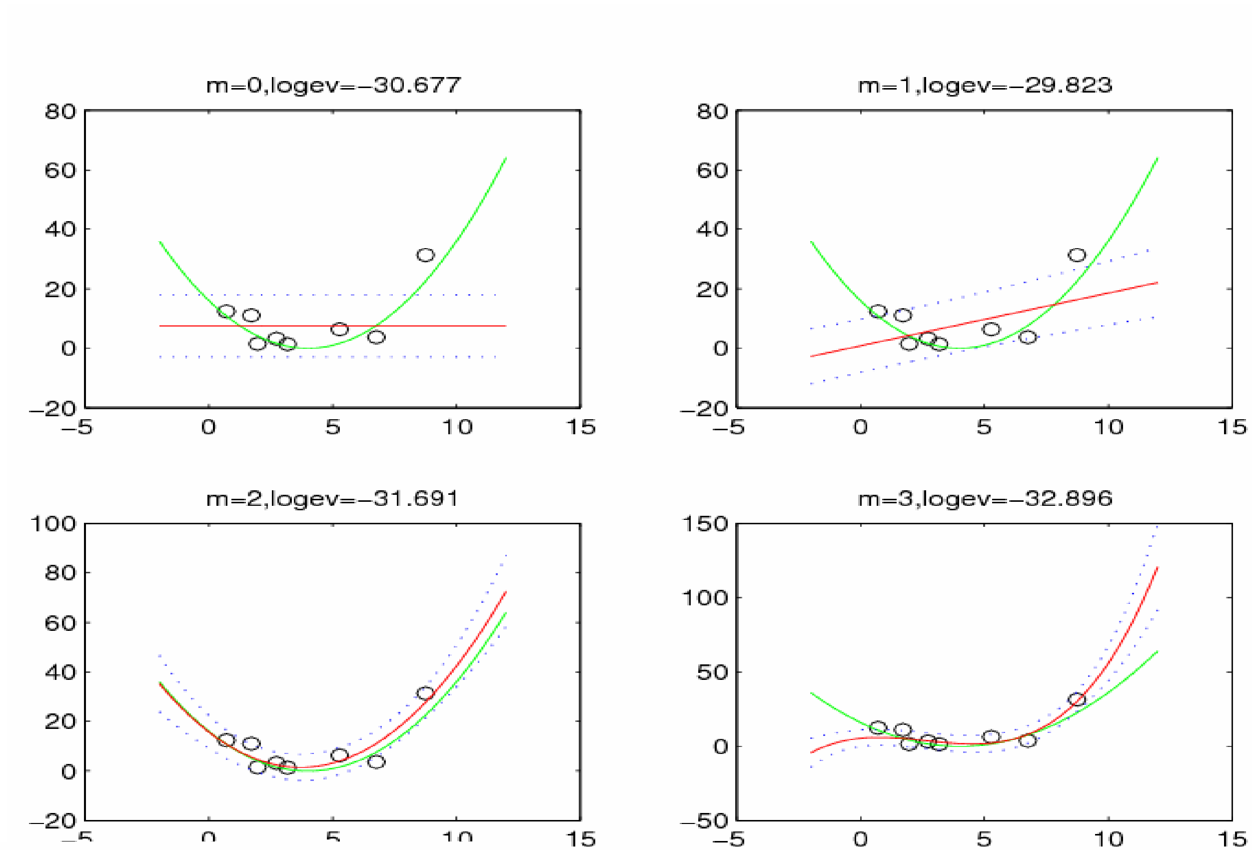
# CS340 Machine learning
# Bayesian model selection

# Bayesian model selection

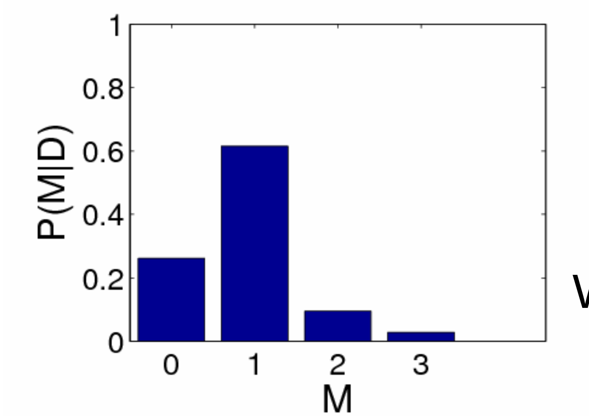
- Suppose we have several models, each with potentially different numbers of parameters.
- Example: M0 = constant, M1 = straight line, M2 = quadratic, M3 = cubic
- The posterior over models is defined using Bayes rule, where p(D|m) is called the marginal likelihood or "evidence" for m

$$
\begin{aligned}
p(m|D) &= \frac{p(m)p(D|m)}{p(D)} \\
p(D|m) &= \int p(D|\theta, m)p(\theta|m)d\theta \\
p(D) &= \sum_{m \in \mathcal{M}} p(D|m)p(m)
\end{aligned}
$$

# Polynomial regression, n=8




truth=quadratic
(green curve)

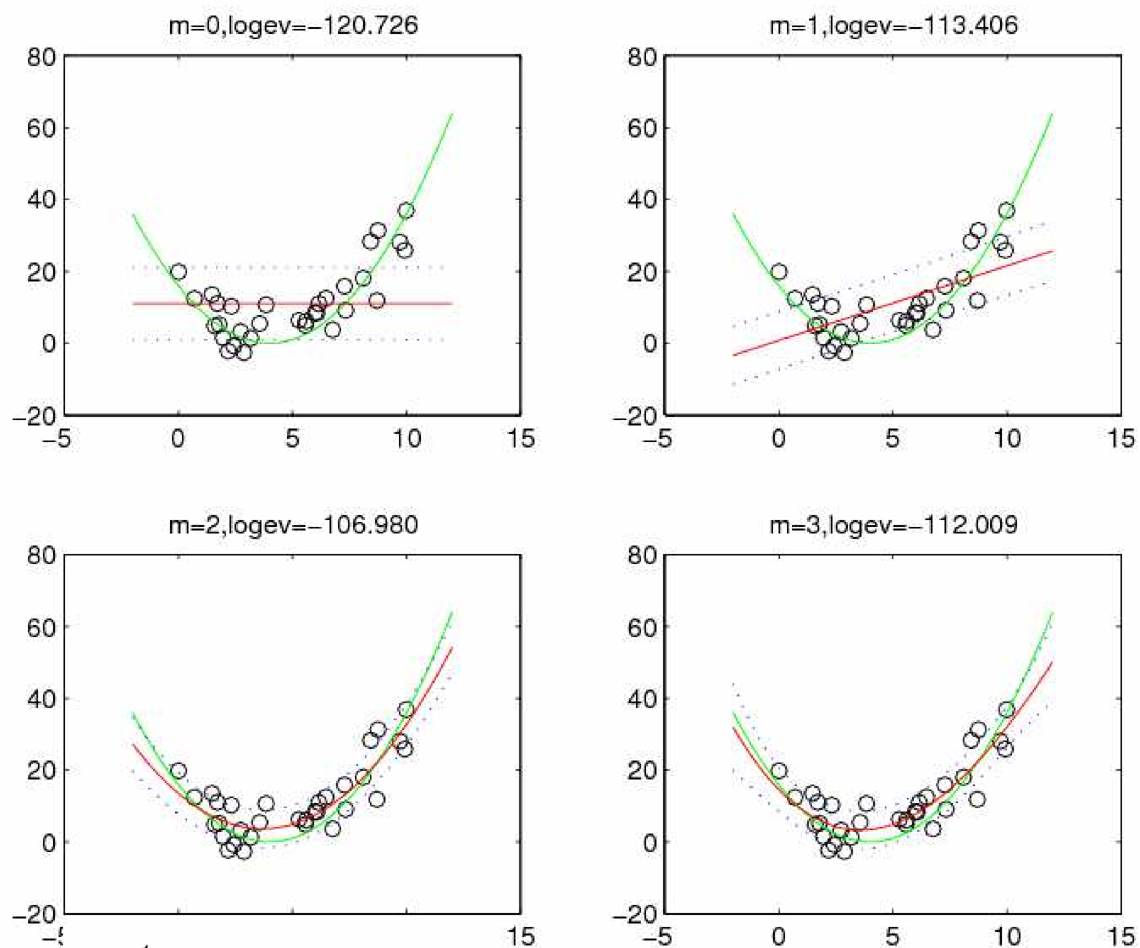



$$\text{logev}(m) = \log p(D|m)$$

$$p(m) = 1/4$$

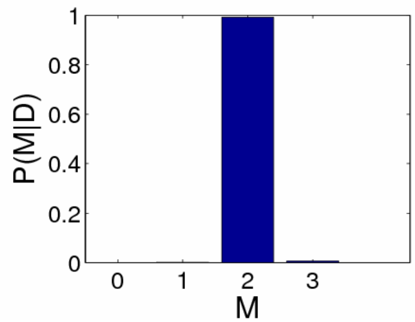
With little data, we choose a simple model

# Polynomial regression, n=32




truth=quadratic
(green curve)

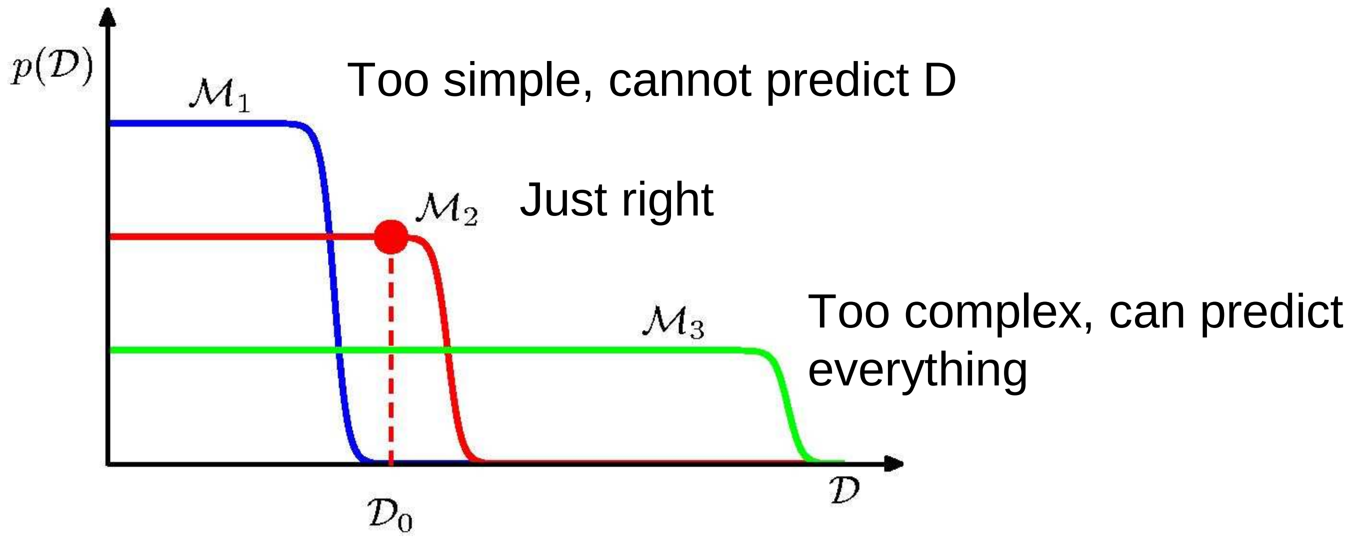
Shape of cubic changes a lot – high variance estimator




With more data, we choose a more complex model

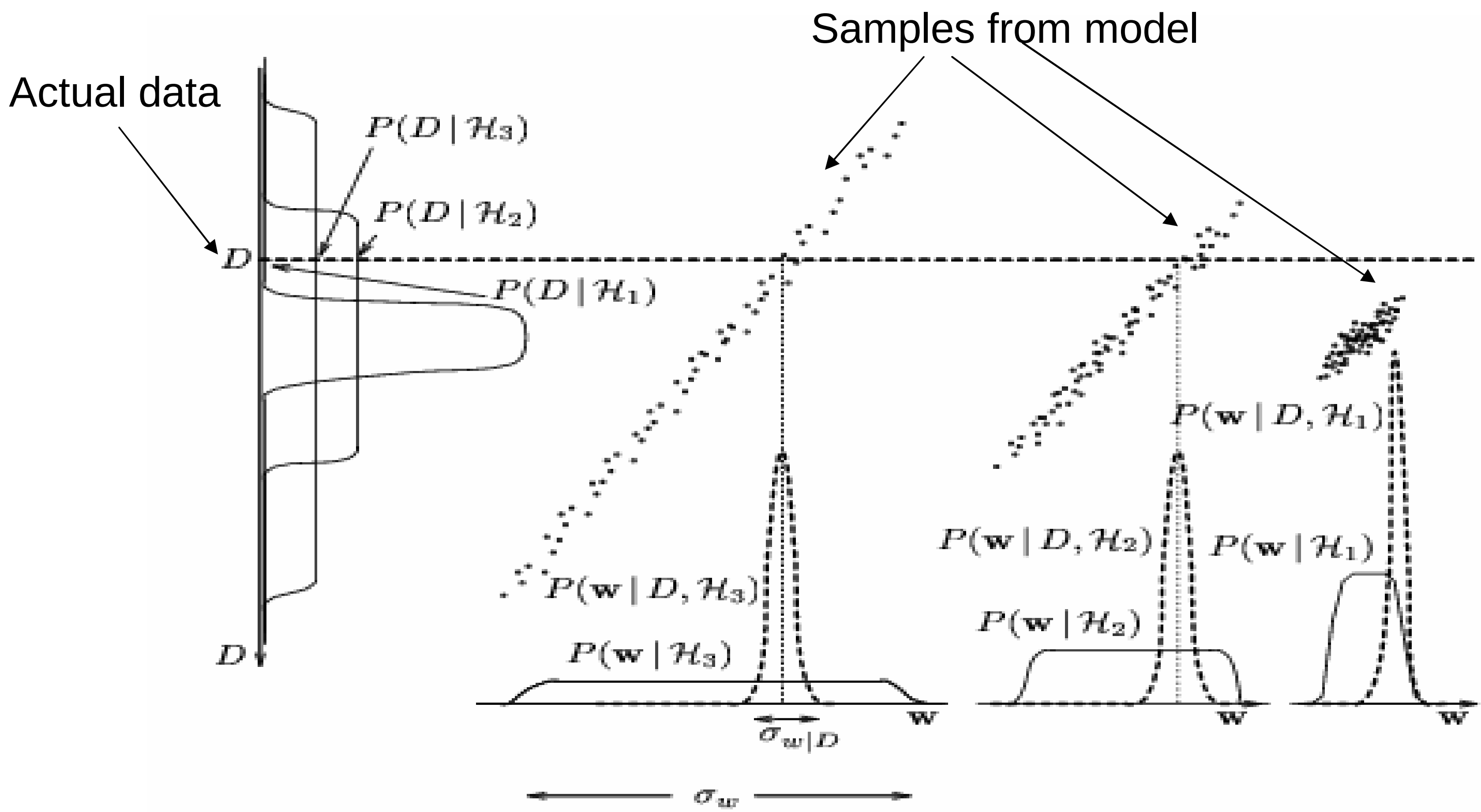
# Bayesian Occam's razor

- The use of the *marginal* likelihood p(D|M) automatically penalizes overly complex models, since they spread their probability mass very widely (predict that everything is possible), so the probability of the actual data is small.




Bishop 3.13

# Bayesian Occam's razor




Model 3 can generate many data sets; prior is broad, posterior is peaked

Model 1 can only generate a few types of data

Mackay 28.6

# Computing marginal likelihoods

- Let p'(D|θ) and p'(θ) be the unnormalized likelihood and prior. Then

$$
\begin{aligned}
p(\theta|D) &= \frac{1}{p(D)} \frac{1}{Z_l} p'(D|\theta) \frac{1}{Z_0} p'(\theta) = \frac{1}{Z_n} p'(\theta|D) \\
\frac{1}{Z_n} &= \frac{1}{p(D)} \frac{1}{Z_l} \frac{1}{Z_0} \\
p(D) &= \frac{Z_n}{Z_0} \frac{1}{Z_l}
\end{aligned}
$$

- Eg. Beta-bernoulli model

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

- Eg. Normal-Gamma-Normal model

$$
p(D) = \frac{\Gamma(\alpha_n) \beta_0^{\alpha_0}}{\Gamma(\alpha_0) \beta_n^{\alpha_n}} \left( \frac{\kappa_0}{\kappa_n} \right)^{1/2} \left( \frac{1}{2\pi} \right)^{n/2}
$$

# Bayesian hypothesis testing

- Suppose we toss a coin N=250 times and observe $N_1$=141 heads and $N_0$=109 tails.
- Consider two hypotheses: $H_0$ that $\theta$=0.5 and $H_1$ that $\theta \neq 0.5$. Actually, we can let $H_1$ be $p(\theta) = U(0,1)$, since $p(\theta=0.5|H_1) = 0$ (pdf).
- For $H_0$, marginal likelihood is

$$p(D|H_0) = 0.5^N$$

- For $H_1$, marginal likelihood is

$$P(D|H_1) = \int_0^1 P(D|\theta, H_1)P(\theta|H_1)d\theta = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)}$$

# Bayes factors

- To compare two models, use posterior odds

$$O_{ij} = \frac{p(M_i|D)}{p(M_j|D)} = \frac{p(D|M_i)p(M_i)}{p(D|M_j)p(M_j)}$$

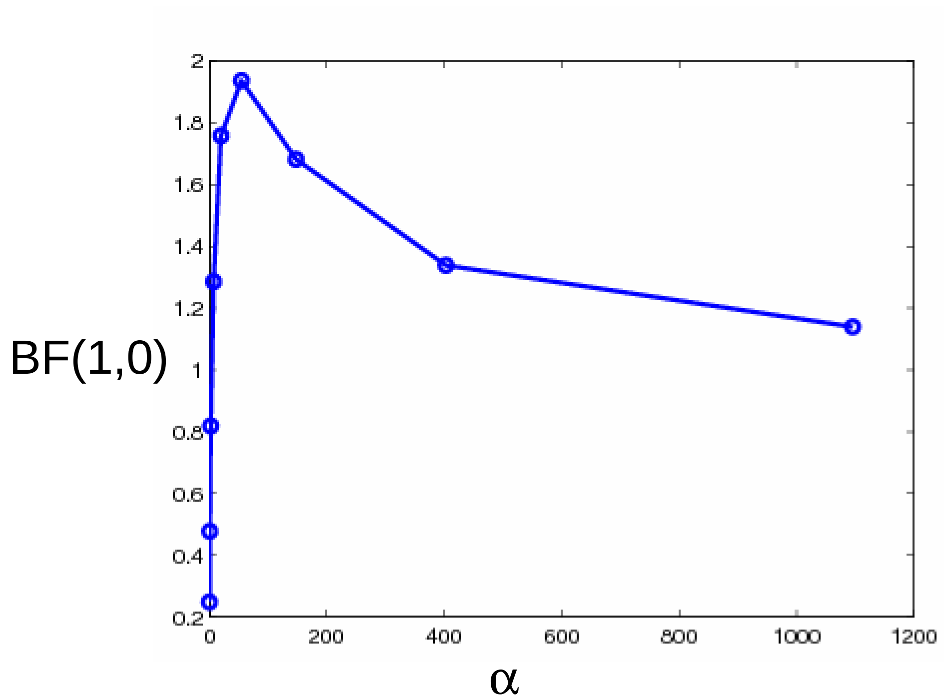
Posterior odds     Bayes factor     Prior odds

- If the priors are equal, it suffices to use the BF.
- The BF is a Bayesian version of a likelihood ratio test, that can be used to compare models of different complexity. If BF(i,j)>>1, prefer model i.
- For the coin example,

$$BF(1,0) = \frac{P(D|H_1)}{P(D|H_0)} = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)} \frac{1}{0.5^N}$$

# Bayes factor vs prior strength

- Let $\alpha_1 = \alpha_0$ range from 0 to 1000.
- The largest BF in favor of H1 (biased coin) is only 2.0, which is only very weak evidence of bias.

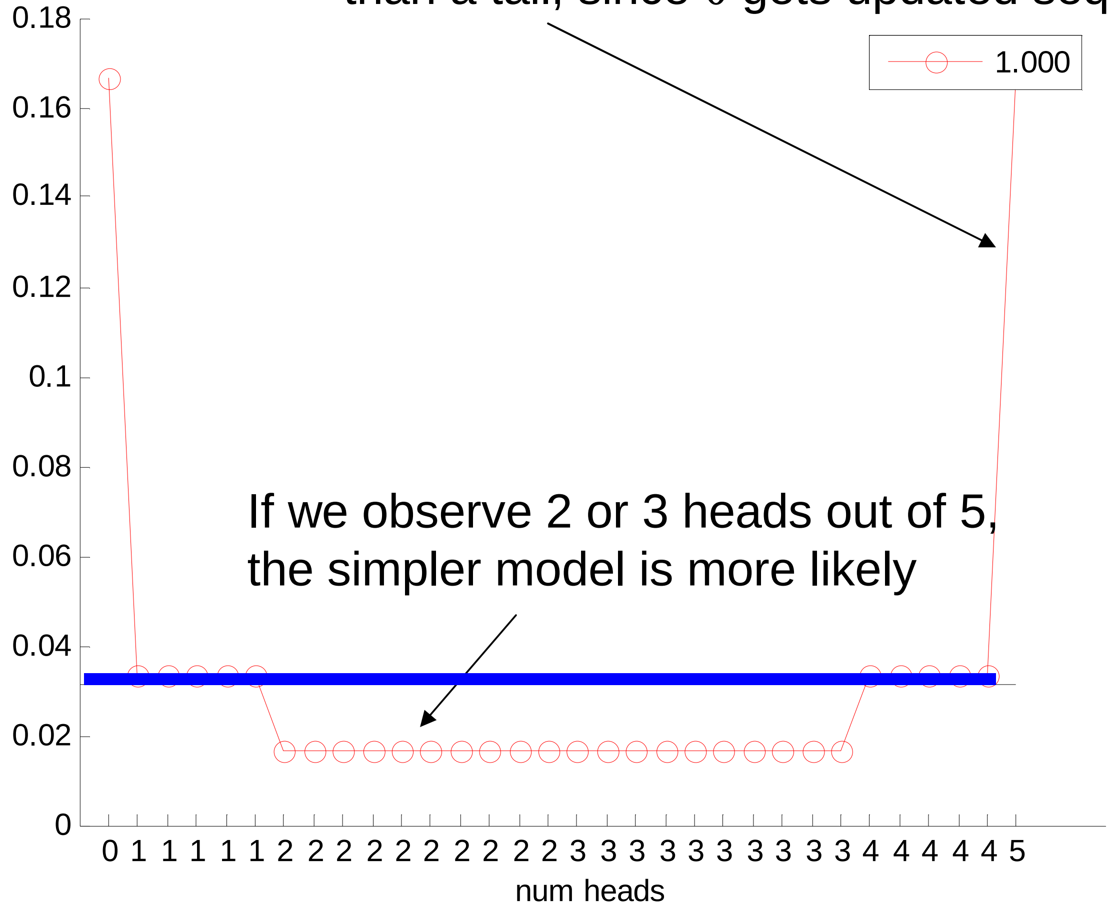


# Bayesian Occam's razor for biased coin

Blue line = $p(D|H_0) = 0.5^N$

Red curve = $p(D|H_1) = \int p(D|\theta)\, Beta(\theta|1,1)\, d\theta$

If we have already observed 4 heads, it is much more likely to observe a 5th head than a tail, since θ gets updated sequentially.




If we observe 2 or 3 heads out of 5, the simpler model is more likely

# CS340 Machine learning
# Frequentist parameter estimation

# Parameter estimation

- We have seen how Bayesian inference offers a principled solution to the parameter estimation problem.
- However, when the number of samples (relative to the number of parameters) is large, we can often approximate the posterior as a delta function centered on the MAP estimate.

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

- An even simpler approximation is to just use the maximum likelihood estimate

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

# Why maximum likelihood?

- Recall that the KL divergence from the true distribution p to the approximation q is

$$
\begin{aligned}
KL(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\
&= \text{const} - \sum_x p(x) \log q(x)
\end{aligned}
$$

- Let p be the empirical distribution

$$p_{emp}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

# ML = min KL to empirical

- KL to the empirical

$$
\begin{aligned}
KL(p_{emp}||q) &= C - \sum_x [\frac{1}{n}\sum_i \delta(x - x_i)] \log q(x) \\
&= C - \frac{1}{n}\sum_i \log q(x_i)
\end{aligned}
$$

- Hence minimizing KL is equivalent to minimizing the average negative log likelihood on the training set

# Computing the Bernoulli MLE

- We maximize the log-likelihood

$$\begin{aligned} \ell(\theta) &= N_1 \log \theta + N_0 \log(1-\theta) \\ \frac{d\ell}{d\theta} &= \frac{N_1}{\theta} - \frac{N - N_1}{1-\theta} \\ &= 0 \\ &\Rightarrow \\ \hat{\theta} &= \frac{N_1}{N} \end{aligned}$$

Empirical fraction of heads eg. 47/100

# Regularization

- Suppose we toss a coin N=3 times and see 3 tails. We would estimate the probability of heads as 0.

$$\hat{\theta} = \frac{0}{3}$$

- Intuitively, this seems unreasonable. Maybe we just haven't seen enough data yet (*sparse data problem*).
- We can add *pseudo counts* $C_0$ and $C_1$ (e.g., 0.1) to the sufficient statistics $N_0$ and $N_1$ to get a better behaved estimate.

$$\hat{\theta} = \frac{N_1 + C_1}{N_0 + N_1 + C_0 + C_1}$$

- This is the MAP estimate using a Beta prior.

# MLE for the multinomial

- If $x_n \in \{1,\ldots,K\}$, the likelihood is

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

- The $N_i$ are the sufficient statistics
- The log-likelihood is

$$\ell(\theta) = \sum_k N_k \log \theta_k$$

# Computing the multinomial MLE

- We maximize L(θ) subject to the constraint $\sum_k \theta_k = 1$.
- We enforce the constraint using a *Lagrange multiplier* λ.

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

- Taking derivatives wrt $\theta_k$

$$\frac{\partial \tilde{\ell}}{\partial \theta_k} = \frac{N_k}{\theta_k} - \lambda = 0$$

- Taking derivatives wrt λ yields the constraint

$$\frac{\partial \tilde{\ell}}{\partial \lambda} = \left(1 - \sum_k \theta_k\right) = 0$$

# Computing the multinomial MLE

- Using the sum-to-one constraint, we get

$$
\begin{aligned}
N_k &= \lambda \theta_k \\
\sum_k N_k &= \lambda \sum_k \theta_k = \lambda \\
\hat{\theta}_k &= \frac{N_k}{\sum_k N_k}
\end{aligned}
$$

Empirical fraction of counts

- Example: $N_1 = 100$ spam, $N_2 = 10$ urgent, $N_3 = 20$ normal, $\theta = (100/130, 10/130, 20/130)$.
- Can add pseudo counts if some classes are rare.

# Computing the Gaussian MLE

- The log likelihood is

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

$$\ell(\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

- The MLE for the mean is the sample mean

$$\frac{\partial \ell}{\partial \mu} = -\frac{2}{2\sigma^2}\sum_n (x_n-\mu) = 0$$

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

# Estimating σ

- The log likelihood is

$$\ell(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

- The MLE for the variance is the sample variance (see handout for proof)

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{1}{2}\sigma^{-4}\sum_n (x_n - \hat{\mu}) - \frac{N}{2\sigma^2} = 0$$

$$\hat{\sigma^2}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

$$= \frac{1}{N}\sum_n x_n^2 - (\hat{\mu})^2$$

# Sampling distribution

- MLE returns a point estimate $\hat{\theta}(D)$
- In frequentist (classical/ orthodox) statistics, we treat D as random and θ as fixed, and ask how the estimate would change if D changed. This is called the *sampling distribution* of the estimator.

$$p(\hat{\theta}(D)|D \sim \theta)$$

- The sampling distribution is often approximately Gaussian.
- In Bayesian statistics, we treat D as fixed and θ as random, and model our uncertainty with the posterior p(θ|D)

# Unbiased estimators

- The bias of an estimator is defined as

$$\text{bias}(\hat{\theta}) = E\left[\hat{\theta}(D) - \theta | D \sim \theta\right]$$

- An estimator is unbiased if bias=0.
- Eg. MLE for Gaussian mean is unbiased

$$E\hat{\mu} = E\frac{1}{N}\sum_{n=1}^{N} X_n = \frac{1}{N}\sum_n E[X_n] = \frac{1}{N}N\mu = \mu$$

# Estimators for $\sigma^2$

- The MLE for Gaussian variance is biased (HW3)

$$E\hat{\sigma}^2 = \frac{N-1}{N}\sigma^2$$

- It is common to use the following unbiased estimator instead $\hat{\sigma}^2_{N-1} = \frac{N}{N-1}\hat{\sigma}^2$
- This is unbiased

$$E[\hat{\sigma}^2_{N-1}] = E[\frac{N}{N-1}\hat{\sigma}^2] = \frac{N}{N-1}\frac{N-1}{N}\sigma^2 = \sigma^2$$

- In Matlab, var(X) returns $\hat{\sigma}^2_{N-1}$ whereas var(X,1) returns $\hat{\sigma}^2$
- The MLE underestimates the variance (e.g., N=1, no variance) since we use an estimated $\mu$, which is shifted from the true $\mu$ towards the data (see HW3).

# Is being unbiased enough?

- Consider the estimator

$$\tilde{\mu}(x_1, \ldots, x_N) = x_1$$

- This is unbiased

$$E\tilde{\mu}(X_1, \ldots, X_N) = E[X_1] = \mu$$

- But intuitively it is unreasonable since it will not improve, no matter how many samples N we have.

# Consistent estimators

- An estimator is consistent if it converges (in probability) to the true value with enough data

$$P(|\hat{\theta}(D) - \theta| > \epsilon | D \sim \theta) \to 0 \text{ as } |D| \to \infty$$

- MLE is a consistent estimator.

# Bias-variance tradeoff

- Being unbiased is not necessarily desirable! Suppose our loss function is mean squared error

$$MSE = E[\hat{\theta}(D) - \theta)^2 | D \sim \theta]$$

- To minimize MSE, we can either minimize bias or minimize variance. Define

$$\overline{\theta} = E[\hat{\theta}(D) | D \sim \theta]$$

- Then

$$\begin{aligned}
E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \theta)^2 &= E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta} + \overline{\theta} - \theta)^2 \\
&= E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta})^2 + 2(\overline{\theta} - \theta)E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta}) + (\overline{\theta} - \theta)^2 \\
&= E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta})^2 + (\overline{\theta} - \theta)^2 \\
&= V(\hat{\theta}) + \text{bias}^2(\hat{\theta})
\end{aligned}$$

$$E_D(\hat{\theta}(D) - \overline{\theta}) = \overline{\theta} - \overline{\theta} = 0$$

We will frequently use biased estimators!

Not on exam