Homework 2.1

2.1.1

1

$$P(x \mid p) = \prod_{d=1}^{D} p_d^{x_d}(1 - p_d)^{1-x_d}.$$

2

$$P(x^{(i)} \mid \mathbf{p}, \pi) = \sum_{k=1}^{K} P(x^{(i)} \mid p^{(k)})\pi(k).$$

3

The log likelihood is

$$\sum_{i=1}^{n} \log P(x^{(i)} \mid \mathbf{p}, \pi).$$

2.1.2

1

The probability that $p^{(k)}$ was chosen for $x^{(i)}$ to be drawn from is $\pi(k)$. So,

$$P(z^{(i)} \mid \pi) = \sum_{k=1}^{K} \pi(k) z_k^{(i)}.$$

Edit: This can also be written as

$$\prod_{k=1}^{K} \pi(k)^{z_k^{(i)}}.$$

For the second question, the indicator variable tells us that the distribution we are drawing from is

$$p=\sum_{k=1}^{K} z_k^{(i)} p^{(k)}=\langle z^{(i)}, \mathbf{p}\rangle .$$

So, the probability $x^{(i)}$ is selected is

$$P\left(x^{(i)} \mid \langle z^{(i)}, \mathbf{p}\rangle\right)$$

Note that π isn't actually a factor. Edit: This can be written another way as

$$\prod_{k=1}^{K}\left(\prod_{d=1}^{D} p_d^{x_d^{(i)}}\left(1-p_d\right)^{1-x_d^{(i)}}\right)^{z_k} .$$

2

By Bayes' law (and assuming independence of the $x^{(i)}$),

$$\begin{aligned} P(X, Z \mid \pi, \mathbf{p}) &= P(X \mid Z, \pi, \mathbf{p}) P(Z \mid \pi, \mathbf{p}) \\ &= P(X \mid Z, \pi, \mathbf{p}) P(Z \mid \pi) \\ &= \left(\prod_{i=1}^{n} P(x^{(i)} \mid z^{(i)}, \pi, \mathbf{p})\right)\left(\prod_{i=1}^{n} P(z^{(i)} \mid \pi)\right). \end{aligned}$$

3

Since $z^{(i)}$ is an indicator variable,

$$\eta\left(z_k^{(i)}\right)=E\left[z_k^{(i)} \mid x^{(i)}, \pi, \mathbf{p}\right]=P\left(z_k^{(i)}=1 \mid x^{(i)}, \pi, \mathbf{p}\right).$$

By Bayes' theorem,

$$\begin{aligned} P\left(z_k^{(i)}=1 \mid x^{(i)}, \pi, \mathbf{p}\right) &= \frac{P\left(x^{(i)} \mid z_k^{(i)}=1, \pi, \mathbf{p}\right) P(z_k^{(i)}=1 \mid \pi, \mathbf{p})}{P\left(x^{(i)} \mid \pi, \mathbf{p}\right)} \\ &= \frac{P\left(x^{(i)} \mid p^{(k)}\right) \pi_k}{P\left(x^{(i)} \mid \pi, \mathbf{p}\right)} \end{aligned} .$$

All of these are answers we already have. Plugging those in, we get

$$\eta\left(z_k^{(i)}\right)=\frac{\pi_k \prod_{d=1}^{D}\left(p_d^{(k)}\right)^{x_d^{(i)}}\left(1-p_d^{(k)}\right)^{1-x_d^{(i)}}}{\sum_{j=1}^{n} \pi_j \prod_{d=1}^{D}\left(p_d^{(j)}\right)^{x_d^{(i)}}\left(1-p_d^{(j)}\right)^{1-x_d^{(i)}}},$$

as desired.

For the second part, note that we have already calculated $P(X, Z \mid \tilde{\pi}, \tilde{\mathbf{p}})$, so we can just write down

$$\begin{aligned}
\log P(X, Z \mid \tilde{\pi}, \tilde{\mathbf{p}}) &= \log\left(\left(\prod_{i=1}^{n} P(x^{(i)} \mid z^{(i)}, \tilde{\pi}, \tilde{\mathbf{p}})\right)\left(\prod_{i=1}^{n} P(z^{(i)} \mid \tilde{\pi})\right)\right) \\
&= \sum_{i=1}^{n} \log\left(\prod_{k=1}^{K}\left(\prod_{d=1}^{D} \tilde{p}_d^{(k)})^{x_d^{(i)}}(1-\tilde{p}_d^{(k)})^{1-x_d^{(i)}}\right)^{z_k^{(i)}}\right) \\
&\quad + \sum_{i=1}^{n} \log\left(\prod_{k=1}^{K} \tilde{\pi}(k)^{z_k^{(i)}}\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} z_k^{(i)}\left[\log \tilde{\pi}_k + \sum_{d=1}^{D} x_d^{(i)} \log \tilde{p}_d^{(k)} + (1-x_d^{(i)}) \log(1-\tilde{p}_d^{(
\end{aligned}$$

Taking the conditional expectation of this gives the desired result (since $E(z_i \mid X, \pi, \mathbf{p}) = \eta(z_k^{(i)})$ and the other values are constant).

2.1.3

1

The partial derivative with respect to $\tilde{p}_d^{(k)}$ of the E step is

$$\sum_{i=1}^{N} \eta(z_k^{(i)})\left(\frac{x_d^{(i)}}{\tilde{p}_d^{(k)}} - \frac{1-x_d^{(i)}}{1-\tilde{p}_d^{(k)}}\right).$$

To find the maximum, we want this to equal zero. Multiply everything by $\tilde{p}_d^{(k)}(1-\tilde{p}_d^{(k)})$ and collect terms to get that

$$\sum_{i=1}^{N} \eta(z_k^{(i)})(x_d^{(i)} - \tilde{p}_d^{(k)}) = 0.$$

Solving for $\tilde{p}_d^{(k)}$, we get that

$$\tilde{p}_d^{(k)} = \frac{\sum_{i=1}^{N} \eta(z_k^{(i)}) x_d^{(i)}}{N_k},$$

as desired.

2

We want to maximize

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \eta(z_k^{(i)}) \tilde{\pi}_k = \sum_{k=1}^{K} N_k \tilde{\pi}_k.$$

subject to $\sum_{k'} \tilde{\pi}_{k'} = 1$. Using Lagrange multipliers, the above is maximized when

$$\langle N_1, N_2, \ldots, N_K \rangle = \lambda \langle 1, 1, \ldots, 1 \rangle$$

for some constant λ. This allows us to easily see that

$$\tilde{\pi}_k = \frac{N_k}{\sum_{k'} N_k},$$

as desired.

Homework 2.3

2.3.1

1

$$C = \frac{1}{N}\sum_{i=1}^{N}\left(\phi(x_i) - \overline{\phi(x)}\right)\left(\phi(x_i) - \overline{\phi(x)}\right)^T.$$

2

Suppose that v is an eigenvector of C. Then

$$\lambda v = Cv = \frac{1}{N}\sum_{i=1}^{N}\left(\phi(x_i) - \overline{\phi(x)}\right)\left[\left(\phi(x_i) - \overline{\phi(x)}\right)^T v\right].$$

Thus, λv is a linear combination of the list

$$\phi(x_1) - \overline{\phi(x)}, \quad \phi(x_2) - \overline{\phi(x)}, \quad \ldots, \quad \phi(x_N) - \overline{\phi(x)},$$

which means so is v.

3

The jth component of $\tilde{K}\alpha$ is

$$(\tilde{K}\alpha)_j = \sum_{i=1}^{N}\alpha_i\left(\phi(x_j) - \overline{\phi(x)}\right)^T\left(\phi(x_i) - \overline{\phi(x)}\right) = \left(\phi(x_j) - \overline{\phi(x)}\right)^T v.$$

Multiplying by $\tilde{K}$ on the left again, we get that the kth component of $\tilde{K}^2\alpha$ is

$$\sum_{j=1}^{N}\alpha_j\left(\phi(x_k) - \overline{\phi(x)}\right)^T\left(\phi(x_j) - \overline{\phi(x)}\right)\left(\phi(x_j) - \overline{\phi(x)}\right)^T v. \tag{1}$$

Using the fact that

$$C = \frac{1}{N} \sum_{j=1}^{N} \alpha_j \left(\phi(x_j) - \overline{\phi(x)} \right) \left(\phi(x_j) - \overline{\phi(x)} \right)^T,$$

(1) is equal to

$$\sum_{i=1}^{N} \alpha_i \left(\phi(x_k) - \overline{\phi(x)} \right)^T NCv.$$

But $Cv = \lambda v$, so this turns into

$$\sum_{i=1}^{N} \alpha_i \left(\phi(x_k) - \overline{\phi(x)} \right)^T N\lambda v = N\lambda(\tilde{K}\alpha)_k.$$

Thus,

$$\tilde{K}^2 \alpha = N\lambda \tilde{K}\alpha,$$

as desired.

4

Suppose that α is a solution to

$$N\lambda\alpha = \tilde{K}\alpha.$$

Then, since $N\lambda$ is a scalar,

$$N\lambda\tilde{K}\alpha = \tilde{K}(N\lambda\alpha) = \tilde{K}^2\alpha,$$

as desired.

5

First off, ee^T is the $N \times N$ matrix with every entry equal to $1/N$.

For ease of typing this up, let $\mu = \overline{\phi(x)}$.
Note that the component of $ee^T K$ in the ith row and jth column is

$$\frac{1}{N} \sum_{k=1}^{N} \langle \phi(x_k), \phi(x_j) \rangle = \left\langle \frac{1}{N} \sum_{k=1}^{N} \phi(x_k), \phi(x_j) \right\rangle = \langle \mu, \phi(x_j) \rangle.$$

Note that the component of Kee^T in the ith row and jth column is

$$\langle \phi(x_j), \mu \rangle.$$

Finally, the component of $ee^T Kee^T$ in the ith row and jth column is

$$\langle \mu, \mu \rangle.$$

Thus, the component in the i, jth location of

$$(I - ee^T)K(I - ee^T) = K - ee^T K - Kee^T + ee^T Kee^T.$$

equals

$$\langle \phi(x_i), \phi(x_j) \rangle - \langle \phi(x_i), \mu \rangle - \langle \mu, \phi(x_j) \rangle + \langle \mu, \mu \rangle = \langle \phi(x_i) - \mu, \phi(x_j) - \mu \rangle,$$

which is exactly the i, jth entry of $\tilde{K}_{i,j}$, as desired.

6

Note that

$$\langle v, v \rangle = \alpha^T \tilde{K} \alpha.$$

the factor we want is the square root of this.

7

Its position in the normalized new space is

$$\begin{aligned} \phi(x) &= \sum_v \frac{\langle \phi(x), v \rangle v}{\langle v, v \rangle} \\ &= \sum_{(\alpha, v)} \frac{\left(\sum_{i=1}^N \alpha_i \left\langle \phi(x), \phi(x_i) - \overline{\phi(x)} \right\rangle \right)}{\alpha^T \tilde{K} \alpha} v. \\ &= \sum_{(\alpha, v)} \frac{\left(\sum_{i=1}^N \alpha_i k \left(x, \phi(x_i) - \overline{\phi(x)} \right) \right)}{\alpha^T \tilde{K} \alpha} v \end{aligned}$$

So, we only need to know α, v, and k to compute what x is in the new space.

Homework 2.4

2.4.1

1

The minimum occurs at $\lambda = 0$, $\lambda = +\infty$, or when

$$\frac{d}{d\lambda}\left(\lambda + \frac{x^2}{\lambda + d}\right) = 1 - \frac{x^2}{(\lambda + d)^2} = 0.$$

The latter occurs only if $|x| \geq d$, when $\lambda = |x| - d$. Testing out these three values for λ, the minimum is x^2/d if $|x| > d$ and $2|x| - d$ otherwise, with the minimum occurring at $\lambda^* = \max(0, |x| - d)$.

2

Take a second derivative to get

$$\frac{\partial^2 B(x,d)}{\partial x^2} = \begin{cases} \frac{2}{d} & \text{if } |x| \leq d, \\ 0 & \text{if } |x| > d. \end{cases}$$

This is non-negative everywhere, so B is convex in x (for fixed d).

3

$$\nabla B = \left\langle \frac{\partial}{\partial x} B, \frac{\partial}{\partial d} B \right\rangle = \begin{cases} \left\langle \frac{2x}{d}, -\frac{x^2}{d^2} \right\rangle & \text{if } |x| \leq d, \\ \left\langle \frac{2x}{|x|}, -1 \right\rangle & \text{if } |x| > d. \end{cases}$$

4

If d is higher, the penalty term B is comparatively less for more extreme values of x.

2.4.2

1

Suppose that there is a strictly feasible solution for which $\|a\|_1 \leq 1$. Then by the triangle inequality

$$|a^T x| = \left|\sum a_i x_i\right| \leq \sum |a_i||x_i| < \sum |a_i| = 1,$$

where the last inequality comes from the fact that $|x_i| < 1$ for all i. This contradicts the fact that $a^T x = 1$, though. So, all strictly feasible solutions must have $\|a\|_1 > 1$.

2

The Lagrangian is

$$L(x, \mu, \lambda) = \lambda - \sum_{i=1}^{N} \left[\frac{1}{2} d_i x_i^2 + r_i x_i + \frac{1}{2} \mu_i (x_i^2 - 1) + \lambda a_i x_i \right].$$

We have the constraints $\mu_i \geq 0$ for all i.
We want to minimize this Lagrangian for all i. Note that

$$\frac{\partial}{\partial x_i} L(x, \mu, \lambda) = -(d_i x_i + r_i + \mu_i x_i + \lambda a_i).$$

This is equal to zero at

$$x_i^* = -\frac{\lambda a_i + r_i}{\mu_i + d_i}$$

(where the $*$ is there to remind us that this is where x_i is optimized). Plugging this into the Lagrangian (and simplifying a lot) gives us the dual problem

$$\min_{\mu, \lambda} \lambda + \frac{1}{2} \sum_{i=1}^{N} \left[\mu_i + \frac{(\lambda a_i + r_i)^2}{\mu_i + d_i} \right]$$

subject to $\mu_i \geq 0$ for all i.

3

For the tuple (x^*, μ, λ) to satisfy the KKT conditions, we need the following:
Stationary:
For all i,

$$d_i x_i + r_i + \lambda a_i + \mu_i x_i = 0$$

This is how we defined x_i^* earlier, so this is always satisfied at x^*.

Primal feasibility:

$$\sum_{i=1}^{N} a_i x_i^* = 1.$$

and

$$(x_i^*)^2 \leq 1$$

for all i. Equivalently,

$$|\lambda a_i + r_i| \leq |\mu_i + d_i|.$$

Dual feasibility:
This is the constraint $\mu_i \geq 0$ for all i.

Complementary slackness:

$$\sum_{i=1}^{N} \mu_i((x_i^*)^2 - 1) = 0.$$

Slater's condition and part 1 tells us this characterizes the optimal solution when $\|a\|_1 > 1$.

4

It doesn't matter what order we do the minimization, so the dual problem is the same as

$$\min_{\lambda} \min_{\mu} \lambda + \frac{1}{2} \sum_{i=1}^{N} \left[\mu_i + \frac{(\lambda a_i + r_i)^2}{\mu_i + d_i} \right] = \min_{\lambda} \lambda + \frac{1}{2} \sum_{i=1}^{N} B(r_i + \lambda a_i, d_i),$$

since that's how $B(r_i + \lambda a_i, d_i)$ is defined. Of course, we then have to make sure our constraints our satisfied. Since the μ_i is optimized at

$$\mu_i^* = \max(0, |r_i + \lambda a_i| - d_i),$$

this is what we should plug in to calculate x_i^* and make sure the constraints work. In fact, this value of μ_i^* guarantees that $|x_i^*| \leq 1$, so we have one fewer constraint to worry about.

5

Define $A(\lambda) = 1 - \sum a_i x_i^*$, where x_i^* is the optimal value for x_i for the given lambda. We want to find

$$\min_{\lambda} \lambda + \frac{1}{2} \sum_{i=1}^{N} B(r_i + \lambda a_i, d_i)$$

given that $A(\lambda) = 0$.

There are $2N$ critical points for λ for which x_i^* switches from being ± 1 to being in the interval $(-1, 1)$. $A(\lambda)$ is linear between any pair of these critical points, which makes it easy to compute the values of λ for which $A(\lambda) = 0$ by computing $A(\lambda)$ at the critical points. Using this idea, the following algorithm computes the desired minimum in $O(N \log N)$ time:

1. Compute the critical points $\lambda_1, \lambda_2, \ldots, \lambda_{2N}$. These occur when $|r_i + \lambda a_i| = d_i$. This can be done in $O(N)$ time. While you do so, record which x_i^* each critical point corresponds to. This allows you to compute the change of the slope of $A(\lambda)$ at each critical point in $O(1)$ time.
2. Sort your critical points so that $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_{2N}$. This takes $O(N \log N)$ time.
3. Compute $A(\lambda_i) = 1 - \sum a_i x_i^*$ at each of the critical points. This can be done in $O(N)$ time by keeping track of the slope of $A(\lambda)$ between critical points.
4. Loop through the critical points, recording which pairs of consecutive points $(A(\lambda_i), A(\lambda_{i+1}))$ have opposite signs. This can be done in $O(N)$ time.
5. Between each of these pairs of points, calculate the value of λ makes $A(\lambda) = 0$. This can be done in $O(1)$ time per pair.
6. Use a binary search to find which of the above values of λ minimizes

$$f(\lambda) = \lambda + \frac{1}{2} \sum_{i=1}^{N} B(r_i + \lambda a_i, d_i).$$

   (We can use a binary search because this is a convex function.) This takes $O(N \log N)$ time since it takes $O(N)$ time to calculate $f(\lambda)$ for each λ, and we need to calculate $f(\lambda)$ at most $\log_2(N)$ times.

In total, the algorithm takes $O(N \log N)$ time.

6

To recover the primal solution x, use the fact that

$$\mu_i = \max(0, |r_i + \lambda a_i| - d_i)$$

to compute

$$x_i = \frac{\lambda a_i + r_i}{\mu_i + d_i}.$$

This can be combined all together to get

$$x_i = \begin{cases} \frac{\lambda a_i + r_i}{d_i} & \text{if } |r_i + \lambda a_i| \le d_i \\ \frac{\lambda a_i + r_i}{|\lambda a_i + r_i|} & \text{if } |r_i + \lambda a_i| > d_i. \end{cases}$$