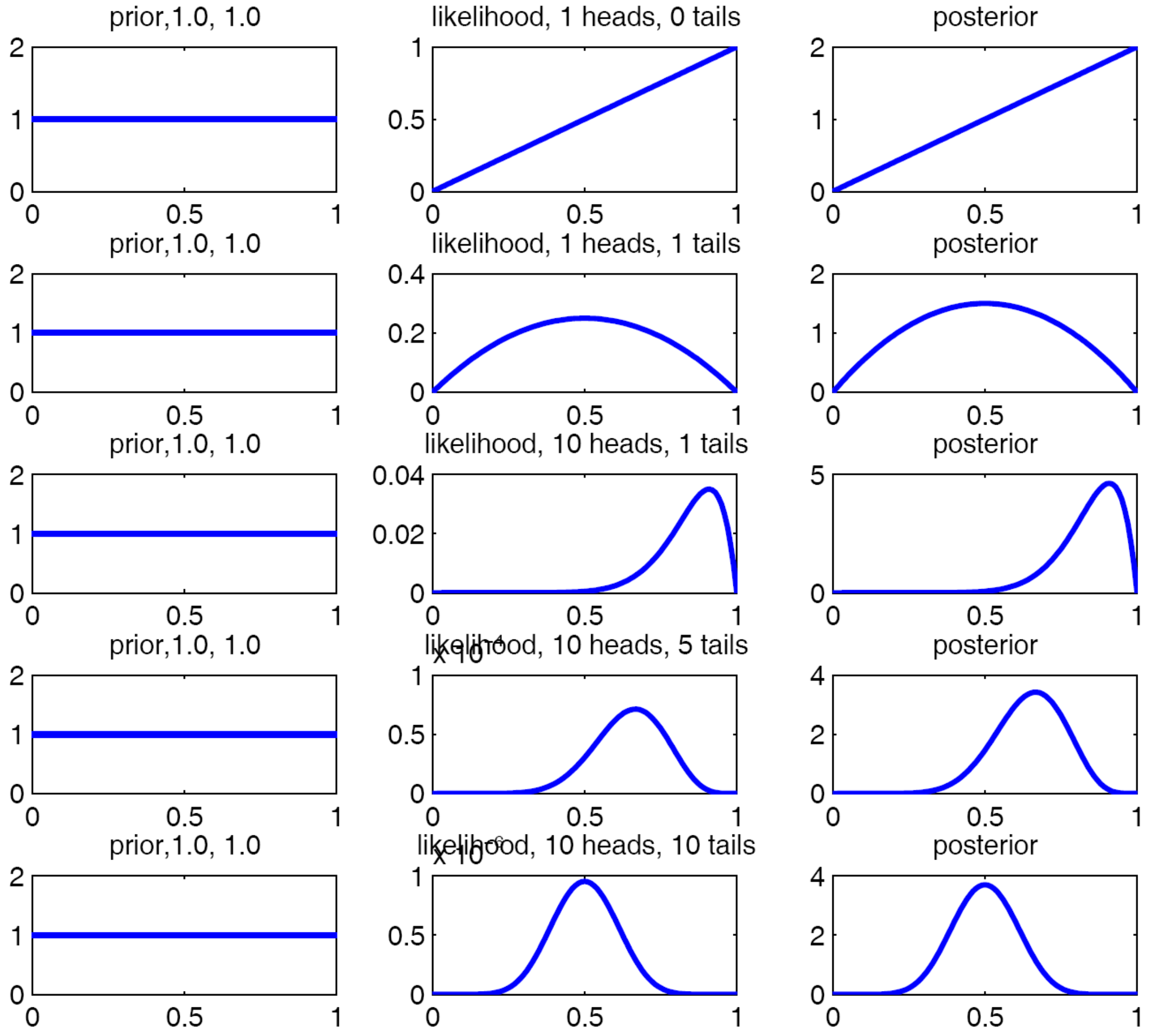
- Strong prior $N' = 20$. Posterior prediction:

$$\frac{3+10}{3+10+7+10} = \frac{13}{30} \approx 0.43$$

- However, if we have enough data, it washes away the prior. e.g., $N_h = 300$, $N_t = 700$. Estimates are $\frac{300+1}{1000+2}$ and $\frac{300+10}{1000+20}$, both of which are close to 0.3
- As $N \rightarrow \infty$, $P(\theta|D) \rightarrow \delta(\theta, \hat{\theta}_{ML})$, so $E[\theta|D] \rightarrow \hat{\theta}_{ML}$.

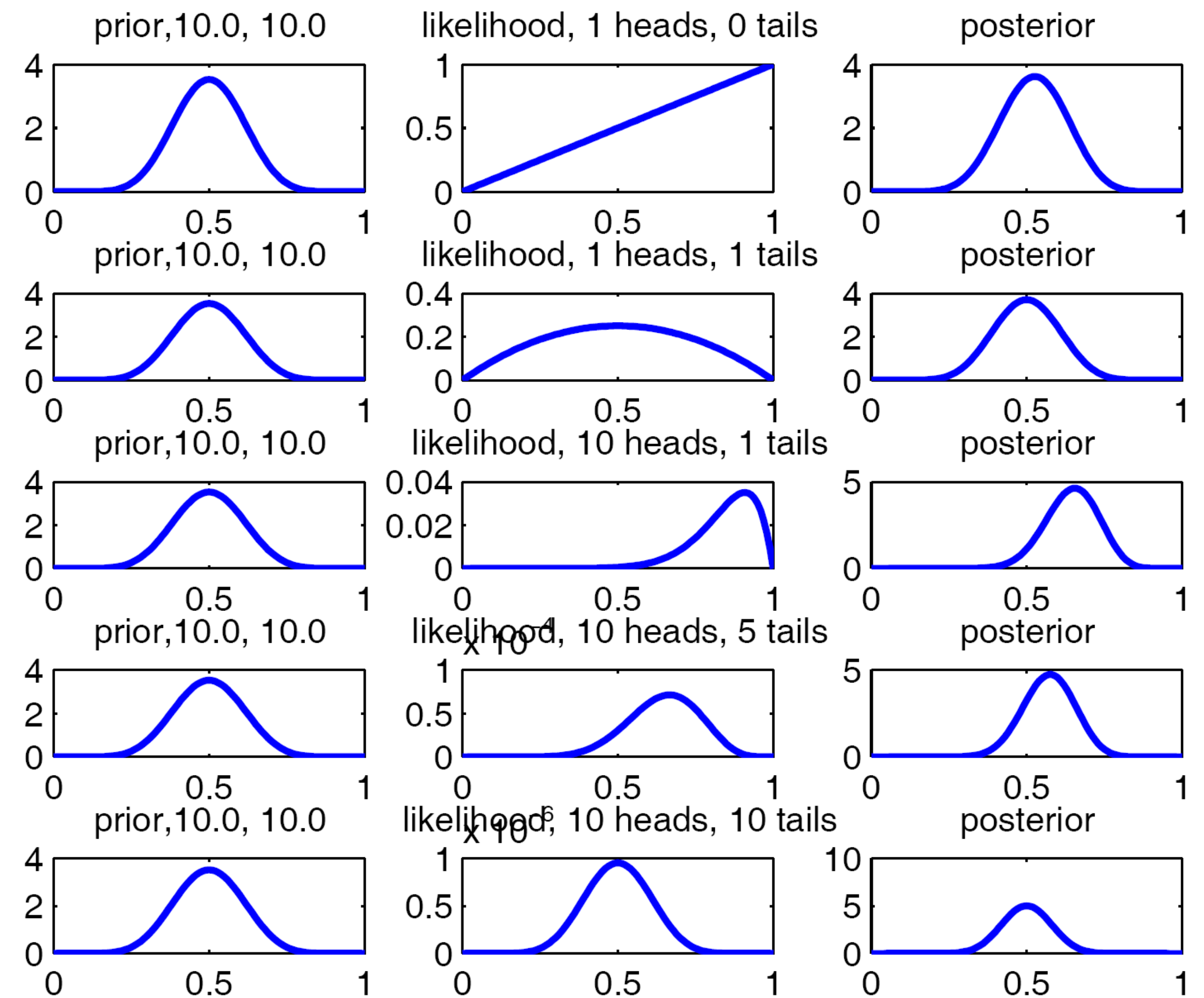
Ref: Kevin Murphy

Parameter Posterior – Small Sample, Uniform Prior



Ref: Kevin Murphy

Parameter Posterior – Small Sample, Strong Prior



Ref: Kevin Murphy

Maximum A Posteriori (MAP) Estimation

- MAP estimation picks the mode of the posterior

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(D|\theta)p(\theta)$$

- If $\theta \sim Be(a, b)$, this is just

$$\hat{\theta}_{MAP} = (a-1)/(a+b-2)$$

- MAP is equivalent to maximizing the penalized maximum log-likelihood

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(D|\theta) - \lambda c(\theta)$$

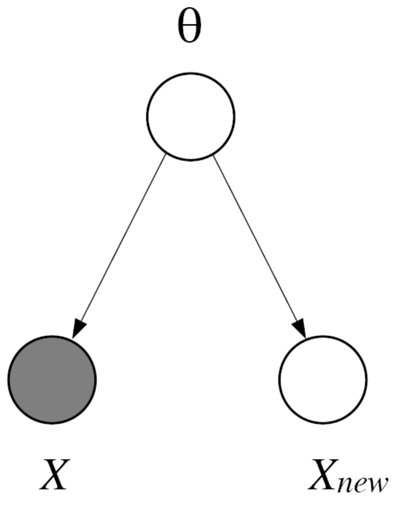
where $c(\theta) = -\log p(\theta)$ is called a *regularization term*. λ is related to the strength of the prior.

Ref: Kevin Murphy

Integrate Out or Optimize

- $\hat{\theta}_{MAP}$ is not Bayesian (even though it uses a prior) since it is a point estimate.
- Consider predicting the future. A Bayesian will integrate out all uncertainty:

$$\begin{aligned} p(x_{new}|X) &= \int p(x_{new}, \theta|X) d\theta \\ &= \int p(x_{new}|\theta, X) p(\theta|X) d\theta \\ &\propto \int p(x_{new}|\theta) p(X|\theta) p(\theta) d\theta \end{aligned}$$



- A frequentist will use a "plug-in" estimator eg ML/MAP:

$$p(x_{new}|X) = p(x_{new}|\hat{\theta}), \quad \hat{\theta} = \arg\max_{\theta} p(X|\theta)$$

Ref: Kevin Murphy

From Coin to Dice

- Suppose we observe N iid die rolls (K-sided): D=3,1,K,2,...
- Let $[x] \in \{0,1\}^K$ be a one-of-K encoding of x eg. if $x = 3$ and $K = 6$, then $[x] = (0,0,1,0,0,0)^T$.
- Multinomial distribution: $p(X = k) = \theta_k \quad \sum_k \theta_k = 1$
- Likelihood

$$\ell(\theta; D) = \log p(D|\theta) = \sum_m \log \prod_k \theta_k^{[x^m=k]}$$
$$= \sum_m \sum_k [x^m = k] \log \theta_k = \sum_k N_k \log \theta_k$$

- We need to maximize this subject to the constraint $\sum_k \theta_k = 1$, so we use a Lagrange multiplier.

Ref: Kevin Murphy

MLE for Multinomial

- Constrained cost function:

$$\tilde{l} = \sum_k N_k \log \theta_k + \lambda \left(1 - \sum_k \theta_k \right)$$

- Take derivatives wrt θ_k:

$$\begin{aligned}
\frac{\partial \tilde{l}}{\partial \theta_k} &= \frac{N_k}{\theta_k} - \lambda = 0 \\
N_k &= \lambda \theta_k \\
\sum_k N_k &= N = \lambda \sum_k \theta_k = \lambda \\
\hat{\theta}_{k,ML} &= \frac{N_k}{N}
\end{aligned}$$

- $\hat{\theta}_{k,ML}$ is the fraction of times k occurs.

Ref: Kevin Murphy

Dirichlet Priors

- Let $X \in \{1, \ldots, K\}$ have a multinomial distribution

$$P(X|\theta) = \theta_1^{I(X=1)}\theta_2^{I(X=2)} \cdots \theta_K^{I(X=k)}$$

- For a set of data $X^1, \ldots, X^N$, the sufficient statistics are the counts $N_i = \sum_n I(X_n = i)$.
- Consider a Dirichlet prior with hyperparameters α

$$p(\theta|\alpha) = \mathcal{D}(\theta|\alpha) = \frac{1}{Z(\alpha)} \cdot \theta_1^{\alpha_1 - 1} \cdot \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1}$$

where $Z(\alpha)$ is the normalizing constant

$$\begin{aligned} Z(\alpha) &= \int \cdots \int \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1} d\theta_1 \cdots d\theta_K \\ &= \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)} \end{aligned}$$

Ref: Kevin Murphy

Gaussian Density in 1-D

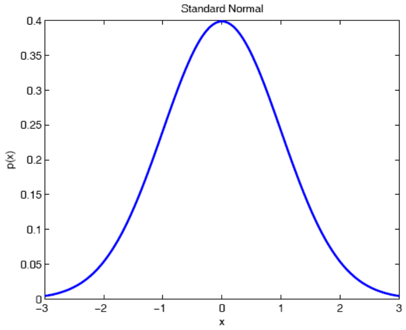
- If $X \sim N(\mu, \sigma^2)$, the probability density function (pdf) of X is defined as

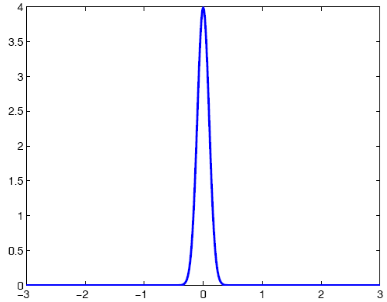
$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

We will often use the precision $\lambda = 1/\sigma^2$ instead of the variance σ^2.

- Note that a density evaluated at a point can be bigger than 1!
- Here is how we plot the pdf in matlab

```
xs=-3:0.01:3;  plot(xs,normpdf(xs,mu,sigma))
```





Ref: Kevin Murphy

Multivariate Gaussian

1-dimensional Gaussian

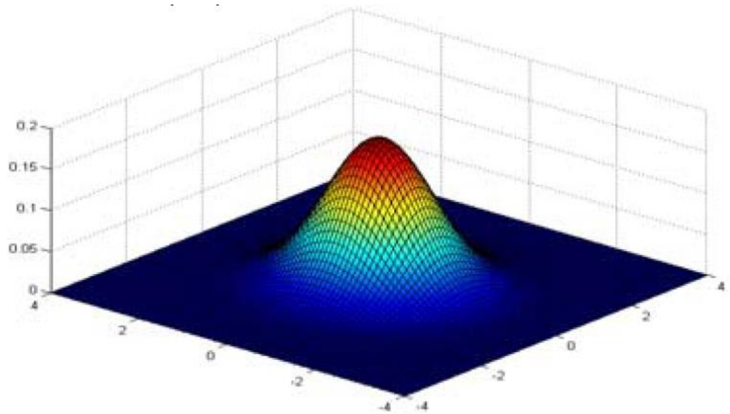
$$p(x|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

2-dimensional Gaussian

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

d-dimensional Gaussian

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



Multivariate Gaussian

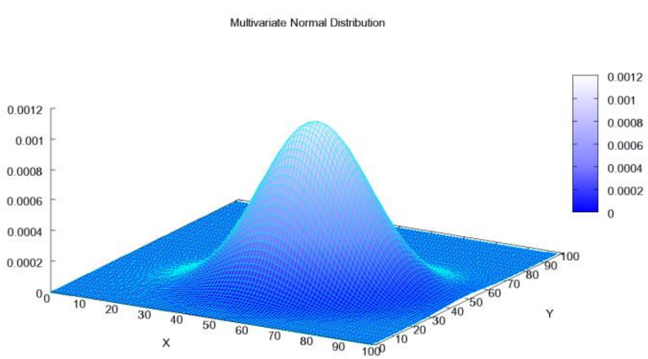
- If $X \in \mathbb{R}^d$ is a jointly gaussian random vector, then its pdf is

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

- The quantity $\Delta^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$ is called the Mahalanobis distance between x and μ.
- The first and second moments are

$$E[X] = \mu, \quad \mathrm{Cov}[X] = \Sigma$$

- Sometimes we will use the precision matrix Σ^{-1} instead of the covariance matrix Σ.



Ref: Kevin Murphy

Conditional Gaussian

- Suppose $x = (x_a, x_b)$ is jointly Gaussian with parameters

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \quad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix},$$

- It can be shown that $P(X_a|x_b) = N(X_a; \mu_{a|b}, \Sigma_{a|b})$ where

$$\begin{aligned} \mu_{a|b} &= \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b) \\ \Sigma_{a|b} &= \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \end{aligned}$$

- Note that the new mean is a linear function of x_b, and the new covariance is independent of x_a.
- Similarly, the marginal $P(X_a) = N(X_a; \mu_a, \Sigma_{aa})$.
- You should memorize these equations!

Ref: Kevin Murphy