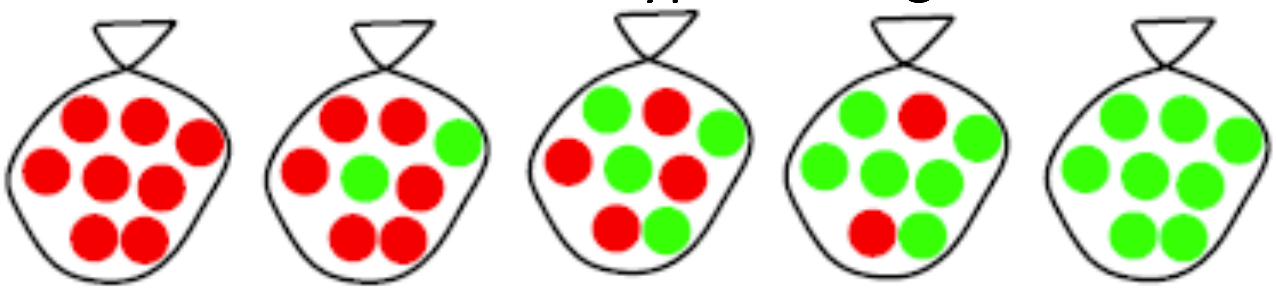
Maximum Likelihood (ML) and Maximum A Posteriori (MAP) estimation

Announcements

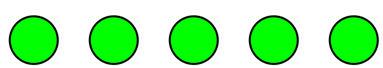
- A2 posted, sign up and create your teams
- Midterm exam 10/26 6:30pm-7:45pm
  - Practice materials posted on Canvas (+tophat)
- Don’t forget the quiz (deadline on Friday)
- A0 grades, submit your regrade requests following instructions in the syllabus
- Next class (Monday) will be via zoom (we will do another coding exercise)

Candy Example

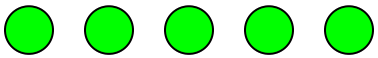
- Candy comes in 2 flavors, cherry and lime
- Manufacturer makes 5 types of bags:

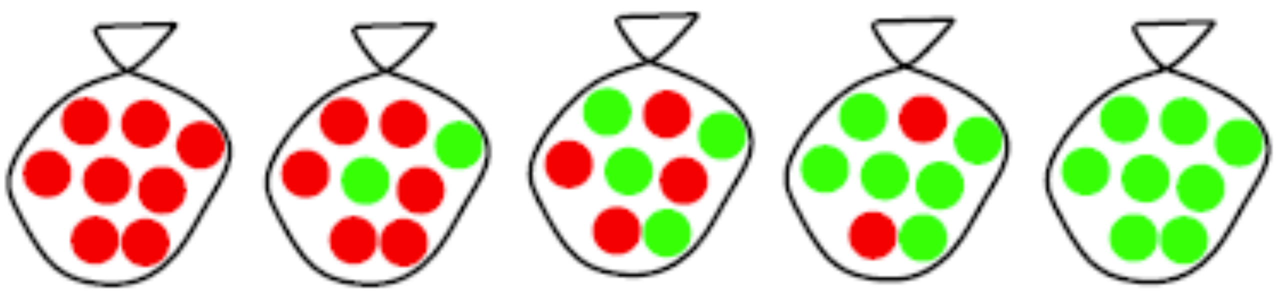


h1	h2	h3	h4	h5
C: 100%	C: 75%	C: 50%	C: 25%	C: 0%
L: 0%	L: 25%	L: 50%	L: 75%	L: 100%

- h1 and h5 are equally common. h2 is twice as common as h1, h4 is twice as common as h5, and h3 is twice as common as h2.
- Suppose we draw 
- Which bag are drawing from?

Bayesian Learning

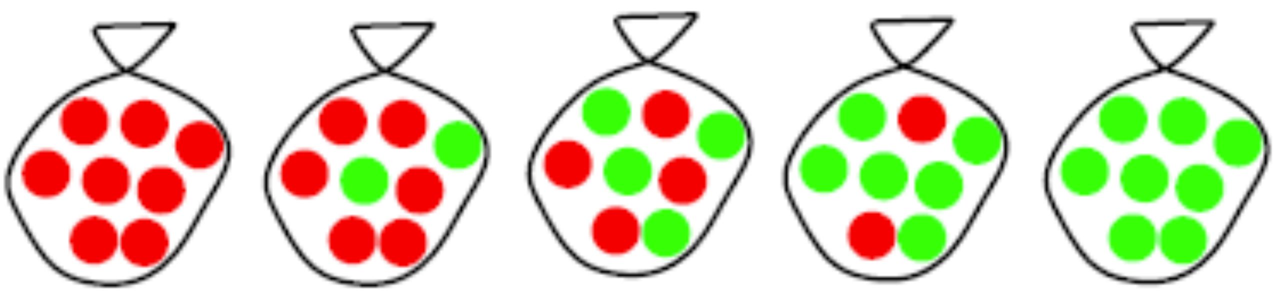
- Main idea: Compute the probability of each hypothesis, given the data
- Data **d**: 
- Hypotheses: $h_1,\ldots,h_5$



h1	h2	h3	h4	h5
C: 100%	C: 75%	C: 50%	C: 25%	C: 0%
L: 0%	L: 25%	L: 50%	L: 75%	L: 100%

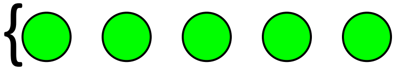
Using Bayes' Rule

- $P(h_i|\mathbf{d}) = \alpha\, P(\mathbf{d}|h_i)\, P(h_i)$ is the **posterior**
  - (Recall, $1/\alpha = P(\mathbf{d}) = \sum_i P(\mathbf{d}|h_i)\, P(h_i)$)
- $P(\mathbf{d}|h_i)$ is the **likelihood**
- $P(h_i)$ is the **hypothesis prior**



h1	**h2**	**h3**	**h4**	**h5**
C: 100%	C: 75%	C: 50%	C: 25%	C: 0%
L: 0%	L: 25%	L: 50%	L: 75%	L: 100%

Computing the Posterior

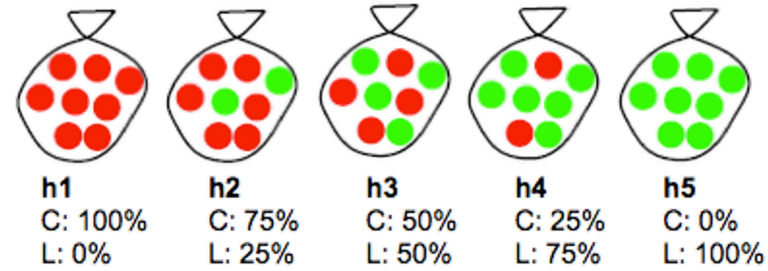
- Assume draws are independent
- Let $P(h_1),\ldots,P(h_5) = (0.1, 0.2, 0.4, 0.2, 0.1)$
- $\mathbf{d} = \{$$\}$

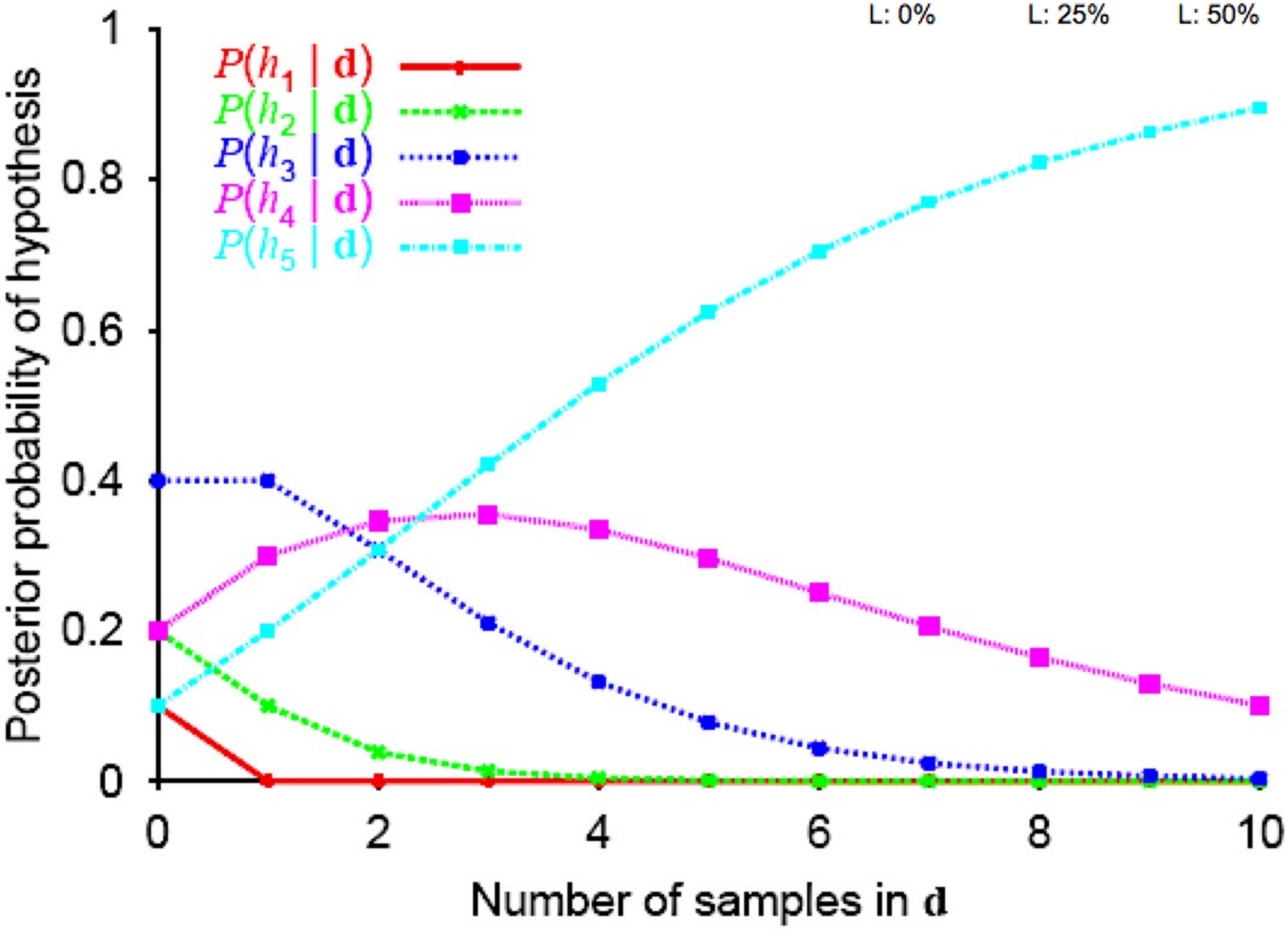
$P(d\|h_1) = 0$	$P(\mathbf{d}\|h_1)P(h_1) \approx 0$	$P(h_1\|\mathbf{d}) = 0$
$P(d\|h_2) = 0.25^5$	$P(\mathbf{d}\|h_2)P(h_2) \approx 1.9\text{e-}4$	$P(h_2\|\mathbf{d}) \approx 1.2\text{e-}3$
$P(d\|h_3) = 0.5^5$	$P(\mathbf{d}\|h_3)P(h_3) \approx 1.2\text{e-}2$	$P(h_3\|\mathbf{d}) \approx 0.078$
$P(d\|h_4) = 0.75^5$	$P(\mathbf{d}\|h_4)P(h_4) \approx 4.7\text{e-}2$	$P(h_4\|\mathbf{d}) \approx 0.29$
$P(d\|h_5) = 1^5$	$P(\mathbf{d}\|h_5)P(h_5) \approx 0.1$	$P(h_5\|\mathbf{d}) \approx 0.62$
	$P(\mathbf{d}) \approx 0.16$	

Posterior Hypotheses

Let $P(h_1),\ldots,P(h_5) = (0.1, 0.2, 0.4, 0.2, 0.1)$

Data: All our samples are limes.





Parameter estimation

- Assume that data is believed to follow some distribution or model (e.g. Gaussian, Poisson, etc.), represented as M
  - Maybe by looking at the histogram of the data we suspect that data is Gaussian or Uniform
  - Maybe we know something about the process that generated the data
- Unfortunately
  - The model has unknown parameter(s) Θ (e.g. model parameters -> μ *or* σ)
  - Observe a random sample from distribution (independent, identically distributed): $X_1,\ldots,X_n$ (i.i.d) ~ $P(X \mid \Theta)$
  - Want to estimate parameter (Θ) from the data, $D = \{X_i : \widehat{\Theta}\},\ i = 1 \ldots N$
- How do we compute the "best" or "optimal" parameter estimates from the data?

Parameter estimation

- Generally, there are two types of parameter estimation approaches:
  - Maximum Likelihood Estimation (MLE)

$$M_{ML} = \underset{M \in \mathcal{M}}{\arg\max} \left\{ p(\mathcal{D}|M) \right\}.$$

  - Maximum a posteriori (MAP) Estimation

$$M_{MAP} = \underset{M \in \mathcal{M}}{\arg\max} \left\{ p(M|\mathcal{D}) \right\}$$

Probability components

- From Bayes Rule

$$p(M|\mathcal{D}) = \frac{p(\mathcal{D}|M) \cdot p(M)}{p(\mathcal{D})},$$

- *P(M|*D) is the **posterior** distribution of the model given the data (or observation)

- P(D|M*)* is the **likelihood** of the data given the model

- *P(M)* is the **prior** distribution of the model

- *P(D)* is the marginal distribution of the data

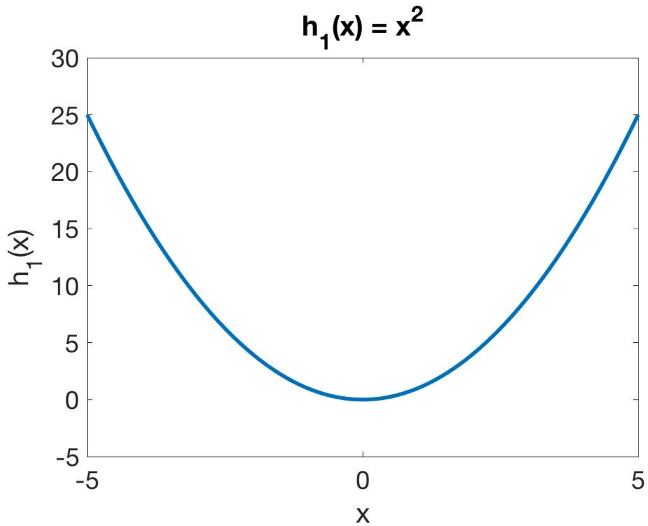
$$p(\mathcal{D}) = \begin{cases} \sum_{f \in \mathcal{F}} p(\mathcal{D}|f)p(f) & f : \text{discrete} \\ \\ \int_{\mathcal{F}} p(\mathcal{D}|f)p(f)df & f : \text{continuous} \end{cases}$$

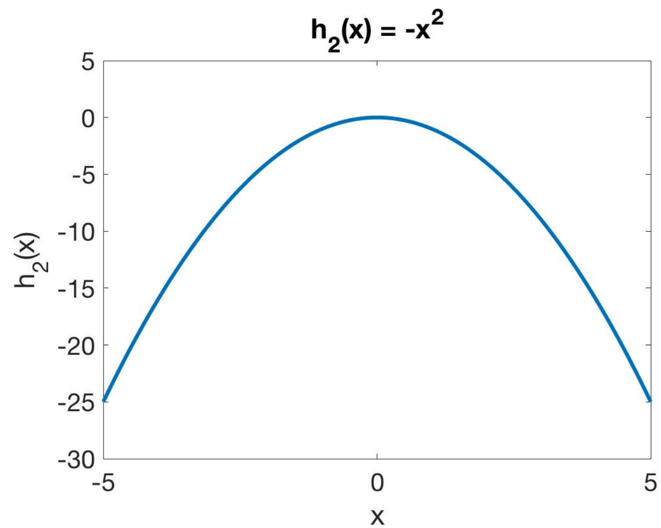
Calculus Review: Function Max or Min

- To find max or min of a function *h(x):*
  1. Take the first and second derivatives of *h(x)*with respect to *x*
  2. Set the first derivative to zero and solve for *x (i.e. "value")*
  3. Evaluate the second derivative of *h(x)* using the solution(s) from step 2
     1. If h''("value") > 0, then minimum at x = "value"
     2. If h''("value") < 0, then maximum at x = "value"

Calculus Review: Function Max or Min

- Examples: $h_1(x) = x^2$ and $\mathbf{h_2(x) = -x^2}$
  - *$h_1'(x) = 2x$, $h_1''(x) = 2$;* ***$h_2'(x) = -2x$, $h_2''(x) = -2$***
  - *$h_1'(x) = 0 \Rightarrow x = 0$…* ***$h_2'(x) = 0 \Rightarrow x = 0$***
  - *For $h_{1:}$ minimum at $x = 0$;* ***For $h_{2:}$ maximum at $x = 0$***
- ***Note***: *maximizing $f(x)$ is equivalent to minimizing $-f(x)$*

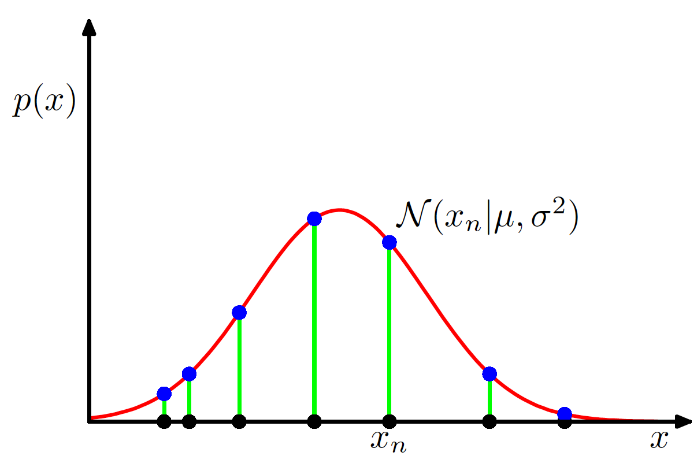




ML Estimation

$$M_{ML} = \arg\max_{M \in \mathcal{M}} \{p(\mathcal{D}|M)\}\,.$$

- MLE chooses Θ to best explain the data (assuming i.i.d)
  - Assumes no knowledge of prior distribution of the model or data



Example: ML Estimation for Poisson distribution

Example 8: Suppose data set $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ is an i.i.d. sample from a Poisson distribution with an unknown parameter λ_t. Find the maximum likelihood estimate of λ_t.

The probability density function of a Poisson distribution is expressed as $p(x|\lambda) = \lambda^x e^{-\lambda}/x!$, with some parameter $\lambda \in \mathbb{R}^+$. We will estimate this parameter as

$$\lambda_{ML} = \arg\max_{\lambda \in (0,\infty)} \{p(\mathcal{D}|\lambda)\}. \tag{2.2}$$

Example: ML Estimation for Poisson distribution

- Poisson Distribution $p(x|\lambda) = \lambda^x e^{-\lambda}/x!$
- We can write the likelihood function as

$$\begin{aligned} p(\mathcal{D}|\lambda) &= p(\{x_i\}_{i=1}^n \,|\lambda) \\ &= \prod_{i=1}^n p(x_i|\lambda) \\ &= \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}. \end{aligned}$$

- To make it easier to find λ that maximizes the likelihood we take a log (since log is a monotonic function it won't affect the result)
- We express the log-likelihood as $ll(D, \lambda) = \ln p(\mathcal{D}|\lambda)$

$$ll(\mathcal{D}, \lambda) = \ln \lambda \sum_{i=1}^n x_i - n\lambda - \sum_{i=1}^n \ln(x_i!)$$

Example: ML Estimation for Poisson distribution

- Next we take the first derivative with respect to λ

$$ll(\mathcal{D},\lambda) = \ln\lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln(x_i!)$$

$$\frac{\partial ll(\mathcal{D},\lambda)}{\partial\lambda} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n$$
$$= 0.$$

- And we find that λ_{ML} is equal to the sample mean

$$\lambda_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- When a distribution is parametrized by its mean (as with Poisson, remember E(p(x)) =λ) it often happens that

$$\hat{\Theta}_{MLE} = \bar{x}$$

MAP Estimation

$$M_{MAP} = \underset{M \in \mathcal{M}}{\arg\max} \{p(M|\mathcal{D})\}$$

$$p(M|\mathcal{D}) = \frac{p(\mathcal{D}|M) \cdot p(M)}{p(\mathcal{D})}$$

- As before have $X_1, \ldots X_n$ (i.i.d) ~P(X|Θ), and want to estimate Θ
- Suppose now we have a prior belief on Θ, expressed as a prob. distribution: Θ ~ P(Θ)
- Using Bayes Rule, can compute the posterior distribution

$$P(\Theta|x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n|\Theta)P(\Theta)}{P(x_1, \ldots, x_n)}$$

This reflects everything that we know about Θ after observation

Most likely Θ given knowledge

- Then, MAP estimate is: $\hat{\Theta}_{MAP} = \underset{\Theta}{\arg\max}\, P(\Theta|x_1, \ldots, x_n)$

MAP Estimation

$$P(\Theta|x_1,...,x_n) = \frac{P(x_1,...,x_n|\Theta)P(\Theta)}{P(x_1,...,x_n)}$$

$$\hat{\Theta}_{MAP} = \arg\max_{\Theta} P(\Theta|x_1,...,x_n)$$

- Since $P(X_1,\ldots,X_n)$ is constant once observed,

$$\hat{\Theta}_{MAP} = \arg\max_{\Theta} P(\Theta|x_1,...,x_n) = \arg\max_{\Theta} P(\Theta)P(x_1,...,x_n|\Theta)$$

MAP Estimation

- The same thing written in a different way

$$M_{MAP} = \arg\max_{M \in \mathcal{M}} \{p(M|\mathcal{D})\}$$

$$p(M|\mathcal{D}) = \frac{p(\mathcal{D}|M) \cdot p(M)}{p(\mathcal{D})}$$
$$\propto p(\mathcal{D}|M) \cdot p(M)$$

$$p(\mathcal{D}) = \begin{cases} \sum_{M \in \mathcal{M}} p(\mathcal{D}|M)p(M) & M : \text{discrete} \\ \int_{\mathcal{M}} p(\mathcal{D}|M)p(M)dM & M : \text{continuous} \end{cases}$$

$$M_{MAP} = \arg\max_{M \in \mathcal{M}} \{p(\mathcal{D}|M)p(M)\}$$

$$M_{ML} = \arg\max_{M \in \mathcal{M}} \{p(\mathcal{D}|M)\}$$

- Where ∝ is the proportionality symbol

Example: MAP Estimation for Poisson distribution (with Gamma prior on parameters)

Example 9: Let $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ again be an i.i.d. sample from Poisson(λ_t), but now we are also given additional information. Suppose the prior knowledge about λ_t can be expressed using a gamma distribution $\Gamma(x|k, \theta)$ with parameters $k = 3$ and $\theta = 1$. Find the maximum a posteriori estimate of λ_t.

First, we write the probability density function of the gamma family as

$$\Gamma(x|k, \theta) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)},$$

$x > 0$, $k > 0$, and $\theta > 0$

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

Example: MAP Estimation for Poisson distribution (with Gamma prior on parameters)

As before, we can write the likelihood function as

$$p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

and the prior distribution as

$$\lambda_{MAP} = \arg\max_{\lambda \in (0,\infty)} \{p(\mathcal{D}|\lambda)p(\lambda)\}$$

$$p(\lambda) = \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)}.$$

Now, we can maximize the logarithm of the posterior distribution $p(\lambda|\mathcal{D})$ using

$$\begin{aligned}\ln p(\lambda|\mathcal{D}) &\propto \ln p(\mathcal{D}|\lambda) + \ln p(\lambda) \\ &= \ln \lambda (k - 1 + \sum_{i=1}^{n} x_i) - \lambda (n + \frac{1}{\theta}) - \sum_{i=1}^{n} \ln x_i! - k \ln \theta - \ln \Gamma(k)\end{aligned}$$

Example: MAP Estimation for Poisson distribution (with Gamma prior on parameters)

$$
\begin{aligned}
\ln p(\lambda|\mathcal{D}) &\propto \ln p(\mathcal{D}|\lambda) + \ln p(\lambda) \\
&= \ln \lambda(k - 1 + \sum_{i=1}^{n} x_i) - \lambda(n + \frac{1}{\theta}) - \sum_{i=1}^{n} \ln x_i! - k \ln \theta - \ln \Gamma(k)
\end{aligned}
$$

After taking partial derivative with respect to λ and equaling to 0, we can solve for λ

$$
\begin{aligned}
\lambda_{MAP} &= \frac{k - 1 + \sum_{i=1}^{n} x_i}{n + \frac{1}{\theta}} \\
&= 5
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{D} &= \{2, 5, 9, 5, 4, 8\} \\
k &= 3 \\
\theta &= 1
\end{aligned}
$$

ML converges to MAP when we have many samples

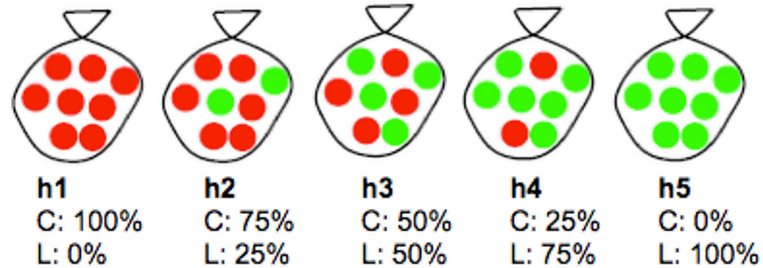
- Notice that with ML estimation we had λ_{ML}=5.5 and with MAP estimation we got λ_{MAP}=5.
- In the limit of infinite samples, both the MAP and ML converge to the same model, M (as long as the prior does not have zero probability on M).
- In other words, large data diminishes the importance of prior knowledge.
- To get some intuition for this result, we will show that the MAP and ML estimates converge to the same solution for the above example with a Poisson distribution.

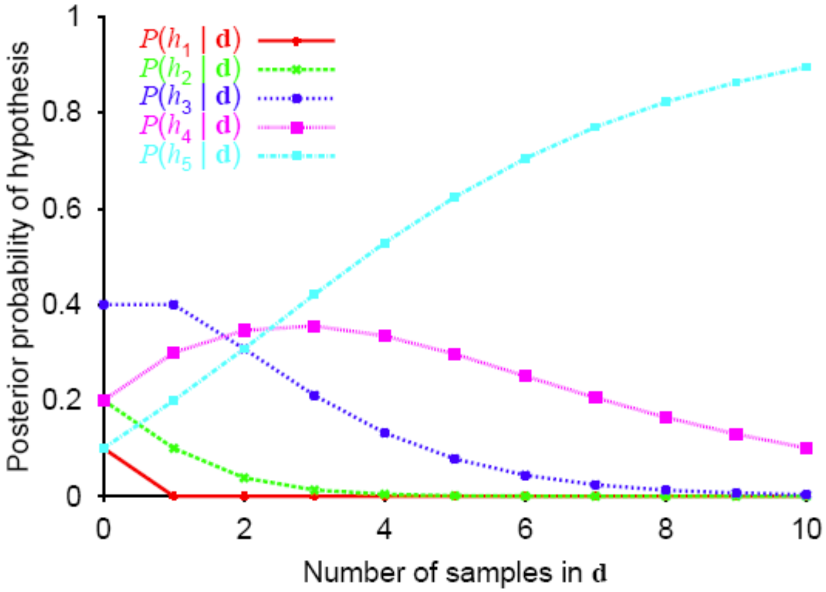
Back to candy

Let $P(h_1),\ldots,P(h_5) = (0.1, 0.2, 0.4, 0.2, 0.1)$

Data: All our samples are limes.

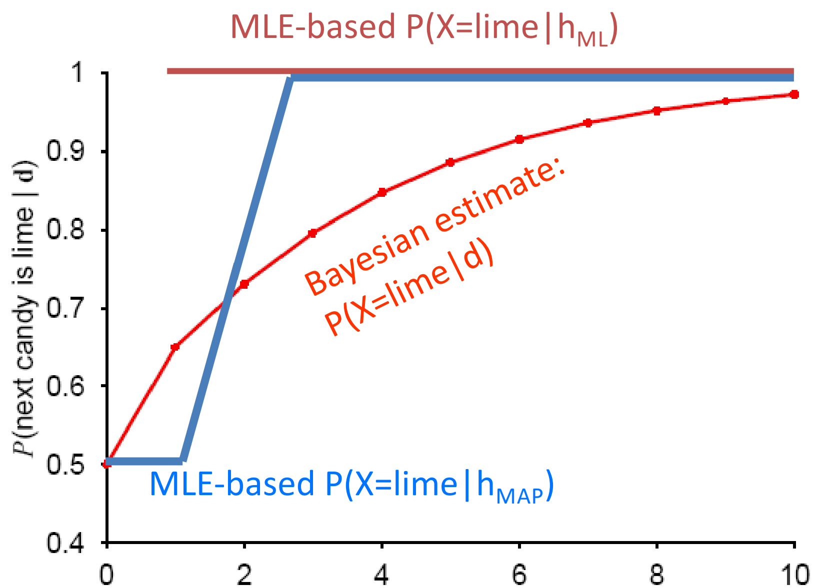
$h_{MAP} = \text{argmax}_i\ P(h_i|\mathbf{d})$ $\quad$ $h_{ML} = \text{argmax}_i\ P(\mathbf{d}|h_i)$





# samples:	0	1	2	3	4	5	6	7	8	9	10
h_{MAP}	h3	h3	h4	h5	h5	h5	h5	h5	h5	h5	h5
h_{ML}		h5	h5	h5	h5	h5	h5	h5	h5	h5	h5

What is probability next candy is lime?



Next class

- Zoom coding exercise on Viterbi algorithm