Tutorial
MLE, MAP, Full Bayesian:
Parameter Estimation

Recommended Reading: Murphy chapter 3

1 The importance of a good prior

We have seen that a prior can help mitigate overfitting of the maximum likelihood estimate. But setting a prior causes *inductive bias*: Certain solutions are preferred over others for subjective reasons. (*Subjective* means they are not motivated purely mathematically. They may be objective or reasonable from our intuitive understanding of the problem.)

Often, this is desired—certain model parameters indeed are more likely “from experience”. In this exercise, however, we will see how a sloppy choice of a prior can impose a harmful inductive bias.

Problem 1: You are visiting Alice's casino. You have been gambling in this casino for years, and you have no doubt in Alice's integrity. Today when you arrive at the casino, she is in hospital and her son Bob has taken over the casino. You don't know Bob much, but him being Alice's offspring, you are sure you can trust him just as much. As an eager student of statistics, who just learned about priors, you decide to test him and walk up to your favourite game: Guess the flip! You like the elegant simplicity of the game: You place a bet on the outcome of a coin flip.

Taking into account Bob's splendid family background, you choose a centred Beta distribution as a prior, i.e., parameters $a = b = n > 0$. What you don't know is that Bob is trying to make most money out of his short intermission as the boss of the casino. Not being the most clever guy, he has decided to use coins that *always* show up tails.

Determine how long in terms of n and N it takes to recover from your overwhelming trust:

- Determine ML, MAP, and fully Bayesian estimates for θ, the probability of tails showing up next.
- Interpret these results. Which estimate takes longest to recover? Why? Is this expected from the results in the lecture?

First of all, for those who think this casino and game are made up: `https://youtu.be/8c1BQkKUsx0`

All the requested estimates are given in the slides on flipping a coin:

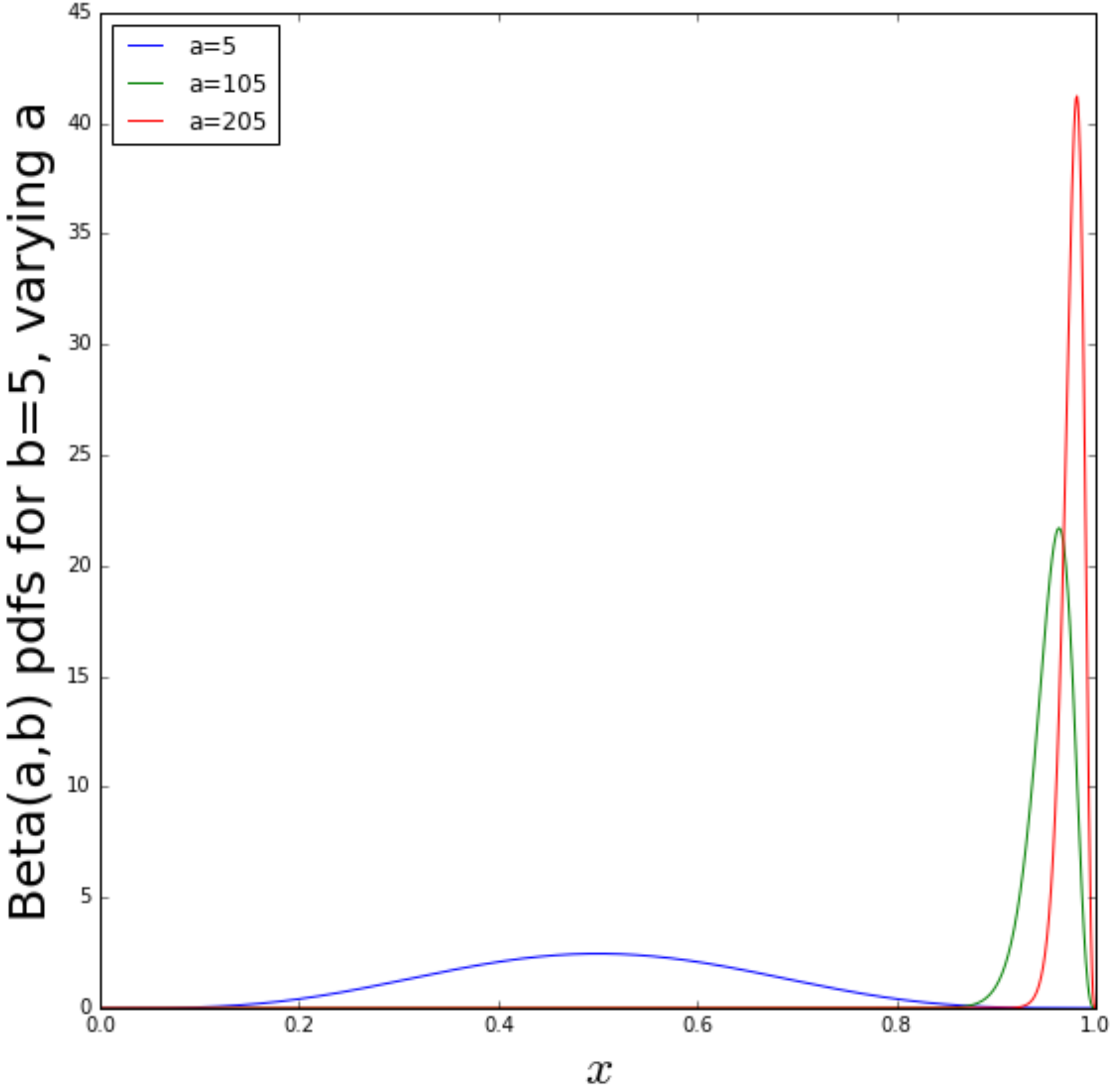
$$\begin{aligned}\theta_{\mathrm{MLE}} &= \frac{N}{N+0} = 1,\\ \theta_{\mathrm{MAP}} &= \frac{N+n-1}{N+2(n-1)} = 1 - \frac{n-1}{N+2(n-1)},\\ \theta_{\mathrm{FB}} &= \frac{N+n}{N+2n} = 1 - \frac{n}{N+2n}.\end{aligned}$$

To see which estimate takes longest to recover, let's look how many flips N it takes to push the estimate from the prior mode (maximum of density) 0.5 to 0.95:

$$\theta_{\mathrm{MAP}} \geq 0.95 \Leftrightarrow \frac{n-1}{N+2(n-1)} \leq 0.05 \Leftrightarrow N \geq 18(n-1),$$

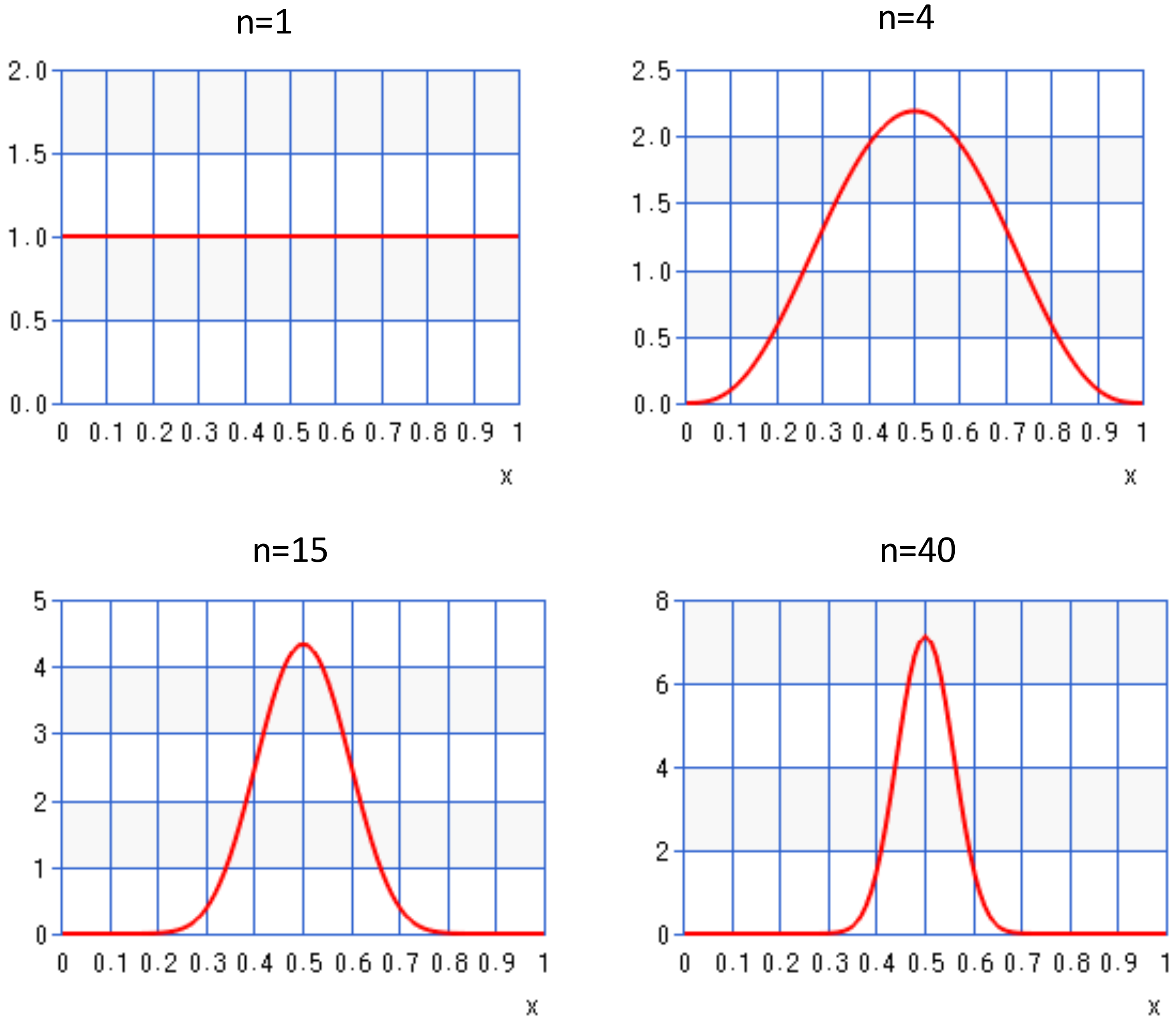
and similarly $N \geq 18n$ for θ_{FB} to recover. In other words: MLE works better than MAP works better than fully Bayes. But that's not what we learned!

Actually it is. For $n > 1$ the prior is far off the true parameter, to the extent that the true parameter almost has no probability. The MAP working better than fully Bayes is caused by the extreme true value: Because the true value is on the margin of the feasible interval, the posterior's tail to the left becomes a problem rather than information we would like to use (as we do in the fully Bayes case). This is a direct manifestation of inductive bias towards (in this case) completely wrong solutions. The following figure illustrates this:


Beta(a,b) pdfs for b=5, varying a
a=5
a=105
a=205
45
40
35
30
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0
x

Beta distributions: varying n:

http://keisan.casio.com/has10/SpecExec.cgi?id=system/2006/1180573226



Notice how even a fairly weak prior like $n = 1.5$ requires 9 and 27 trials, respectively, to recover. Quite some wasted earnings given that Bob's strategy is actually a losing one without adapting the odds.

2 Mark, you, and the coin

Once again, you're sitting in Mark's office. This time however, it is not Mark who sits in the nice leather chair, no, it is his speaking parrot Zucky. Mark himself is having a relaxing bath in his personal spa right next to the office. You can't see him, but because of a funny *bling* noise coming from the spa you know that Mark is engaged in his favourite past time, tossing gold coins. Mark shouts: "Dude, last time we talked, you seemed to have a knack for golden coin tossing, that's why you are here. I wonder if this coin here is biased. Let me flip it a couple of times for you and you tell me what you think!" He starts tossing (according to the *blings*) so you shout back: "Yo, Mark, you know, would be nice if I could *see* every toss… Maybe you can shout what every toss results in?" Mark starts: "Ok, ehrrr, no, sorry, need to call my friend Larry. Let's do it like this then: I toss, and Zucky will tell you the result. Oh, wait, right, Zucky finds it funny to tell sometimes, ehrrr, not the truth, don't know where he picked that habit. Check out the sample run I did with him yesterday, it is the paper lying right next to you. I'll start tossing when you're done with your math magic, just let me know, Zucky and I are waiting. Yo, Larry, …"

You sort your thoughts and start modelling: Denote with f the result of a coin flip ($f = 0$ is heads, $f = 1$ is tails). Model the bias of the coin with θ_1 and use θ_2 for Zucky's *truthfulness*. Zucky's answer is denoted by z. Furthermore, assume that θ_2 is independent of f and θ_1. Thus, $p(z \mid f, \theta_2)$ is given as:

	$z = 0$	$z = 1$
$f = 0$	θ_2	$1 - \theta_2$
$f = 1$	$1 - \theta_2$	θ_2

Problem 2: Make a *similar* 2×2 table for the joint probability distribution $p(f, z \mid \boldsymbol{\theta})$ in terms of $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Show your work. Note that the likelihood function $p(f, z \mid \theta_1, \theta_2)$ factorises and simplifies under our independence assumptions, i.e.,

Problem 2: Make a *similar* 2 × 2 table for the joint probability distribution $p(f, z \mid \boldsymbol{\theta})$ in terms of $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Show your work. Note that the likelihood function $p(f, z \mid \theta_1, \theta_2)$ factorises and simplifies under our independence assumptions, i.e.,

$$p(f, z \mid \theta_1, \theta_2) = p(z \mid f, \theta_2) p(f \mid \theta_1).$$

We can reuse the above table: The first factor in the joint distribution is the previous table. Each cell only needs to be multiplied with the second factor $p(f \mid \theta_1)$.

	$z = 0$	$z = 1$
$f = 0$	$\theta_2(1 - \theta_1)$	$(1 - \theta_2)(1 - \theta_1)$
$f = 1$	$(1 - \theta_2)\theta_1$	$\theta_2\theta_1$

Problem 3: The sample run on the paper looks like this:

f	1	1	0	1	1	0	0
z	1	0	0	0	1	0	1

What are the maximum likelihood estimates for θ_1 and θ_2? Justify your answer.

Problem 3: The sample run on the paper looks like this:

f	1	1	0	1	1	0	0
z	1	0	0	0	1	0	1

What are the maximum likelihood estimates for θ_1 and θ_2? Justify your answer.

Assuming i.i.d. data, we can determine the likelihood by multiplying the factors from the table from the previous exercise for each data point (column on the paper).

The log likelihood of the dataset is $4\log\theta_2 + 3\log(1-\theta_2) + 3\log(1-\theta_1) + 4\log\theta_1$. Thus the MLE of θ_1 is $4/7$ and of θ_2 is $4/7$.

Notice that we could have arrived at this result by doing MLE individually, since we assumed independence of Zucky's truthfulness from the outcome. However, this crucially relies on the *pre-processing* of data on the piece of paper, where Zucky's answer (raw data) has already been converted into a truthfulness value.

3 The probabilistic coin game

In the following we are considering a more involved version of predicting coin tosses. Instead of one coin that we observe tosses from, two coins with different characteristics exist. At the beginning of a series of N coin flips, one of the two coins is drawn randomly and with this coin the observed tosses are performed. After N tosses the goal is to predict the outcome of the next flip with this coin.

One of the two coins is drawn randomly and 10 coin tosses are made: 7 heads and 3 tails.

Assume for coin number 1 a prior of $p(\theta \mid c = 1) = \text{Beta}(\theta \mid 4, 4)$ and for coin number 2 a prior of $p(\theta \mid c = 2) = \text{Beta}(\theta \mid 6, 2)$. The overall prior for a randomly drawn coin should be $p(\theta) = 0.5p(\theta \mid c = 1) + 0.5p(\theta \mid c = 2)$.

Problem 4: Why is this overall prior a valid assumption? Argue in 2–3 sentences.

The Beta distribution is the conjugate prior for the parameter of a Bernoulli random variable, so it makes sense as a prior for each variable. One of the coins is drawn randomly, so both coins have weight 0.5.

Problem 5: Compute $p(\theta \mid \mathcal{D})$ where $\mathcal{D}$ denotes the observed data. Show your work! Use the following steps:

1. Write $p(\theta \mid \mathcal{D})$ in terms of $p(\theta, c \mid \mathcal{D})$ for $c = 1$ and $c = 2$.
2. Find an expression that involves the class-dependent posterior of θ, $p(\theta \mid c, \mathcal{D})$ for $c = 1, 2$. Why is this advantageous?
3. Compute an easier expression for this posterior via Bayes' Rule.
4. Why is $p(\mathcal{D} \mid \theta, c) \equiv p(\mathcal{D} \mid \theta)$, i.e., why is the likelihood independent of the class? What will be the posterior distribution?
5. Determine the missing components from step 2, i.e., the factors that are not the class-dependent posterior. If you get stuck, inspect your results from steps 3 and 4 closely to get to a solution.
6. Put the pieces together and determine the posterior distribution $p(\theta \mid \mathcal{D})$.

Steps 1 & 2:

$$p(\theta \mid \mathcal{D}) = \sum_c p(\theta, c \mid \mathcal{D}) = \sum_c p(\theta \mid \mathcal{D}, c) p(c \mid \mathcal{D})$$

For a fixed class c, the expression $p(\theta \mid \mathcal{D}, c)$ is exactly the same thing as in the lecture, where we only had a single coin. We can reuse our knowledge from there.

Steps 3 & 4: As in the lecture:

$$p(\theta \mid \mathcal{D}, c) = \frac{p(\mathcal{D} \mid c, \theta) p(\theta \mid c)}{p(\mathcal{D} \mid c)} \equiv \frac{p(\mathcal{D} \mid \theta) p(\theta \mid c)}{p(\mathcal{D} \mid c)} \tag{1}$$

The latter equivalence is valid because θ uniquely determines the Bernoulli sequence. θ itself is influenced by c, but given θ, the data sequence $\mathcal{D}$ is independent of c. Notice that $\mathcal{D}$ and c are by no means independent. They are just *conditionally independent* given θ.

From the lecture, we know that the individual class posteriors will be Beta distributions. As $p(c \mid \mathcal{D})$ is independent of θ, these are just mixture weights, so that we end up with a mixture of Betas.

Step 5: We have to determine the *mixture weights* $p(c \mid \mathcal{D})$. Once again, we apply Bayes' formula:

$$p(c \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid c)p(c)}{p(\mathcal{D})},$$

where

$$p(\mathcal{D}) = \sum_c p(\mathcal{D} \mid c)p(c).$$

$p(c)$ is easy. $p(\mathcal{D} \mid c)$ is hard, though. Following the hint, we observe that $p(\mathcal{D} \mid c)$ is the normalising constant of the class-dependent posterior $p(\theta \mid \mathcal{D}, c)$ from eq. (1).

Notice: There is a dangerous pitfall here. You might think that you can just take the constant term of the Beta distribution in (1)—but part of this constant is already baked into the numerator. We have to take care of this to be allowed to use our beloved reverse-engineering trick. In the end, we get that

$$p(\mathcal{D} \mid c) = \frac{\text{constant factor of } p(\theta \mid \mathcal{D}, c)}{\text{constant factor of } p(\theta \mid c)}.$$

We know both of these terms. We just need to avoid the fallacy of just using the numerator.

Step 6: Steps 3 & 4, and inserting the values from the assignment give us

$$p(\theta \mid \mathcal{D}, c = 1) = \text{Beta}(\theta \mid 11, 7),$$
$$p(\theta \mid \mathcal{D}, c = 2) = \text{Beta}(\theta \mid 13, 5).$$

Avoiding the pitfall as described in Step 5, the mixture weights are

$$p(\mathcal{D} \mid c = 1) = \frac{\text{constant factor of } p(\theta \mid \mathcal{D}, c = 1)}{\text{constant factor of } p(\theta \mid c = 1)} = \frac{\Gamma(7)\Gamma(11)}{\Gamma(18)} \frac{\Gamma(8)}{\Gamma(4)\Gamma(4)} = \frac{5}{4862},$$
$$p(\mathcal{D} \mid c = 2) = \frac{\text{constant factor of } p(\theta \mid \mathcal{D}, c = 2)}{\text{constant factor of } p(\theta \mid c = 2)} = \frac{\Gamma(5)\Gamma(13)}{\Gamma(18)} \frac{\Gamma(8)}{\Gamma(2)\Gamma(6)} = \frac{3}{2210}.$$

Using $p(c) = 0.5$ (as we draw the coin fairly), we get

$$p(c = 1 \mid \mathcal{D}) = \frac{25}{58}, \qquad p(c = 2 \mid \mathcal{D}) = \frac{33}{58}.$$

Putting pieces together, we end up with

$$p(\theta \mid \mathcal{D}) = \frac{25}{58}\text{Beta}(\theta \mid 11, 7) + \frac{33}{58}\text{Beta}(\theta \mid 13, 5).$$

Problem 6: Sketch in one or two sentences how you then can use the computed posterior in this prediction game. (No computations are required!)

Problem 6: Sketch in one or two sentences how you then can use the computed posterior in this prediction game. (No computations are required!)

The posterior will be used to predict the result of the next coin flip. Using a fully bayesian approach, one can integrate over all possible values of θ and thus make a prediction for the outcome. Also possible, but less valuable, because a point estimate: MAP, the maximum value of the posterior.

more problems ☺

just for reference:

$$\mathrm{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \qquad B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$\mathrm{mean} = \frac{a}{a+b}, \ \ \mathrm{mode} \ = \frac{a-1}{a+b-2}$$

is conjugate to $Ber(x|\theta) = \theta^{I(x=1)} \ (1-\theta)^{I(x=0)}$

as well as $Bin(k|n,\theta) = \binom{n}{k} \ \ \theta^k \ \ (1-\theta)^{n-k}$

just for reference:

$$S_K = \{\mathrm{x} : 0 \leq x_k \leq 1, \sum_{k=1}^{K} x_k = 1\}$$

$$\mathrm{Dir}(\mathrm{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(\mathrm{x} \in S_K)$$

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

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

mean = $\mathbb{E}[x_k] = \frac{\alpha_k}{\alpha_0}, \ \mathrm{mode}[x_k] = \frac{\alpha_k - 1}{\alpha_0 - K}$

is conjugate to $\quad Cat(\mathbf{x}|\boldsymbol{\theta}) = \mathrm{Mu}(\mathrm{x}|1, \theta) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j = 1)}$

as well as $\quad \mathrm{Mu}(\mathrm{x}|n, \boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$

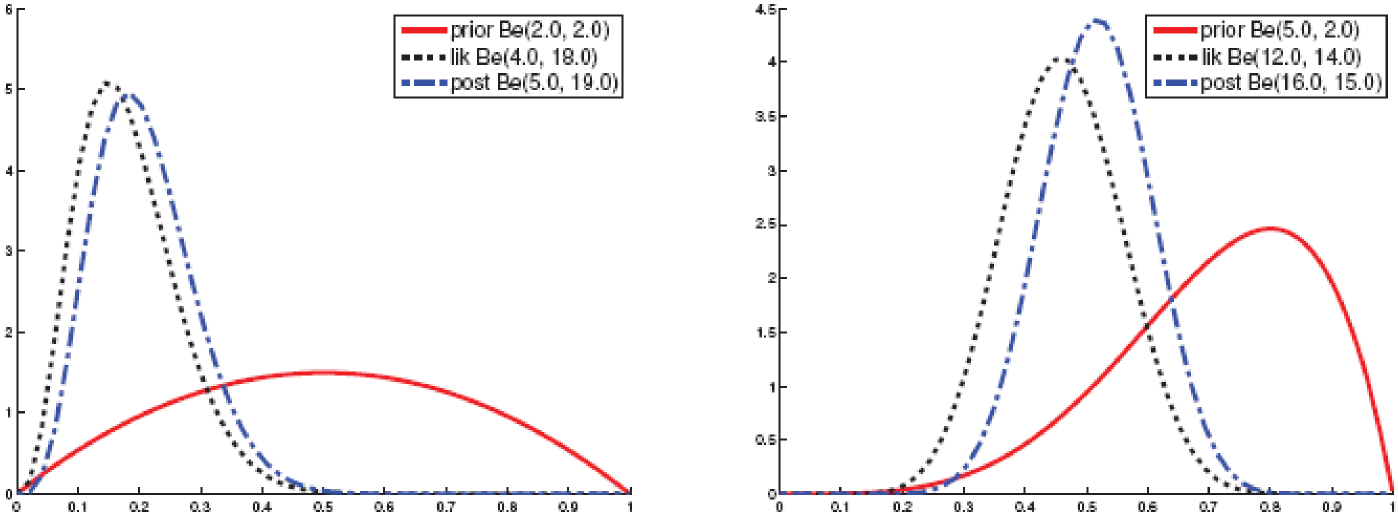
Coin: MLE, MAP for Θ

$$\left.\begin{aligned} p(D|\Theta) &= \prod_{i=1}^{N} \Theta^{\mathbb{1}(F_i=H)}(1-\Theta)^{\mathbb{1}(F_i=T)} \\ &= \Theta^{N_H}(1-\Theta)^{N-N_H} \end{aligned}\right\} \text{likelihood}$$

$$argmax_{\Theta} p(D|\Theta) = \frac{N_H}{N} = \Theta_{MLE}$$ („mode“ of the likelihood of Θ) MLE

$$\left.\begin{aligned} p(\Theta|D) &\propto p(D|\Theta)p(\Theta|a,b) \\ &\propto \Theta^{N_H}(1-\Theta)^{N-N_H} Beta(\Theta|a,b) \\ &\propto Beta(\Theta|N_H+a, N-N_H+b) \end{aligned}\right\} \text{posterior}$$

$$argmax_{\Theta} p(\Theta|D) = \frac{N_H+a-1}{N+a+b-2} = \Theta_{MAP}$$ („mode“ of the posterior of Θ) MAP



(a) (b)

source: Murphy, p 75

Figure 3.6 (a) Updating a Beta(2, 2) prior with a Binomial likelihood with sufficient statistics $N_1 = 3, N_0 = 17$ to yield a Beta(5,19) posterior. (b) Updating a Beta(5, 2) prior with a Binomial likelihood with sufficient statistics $N_1 = 11, N_0 = 13$ to yield a Beta(16, 15) posterior. Figure generated by `binomialBetaPosteriorDemo`.

$$Bin(N_1|\theta, N_0) \;\propto\; \theta^{N_1}(1-\theta)^{N_0} \;\propto\; Be(\theta \mid N_1+1, N_0+1) \;\propto\; \theta^{N_1+1-1}(1-\theta)^{N_0+1-1}$$

(regard that $\binom{N_0+N_1}{N_1} = \frac{(N_0+N_1)!}{N_0!N_1!} \neq \frac{\Gamma(N_1+1+N_0+1)}{\Gamma(N_1+1)\Gamma(N_0+1)} = \frac{(N_0+N_1+1)!}{N_0!N_1!}$)
(←→ regard that $Bin(N_1|\theta, N_0)$ is a discrete prob. dist. in N_1 and $Be(\theta \mid N_1+1, N_0+1)$ is a continuous prob. dist. in θ)
.

Updating a Beta-prior for θ with a Beta-Likelihood of θ (a Binomial distr. in N_1) :
$Be(\theta|a_1,b_1)\, Be(\theta|a_2,b_2) \propto \theta^{a_1-1+a_2-1}(1-\theta)^{b_1-1+b_2-1} = Be(\theta|a_1+a_2-1, b_1+b_2-1)$

Coin: posterior predictive distribution

$$p(F|\Theta_{MLE}) = \Theta_{MLE}^{\mathbb{1}(F=H)}(1-\Theta_{MLE})^{\mathbb{1}(F=T)} \qquad \text{MLE}$$

$$p(F|\Theta_{MAP}) = \Theta_{MAP}^{\mathbb{1}(F=H)}(1-\Theta_{MAP})^{\mathbb{1}(F=T)} \qquad \text{MAP}$$

$$\left.\begin{aligned}
p(F=H|\mathcal{D}) &\overset{\boxed{F \perp\!\!\!\perp \mathcal{D}|\Theta}}{=} \int_0^1 p(F=H|\Theta)p(\Theta|D)d\Theta \\
&= \int_0^1 Ber(F=H|\Theta)Beta(\Theta|N_H+a, N-N_H+b)d\Theta \\
&= \int_0^1 \Theta Beta(\Theta|N_H+a, N-N_H+b)d\Theta \\
&= E_{Beta(\Theta|N_H+a, N-N_H+b)}[\Theta] \\
&= \frac{N_H+a}{N+a+b}
\end{aligned}\right\} \text{Full Bayesian approach}$$

mean

Problem 1: using a different data representation

What do we need to change in the previous calculations if

- we switch to another representation of the data $\mathcal{D} := \mathcal{D}_N := X_N : \Omega = \{H, T\}^N \to \mathbb{N}$ (where $X_N = N_H$ means N_H heads have occurred in N tosses) and
- we are interested in predicting the number of heads $X := X_{N+1}$ after a new toss?

Solution to problem 1

We don't have to change anything fundamental, since (N_H, N) is a sufficient statistics of the problem for both cases. equation 2 would change only by multiplication with the constant term $\binom{N}{N_H}$

$$p(\mathcal{D}|\Theta) = p(X_N|\Theta) = \binom{N}{N_H}\Theta^{N_H}(1-\Theta)^{N-N_H}$$

so that the MLE estimation for Θ is unchanged.

For the MAP estimate there is also no relevant change:

$$\begin{aligned} p(\Theta|D) = p(\Theta|X_N) = &\propto p(X_N|\Theta)p(\Theta|a,b) \\ &\propto \binom{N}{N_H}\Theta^{N_H}(1-\Theta)^{N-N_H} Beta(\Theta|a,b) \\ &\propto Beta(\Theta|N_H+a, N-N_H+b) \end{aligned}$$

so that the MAP estimation for Θ is also unchanged.

However for the posterior predictive distributions using Θ_{MLE} and Θ_{MAP} and the full Bayesian approach, we have to revert to the posterior predictive distribution $p(F|\Theta)$ for F (the next coin toss) and use this immediately to calculate $p(X_{N+1}|\Theta)$: If the known data is $X_N = N_H$, we have for the MLE and MAP cases:

$$p(X_{N+1} = x|\Theta) = \begin{cases} p(F = H|\Theta) & \text{if } x = N_H + 1 \\ p(F = T|\Theta) & \text{if } x = N_H \\ 0 & \text{else} \end{cases} \tag{14}$$

For the full Bayesian case we cannot naively use something like

$$\begin{aligned} p(X_{N+1}|X_N) &= \int_0^1 p(X_{N+1}|\Theta)p(\Theta|X_N)d\Theta \\ &\propto \int_0^1 Bin(X_{N+1}|\Theta, N+1)Beta(\Theta|X_N + a, N - X_N + b)d\Theta \\ &= \dots \end{aligned}$$

because $X_{N+1} \not\perp\!\!\!\perp X_N|\Theta$. (The number of heads after $N+1$ tosses is not conditionally independent from the number of heads after N tosses, (even) if we know the model for the coin (know Θ)).

We must rather calculate

$$\begin{aligned}
p(X_{N+1} = N_H + 1|\mathcal{D}) &= p(X_{N+1} = N_H + 1|X_N = N_H) && (15)\\
&= p(F = H|X_N = N_H) && (16)\\
&= \int_0^1 p(F = H|\Theta)p(\Theta|X_N = N_H)d\Theta && (17)\\
&= \int_0^1 Ber(F = H|\Theta)Beta(\Theta|N_H + a, N - N_H + b)d\Theta && (18)\\
&= \int_0^1 \Theta Beta(\Theta|N_H + a, N - N_H + b)d\Theta && (19)\\
&= E_{Beta(\Theta|N_H+a,N-N_H+b)}[\Theta] && (20)\\
&= \frac{N_H + a}{N + a + b} && (21)
\end{aligned}$$

(second to third line because $F \perp\!\!\!\perp X_N|\Theta$) which is no change as well.

Problem 2: Dirichlet-multinomial model

Generalize the model for a two sided coin from the lecture to a K-sided dice! Compute the MLE and MAP point estimations for the model parameters! What is the posterior predictive distribution for MLE, MAP, and the full Bayesian approach?

Solution to problem 2

Observing N dice rolls, we have $\mathcal{D} = (x_1, x_2, \ldots, x_N)$ with $x_i \in \{1, 2, \ldots; K\}$. Assuming iid, we have for the likelihood

$$p(\mathcal{D}|\Theta) = \prod_{k=1}^{K} \Theta_k^{N_k} \propto Mu(N_1, \ldots, N_K|\Theta_1, \ldots, \Theta_k)$$

with $\sum_k \Theta_k = 1$, and N_k being the number of occurrences of side k.

prior: $$p(\Theta|\alpha) \sim Dir(\Theta|\alpha) \propto \prod_{k=1}^{K} \Theta_k^{\alpha_k - 1}$$

posterior: $$\begin{aligned} p(\Theta|D) &\propto p(D|\Theta)p(\Theta|\alpha) \\ &\propto \prod_{k=1}^{K} \Theta_k^{N_k} \Theta_k^{\alpha_k - 1} \\ &\propto \prod_{k=1}^{K} \Theta_k^{N_k + \alpha_k - 1} \\ &\propto Dir(\Theta|(\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K) \end{aligned}$$

MLE:

$$\left.\begin{aligned} &\frac{\partial}{\partial\lambda}(p(\mathcal{D}|\Theta) + \lambda(1 - \sum_k \Theta_k)) \overset{!}{=} 0 \\ &\frac{\partial}{\partial\Theta_i}(p(\mathcal{D}|\Theta) + \lambda(1 - \sum_k \Theta_k)) \overset{!}{=} 0 \end{aligned}\right\}$$

→ result: $\Theta_k^{MLE} = \frac{N_k}{N}$

MAP:

$$\left.\begin{aligned} &\frac{\partial}{\partial\lambda}(p(\Theta|\mathcal{D}) + \lambda(1 - \sum_k \Theta_k)) \overset{!}{=} 0 \\ &\frac{\partial}{\partial\Theta_i}(p(\Theta|\mathcal{D}) + \lambda(1 - \sum_k \Theta_k)) \overset{!}{=} 0 \end{aligned}\right\}$$

→ result: $\Theta_k^{MAP} = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}$

Posterior Predictive Distribution:

MLE / MAP

$$p(F|\Theta_{MLE/MAP}) = \prod_{k=1}^{K} \Theta_{k MLE/MAP}{}^{\mathbb{1}(F=k)}$$

or

$$p(F = k|\Theta_{MLE/MAP}) = \Theta_{k MLE/MAP}$$

Full Bayesian approach

$$\begin{aligned} p(F = k|\mathcal{D}) &= \int_{S_K} p(F = k|\Theta)p(\Theta|D)d\Theta \\ &= \int_0^1 p(F = k|\Theta)p(\Theta|D)d\Theta_k \\ &= \int_0^1 \Theta_k Dir(\Theta|(\alpha_1 + N_1, \alpha_2 + N_2, \dots, \alpha_K + N_K)d\Theta_k \\ &= \frac{N_k + \alpha_k}{N + \alpha_0} \end{aligned}$$

where the second line follows from “integrating out” the other $\Theta_j, j \neq k$.

Problem 3: Classification with a Naive Bayes classifier

Assume that patterns e.g. are word-counts for a fixed vocabulary of size D. Thus a text-document corresponds to a pattern vector $x \in \mathbb{N}^D$. Naive Bayes models generally assume, that the features $x_j \in \mathbb{N}, j \in 1, \ldots, D$ are conditionally independent given the class, so that we have for the (class-)likelihood

$$p(x|y = c, \Theta) = \prod_{j=1}^{D} p(x_j|y = c, \Theta)$$

Let Θ denote the set of all the parameters: $\Theta = \{\pi, \theta\} = \{\pi_c, \theta_{jc}|c \in \{1, \ldots, C\}, j \in \{1, \ldots, D\}\}$.

So for each class, we have a class-specific set of parameters $\Theta_c = \{\theta_{jc}, \pi_c|j \in \{1, \ldots, D\}\}$.

For text-classification, θ_{jc} is the probability for a word j occurring in that class c. As training data $\mathcal{D}$, we have a number of known vectors $x^{(i)}$ and their known class labels $y^{(i)}$: $\mathcal{D} = ((x^{(1)}, y^{(1)}), (x^{(1)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)}))$.

$p(y^{(i)} = c|\pi)) = \pi_c$ denotes respective element of the vector π of prior probabilities of the classes .

- Derive the general expression for the joint probability $p(x^{(i)}, y^{(i)}|\Theta)$ of a single pattern-vector and its class-label!

- Derive a general expression for the log-likelihood $\log p(\mathcal{D}|\Theta)$!

- Derive a general expressions for the predictive posterior distribution for MLE, MAP, and the full Bayesian approach!

$$
\begin{aligned}
p(x, y|\Theta) &= p(x|y, \Theta)p(y|\Theta) \\
&= p(x|y, \theta, \pi)p(y|\theta, \pi) \\
&= p(x|y, \theta)p(y|\pi) \\
&= \prod_{j=1}^{D} p(x_j|y, \theta)p(y|\pi) \\
&= \prod_{c=1}^{C} \prod_{j=1}^{D} p(x_j|\theta_{jc})^{\mathbb{1}(y=c)} \prod_{c'=1}^{C} \pi_{c'}^{\mathbb{1}(y=c')}
\end{aligned}
$$

$$
\begin{aligned}
p(\mathcal{D}|\Theta) &= \prod_{i=1}^{N} p(x^{(i)}, y^{(i)}|\theta, \pi) \\
&= \prod_{i=1}^{N} p(x^{(i)}|y^{(i)}, \theta)p(y^{(i)}|\pi) \\
&= \prod_{i=1}^{N} \prod_{j=1}^{D} p(x_j^{(i)}|y^{(i)}, \theta)p(y^{(i)}|\pi) \\
&= \prod_{i=1}^{N} \prod_{c=1}^{C} \prod_{j=1}^{D} p(x_j^{(i)}|\theta_{jc})^{\mathbb{1}(y^{(i)}=c)} \prod_{c'=1}^{C} \pi_{c'}^{\mathbb{1}(y^{(i)}=c')}
\end{aligned}
$$

Then the log likelihood is

$$logp(\mathcal{D}|\Theta) = \sum_{c=1}^{C} \sum_{j=1}^{D} \sum_{\{i|y^{(i)}=c\}} log\, p(x_j^{(i)}|\theta_{jc}) + \sum_{c'=1}^{C} N_{c'} log\, \pi_{c'}$$

from which we can derive the MLE estimation Θ_{MLE} for the parameters via $\partial/\partial\Theta_{jc}\, log\, p(\mathcal{D}|\theta_{jc}) \overset{!}{=} 0$ and $\partial/\partial\pi_c\, log\, p(\mathcal{D}|\Theta) \overset{!}{=} 0$

We can also derive a MAP estimate by incorporating suitable conjugate priors $p(\Theta|\alpha,\beta) = p(\theta|\beta)p(\pi|\alpha)$:

$$p(\Theta|\mathcal{D}) \propto p(\mathcal{D}|\Theta)p(\Theta|\alpha,\beta)$$

(as always: posterior(Θ) ∝ likelihood(Θ) * prior(Θ))

and taking the logarithm and computing the argmax.

Posterior Predictive Distribution:

MLE / MAP

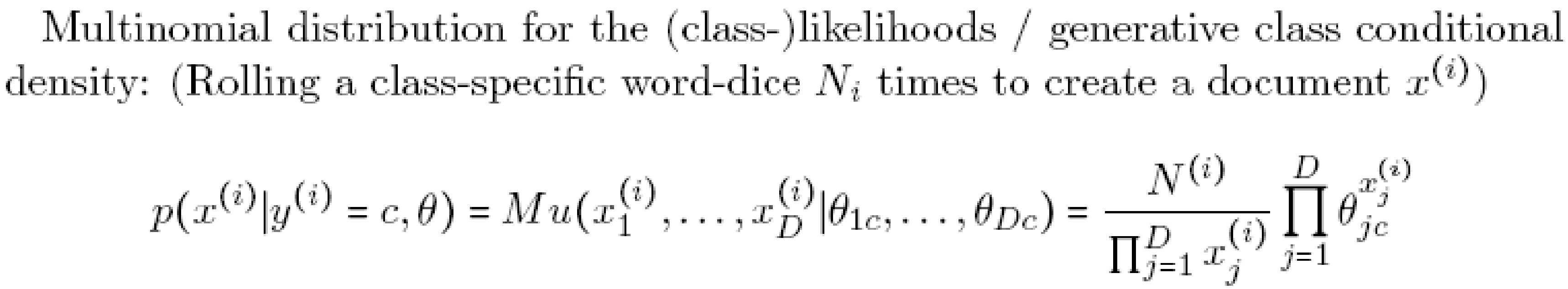
$$p(y = c|x, \Theta_{MAP/MLE}) \propto p(x|y = c, \Theta_{MAP/MLE}) * p(y|\Theta_{MAP/MLE})$$

(class-posterior $\propto$ class-likelihood * class-prior)   (all using the MAP/MLE estimate for Θ)

Full Bayesian approach

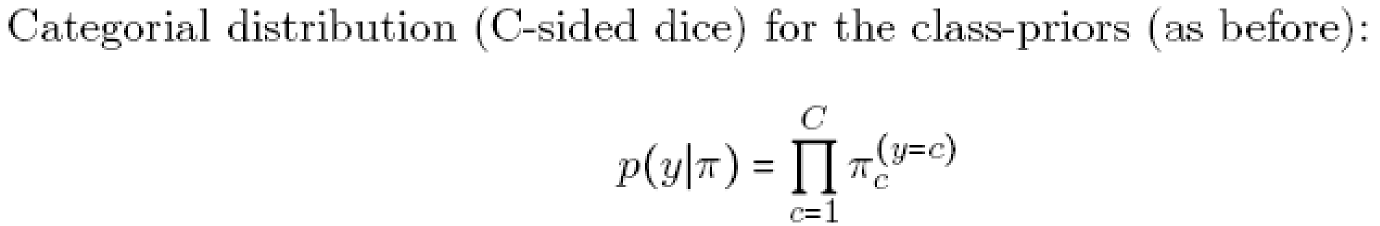
$$\begin{aligned}
p(y = c|x, \mathcal{D}) &\propto \int p(x|y = c, \Theta) * p(y|\Theta)p(\Theta|\mathcal{D})d\Theta \\
&\propto \int p(x|y = c, \theta, \pi) * p(y|\theta, \pi)p(\theta, \pi|\mathcal{D})d\theta d\pi \\
&\propto \int p(x|y = c, \theta, \pi) * p(y|\theta, \pi)p(\theta|\mathcal{D})p(\pi|\mathcal{D})d\theta d\pi \\
&\propto \int p(x|y = c, \theta) * p(\theta|\mathcal{D})d\theta \int p(y|\pi)p(\pi|\mathcal{D})d\pi
\end{aligned}$$

reasonable choices for class-likelihood and class-prior:



Multinomial distribution for the (class-)likelihoods / generative class conditional density: (Rolling a class-specific word-dice N_i times to create a document $x^{(i)}$)

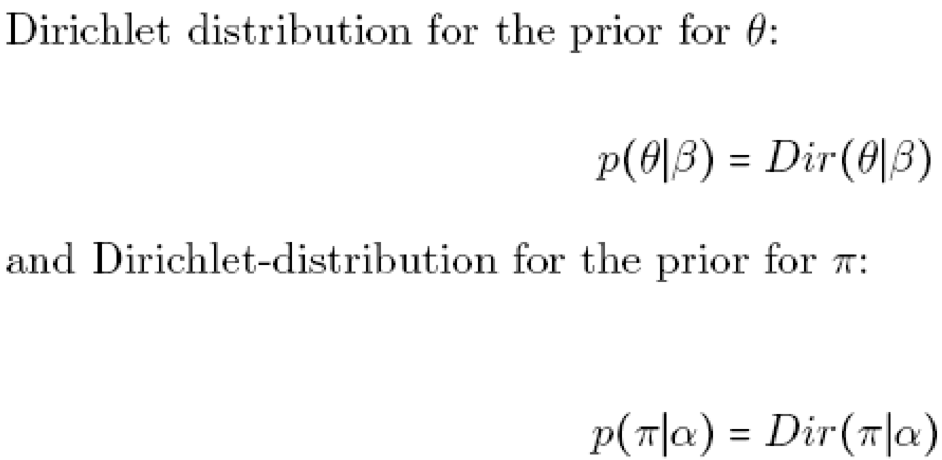
$$p(x^{(i)}|y^{(i)} = c,\theta) = Mu(x_1^{(i)},\dots,x_D^{(i)}|\theta_{1c},\dots,\theta_{Dc}) = \frac{N^{(i)}}{\prod_{j=1}^{D} x_j^{(i)}} \prod_{j=1}^{D} \theta_{jc}^{x_j^{(i)}}$$



Categorial distribution (C-sided dice) for the class-priors (as before):

$$p(y|\pi) = \prod_{c=1}^{C} \pi_c^{(y=c)}$$

reasonable choices for Θ-prior:



Dirichlet distribution for the prior for θ:

$$p(\theta|\beta) = Dir(\theta|\beta)$$

and Dirichlet-distribution for the prior for π:

$$p(\pi|\alpha) = Dir(\pi|\alpha)$$

and Dirichlet-distribution for the prior for π:

$$\begin{aligned} p(\mathcal{D}|\pi) &\propto p(\pi|\mathcal{D})p(\pi|\psi) \\ &= \prod_{c'=1}^{C} \pi_{c'}^{\mathbb{1}(y^{(i)}=c')} p(\pi|\psi) \\ &= \prod_{c'=1}^{C} \pi_{c'}^{\mathbb{1}(y^{(i)}=c')} Dir(\pi|\psi) \end{aligned}$$