Solutions by **Andrew Lys** Collaborated with **Sam Fine** `andrewlys (at) u.e.`

Gaussian Mixtures

1. .

(a) **Parameter Estimation** Our unknown parameters are $\theta = \{p_+, \mu_-, \mu_+, \operatorname{diag}\Sigma_-, \operatorname{diag}\Sigma_+\}$.

First we determine the log likelihood of a given sample S. We denote the indicator function to be

$$[[y_i = 1]] = (1 + y_i)/2$$

and

$$[[y_i = -1]] = (1 - y_i)/2$$

Additionally, we denote the density of a multivariate Gaussian with mean μ and covariance Σ to be

$$f(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal\Sigma^{-1}(x-\mu)\right)$$

We derive the log-likelihood as follows:

$$\begin{aligned}
\ell(\theta|S) &= \log P(S|\theta) = \log\prod_{i=1}^m P(x_i, y_i|\theta) = \log\prod_{i=1}^m P(y_i|\theta)P(x_i|y_i,\theta) \\
&= \sum_{i=1}^m \log(P(y_i|\theta)) + \sum_{i=1}^m \log(P(x_i|y_i,\theta)) \\
&= \sum_{i=1}^m [[y_i = 1]]\log(p_+) + [[y_i = -1]]\log(1-p_+) + \sum_{i=1}^m [[y_i = 1]]\log f(x_i|\mu_+, \Sigma_+) + [[y_i = -1]]\log f(x_i|\mu_-, \Sigma_-) \\
&= \sum_{i=1}^m [[y_i = 1]](\log(p_+) + \log f(x_i|\mu_+, \Sigma_+)) + [[y_i = -1]](\log(1-p_+) + \log f(x_i|\mu_-, \Sigma_-)) \\
&= \sum_{i=1}^m [[y_i = 1]](\log(p_+) - \frac{1}{2}(x_i-\mu_+)^\intercal\Sigma_+^{-1}(x_i-\mu_+) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_+|) \\
&+ [[y_i = -1]](\log(1-p_+) - \frac{1}{2}(x_i-\mu_-)^\intercal\Sigma_-^{-1}(x_i-\mu_-) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_-|)
\end{aligned}$$

From here, we can take the derivatives with respect to each parameter.

(a) p_+ is the probability of a positive sample. We then take the derivative of the log likelihood w.r.t. p_+ and set it to 0, which yields

$$\begin{aligned}
\frac{\partial\ell}{\partial p_+} &= \sum_{i=1}^m [[y_i = +1]]\frac{1}{p_+} - \sum_{i=1}^m [[y_i = -1]]\frac{1}{1-p_+} = 0 \\
\implies \frac{p_+}{1-p_+} &= \frac{\sum_{i=1}^m [[y_i = +1]]}{\sum_{i=1}^m [[y_i = -1]]} \\
\implies p_+ &= \frac{\sum_{i=1}^m [[y_i = +1]]}{\sum_{i=1}^m [[y_i = +1]] + [[y_i = -1]]} \\
\hat{p}_+ &= \frac{\sum_{i=1}^m [[y_i = +1]]}{m}
\end{aligned}$$

(b) To find μ_+, we take the gradient with respect to μ_+ and set it to 0.

$$\nabla_{\mu_+}\ell = \sum_{i=1}^m [[y_i = 1]](-1)(\Sigma_+^{-1} + \Sigma_+^{-1\intercal})(x_i - \mu_+) = 0$$

Since Σ_+ is a diagonal matrix, the inverse is symmetric.

$$
\begin{aligned}
0 &= \sum_{i=1}^{m} [[y_i = 1]] \Sigma_+^{-1} (x_i - \mu_+) \\
\implies \sum_{i=1}^{m} [[y_i = 1]] x_i &= \mu_+ \sum_{i=1}^{m} [[y_i = 1]] \\
\hat{\mu}_+ &= \frac{\sum_{i=1}^{m} [[y_i = 1]] x_i}{\sum_{i=1}^{m} [[y_i = 1]]}
\end{aligned}
$$

(c) The process to find μ_- is the same as above, so we have

$$
\hat{\mu}_- = \frac{\sum_{i=1}^{m} [[y_i = -1]] x_i}{\sum_{i=1}^{m} [[y_i = -1]]}
$$

(d) In the cases of Σ_+ and Σ_- we thankfully rely on the fact that Σ is diagonal,

$$
\begin{aligned}
\frac{\partial}{\partial \Sigma_+} \ell(\theta | S) &= -\frac{1}{2} \sum_{i=1}^{m} [[y_i = 1]] \frac{\partial}{\partial \Sigma_+} \left((x_i - \mu_+)^\intercal \Sigma_+^{-1} (x_i - \mu_+) + \log |\Sigma_+| \right) \\
&= -\frac{1}{2} \sum_{i=1}^{m} [[y_i = 1]] \left(-\Sigma_+^{-\intercal} (x_i - \mu_+)(x_i - \mu_+)^\intercal \Sigma_+^{-\intercal} + \Sigma_+^{-1} \right) \\
\implies \sum_{i=1}^{m} [[y_i = 1]] \Sigma_+^{-1} &= \sum_{i=1}^{m} [[y_i = 1]] \Sigma_+^{-1} (x_i - \mu_+)(x_i - \mu_+)^\intercal \Sigma_+^{-1}
\end{aligned}
$$

These derivatives are elementary[1] matrix calculus operations[2]. From here, we simplify further.

$$
\begin{aligned}
\sum_{i=1}^{m} [[y_i = 1]] I_d &= \sum_{i=1}^{m} [[y_i = 1]] \Sigma_+^{-1} (x_i - \mu_+)(x_i - \mu_+)^\intercal \\
\Sigma_+ \sum_{i=1}^{m} [[y_i = 1]] &= \sum_{i=1}^{m} [[y_i = 1]] (x_i - \mu_+)(x_i - \mu_+)^\intercal \\
\hat{\Sigma}_+ &= \frac{\sum_{i=1}^{m} [[y_i = 1]] (x_i - \mu_+)(x_i - \mu_+)^\intercal}{\sum_{i=1}^{m} [[y_i = 1]]}
\end{aligned}
$$

(e) The process to find Σ_- is the same as above, so we have

$$
\hat{\Sigma}_- = \frac{\sum_{i=1}^{m} [[y_i = -1]] (x_i - \mu_-)(x_i - \mu_-)^\intercal}{\sum_{i=1}^{m} [[y_i = -1]]}
$$

To summarize, our MLE estimators are:

$$
\begin{aligned}
\hat{p}_+ &= \frac{\sum_{i=1}^{m} [[y_i = +1]]}{m} \\
\hat{\mu}_+ &= \frac{\sum_{i=1}^{m} [[y_i = +1]] x_i}{\sum_{i=1}^{m} [[y_i = +1]]} \\
\hat{\mu}_- &= \frac{\sum_{i=1}^{m} [[y_i = -1]] x_i}{\sum_{i=1}^{m} [[y_i = -1]]} \\
\hat{\Sigma}_+ &= \frac{\sum_{i=1}^{m} [[y_i = 1]] (x_i - \mu_+)(x_i - \mu_+)^\intercal}{\sum_{i=1}^{m} [[y_i = 1]]} \\
\hat{\Sigma}_- &= \frac{\sum_{i=1}^{m} [[y_i = -1]] (x_i - \mu_-)(x_i - \mu_-)^\intercal}{\sum_{i=1}^{m} [[y_i = -1]]}
\end{aligned}
$$

[1] Wikipedia matrix calculus

[2] MSE post differentiating quadratic form

(b) **Prediction**

$$P(Y=1|x) = \frac{P(X=x|Y=1)P(Y=1)}{P(X=x)} = \frac{P(X=x|Y=1)p_+}{P(X=x|Y=1)P(Y=1)+P(X=x|Y=0)P(Y=0)}$$
$$= \frac{1}{1+\frac{P(X=x|Y=0)P(Y=0)}{P(X=x|Y=1)P(Y=1)}} = \frac{1}{1+\frac{1-p_+}{p_+}\frac{f(x|\mu_-,\Sigma_-)}{f(x|\mu_+,\Sigma_+)}}$$

We obtain the following discriminant:

$$r(x) = \log\left(\frac{p_+}{1-p_+}\right) + \log\left(\frac{f(x|\mu_+,\Sigma_+)}{f(x|\mu_-,\Sigma_-)}\right)$$
$$= \log\left(\frac{p_+}{1-p_+}\right) + \log\left(\frac{\sqrt{|\Sigma_-|}}{\sqrt{|\Sigma_+|}}\right) - \frac{1}{2}(x-\mu_+)^\intercal\Sigma_+^{-1}(x-\mu_+) + \frac{1}{2}(x-\mu_-)^\intercal\Sigma_-^{-1}(x-\mu_-)$$
$$= \log\left(\frac{p_+}{1-p_+}\right) + \frac{1}{2}\log\left(\frac{|\Sigma_-|}{|\Sigma_+|}\right) + \frac{1}{2}(\mu_+^\intercal\Sigma_+^{-1}\mu_+ - \mu_-^\intercal\Sigma_-^{-1}\mu_-) + \frac{1}{2}x^\intercal(\Sigma_-^{-1}-\Sigma_+^{-1})x + x^\intercal(\Sigma_+^{-1}\mu_+ - \Sigma_-^{-1}\mu_-)$$

The Bayes predictor is simply

$$h(x) = \operatorname{sign}(r(x))$$

Since, when $r(x) > 0$, we have $P(Y=1|x) > \frac{1}{2}$, and when $r(x) < 0$, we have $P(Y=1|x) < \frac{1}{2}$.

(c) **As a Linear Predictor** Letting

$$b = \log\left(\frac{p_+}{1-p_+}\right) + \frac{1}{2}\log\left(\frac{|\Sigma_-|}{|\Sigma_+|}\right) + \frac{1}{2}(\mu_+^\intercal\Sigma_+^{-1}\mu_+ - \mu_-^\intercal\Sigma_-^{-1}\mu_-)$$
$$\operatorname{diag}(a_1,\dots,a_d) = \frac{1}{2}(\Sigma_-^{-1}-\Sigma_+^{-1})$$
$$v = \Sigma_+^{-1}\mu_+ - \Sigma_-^{-1}\mu_-$$

We can write our discriminant as

$$r(x) = b + x^\intercal Ax + x^\intercal v$$

Let $v = (v_1,\dots,v_d)^\intercal$. Then we can write

$$r(x) = b + \sum_{i=1}^d a_i x_i^2 + \sum_{i=1}^d v_i x_i$$

Thus, it is clear that with the feature map:

$$\phi : x \mapsto (1, x_1, \dots, x_d, x_1^2, \dots, x_d^2)^\intercal$$

r is a linear predictor. Namely:

$$r(x) = \langle w, \phi(x)\rangle$$
$$w = (b, v_1, \dots, v_d, a_1, \dots, a_d)^\intercal$$

This shows that $D = 2d+1$ is good enough.

(d) Given

$$w = (b, v_1, \dots, v_d, a_1, \dots, a_d)^\intercal$$

Note that we have $4d+1$ parameters in our model. First, let us write b, A and v in terms of μ_+, μ_-, Σ_+ and Σ_-. Let

$$\mu_y = (\mu_y[1], \dots, \mu_y[d])^\intercal$$
$$\Sigma_y = \operatorname{diag}(s_y[1], \dots, s_y[d])^\intercal$$

Then we have:

$$v = \operatorname{diag}(s_+[1]^{-1}, \dots, s_+[d]^{-1})\mu_+ - \operatorname{diag}(s_-[1]^{-1}, \dots, s_-[d]^{-1})\mu_-$$

$$
\begin{aligned}
&= \sum_{i=1}^{d} \frac{\mu_+[i]}{s_+[i]} e_i - \sum_{i=1}^{d} \frac{\mu_-[i]}{s_-[i]} e_i \\
\implies v_i &= \frac{\mu_+[i]}{s_+[i]} - \frac{\mu_-[i]}{s_-[i]} \\
\operatorname{diag}(a_1, \dots, a_d) &= \frac{1}{2}(\operatorname{diag}(s_-[1]^{-1}, \dots, s_-[d]^{-1}) - \operatorname{diag}(s_+[1]^{-1}, \dots, s_+[d]^{-1})) \\
&= \operatorname{diag}\left(\frac{1}{2}\left(s_-[1]^{-1} - s_+[1]^{-1}\right), \dots, \frac{1}{2}\left(s_-[d]^{-1} - s_+[d]^{-1}\right)\right) \\
\implies a_i &= \frac{1}{2}\left(s_-[i]^{-1} - s_+[i]^{-1}\right) \\
b &= \log\left(\frac{p_+}{1-p_+}\right) + \frac{1}{2}\log\left(\frac{|\Sigma_-|}{|\Sigma_+|}\right) + \frac{1}{2}(\mu_+^\intercal \Sigma_+^{-1} \mu_+ - \mu_-^\intercal \Sigma_-^{-1} \mu_-) \\
\frac{|\Sigma_-|}{|\Sigma_+} &= \prod_{i=1}^{d} \frac{s_-[i]}{s_+[i]} \implies \frac{1}{2}\log\frac{|\Sigma_-|}{|\Sigma_+|} = \frac{1}{2}\sum_{i=1}^{d} s_-[i] - s_+[i] \\
\mu_+^\intercal \Sigma_+^{-1} \mu_+ &= \sum_{i=1}^{d} \frac{\mu_+[i]^2}{s_+[i]} \qquad \mu_-^\intercal \Sigma_-^{-1} \mu_- = \sum_{i=1}^{d} \frac{\mu_-[i]^2}{s_-[i]} \\
b &= \log\left(\frac{p_+}{1-p_+}\right) + \frac{1}{2}\sum_{i=1}^{d} s_-[i] - s_+[i] + \frac{\mu_+[i]^2}{s_+[i]} - \frac{\mu_-[i]^2}{s_-[i]}
\end{aligned}
$$

Let us make the simplifying assumption that $s_-[i] = 1$ when $a_i < 0$ and $s_+[i] = 1$ when $a_i > 0$. Suppose $a_i < 0$. Then we have:

$$
\begin{aligned}
a_i &= \frac{1}{2} - \frac{1}{2} s_+[i]^{-1} \\
\implies s_+[i] &= \frac{1}{1-2a_i} > 0
\end{aligned}
$$

Let us make the simplifying assumption that $\mu_+[i] = 0$ when $a_i < 0$ and $\mu_-[i] = 0$ when $a_i > 0$. Suppose $a_i < 0$. Then we have:

$$
\begin{aligned}
v_i &= -\frac{\mu_-[i]}{s_+[i]} = -\frac{\mu_-[i]}{1-2a_i} \\
\implies \mu_-[i] &= -v_i(1-2a_i)
\end{aligned}
$$

When $a_i > 0$, then $s_+[i] = 1$ and $\mu_+[i] = 0$. Thus, we have:

$$
\begin{aligned}
a_i &= \frac{1}{2}(s_-[i]^{-1} - 1) \\
\implies s_-[i] &= \frac{1}{1+2a_i} \\
v_i &= \frac{\mu_+[i]}{s_+[i]} = (1+2a_i)\mu_+[i] \\
\mu_+[i] &= \frac{v_i}{1+2a_i}
\end{aligned}
$$

To summarize:

$$
\begin{aligned}
\mu_+[i] &= [[a_i > 0]] \frac{v_i}{1+2a_i} \\
\mu_-[i] &= [[a_i < 0]](-v_i(1-2a_i)) \\
s_+[i] &= (1-2a_i)^{-[[a_i<0]]}
\end{aligned}
$$

$$s_-[i] = (1+2a_i)^{-[[a_i>0]]}$$

Now we can solve for p_+.

$$\log\frac{p_+}{1-p_+} + \frac{1}{2}\sum_{i=1}^{d} \frac{1}{(1+2a_i)}^{[[a_i>0]]} - \frac{1}{(1-2a_i)}^{-[[a_i<0]]} + [[a_i>0]]\frac{v_i^2(1-2a_i)^{[[a_i<0]]}}{(1+2a_i)^2} - [[a_i<0]]v_i^2(1-2a_i)^2(1+2a_i)^{[[a_i>0]]}$$

$$b = \log\frac{p_+}{1-p_+} + \frac{1}{2}\sum_{i=1}^{d}(1+2a_i)^{-[[a_i>0]]} - (1-2a_i)^{-[[a_i<0]]} + [[a_i>0]]\frac{v_i^2}{(1+2a_i)^2} - [[a_i<0]]v_i^2(1-2a_i)^2$$

$$\frac{p_+}{1-p_+} = \exp\left(b - \frac{1}{2}\left(\sum_{i=1}^{d}(1+2a_i)^{-[[a_i>0]]} - (1-2a_i)^{-[[a_i<0]]} + [[a_i>0]]\frac{v_i^2}{(1+2a_i)^2} - [[a_i<0]]v_i^2(1-2a_i)^2\right)\right)$$

$$p_+ = \frac{1}{1+\exp\left(-b + \frac{1}{2}\left(\sum_{i=1}^{d}(1+2a_i)^{-[[a_i>0]]} - (1-2a_i)^{-[[a_i<0]]} + [[a_i>0]]\frac{v_i^2}{(1+2a_i)^2} - [[a_i<0]]v_i^2(1-2a_i)^2\right)\right)}$$

(e) The decision boundary is a hyperplane in the feature space given by

$$x \mapsto (x_1,\ldots,x_d,x_1^2,\ldots,x_d^2)$$

We can write the discriminant as:

$$\begin{aligned} r(x) &= b + \sum_{i=1}^{d} a_i x_i^2 + \sum_{i=1}^{d} v_i x_i \\ &= b - \sum \frac{v_i^2}{4a_i} + \sum_{i=1}^{d} a_i\left(x_i + \frac{v_i}{2a_i}\right)^2 \end{aligned}$$

So the decision boundary is determined by an ellipsoid, i.e.

$$r(x) = 0 \implies \sum_{i=1}^{d} a_i\left(x_i + \frac{v_i}{2a_i}\right)^2 = \sum_{i=1}^{d}\frac{v_i^2}{4a_i} - b$$

Modeling Text Documents

2. **A Simple Model.**

(a) Let Y be the random variable with the topics as outputs, and the distribution given by p_{topic}. We shall denote p_{topic} as p, a two dimensional vector, where $p(y) = P(Y=y)$

Let X be the random variable with the documents as outputs. X is a N-dimensional vector, where $X[i]$ is the ith word in the document. By the conditional independence assumption, we have

$$P(X[i]=x, X[j]=x'|Y=y) = P(X[i]=x|Y=y)P(X[j]=x'|Y=y)$$

Further, we assume that $X[i]$ is drawn identically from p_y, so for every i, we let

$$p(x|y) := P(X[i]=x|Y=y)$$

Let $p_y := p(\cdot|y) \in \mathbb{R}^D$. For simplicity, we assume that $Y \in \{0,1\}$ and $X[i] \in [D]$, for every i. We denote the indicator function to be $[[*]]$

Given a sample

$$S = \{(x_1,y_1),\ldots,(x_n,y_n)\}$$

We define the following sample statistics. For $x \in [D]$, $y \in \{0,1\}$

$$n_j(x,y) = |\{i : (x_i,y_i) \in S, x_i[j] = x, y_i = y\}|$$

$$n(x,y) = \sum_{j=1}^{N} n_j(x,y)$$

$$n(y) = |\{i : (x_i, y_i) \in S, y_i = y\}|$$

We want to find estimators for p and for

$$P(X[1] = x_1, \ldots, X[N] = x_N | Y = y)$$

Since $X[i]|y$ is i.i.d., we can simplify this expression:

$$P(X[1] = x_1, \ldots, X[N] = x_N | Y = y) = \prod_{i=1}^{N} P(X[i] = x_i | Y = y) = \prod_{i=1}^{N} p(x_i|y)$$

Thus, we can focus on estimators of p and $p(x|y)$. We should expect our MLEs for p and $p(x|y)$ to be the sample means, i.e.

$$\hat{p} = \frac{n(1)}{n}$$
$$\hat{p}(x|y) = \frac{n(x,y)}{n(y)}$$

Let $\theta = (p, \{p_y\})$. We define our likelihood function as:

$$L(\theta, S) = P((X = x_1, Y = y_1), \ldots, (X = x_n, Y = y_n)|\theta)$$

Since S was drawn i.i.d., we can simplify:

$$L(\theta, S) = \prod_{i=1}^{n} P(X = x_i, Y = y_i|\theta)$$

By the monotonicity of log, we can optimize over the log-likelihood, ℓ. Given that $X[i]|Y$ is i.i.d., we can simplify.

$$\begin{aligned}
\ell(\theta|S) &= \sum_{i=1}^{n} \log(P(X[1] = x_i[1], \ldots, X[N] = x_i[N]|Y = y_i)P(Y = y_i)) \\
&= \sum_{i=1}^{n} \log(P(Y = y_i) \prod_{j=1}^{N} P(X[j] = x_i[j]|Y = y_i)) \\
&= \sum_{i=1}^{n} \log P(Y = y_i) + \sum_{j=1}^{N} \log P(X[j] = x_i[j]|Y = y_i) \\
&= \sum_{i=1}^{n} \log p(y_i) + \sum_{i=1}^{n} \sum_{j=1}^{N} \log p(x_i[j]|y_i) \\
&= \sum_{i=1}^{n} \sum_{y \in \{0,1\}} [[y_i = y]] \log p(y) + \sum_{i=1}^{n} \sum_{j=1}^{N} \sum_{x \in [D]} [[x_i[j] = x]] \log p(x|y_i) \\
&= \sum_{i=1}^{n} \sum_{y \in \{0,1\}} [[y_i = y]] \log p(y) + \sum_{i=1}^{n} \sum_{j=1}^{N} \sum_{x \in [D]} \sum_{y \in \{0,1\}} [[x_i[j] = x \wedge y_i = y]] \log p(x|y) \\
&= \sum_{y \in \{0,1\}} n(y) \log p(y) + \sum_{x \in [D]} \sum_{y \in \{0,1\}} \sum_{j=1}^{N} n_j(x,y) \log p(x|y) \\
&= \sum_{y \in \{0,1\}} n(y) \log p(y) + \sum_{x \in [D]} \sum_{y \in \{0,1\}} n(x,y) \log p(x|y)
\end{aligned}$$

To solve for the minmimum of $\ell(\theta|S)$, we use the method of Lagrange multipliers. We have the following constraints:

$$\sum_{y \in \{0,1\}} p(y) = 1$$

$$\sum_{x \in [D]} p(x|y) = 1 \qquad \forall y \in \{0,1\}$$

Then we have the following Lagrangian:

$$\mathcal{L} = \sum_{y \in \{0,1\}} n(y) \log p(y) + \sum_{x \in [D]} \sum_{y \in \{0,1\}} n(x,y) \log p(x|y) - \mu \left(\sum_{y \in \{0,1\}} p(y) - 1 \right) - \sum_{y \in \{0,1\}} \lambda_y \left(\sum_{x \in [D]} p(x|y) - 1 \right)$$

Taking critical points:

$$[p(y)] : \frac{n(y)}{p(y)} = \mu$$
$$[p(x|y)] : \frac{n(x,y)}{p(x|y)} = \lambda_y$$
$$[\mu] : \sum_{y \in \{0,1\}} p(y) = 1$$
$$[\lambda_y] : \sum_{x \in [D]} p(x|y) = 1$$

We solve for $p(y)$:

$$\begin{aligned}
\frac{n(1)}{p(1)} &= \frac{n(0)}{p(0)} \\
p(0) &= p(1) \frac{n(0)}{n(1)} \\
1 &= p(0) + p(0) \frac{n(0)}{n(1)} \\
n(1) &= p(1)n(1) + p(1)n(0) \\
p(1) &= \frac{n(1)}{n} \\
p(0) &= \frac{n(0)}{n}
\end{aligned}$$

We solve for $p(x|y)$:

$$\begin{aligned}
\frac{n(x,y)}{p(x|y)} &= \frac{n(x',y)}{p(x'|y)} \\
p(x'|y) &= p(x|y) \frac{n(x',y)}{n(x,y)} \\
1 &= p(x|y) + \sum_{x' \neq x} p(x|y) \frac{n(x',y)}{n(x,y)} \\
n(x,y) &= p(x|y) \sum_{x' \in [D]} n(x',y) = p(x|y) n(y) \\
p(x|y) &= \frac{n(x,y)}{n(y)}
\end{aligned}$$

(b) Using Baye's Law, and conditional independence we have:

$$\begin{aligned}
P(Y=1|X=x) &= \frac{P(X=x|Y=1)P(Y=1)}{P(X=x)} \\
&= \frac{P(X[1]=x[1], \ldots, X[N]=x[N]|Y=1)P(Y=1)}{P(X[1]=x[1], \ldots, X[N]=x[n])}
\end{aligned}$$

$$= \frac{P(Y=1)\prod_{i=1}^N P(X[i]=x[i]|Y=1)}{P(X[1]=x[1],\ldots,X[N]=x[n]|Y=1)P(Y=1)+P(X[1]=x[1],\ldots,X[n]=x[n]|Y=0)P(Y=0)}$$
$$= \frac{p(1)\prod_{i=1}^N p(x[i]|1)}{p(1)\prod_{i=1}^N p(x[i]|1)+p(0)\prod_{i=1}^N p(x[i]|0)}$$

Now we can reduce this into the form of a logistic function.

$$P(Y=1|X=x) = \frac{1}{1+\frac{p(0)}{p(1)}\prod_{i=1}^N \frac{p(x[i]|0)}{p(x[i]|1)}}$$
$$= \frac{1}{1+\exp-\left(\log\frac{p(1)}{p(0)}+\sum_{i=1}^N \log\frac{p(x[i]|1)}{p(x[i]|0)}\right)}$$

Therefore, we can get our discriminant as follows:

$$r(x) = \log\frac{p(1)}{p(0)} + \sum_{i=1}^N \log\frac{p(x[i]|1)}{p(x[i]|0)}$$

(c) We can simplify the discriminant by noting

$$p(x|y) = \prod_{x'\in[D]} p(x'|y)^{[[x'=x]]}$$

Giving us

$$r(x) = \log\frac{p(1)}{p(0)} + \sum_{i=1}^N \log(p(x[i]|1)) - \sum_{i=1}^N \log(p(x[i]|0))$$
$$= \log\frac{p(1)}{p(0)} + \sum_{i=1}^N \log\prod_{x'\in[D]} p(x'|1)^{[[x[i]=x']]} - \sum_{i=1}^N \log\prod_{x'\in[D]} p(x'|0)^{[[x[i]=x']]}$$
$$= \log\frac{p(1)}{p(0)} + \sum_{i=1}^N \sum_{x'\in[D]} [[x[i]=x']]\log\frac{p(x'|1)}{p(x'|0)}$$
$$= \log\frac{p(1)}{p(0)} + \sum_{x'\in[D]} \log\frac{p(x'|1)}{p(x'|0)}\sum_{i=1}^N [[x[i]=x']]$$

This leads us to consider a bag of words, with a bias term, feature map for x. Specifically, we define $\phi : \mathcal{X} \to \mathbb{R}^{D+1}$ as follows:

$$\phi : x \mapsto \left(1, \sum_{i=1}^N [[x[i]=1]], \ldots, \sum_{i=1}^N [[x[i]=D]]\right)$$

Thus, we define our vector w as follows:

$$w = \left(\log\frac{p(1)}{p(0)}, \log\frac{p(1|1)}{p(1|0)}, \ldots, \log\frac{p(D|1)}{p(D|0)}\right)$$

We can see that $r(x) = \langle w, \phi(x)\rangle$.

(d) The log odds term in the bias has a simple interpretation.

$$\frac{\hat{p}(1)}{\hat{p}(0)} = \frac{n(1)/n}{n(0)/n} = \frac{n(1)}{n(0)} \implies \log\frac{\hat{p}(1)}{\hat{p}(0)} = \log\frac{n(1)}{n(0)}$$

We simplify the other terms.

$$\log \frac{\hat{p}(x|1)}{\hat{p}(x|0)} = \log \frac{n(x,1)/n(1)}{n(x,0)/n(0)}$$
$$= \log \frac{n(x,1)}{n(x,0)} - \log \frac{n(1)}{n(0)}$$

So we have the following simplification for w:

$$w = \left(\log \frac{n(1)}{n(0)}, \log \frac{n(1,1)}{n(1,0)} - \log \frac{n(1)}{n(0)}, \dots, \log \frac{n(D,1)}{n(D,0)} - \log \frac{n(1)}{n(0)}\right)$$

3. **Adding a Prior.**

(a) The MAP estimate is defined as follows:

$$\hat{\theta} = \arg\max_{\theta} p(\theta|S)$$

In our case,

$$\theta = (p, \{p_y\})$$

Where we define:

$$p(y) = P(Y = y)$$
$$p_i(x|y) = P(X[i] = x|Y = y)$$

However, note that by the conditional independence assumption, we have that $p_i(x|y) = p_j(x|y)$ for all $i, j \in [N]$. Therefore, we can just define $p_y = p(\cdot|y) = p_i(\cdot|y)$ for all $i \in [N]$, and treat p_y as a D dimensional vector. Let S be a sample of n i.i.d. points.

$$S = ((x_1, y_1), \dots, (x_n, y_n))$$

Our posterior distribution, $p(\theta|S)$ is given by:

$$P(\theta|S) = \frac{P(S|\theta)P(\theta)}{P(S)}$$
$$= \frac{P(X|Y,\theta)P(Y|\theta)P(\theta)}{P(S)}$$
$$= \frac{P(X|Y,\{p_y\})P(Y|p)P(\theta)}{P(X|Y)P(Y)}$$

where X is the vector of x_i's and Y is the vector of y_i's.

Note that, we are not conditioning the denominator with respect to the parameters we are optimizing over. The denominator is the intergral over the distributions of p and $\{p_y\}$, meaning the values of p and p_y that we end up choosing do not impact its value. Therefore, we can ignore it in the optimization problem.

$$\hat{\theta} = \arg\max_{p,\{p_y\}} P(X|Y,\{p_y\})P(Y|p)P(p,\{p_y\})$$

We break this expression down, term by term, first focusing on the last term.

$$P(p,\{p_y\}) = P(p)P(\{p_y\}) = f_{Dir(1)}(p)P(p_1)P(p_0)$$
$$= f_{Dir(\alpha)}(p_1)f_{Dir(\alpha)}(p_0)$$
$$= \frac{1}{Z(\alpha)^2} \prod_{x\in[D]} p(x|1)^{\alpha-1}p(x|0)^{\alpha-1}$$

Since $Z(\alpha)^2$ is fixed, we can ignore it in the expression for $\hat{\theta}$. Now we focus on the second term.

$$P(Y|p) = P(Y_1 = y_1, \dots, Y_n = y_n|p) = \prod_{i=1}^{n} P(Y_i = y_i|p)$$

$$= \prod_{i=1}^{n} \prod_{y \in \{0,1\}} p(y)^{[[y_i = y]]}$$

Now we focus on the first term.

$$\begin{aligned}
P(X|Y, \{p_y\}) &= P(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_n = y_n, \{p_y\}) \\
&= \prod_{i=1}^{n} P(X_i = x_i | Y_1 = y_1, \ldots, Y_n = y_n, \{p_y\}) \\
&= \prod_{i=1}^{n} P(X_i = x_i | Y_i = y_i, \{p_y\}) \\
&= \prod_{i=1}^{n} P(X_i[1] = x_i[1], \ldots, X_i[N] = x_i[N] | Y_i = y_i, \{p_y\}) \\
&= \prod_{i=1}^{n} \prod_{j=1}^{N} P(X_i[j] = x_i[j] | Y_i = y_i, \{p_y\}) \\
&= \prod_{i=1}^{n} \prod_{j=1}^{N} p(x_i[j] | y_i)
\end{aligned}$$

Since log is monotone, we can take the log of our expression to get the arg max.

$$\hat{\theta} = \arg\max_{p, \{p_y\}} \sum_{i=1}^{n} \sum_{j=1}^{N} \log p(x_i[j]|y_i) + \sum_{i=1}^{n} y_i \log(p) + (1 - y_i) \log(1 - p) + \sum_{x \in [D]} \sum_{y \in \{0,1\}} (\alpha - 1) \log p(x|y)$$

First, we get $\hat{p}$ by differentiating with respect to p and setting it to zero.

$$\begin{aligned}
\frac{d}{dp}\hat{\theta} &= \sum_{i=1}^{n} \frac{y_i}{p} - \frac{1 - y_i}{1 - p} = 0 \\
\sum_{i=1}^{n} \frac{y_i}{p} &= \sum_{i=1}^{n} \frac{1 - y_i}{1 - p} \\
\frac{1 - p}{p} &= \frac{\sum_{i=1}^{n} 1 - y_i}{\sum_{i=1}^{n} y_i} \\
p &= \frac{\sum_{i=1}^{n} y_i}{n} = \frac{n(1)}{n}
\end{aligned}$$

Where $n(y)$ is the number of y_i's that are equal to y. Before we try and solve for $p(x|y)$, we can do a better job at simplifying the first term.

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{N} \log p(x_i[j]|y_i) &= \sum_{i=1}^{n} \sum_{j=1}^{N} \sum_{x \in [D]} [[x_i[j] = x]] \log(p(x|y_i)) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{N} \sum_{x \in [D]} \sum_{y \in \{0,1\}} [[x_i[j] = x \wedge y_i = y]] \log(p(x|y)) \\
&= \sum_{y \in \{0,1\}} \sum_{x \in [D]} \sum_{j=1}^{N} \log(p(x|y)) \sum_{i=1}^{n} [[x_i[j] = x \wedge y_i = y]] \\
&= \sum_{y \in \{0,1\}} \sum_{x \in [D]} \sum_{j=1}^{N} \log(p(x|y)) n_j(x, y)
\end{aligned}$$

$$= \sum_{y\in\{0,1\}} \sum_{x\in[D]} n(x,y)\log(p(x|y))$$

Where $n_j(x,y)$ is the number of (x_i,y_i)'s such that $x_i[j]=x$ and $y_i=y$, and $n(x,y)$ is the number of (x_i,y_i)'s such that $x_i[j]=x$ and $y_i=y$, for some $j\in[N]$. We then have the following Lagrangian for $p(x|y)$.

$$\begin{aligned}
\mathcal{L} &= \sum_{y\in\{0,1\}} \sum_{x\in[D]} \log(p(x|y))n(x,y) + \sum_{y\in\{0,1\}} \sum_{x\in[D]} (\alpha-1)\log(p(x|y)) - \sum_{y\in\{0,1\}} \lambda_y \left(\sum_{x\in[D]} p(x|y) - 1\right) \\
&= \sum_{y\in\{0,1\}} \sum_{x\in[D]} (n(x,y)+(\alpha-1))\log p(x|y) - \sum_{y\in\{0,1\}} \lambda_y \left(\sum_{x\in[D]} p(x|y) - 1\right)
\end{aligned}$$

We take critical points.

$$\begin{aligned}
[p(x|y)] &: \frac{n(x,y)+(\alpha-1)}{p(x|y)} = \lambda_y \\
[\lambda_y] &: \sum_{x\in[D]} p(x|y) = 1
\end{aligned}$$

We solve this system of equations.

$$\begin{aligned}
\frac{n(x,y)+(\alpha-1)}{p(x|y)} &= \frac{n(x',y)+(\alpha-1)}{p(x'|y)} \\
p(x'|y) &= p(x|y)\frac{n(x',y)+(\alpha-1)}{n(x,y)+(\alpha-1)} \\
1 &= p(x|y) + \sum_{x'\neq x} p(x|y)\frac{n(x',y)+(\alpha-1)}{n(x,y)+(\alpha-1)} \\
n(x,y)+(\alpha-1) &= p(x|y)\sum_{x'\in[D]} n(x',y)+(\alpha-1) \\
\implies p(x|y) &= \frac{n(x,y)+(\alpha-1)}{\sum_{x'\in[D]} n(x',y)+(\alpha-1)}
\end{aligned}$$

(b) Recall that the discriminant is given by

$$r(x) = \log\frac{p(1)}{p(0)} + \sum_{x'\in[D]} \log\frac{p(x'|1)}{p(x'|0)} \sum_{i=1}^{N} [[x[i]=x']]$$

We can take the same feature map as from before.

$$\phi : x \mapsto \left(1, \sum_{i=1}^{N}[[x[i]=1]], \ldots, \sum_{i=1}^{N}[[x[i]=D]]\right)$$

We have the following $w\in\mathbb{R}^{D+1}$ such that $r(x)=\langle w,\phi(x)\rangle$:

$$\begin{aligned}
w[1] &= \log\frac{p(1)}{p(0)} \\
w[1+x] &= \log\frac{p(x|1)}{p(x|0)}
\end{aligned}$$

Now we can easily plug in our MAP estimators.

$$
\begin{aligned}
w[1] &= \log \frac{n(1)}{n(0)} \\
w[1+x] &= \log \frac{n(x,1)+(\alpha-1)}{\sum_{x'\in[D]} n(x',1)+(\alpha-1)} \frac{\sum_{x'\in[D]} n(x',0)+(\alpha-1)}{n(x,0)+(\alpha-1)} \\
&= \log \frac{n(x,1)+(\alpha-1)}{n(x,0)+(\alpha-1)} - \log \frac{\sum_{x'\in[D]} n(x',1)+(\alpha-1)}{\sum_{x'\in[D]} n(x',0)+(\alpha-1)}
\end{aligned}
$$

4. **Multiple Classes.**

(a) Let $p(k) = P(Y=k)$ and $p_i(x|y) = P(X[i]=x|Y=y)$. As before, we assume that $X[i]$ is drawn identically from p_y, so for every i, we let $p_y = p(\cdot|y) = p_i(\cdot|y)$. We derive the posterior distribution.

$$
\begin{aligned}
P(Y=y|X=x) &= \frac{P(X=x|Y=y)P(Y=y)}{P(X=x)} = \frac{P(X[1]=x[1],\dots,X[N]=x[n]|Y=y)p(y)}{P(X=x|Y=1)P(Y=1)+\ldots+P(X=x|Y=k)P(Y=k)} \\
&= \frac{p(y)\prod_{i=1}^N p(x[i]|y)}{\sum_{y'\in\mathcal{Y}} p(y')\prod_{i=1}^N p(x[i]|y')}
\end{aligned}
$$

(b) From part (a), we solve the equation:

$$
\exp(\langle w_y, \phi(x)\rangle) = p(y)\prod_{i=1}^N p(x[i]|y)
$$

We take the log of both sides and simplify.

$$
\begin{aligned}
\langle w_y, \phi(x)\rangle &= \log p(y) + \sum_{i=1}^N \log p(x[i]|y) \\
&= \log p(y) + \sum_{i=1}^N \sum_{x'\in[D]} [[x[i]=x']] \log p(x'|y) \\
&= \log p(y) + \sum_{x'\in[D]} \log p(x'|y) \sum_{i=1}^N [[x[i]=x']]
\end{aligned}
$$

We let the feature map be

$$
\phi : x \mapsto \left(1, \sum_{i=1}^N [[x[i]=1]], \dots, \sum_{i=1}^N [[x[i]=D]]\right)
$$

And we let $w_y \in \mathbb{R}^{N+1}$ be

$$
w_y = (\log p(y), \log p(1|y), \dots, \log p(D|y))
$$

(c) The process for computing the MAP estimators for the parameters is almost entirely the same as before. Our objective function, for a sample $S = X \times Y$ is as follows:

$$
\begin{aligned}
P(X|Y,\{p_y\})P(Y|p)P(p,\{p_y\}) &= \prod_{i=1}^n P(X=x_i|Y=y_i,\{p_y\})P(Y=y_i|p)P(p)P(\{p_y\}) \\
&= \prod_{i=1}^n \prod_{j=1}^N \prod_{y=1}^k \prod_{x=1}^D P(X[j]=x_i[j]|Y_i=y_i,\{p_y\})P(Y=y_i|p)\frac{1}{Z(1)}p(y)^{1-1}\frac{1}{Z(\alpha)^k}p_y(x)^{\alpha-1} \\
&= \frac{1}{Z(\alpha)^k Z(1)} \prod_{i=1}^n \prod_{j=1}^N \prod_{y=1}^k \prod_{x=1}^D p(x_i[j]|y_i)p(y_i)p_y(x)^{\alpha-1}
\end{aligned}
$$

Again, we can disregard the constant, and take log, to get the following:

$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{N}\log p(x_i[j]|y_i) + \sum_{i=1}^{n}\log p(y_i) + \sum_{y=1}^{k}\sum_{x=1}^{D}(\alpha-1)\log p_y(x) \\
&\quad= \sum_{y=1}^{k}\sum_{x=1}^{D}\sum_{i=1}^{n}\sum_{j=1}^{N}[[x_i[j]=x \wedge y_i=y]]\log p(x|y) + \sum_{i=1}^{n}\sum_{y=1}^{k}[[y_i=y]]\log p(y) + \sum_{y=1}^{k}\sum_{x=1}^{D}(\alpha-1)\log p_y(x) \\
&\quad= \sum_{y=1}^{k}\sum_{x=1}^{D}n(x,y)\log p(x|y) + \sum_{y=1}^{k}n(y)\log p(y) + \sum_{y=1}^{k}\sum_{x=1}^{D}(\alpha-1)\log p_y(x) \\
&\quad= \sum_{y=1}^{k}\sum_{x=1}^{D}(n(x,y)+(\alpha-1))\log p(x|y) + \sum_{y=1}^{k}n(y)\log p(y)
\end{aligned}
$$

We have a constrained optimization problem with the following constraints:

$$\sum_{y=1}^{k}p(y)=1$$

$$\sum_{x=1}^{D}p(x|y)=1$$

We write the Lagrangian.

$$\mathcal{L} = \sum_{y=1}^{k}\sum_{x=1}^{D}(n(x,y)+(\alpha-1)\log p(x|y)) + \sum_{y=1}^{k}n(y)\log p(y) - \sum_{y=1}^{k}\lambda_y\left(-1+\sum_{x=1}^{D}p(x|y)\right) - \mu\left(-1+\sum_{y=1}^{k}p(k)\right)$$

This gives us the following critical points:

$$
\begin{aligned}
[p(x|y)] &: \frac{n(x,y)+(\alpha-1)}{p(x|y)} = \lambda_y \\
[\lambda_y] &: \sum_{x=1}^{D}p(x|y)=1 \\
[p(y)] &: \frac{n(y)}{p(y)} = \mu \\
[\mu] &: \sum_{y=1}^{k}p(y)=1
\end{aligned}
$$

We solve for $p(y)$.

$$
\begin{aligned}
\frac{n(y)}{p(y)} &= \frac{n(y')}{p(y')} \\
p(y') &= p(y)\frac{n(y')}{n(y)} \\
1 &= p(y) + \sum_{y'\neq y}p(y)\frac{n(y')}{n(y)} \\
n(y) &= p(y)\sum_{y'\in[k]}n(y') = p(y)n \\
p(y) &= \frac{n(y)}{n}
\end{aligned}
$$

We solve for $p(x|y)$.

$$\begin{aligned}
\frac{n(x,y)+(\alpha-1)}{p(x|y)} &= \frac{n(x',y)+(\alpha-1)}{p(x'|y)} \\
p(x'|y) &= p(x|y)\frac{n(x',y)+(\alpha-1)}{n(x,y)+(\alpha-1)} \\
1 &= p(x|y) + \sum_{x'\neq x} p(x|y)\frac{n(x',y)+(\alpha-1)}{n(x,y)+(\alpha-1)} \\
n(x,y)+(\alpha-1) &= p(x|y)\sum_{x'\in[D]} n(x',y)+(\alpha-1) \\
\implies p(x|y) &= \frac{n(x,y)+(\alpha-1)}{\sum_{x'\in[D]} n(x',y)+(\alpha-1)}
\end{aligned}$$

Now we can plug in our MAP estimators.

$$\begin{aligned}
w_y[1] &= \log\frac{n(y)}{n} \\
w_y[1+x] &= \log\frac{n(x,y)+(\alpha-1)}{\sum_{x'\in[D]} n(x',y)+(\alpha-1)}
\end{aligned}$$

(d) We write $-\log P(\{y_i\}|\{x_i\},\{w_y\}) = -\log P(Y|X,W)$. Recall the posterior:

$$P(Y=y|x) = \frac{\exp(r_y(x))}{\sum_{y'=1}^k \exp(r_{y'}(x))} = \frac{\exp(\langle w_y,\phi(x)\rangle)}{\sum_{y'=1}^k \exp(\langle w_{y'},\phi(x)\rangle)}$$

Noting that (x_i, y_i) are i.i.d., we have:

$$\begin{aligned}
P(Y|X,w) &= \prod_{i=1}^n P(Y_i=y_i|X_i=x_i,W) = \prod_{i=1}^n \frac{\exp(\langle w_{y_i},\phi(x_i)\rangle)}{\sum_{l=1}^k \exp(\langle w_l,\phi(x_i)\rangle)} \\
-\log P(Y|X,w) &= -\left(\sum_{i=1}^n \langle w_{y_i},\phi(x_i)\rangle - \sum_{i=1}^n \log\sum_{y=1}^k \exp(\langle w_y,\phi(x_i)\rangle)\right)
\end{aligned}$$

(e)

$$\begin{aligned}
-\log P(Y|X,w) &= -\left(\sum_{i=1}^n \langle w_{y_i},\phi(x_i)\rangle - \sum_{i=1}^n \log\sum_{y=1}^k \exp(\langle w_y,\phi(x_i)\rangle)\right) \\
&= -\sum_{i=1}^n \log\sum_{y=1}^k \frac{\exp(\langle w_{y_i},\phi(x_i)\rangle)}{\exp(\langle w_y,\phi(x_i)\rangle)} \\
&= -\sum_{i=1}^n \log\sum_{y=1}^k \exp(\langle w_{y_i}-w_y,\phi(x_i)\rangle)
\end{aligned}$$

Therefore, we have the following loss function:

$$\ell(y_i; r_1(x),\ldots,r_k(x)) = -\log\sum_{y=1}^k \exp(r_{y_i}(x) - r_y(x))$$

Or

$$\ell(y_i; r_1(x),\ldots,r_k(x)) = -r_{y_i}(x)\log\sum_{y=1}^k \exp(-r_y(x))$$

5. **Adding Dependencies: A Markov Model.**

(a) Let $S = \{(x_1, y_1), \ldots (x_n, y_n)\}$, where $x_i \in [D]^N$ and $y_i \in [k]$. Let $(p_{\text{topic}})_{\text{topic} \in \mathcal{Y}} = p \in [0,1]^k$. Let $p_{y,\text{init}} = p_{y,0} \in [0,1]^D$, and let $p_{y,\text{tran}} = p_y \in [0,1]^{D \times D}$. Let $\theta = (p, \{p_{y,0}\}, \{p_y\})$. We also need to define the following summary statistic. For indices $I = \{i_1, \ldots, i_m\} \subset N$, $x \in \{0,1\}^N$, and $y \in [k]$:

$$n_{i_1,\ldots,i_m}(x_{i_1}, \ldots, x_{i_m}|y) = |\{(x_j, y_j) \in S | x_j \restriction I \equiv x \restriction I, y_j = y\}|$$

We let

$$n(x, x', y) = \sum_{j=1}^{N-1} n_{j+1,j}(x, x'|y)$$

We have the following log likelihood:

$$\ell(\theta|S) = \sum_{i=1}^{n} \log P(X = x_i, Y = y_i|\theta)$$

We can simplify this.

$$\log P(X = x_i, Y = y_i|\theta) = \log P(X = x_i|Y = y_i, \theta) + \log P(Y = y_i|\theta) = \log P(X = x_i|Y = y_i, \theta) + \sum_{y=1}^{k} [[y_i = y]] \log p(y)$$

We can simplify the first term.

$$\begin{aligned}
P(X = x_i|Y = y_i, \theta) &= P(X[1] = x_i[1], \ldots, X[N] = x_i[N]|Y = y_i, \theta) \\
&= P(X[1] = x_i[1], \ldots, X[N] = x_i[N]|X[1] = x_i[1], \ldots, X[N-1] = x_i[N-1], Y = y_i, \theta) \\
&\quad \cdot P(X[1] = x_i[1], \ldots, X[N-1] = x_i[N-1]|Y = y_i, \theta) \\
&= P(X[N] = x_i[N]|X[N-1] = x_i[N-1], Y = y_i, \theta) \\
&\quad \cdot P(X[1] = x_i[1], \ldots, X[N-1] = x_i[N-1]|Y = y_i, \theta) \\
&= p_{y_i}(x_i[N]|x_i[N-1]) P(X[1] = x_i[1], \ldots, X[N-1] = x_i[N-1]|Y = y_i, \theta) \\
&= p_{y_i}(x_i[N]|x_i[N-1]) \cdot \ldots \cdot p_{y_i}(x_i[2]|x_i[1]) p_{y_i,0}(x_i[1])
\end{aligned}$$

$$\implies \log P(X = x_i|Y = y_i, \theta) = \log p_{y_i,0}(x_i[1]) + \sum_{j=1}^{N-1} \log p_{y_i}(x_i[j+1]|x_i[j])$$

This gives us the following log likelihood:

$$\begin{aligned}
\ell(\theta|S) &= \sum_{i=1}^{n} \log p_{y_i,0}(x_i[1]) + \sum_{i=1}^{n} \sum_{j=1}^{N-1} \log p_{y_i}(x_i[j+1]|x_i[j]) + \sum_{i=1}^{n} \sum_{y=1}^{k} [[y_i = y]] \log p(y) \\
&= \sum_{i=1}^{n} \sum_{y=1}^{k} \sum_{x \in [D]} [[y_i = y \wedge x_i[1] = x]] \log p_{l,0}(x) \\
&+ \sum_{i=1}^{n} \sum_{j=1}^{N-1} \sum_{y=1}^{k} \sum_{x,x' \in [D]} [[y_i = y \wedge x_i[j] = x \wedge x_i[j+1] = x']] \log p_l(x'|x) \\
&+ \sum_{i=1}^{n} \sum_{y=1}^{k} [[y_i = y]] \log p(y) \\
&= \sum_{y=1}^{k} \sum_{x \in [D]} n_1(x, y) \log p_{l,0}(x) + \sum_{y=1}^{k} \sum_{x,x' \in [D]} \log p_y(x'|x) \sum_{j=1}^{N-1} n_{j,j+1}(x, x'|y) + \sum_{y=1}^{k} n(y) \log p(y) \\
&= \sum_{y=1}^{k} \sum_{x \in [D]} n_1(x, y) \log p_{l,0}(x) + \sum_{y=1}^{k} \sum_{x,x' \in [D]} \log p_y(x'|x) n(x, x'|y) + \sum_{y=1}^{k} n(y) \log p(y)
\end{aligned}$$

We have the following Lagrangian:

$$\begin{aligned}\mathcal{L} &= \sum_{y=1}^{k} n(y)\log p(y) - \lambda_1\left(-1+\sum_{y=1}^{k} p(y)\right)\\ &+ \sum_{y=1}^{k}\sum_{x,x'\in[D]} \log p_y(x'|x)n(x,x'|y) - \sum_{y=1}^{k}\sum_{x=1}^{D}\lambda_2(x,y)\left(-1+\sum_{x'=1}^{D} p_y(x'|x)\right)\\ &+ \sum_{y=1}^{k} n_1(x,y)\log p_{y,0}(x) - \sum_{y=1}^{k}\lambda_3(y)\left(-1+\sum_{x=1}^{D} p_{y,0}(x)\right)\end{aligned}$$

We have the following critical points:

$$[p(y)]: \frac{n(y)}{p(y)} = \lambda_1 \tag{1}$$

$$[\lambda_1]: \sum_{y=1}^{k} p(y) = 1 \tag{2}$$

$$[p_y(x'|x)]: \frac{n(x,x'|y)}{p_y(x'|x)} = \lambda_2(x,y) \tag{3}$$

$$[\lambda_2(x,y)]: \sum_{x'\in[D]} p_y(x'|x) = 1 \tag{4}$$

$$[p_{y,0}(x)]: \frac{n_1(x,y)}{p_{y,0}(x)} = \lambda_3(y) \tag{5}$$

$$[\lambda_3(y)]: \sum_{x\in[D]} p_{y,0}(x) = 1 \tag{6}$$

As before, (1) and (2) give us:

$$p(y) = \frac{n(y)}{n}$$

(3) and (4) give us:

$$\begin{aligned}\frac{n(x,x'|y)}{p_y(x'|x)} &= \frac{n(x,x''|y)}{p_y(x''|x)}\\ p_y(x''|x) &= p_y(x'|x)\frac{n(x,x''|y)}{n(x,x'|y)}\\ 1 &= p_y(x'|x) + \sum_{x''\neq x'} p_y(x'|x)\frac{n(x,x''|y)}{n(x,x'|y)}\\ n(x,x'|y) &= p_y(x'|x)\sum_{x''\in[D]} n(x,x''|y)\\ \implies p_y(x'|x) &= \frac{n(x,x'|y)}{\sum_{x''\in[D]} n(x,x''|y)} = \frac{n(x,x'|y)}{n(x,y)}\end{aligned}$$

(5) and (6) give us:

$$\begin{aligned}\frac{n_1(x,y)}{p_{y,0}(x)} = \frac{n_1(x',y)}{p_{y,0}(x')} &\implies p_{y,0}(x') = p_{y,0}(x)\frac{n_1(x',y)}{n_1(x,y)}\\ 1 &= p_{y,0}(x) + \sum_{x'\neq x} p_{y,0}(x)\frac{n_1(x',y)}{n_1(x,y)}\\ n_1(x,y) &= p_{y,0}(x)\sum_{x'\in[D]} n_1(x',y)\end{aligned}$$

$$\implies p_{y,0}(x) = \frac{n_1(x,y)}{\sum_{x' \in [D]} n_1(x',y)} = \frac{n_1(x,y)}{n(y)}$$

To summarize, we have:

$$\begin{aligned} p(y) &= \frac{n(y)}{n} \\ p_y(x'|x) &= \frac{n(x,x'|y)}{n(x,y)} \\ p_{y,0}(x) &= \frac{n_1(x,y)}{n(y)} \end{aligned}$$

(b) The calculations for $P(Y|p)$ are the same as before. We calculate $P(X|Y, \{p_y\}, \{p_{y,0}\})$.

$$\begin{aligned} P(X = x|Y = y, \{p_y\}, \{p_{y,0}\}) &= P(X[1] = x[1], \ldots, X[N] = x[N]|Y, \{p_y\}, \{p_{y,0}\}) \\ &= P(X[1] = x[1], \ldots, X[N] = x[N]|Y, \{p_y\}, \{p_{y,0}\}) \\ &= P(X[1] = x[1], \ldots, X[N] = x[N]|X[1] = x[1], \ldots, X[N-1] = x[N-1], Y, \{p_y\}, \{p_{y,0}\}) \\ &\quad \cdot P(X[1] = x[1], \ldots, X[N-1] = x[N-1]|Y, \{p_y\}, \{p_{y,0}\}) \\ &= P(X[N] = x[N]|X[N-1] = x[N-1], Y, \{p_y\}, \{p_{y,0}\}) \\ &\quad \cdot P(X[1] = x[1], \ldots, X[N-1] = x[N-1]|Y, \{p_y\}, \{p_{y,0}\}) \\ &= p_y(x[N]|x[N-1])P(X[1] = x[1], \ldots, X[N-1] = x[N-1]|Y, \{p_y\}, \{p_{y,0}\}) \\ &= p_y(x[N]|x[N-1]) \cdots p_y(x[2]|x[1])p_{y,0}(x[1]) \end{aligned}$$

Thus, applying this to the sample, we have:

$$\begin{aligned} P(X|Y, \{p_y\}, \{p_{y,0}\}) &= \prod_{i=1}^{n} P(X = x_i|Y = y_i, \{p_y\}, \{p_{y,0}\}) \\ &= \prod_{i=1}^{n} p_{y_i}(x_i[N]|x_i[N-1]) \cdots p_{y_i}(x_i[2]|x_i[1])p_{y_i,0}(x_i[1]) \\ &= \prod_{i=1}^{n} p_{y_i,0}(x_i[1]) \prod_{j=1}^{N-1} p_{y_i}(x_i[j+1]|x_i[j]) \end{aligned}$$

We calculate $P(p, \{p_y\}, \{p_{y,0}\})$.

$$\begin{aligned} P(p, \{p_y\}, \{p_{y,0}\}) &= P(p)P(\{p_y\})P(\{p_{y,0}\}) \\ P(p) &= \frac{1}{Z(1)} \prod_{y=1}^{k} p(y)^{\alpha-1} = \frac{1}{Z(1)} \\ P(\{p_y\}) &= \prod_{y=1}^{k} P(p_y) \\ P(p_y(\cdot|x)) &= \frac{1}{Z(\alpha)} \prod_{x'=1}^{D} p_y(x'|x)^{\alpha-1} \\ \implies P(p_y) &= \frac{1}{Z(\alpha)^D} \prod_{x,x' \in [D]} p_y(x'|x)^{\alpha-1} \\ \implies P(\{p_y\}) &= \frac{1}{Z(\alpha)^{kD}} \prod_{y=1}^{k} \prod_{x,x' \in [D]} p_y(x'|x)^{\alpha-1} \\ P(\{p_{y,0}\}) &= \prod_{y=1}^{k} P(p_{y,0}) = \frac{1}{(Z(\alpha)^k)} \prod_{y=1}^{k} \prod_{x \in [D]} p_{y,0}(x)^{\alpha-1} \end{aligned}$$

As before, we may neglect the constants and take the log.

$$\sum_{i=1}^{n} \log(p_{y_i,0})(x_i[1]) + \sum_{i=1}^{n}\sum_{j=1}^{N-1} \log p_{y_i}(x_i[j+1]|x_i[j])$$
$$+ \sum_{i=1}^{n} \log p(y_i) + \sum_{y=1}^{k}\sum_{x,x'\in[D]} (\alpha-1)(\log p_y(x'|x) + \log p_{y,0}(x))$$

We can simplify this expression in the usual way.

$$= \sum_{y=1}^{k}\sum_{x\in[D]} n_1(x,y)\log p_{y,0}(x) + \sum_{y=1}^{k}\sum_{x,x'\in[D]} n(x,x'|y)\log p_y(x'|x)$$
$$+ \sum_{y=1}^{k} n(y)\log p(y) + \sum_{y=1}^{k}\sum_{x,x'\in[D]} (\alpha-1)(\log p_y(x'|x) + \log p_{y,0}(x))$$
$$= \sum_{y=1}^{k}\sum_{x\in[D]} (n_1(x,y) + (\alpha-1))\log p_{y,0}(x) + \sum_{y=1}^{k}\sum_{x,x'\in[D]} (n(x,x'|y) + (\alpha-1))\log p_y(x'|x) + \sum_{y=1}^{k} n(y)\log p(y)$$

We solve the following Lagrangian

$$\mathcal{L} = \sum_{y=1}^{k}\sum_{x\in[D]} (n_1(x,y) + (\alpha-1))\log p_{y,0}(x) - \sum_{y=1}^{k}\lambda_1(y)\left(-1 + \sum_{x\in[D]} p_{y,0}(x)\right)$$
$$+ \sum_{y=1}^{k}\sum_{x,x'\in[D]} (n(x,x'|y) + (\alpha-1))\log p_y(x'|x) - \sum_{y=1}^{k}\sum_{x\in[D]}\lambda_2(x,y)\left(-1 + \sum_{x'\in[D]} p_y(x'|x)\right)$$
$$+ \sum_{y=1}^{k} n(y)\log p(y) - \lambda_3\left(-1 + \sum_{y=1}^{k} p(y)\right)$$

We take critical points.

$$[p_{y,0}(x)] : \frac{n_1(x,y) + (\alpha-1)}{p_{y,0}(x)} = \lambda_1(y)$$
$$[\lambda_1(y)] : \sum_{x\in[D]} p_{y,0}(x) = 1$$
$$[p_y(x'|x)] : \frac{n(x,x'|y) + (\alpha-1)}{p_y(x'|x)} = \lambda_2(x,y)$$
$$[\lambda_2(x,y)] : \sum_{x'\in[D]} p_y(x'|x) = 1$$
$$[p(y)] : \frac{n(y)}{p(y)} = \lambda_3$$
$$[\lambda_3] : \sum_{y=1}^{k} p(y) = 1$$

We solve for $p_{y,0}(x)$.

$$\begin{aligned}
\frac{n_1(x,y)+(\alpha-1)}{p_{y,0}(x)} &= \frac{n_1(x',y)+(\alpha-1)}{p_{y,0}(x')} \implies p_{y,0}(x') = p_{y,0}(x)\frac{n_1(x',y)+(\alpha-1)}{n_1(x,y)+(\alpha-1)} \\
1 &= p_{y,0}(x) + \sum_{x'\neq x} p_{y,0}(x)\frac{n_1(x',y)+(\alpha-1)}{n_1(x,y)+(\alpha-1)} \\
n_1(x,y)+(\alpha-1) &= p_{y,0}(x)\sum_{x'\in[D]} n_1(x',y)+(\alpha-1) \\
\implies p_{y,0}(x) &= \frac{n_1(x,y)+(\alpha-1)}{\sum_{x'\in[D]} n_1(x',y)+(\alpha-1)} = \frac{n_1(x,y)+(\alpha-1)}{n(y)+(\alpha-1)}
\end{aligned}$$

We solve for $p_y(x'|x)$.

$$\begin{aligned}
\frac{n(x,x'|y)+(\alpha-1)}{p_y(x'|x)} &= \frac{n(x,x''|y)+(\alpha-1)}{p_y(x''|x)} \implies p_y(x''|x) = p_y(x'|x)\frac{n(x,x''|y)+(\alpha-1)}{n(x,x'|y)+(\alpha-1)} \\
1 &= p_y(x'|x) + \sum_{x''\neq x'} p_y(x'|x)\frac{n(x,x''|y)+(\alpha-1)}{n(x,x'|y)+(\alpha-1)} \\
n(x,x'|y)+(\alpha-1) &= p_y(x'|x)\sum_{x''\in[D]} n(x,x''|y)+(\alpha-1) \\
\implies p_y(x'|x) &= \frac{n(x,x'|y)+(\alpha-1)}{\sum_{x''\in[D]} n(x,x''|y)+(\alpha-1)} = \frac{n(x,x'|y)+(\alpha-1)}{n(x,y)+(\alpha-1)}
\end{aligned}$$

We solve for $p(y)$.

$$\begin{aligned}
\frac{n(y)}{p(y)} &= \frac{n(y')}{p(y')} \implies p(y') = p(y)\frac{n(y')}{n(y)} \\
1 &= p(y) + \sum_{y'\neq y} p(y)\frac{n(y')}{n(y)} \\
n &= p(y)\sum_{y'\in[k]} n(y') = p(y)n \\
p(y) &= \frac{n(y)}{n}
\end{aligned}$$

(c) We compute the posterior as follows:

$$\begin{aligned}
P(Y=1|X=x) &= \frac{P(X=x|Y=1)P(Y=1)}{P(X=x)} \\
&= \frac{P(X=x|Y=1)P(Y=1)}{P(X=x|Y=1)P(Y=1)+P(X=x|Y=0)P(Y=0)}
\end{aligned}$$

We compute $P(X=x|Y=y)$ using part (a).

$$P(X=x|Y=y) = p_{y,0}(x[1])\prod_{i=1}^{N-1} p_y(x[i+1]|x[i])$$

Thus, we have our posterior distribution as follows:

$$P(Y=1|X=x) = \frac{p_{1,0}(x[1])\prod_{i=1}^{N-1} p_1(x[i+1]|x[i])P(Y=1)}{p_{1,0}(x[1])\prod_{i=1}^{N-1} p_1(x[i+1]|x[i])P(Y=1) + p_{0,0}(x[1])\prod_{i=1}^{N-1} p_0(x[i+1]|x[i])P(Y=0)}$$

$$= \frac{1}{1 + \frac{p_{0,0}(x[1])}{p_{1,0}(x[1])} \frac{p(0)}{p(1)} \prod_{i=1}^{N-1} \frac{p_0(x[i+1]|x[i])}{p_1(x[i+1]|x[i])}}$$

We solve the following equation:

$$\exp(-r(x)) = \frac{p_{0,0}(x[1])}{p_{1,0}(x[1])} \frac{p(0)}{p(1)} \prod_{i=1}^{N-1} \frac{p_0(x[i+1]|x[i])}{p_1(x[i+1]|x[i])}$$
$$\implies r(x) = \log \frac{p(1) p_{1,0}(x[1])}{p(0) p_{0,0}(x[1])} + \sum_{i=1}^{N-1} \log \frac{p_1(x[i+1]|x[i])}{p_0(x[i+1]|x[i])}$$

(d)

$$r(x) = \log \frac{p(1)}{p(0)} + \sum_{x=1}^{D} [[x[1] = x]] \log \frac{p_{1,0}(x)}{p_{0,0}(x)} + \sum_{i=1}^{N-1} \sum_{x,x'=1}^{D} [[x[i] = x \wedge x[i+1] = x']] \log \frac{p_1(x'|x)}{p_0(x'|x)}$$
$$= \log \frac{p(1)}{p(0)} + \sum_{x,x'=1}^{D} n_1(x) \log \frac{p_{1,0}(x)}{p_{0,0}(x)} + n(x,x') \log \frac{p_1(x'|x)}{p_0(x'|x)}$$

This results in the following feature map $\phi : \mathcal{X} \to \mathbb{R}^{D^2+D+1}$.

$$\phi : x \mapsto (1, n_1(1), \ldots, n_1(D), n(1,1), \ldots, n(D,D))$$

Thus results in the following vector w:

$$w = \left(\log \frac{p(1)}{p(0)}, \log \frac{p_{1,0}(1)}{p_{0,0}(1)}, \ldots, \log \frac{p_{1,0}(D)}{p_{0,0}(D)}, \log \frac{p_1(1|1)}{p_0(1|1)}, \ldots, \log \frac{p_1(D|D)}{p_0(D|D)} \right)$$

(e) We plug in the MAP estimators into the components.

$$\log \frac{p(1)}{p(0)} = \log \frac{n(1)}{n(0)}$$
$$\log \frac{p_{1,0}(x)}{p_{0,0}(x)} = \log \frac{n_1(x,1) + (\alpha - 1)}{n_1(x,0) + (\alpha - 1)} - \log \frac{n(1) + (\alpha - 1)}{n(0) + (\alpha - 1)}$$
$$\log \frac{p_1(x'|x)}{p_0(x'|x)} = \log \frac{n(x,x'|1) + (\alpha - 1)}{n(x,x'|0) + (\alpha - 1)} - \log \frac{n(x,1) + (\alpha - 1)}{n(x,0) + (\alpha - 1)}$$

Thus, we have:

$$w = \Bigg(\log \frac{n(1)}{n(0)}, \log \frac{n_1(1,1) + (\alpha - 1)}{n_1(1,0) + (\alpha - 1)} - \log \frac{n(1) + (\alpha - 1)}{n(0) + (\alpha - 1)},$$
$$\ldots, \log \frac{n_1(D,1) + (\alpha - 1)}{n_1(D,0) + (\alpha - 1)} - \log \frac{n(1) + (\alpha - 1)}{n(0) + (\alpha - 1)}, \log \frac{n(1,1|1) + (\alpha - 1)}{n(1,1|0) + (\alpha - 1)} - \log \frac{n(1,1) + (\alpha - 1)}{n(1,0) + (\alpha - 1)},$$
$$\ldots, \log \frac{n(D,D|D) + (\alpha - 1)}{n(D,D|0) + (\alpha - 1)} - \log \frac{n(D,D) + (\alpha - 1)}{n(D,0) + (\alpha - 1)} \Bigg)$$