EM: Expectation Maximization

🎯 Goal
- We have a generative process as follows: $\theta \rightarrow Z \rightarrow X$
- Here, X is observed and Z is latent
- Estimate θ that maximizes the likelihood of observed data.

▼ Look at an example to better understand the problem

🟢 Coin toss example:
- Someone has K coins with biases $\theta_{1:K}$ … the person picks a coin z_i at random and records the outcome of the toss as x_i, where x_i is either H or T
- N such identical and independent coin toss results are published: $[x_1 \;\; x_2 \;\; ... \;\; x_n]$
- Your goal is to estimate $\theta_{1:k}$ that maximizes the likelihood of the published observations, i.e., $\underset{\theta}{argmax} \;\; p(X|\theta)$
- In fact, I would like you to also estimate $Z = [z_1, z_2, ..., z_n]$ in the process?

▼ Why is this not just a maximum likelihood estimation problem?

🟢 Consider the term $\underset{\theta}{argmax} \;\; p(x_1, x_2, ..., x_n|\theta_{1:K}) = \underset{\theta}{argmax} \; \prod_i \; p(x_i|\theta_{1:K})$
- But how do you compute this individual term?
- You cannot because you don't know which coin this x_i came from.
- One possibility is to pretend x_i came from coin θ_j … and compute the likelihood.
- But then it becomes a huge — exponentially growing — combinatorics problem since there are K^n possible assignments for n tosses with K coins.
- This is crazy.

🟢 But observe that if you can select some possible assignments … and associate probabilities to them … you can compute the expectation of the likelihood.
- In other words, since you don't know how to compute the likelihood function, you are setting up the average likelihood over multiple possible assignments.
- Now, optimize that average likelihood.
- That's what EM wants to go towards.

▼ The insight in plain language

🟢 1. Write out what you need, i.e., $L = p(X, Z|\theta)$ … and this breaks into: $L = p(X|Z, \theta)p(Z|\theta)$, i.e., the likelihood and prior.
2. If you pretend to know $Z = k,\ k \in [1, 2, ..., K]$ … then you can compute the likelihood as a function of (Z, θ) … and the prior can be assumed as equal or some other distribution from domain knowledge.
3. You can simply perform MLE by differentiating w.r.t. θ

4. But you don't know Z … so what do you do?
5. What if you pretend to know the distribution for Z? Then, you can sample some $Z = z_k$, plug that z_k into the likelihood and prior, and since picking z_k is associated with some probability, you multiply that L with $p(Z = z_k)$
6. If you do this for each z_k, you are essentially taking the expectation of L over Z.
7. In other words, L can be viewed as the function of the random variable, $g(Z)$, and $p(Z = z_k)$ is the distribution on Z … and now, the equation $\sum g(Z)p(Z = z_k)$ forms the expectation of the function of the RV.
8. Since $g(Z)$ is the likelihood, we are actually computing the expectation of the likelihood.

9. Now, how do you get this distribution for Z?
10. Why not compute the posterior distribution for Z based on the observed data and some initial guess of $\theta_{1:K}$
11. This won't be right initially … but it can improve the likelihood L which can then improve the posterior … which can then improve L … until convergence.
12. Convergence happens when $\hat{\theta}(t+1) - \hat{\theta(t)} \leq \epsilon$

▼ Set up the likelihood function

🟢 - $L = p(X, Z|\theta) = p(X|Z, \theta)p(Z|\theta) = p(x_1, x_2, ..., x_n|z_1, z_2, ..., z_n; \theta_{1:K})p(z_1, z_2, ..., z_n|\theta_{1:K}) = \prod_i p(x_i|z_i; \theta_i)p(z_i|\theta_i)$
- Since θ_i does not influence the choice of z_i … $p(z_i|\theta_i) = p(z_i)$ … which is the prior for picking z_i. Let's denote this prior with $\pi_{z_i} = p(z_i)$

- The log likelihood is hence: $log\ p(X, Z|\theta_{1:K}) = \sum_{i=1}^{n}\{log\ p(x_i|\theta_{z_i}) + log\ \pi_{z_i}\}$

- Now, z_i can take any one value between $[1:K]$ …
- so let's model that by summing over all K possible values but picking only one term from the sum using a DELTA function → nice trick

- $$log\ p(X, Z|\theta_{1:K}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\delta_k(z_i)\{log\ p(x_i|\theta_k) + log\ \pi_k\}$$

- Here, $\delta_k(z_i) = 1$ when $z_i = k$ … otherwise, $\delta_k(z_i) = 0$

- Importantly, z_i is the only variable we don't know above.

▼ Expectation step → create the posterior distribution

🟢 - The posterior is: $p(Z|X; \theta) = \prod_i p(z_i|x_i; \theta)$

- Using Bayes rule, we have: $p(Z|X; \theta) = \prod_i p(z_i|x_i; \theta) = \prod_i \{\ p(x_i|z_i; \theta)p(z_i; \theta)\ /\ p(x_i)\ \}$

$$= \prod_i \{\ p(x_i|z_i; \theta)p(z_i; \theta)\ /\ \sum_{i=1}^{K} p(x_i|z_i; \theta)p(z_i; \theta)\ \}$$

- Now, we can shorten this by picking the correct θ based on z_i, so we can write this equation as:

$$= \prod_i \{\ p(x_i|\theta_{z_i})\pi_{z_i}\ /\ \sum_{k=1}^{K} p(x_i|\theta_k)\pi_k\ \}$$

- Let's call this posterior distribution on variable Z as $q(Z)$ … we have $q(Z) = \prod_i q_i(z_i)$

▼ Maximization step → take the expectation of the likelihood over the posterior on Z

🟢 - Recall the log likelihood: $log\ p(X, Z|\theta_{1:K}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_k(z_i)\{log\ p(x_i|\theta_k) + log\ \pi_k\}$ … where z_i is the only missing variable.

- Also, we have the posterior on z_i … and that posterior does not have any missing variable … so it's a complete distribution.
- This means we can get a probability value for $z_i = 1, z_i = 2,$ … $z_i = K$

- So we can compute the expectation of the log likelihood … where the expectation is computed over the posterior distribution of z_i , i.e., $q_i(z_i)$

$$E_{q_i}\left[log\ p(X, Z|\theta_{1:K})\right] = \sum_{i=1}^{n} \sum_{k=1}^{K} \ \ E_{q_i}\left[\delta_k(z_i)\{log\ p(x_i|\theta_k) + log\ \pi_k\}\right]$$

- Observe that $q_i(z_i)$ is essentially K probabilities values, one for each $z_i, i \in [1, K]$

- Hence, the expectation can be written as the expectation of only the δ functions, since that's the only one with z_i in it

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \ \ E_{q_i}\left[\delta_k(z_i)\right] \ \ \{log\ p(x_i|\theta_k) + log\ \pi_k\}$$

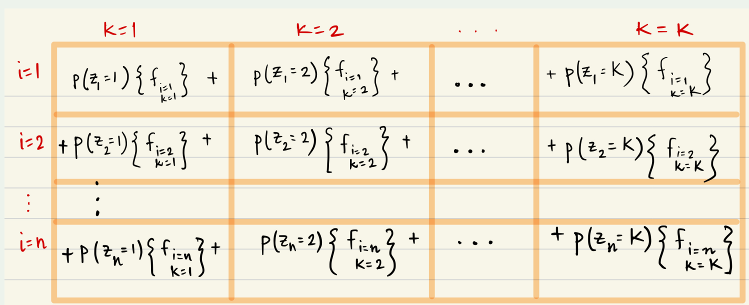
- Now, consider this term: $E_{q_i}\left[\delta_k(z_i)\right]$

- Inside the double summation, take the first term, $i = 1$ and $k = 1$. This first term's expectation can be written as:

$$E_{q_i}\left[\delta_k(z_i)\right] \ = \ p(z_1 = 1)\delta_1(z_1 = 1) + p(z_1 = 2)\delta_1(z_1 = 2) + \ldots \ + p(z_1 = K)\delta_1(z_1 = K)$$

$$E_{q_i}\left[\delta_k(z_i)\right] \ = \ p(z_1 = 1)1 + p(z_1 = 2)0 + \ldots \ + p(z_1 = K)0$$
$$E_{q_i}\left[\delta_k(z_i)\right] \ = \ p(z_1 = 1)$$

- See how the expectation for the first term reduced to only a single probability … since the delta functions zero-forced all other probabilities.
- Now, across all the terms of the double summation (which is a $n \times K$ matrix), we have the following summation:



- Now we can rewrite the expectation of the whole log likelihood function as:

$$E_{q_i}\left[log\ p(X, Z|\theta_{1:K})\right] = \sum_{i=1}^{n} \sum_{k=1}^{K} \ \ q_i(k) \ \ \{log\ p(x_i|\theta_k) + log\ \pi_k \ \ \}$$

🟢 The final step: let's maximize the expected likelihood function.

- Observe we need to maximize the first portion w.r.t. $\theta_k, \ k \in [1, K]$ … and the second portion w.r.t. π_k
- At the end, we want 2 vectors: $\hat{\theta}_{1:K}$ and prior probabilities, $\pi_{1:K}$

🟢 Let's optimize the first term:

- Note that θ_k only occurs in the k^{th} column of the matrix above, so when we optimize for θ_k, we sum over all data points $i = 1 : n$

$$\hat{\theta}_k = \arg\max_{\theta} \sum_{i=1}^{n} q_i(k)\ log\ p(x_i|\theta_k)$$

- This is a standard MLE problem, except that each log term is weighted with a probability value. But differentiation will solve this.

🟢 Let's now optimize the second term:

- We basically need to find $[\pi_1,\ \pi_2,\ ...\ \pi_K]$ that maximizes $\sum_{i=1}^{n} \sum_{k=1}^{K} q_i(k)\ \ log\ \pi_k$

- Note that we can group by column and write this as: $\sum_{i=1}^{n} q_i(1)\ log\ \pi_1 \ + \sum_{i=1}^{n} q_i(2)\ log\ \pi_2\ \ ...\ \ + \sum_{i=1}^{n} q_i(K)\ log\ \pi_K$

- And since all terms are negative (log of a probability is negative), we can optimize each term individually while satisfying the constraints that:

$$\sum_{k=1}^{K} \pi_k = 1 \quad \text{and} \quad \pi_k \geq 0,\ \ \forall k = 1, 2, ..., K$$

- So finding the optimal $\pi_{1:K}$ is equivalent to:

$$\hat{\pi}_{1:K} = \arg\max_{\pi} \sum_{i=1}^{n} q_i(k)\ log\ \pi_k \qquad \text{s.t. } \sum_{k=1}^{K} \pi_k = 1,\ \pi_k \geq 0,\ \ \forall k = 1, 2, ..., K$$

- This is a straight application of Lagrange multipliers → and we get the optimal π_k as the average of the $q_i(k)$ coefficients:

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} q_i(k)$$

▼ The final iteration

🟢 Putting everything together:

- E step: The posterior $q_i(z_i)$ is a function of a single data point x_i … all the $\theta_{1:K}$ calculated till now … and all the $\pi_{1:K}$ estimated till now

- M step: The expected log-likelihood needs to be calculated …
where the log-likelihood depends on all data samples $x_{1:n}$ … all the $\theta_{1:K}$ calculated till now … and all the $\pi_{1:K}$ estimated till now
and the expectation over the posterior also needs $q_i(z_i)$

- So, start the E^0 step with an initial guess on $[\theta^0_{1:K},\ \pi^0_{1:K}]$ … and compute $q^0_i(z_i)$

- Then, compute $[\theta^1_{1:K},\ \pi^1_{1:K}]$ using $q^0_i(z_i)$ and $[\theta^0_{1:K},\ \pi^0_{1:K}]$

- Then iterate as: $[\theta^1_{1:K},\ \pi^1_{1:K}] \ \rightarrow\ q^1_i \ \rightarrow\ [\theta^2_{1:K},\ \pi^2_{1:K}] \ \rightarrow\ q^2_i\ \ ...\ \ [\theta^t_{1:K},\ \pi^t_{1:K}] \ \rightarrow\ q^{t+1}_i$

- Terminate when $|\theta^t - \theta^{t-1}| \leq \epsilon$

Tutorial by Dahua Lin (MIT): here