Sharif University of Technology

Masoud Tahmasbi Fard

Student ID: 402200275

CE957: Reinforcement Learning

Assignment #1

March 13, 2024

Table of Contents

1 Information Theory

1.1 Mutual Information

1.1.1

Suppose a Bernoulli random variables, X and two random variables Y and Z such that:

$$X = Y = Z$$

Therefore:

Conditional entropy, given any of aforementioned random variables, equals to zero.

$I(X;Y) = H(X) - H(X|Y) = H(X) = 1$

$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = 0$

$\implies \boxed{I(X;Y|Z) < I(X;Y)}$

1.1.2

Let's simplify the inequality:

$$I(X;Y|Z) = I(X;Y) - (I(X;Z) - I(X;Z|Y)) \Rightarrow$$

The term $I(X;Y|Z) - I(X;Y)$ is called Interaction Information. Therefore, for $I(X;Y|Z) > I(X;Y)$ the following inequality should be held (negative interaction information):

$$I(X;Z) < I(X;Z|Y)$$

Suppose three pairwise independent random variables, X, Y, Z and Z, such that:

$$X \sim Bernoulli(0.5)$$

$$Z \sim Bernoulli(0.5)$$

$$Y = \begin{cases} X & \text{if } Z = 0 \\ 1 - X & \text{if } Z = 0 \end{cases}$$

Then:

$$\begin{aligned}
&I(X;Z) = H(X) - H(X|Z) \Rightarrow \\
&H(X) = -\sum_x p(x)\log p(x) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1 \\
&H(X|Z) = -\sum\sum p(x,z)\log p(x|z) \overset{indep.}{=} -\sum\sum p(x)p(z)\log p(x) \\
&\qquad\quad = -\sum p(z)\sum p(x)\log p(x) = H(X) = 1 \Longrightarrow \\
&\boxed{I(X;Z) = 1 - 1 = 0}
\end{aligned}$$

$I(X;Z|Y) = H(X|Y) - H(X|Z,Y)$

Given Z and Y, X can be determined uniquely, therefore: $H(X|Z,Y) = 0$

$H(X|Y) = -\sum_y p(y)H(X|Y=y)$

$= p(Y=0)H(X|Y=0) + p(Y=1)H(X|Y=1)$

$p(Y=1) = p(Y=1|Z=1)p(Z=1) + p(Y=1|Z=0)p(Z=0) = \frac{1}{2}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{2} = \frac{1}{2} = p(Y=0) \Rightarrow$

$H(X|Y) = \frac{1}{2}(H(X|Y=0) + H(X|Y=1))$

$H(X|Y=0) = -\sum p(X=x|Y=0)\log p(X=x|Y=0) = 1 = H(X|Y=0)$

$\Longrightarrow \boxed{I(X;Z|Y) = 1}$

So we proved that in this setting, $I(X;Z) < I(X;Z|Y)$ which leads to: $\boxed{I(X;Y|Z) > I(X;Y)}$

1.2 Some Inequalities!

1.2.1

We shall prove that:

$$H(X,Y,Z) - H(X,Y) \le H(X,Z) - H(X)$$

Let's use the chain rule:

$$H(X,Y,Z) = H(X,Y) + H(Z|X,Y) \Rightarrow H(X,Y,Z) - H(X,Y) = H(Z|X,Y) \tag{1}$$

Again, applying the chain rule gives:

$$H(X,Z) = H(X) + H(Z|X) \Rightarrow H(X,Z) - H(X) = H(Z|X) \leq H(Z|X,Y) \tag{2}$$

(Note: Conditioning doesn't decrease entropy!)

Merging Eq.1 and Eq.2:

$$H(X,Y,Z) – H(X,Y) \leq H(X,Z) – H(X)$$

Equality happens when Z and Y are conditionally independent given X.

1.2.2

We shall prove that:

$$I(X;Z|Y) I(Z;Y|X) – I(Z;Y) + I(X;Z)$$

This time we apply the chain rule for mutual information:

$$I(X,Y;Z) = I(Z;X) + I(Y;Z|X)$$

$$I(X,Y;Z) = I(Z;Y) + I(X;Z|Y)$$

$$\Rightarrow I(Z;Y) + I(X;Z|Y) = I(Z;X) + I(Y;Z|X) \Longrightarrow I(X;Z|Y) = I(Z;Y|X) – I(Z;Y) + I(X;Z)$$

Which means that the given inequality is actually an equality!

1.3 Treasure!

1.3.1

$$I(Z;Y) = H(Z) - H(Z|Y)$$

$$\Rightarrow H(Z) = -\sum p(Z)\log p(Z) = -4 \times \frac{1}{4} \times \log\frac{1}{4} = 2 \Rightarrow \boxed{H(Z) = 2}$$

$$\Rightarrow H(Z|Y) = \sum_{y} H(Z|Y=y)p(Y=y)$$

We suppose that Y takes values of $\{1,2\}$ with equal probabilities.

$$\Rightarrow H(Z|Y) = \frac{1}{2}(H(Z|Y=1) + H(Z|Y=2))$$

$$\Rightarrow H(Z|Y=1) = -\sum_{z} p(Z=z|Y=1)\log p(Z=z|Y=1)$$

$$= -p(Z=A|Y=1)\log p(Z=A|Y=1) - p(Z=B|Y=1)\log p(Z=B|Y=1)$$

$$- p(Z=C|Y=1)\log p(Z=C|Y=1) - p(Z=D|Y=1)\log p(Z=D|Y=1)$$

For the sake of simplicity, we define $\alpha \coloneqq \frac{1}{1+e^2}$

$$H(Z|Y=1) = -\alpha\log\alpha - (\frac{1-\alpha}{2})\log(\frac{1-\alpha}{2}) - (\frac{1-\alpha}{6})\log(\frac{1-\alpha}{6}) - (\frac{1-\alpha}{3})\log(\frac{1-\alpha}{3})$$

$$= 1.81$$

Similarly, for $H(Z|Y=2)$, we define $\alpha \coloneqq \frac{1}{1+e^{1.5}}$

$$H(Z|Y=2) = -\beta\log\beta - (\frac{1-\beta}{2})\log(\frac{1-\beta}{2}) - (\frac{1-\beta}{6})\log(\frac{1-\beta}{6}) - (\frac{1-\beta}{3})\log(\frac{1-\beta}{3})$$

$$= 1.88$$

$$\Longrightarrow I(Z;Y) = 2 - 0.5 \times 1.81 - 0.5 \times 1.88 = \boxed{0.155}$$

1.3.2

$I(Y;Z)$ represents the mutual information between Y and Z.

Mutual information is a measure of the amount of information that one random variable contains about another random variable. It quantifies the reduction in uncertainty about one random variable due to the knowledge of the other random variable.

In the MDP setting, $I(Y;Z)$ measures the amount of information that the optimal action Z provides about the reward Y, or vice versa. In other words, it quantifies how much knowing the optimal action reduces the uncertainty about the reward, or how much knowing the reward reduces the uncertainty about the optimal action.

A high value of $I(Y;Z)$ means that the optimal action Z and the reward Y are highly dependent or strongly correlated. Knowing the optimal action provides significant information about the reward, and vice versa. This implies that the optimal policy is well-aligned with maximizing the expected reward.

On the other hand, a low value of $I(Y;Z)$ means that the optimal action Z and the reward Y are relatively independent or weakly correlated. Knowing the optimal action does not provide much information about the reward, and vice versa. This could suggest that the optimal policy is not well-aligned with maximizing the expected reward, or that there are other factors influencing the reward that are not captured by the optimal action alone.

In the context of MDPs, a high mutual information between the optimal action and the reward is generally desirable, as it indicates that the optimal policy is effective in maximizing the expected reward. However, it's important to note that mutual information alone does not fully characterize the relationship between actions and rewards, and other factors, such as the specific reward structure and the complexity of the MDP, should also be considered.

1.4 KL Divergence

$$\begin{aligned}
KL(p(x,y,z)||p(x)p(y)p(z)) &= \iiint p(x,y,z)\log\frac{p(x,y,z)}{p(x)p(y)p(z)}\,dx\,dy\,dz \\
&= \iiint p(x,y,z)\log p(x,y,z)\,dx\,dy\,dz - \iiint p(x,y,z)\log p(x)\,dx\,dy\,dz \\
&\quad - \iiint p(x,y,z)\log p(y)\,dx\,dy\,dz - \iiint p(x,y,z)\log p(z)\,dx\,dy\,dz \\
&= -H(X,Y,Z) - \iint p(x,y,z)dy\,dz\int\log p(x)dx \\
&\quad - \iint p(x,y,z)dx\,dz\int\log p(y)dy - \iint p(x,y,z)dx\,dy\int\log p(z)dz \\
&= -H(X,Y,Z) + H(X) + H(Y) + H(Z) \qquad \square
\end{aligned}$$

1.5 Markov Inequality!

$$\begin{aligned}
&I\left(X_1;X_4\right)+I\left(X_2;X_3\right)-I\left(X_1;X_3\right)-I\left(X_2;X_4\right) \\
&\quad= \quad H\left(X_1\right)-H\left(X_1\mid X_4\right)+H\left(X_2\right)-H\left(X_2\mid X_3\right)-\left(H\left(X_1\right)-H\left(X_1\mid X_3\right)\right) \\
&\qquad\quad -\left(H\left(X_2\right)-H\left(X_2\mid X_4\right)\right) \\
&\quad= \quad H\left(X_1\mid X_3\right)-H\left(X_1\mid X_4\right)+H\left(X_2\mid X_4\right)-H\left(X_2\mid X_3\right) \\
&\quad= \quad H\left(X_1,X_2\mid X_3\right)-H\left(X_2\mid X_1,X_3\right)-H\left(X_1,X_2\mid X_4\right)+H\left(X_2\mid X_1,X_4\right) \\
&\qquad +H\left(X_1,X_2\mid X_4\right)-H\left(X_1\mid X_2,X_4\right)-H\left(X_1,X_2\mid X_3\right)+H\left(X_1\mid X_2,X_3\right) \\
&\quad= -H\left(X_2\mid X_1,X_3\right)+H\left(X_2\mid X_1,X_4\right)-H\left(X_2\mid X_1,X_4\right)+H\left(X_2\mid X_1,X_3,X_4\right) \\
&\quad= \quad I\left(X_2;X_3\mid X_1,X_4\right) \\
&\quad\geq 0
\end{aligned}$$

2 Optimization

2.1 Critical Points

If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and differentiable function, then for a point $x \in \mathbb{R}^n, f(x) \leq f(y) \forall y \in \mathbb{R}^n$ iff $\nabla f(x) = 0$. In other words, a point $x \in \mathbb{R}^n$ is a global minimizer of a convex differrentiable function f if and only if gradient of f at x is 0 or x is a critical point of f.

To prove this, let x be a the global minimizer of f. Then $\forall v \in \mathbb{R}^n$ and $\forall t \in \mathbb{R}$, we have $f(x + tv) \geq f(x)$, we have,

$$0 \leq \frac{f(x + tv) - f(x)}{t}$$

Taking the following limit on both sides we get

$$0 \leq \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t} = \langle \nabla f(x), v \rangle$$

This shows that the inner product of $\nabla f(x)$ and v is non negative $\forall v \in \mathbb{R}^n$. Choosing $v = -\nabla f(x)$, we get,

$$0 \leq \langle \nabla f(x), -\nabla f(x) \rangle = -\|\nabla f(x)\|^2$$

This implies that $\nabla f(x) = 0$, because the negative of the square norm of a quantity is 0 only when that quantity is 0 . This completes the first direction of the proof. Now, for the second direction, assume $\nabla f(x) = 0$. Since f is convex, we can say that $\forall y \in \mathbb{R}^n$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle = f(x) (\because \nabla f(x) = 0)$$

which completes the other direction of the proof.

2.2 Max and Min!

$$\mathcal{L} = f(x, y, z) + \lambda \left(z^2 - x^2 - y^2\right) + \mu (z - x - y - 1) = 0$$

$$\Rightarrow \begin{cases} \frac{\partial \mathcal{L}}{\partial x} = 0 \rightarrow 2x - 2\lambda x - \mu = 0 \\ \frac{\partial \mathcal{L}}{\partial y} = 0 \rightarrow 2y - 2\lambda y + \mu = 0 \\ \frac{\partial \mathcal{L}}{\partial z} = 0 \rightarrow 2z + 2\lambda z + \mu = 0 \\ z^2 - x^2 - y^2 = 0 \\ z - x - y - 1 = 0 \end{cases}$$

$$\Rightarrow x = y; z = 2x + 1 \Rightarrow (2y+1)^2 - y^2 - y^2 = 0$$

$$\Rightarrow 2y^2 + 4y + 1 = 0 \Rightarrow y = \frac{-2 \pm \sqrt{2}}{2}$$

$$(x, y, z) = (-1 + \frac{\sqrt{2}}{2}, -1 + \frac{\sqrt{2}}{2}, \sqrt{2} - 1)$$

$$(x, y, z) = (-1 - \frac{\sqrt{2}}{2}, -1 + -\frac{\sqrt{2}}{2}, -\sqrt{2} - 1)$$

$$\Longrightarrow \boxed{\max f(x, y, z) = 11.66}, \boxed{\min f(x, y, z) = 0.34}$$

2.3 No Inequality!

Let $\mu \in \mathbb{R}^m, \lambda \in \mathbb{R}_+^k = [0, \infty)^k$, and define $L(x, \mu, \lambda) := f(x) + \mu^T g(x) + \lambda^T h(x)$ and $L(\mu, \lambda) = \inf_x \left\{ f(x) + \mu^T g(x) + \lambda^T h(x) \right\} = \inf_x \left\{ L(x, \mu, \lambda) \right\}$.

Let $L(\mu, \lambda) := \min_{x \in \mathbb{R}^n} L(x, \mu, \lambda)$ denote the unconstrained optimization over x for a fixed μ, λ. denote the optimal solution by x^*. We will show that $L(\mu, \lambda) \leq f\left(x^*\right)$ for any $\mu \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}_+^k$ where $f(x^*) = \min_x f(x)$.

Let $x(\mu, \lambda) := \arg\min_{x \in \mathbb{R}^n} L(x, \mu, \lambda)$, and then $L(\mu, \lambda) = L(x(\mu, \lambda), \mu, \lambda) \leq L\left(x^*, \mu, \lambda\right)$ since $x(\mu, \lambda)$ is the minimizer of $L(\cdot, \mu, \lambda)$ given (μ, λ). Since x^* is feasible, we have $g(x^*) = 0$ and $h_i\left(x^*\right) \leq 0$ for $i = 1, \ldots, k$. We also have $\lambda \in \mathbb{R}_+^k$. Hence

$$\begin{aligned} L(\mu, \lambda) &\leq L\left(x^*, \mu, \lambda\right) \\ &= f\left(x^*\right) + \mu^T \cdot 0 + \lambda^T h\left(x^*\right) \\ &\leq f\left(x^*\right). \end{aligned}$$

2.4 Minimization!

Introducing Lagrange multipliers $\lambda^\star \in \mathbb{R}^n$ for the inequality constraints $x^\star \succeq 0$, and a multiplier $\nu^\star \in \mathbb{R}$ for the equality constraint $\mathbf{1}^T x = 1$, we obtain the KKT conditions

$$x^\star \succeq 0, \quad \mathbf{1}^T x^\star = 1, \quad \lambda^\star \succeq 0, \quad \lambda_i^\star x_i^\star = 0, \quad i = 1, \ldots, n,$$

$$-1/\left(\alpha_i + x_i^\star\right) - \lambda_i^\star + \nu^\star = 0, \quad i = 1, \ldots, n.$$

We can directly solve these equations to find $x^\star, \lambda^\star$, and $\nu^\star$. We start by noting that $\lambda^\star$ acts as a slack variable in the last equation, so it can be eliminated, leaving

$$x^\star \succeq 0, \quad \mathbf{1}^T x^\star = 1, \quad x_i^\star \left(\nu^\star - 1/\left(\alpha_i + x_i^\star\right)\right) = 0, \quad i = 1, \ldots, n,$$

$$\nu^\star \geq 1/\left(\alpha_i + x_i^\star\right), \quad i = 1, \ldots, n.$$

If $\nu^\star < 1/\alpha_i$, this last condition can only hold if $x_i^\star > 0$, which by the third condition implies that $\nu^\star = 1/\left(\alpha_i + x_i^\star\right)$. Solving for $x_i^\star$, we conclude that $x_i^\star = 1/\nu^\star - \alpha_i$ if $\nu^\star < 1/\alpha_i$. If $\nu^\star \geq 1/\alpha_i$, then $x_i^\star > 0$ is impossible, because it would imply $\nu^\star \geq 1/\alpha_i > 1/\left(\alpha_i + x_i^\star\right)$, which violates the complementary slackness condition. Therefore, $x_i^\star = 0$ if $\nu^\star \geq 1/\alpha_i$. Thus we have

$$x_i^\star = \begin{cases} 1/\nu^\star - \alpha_i & \nu^\star < 1/\alpha_i \\ 0 & \nu^\star \geq 1/\alpha_i, \end{cases}$$

or, put more simply, $x_i^\star = \max\left\{0, 1/\nu^\star - \alpha_i\right\}$. Substituting this expression for $x_i^\star$ into the condition $\mathbf{1}^T x^\star = 1$ we obtain

$$\sum_{i=1}^n \max\left\{0, 1/\nu^\star - \alpha_i\right\} = 1.$$

The lefthand side is a piecewise-linear increasing function of $1/\nu^\star$, with breakpoints at α_i, so the equation has a unique solution which is readily determined.

2.5 Maximum Entropy

$$\arg\max_{\mathbb{P}}\{-\int \mathbb{P}(x) \log \mathbb{P}(x)\, dx\};\ \mathbb{E}_{\mathbb{P}}\{X\} = \mu,\ \mathbb{V}_{\mathbb{P}}\{X\} = \sigma^2$$

Recall that:

$$\mathbb{V}_{\mathbb{P}}\{X\} = \mathbb{E}\{(X-\mu)^2\}$$

So we can employ a single Lagrangian multiplier since the mean value constraint is incorporated into the variance constraint. Furthermore, for a final solution in a comparable form, and without losing generality, we define the Lagrangian multiplier as $\frac{1}{2\lambda}$ instead of λ.

$$\begin{aligned}
&\mathcal{L} = -\int \mathbb{P}(x)\log\mathbb{P}(x)\,dx + \frac{1}{2\lambda}\left(\sigma^2 - \int (x-\mu)^2\mathbb{P}(x)\,dx\right) \\
\rightarrow &\frac{\partial\mathcal{L}}{\partial\mathbb{P}(x)} = 0 \\
\rightarrow &-\log\mathbb{P}(x) - 1 - \frac{(x-\mu)^2}{2\lambda} = 0 \rightarrow \mathbb{P}(x) = Ke^{-\frac{(x-\mu)^2}{2\lambda}}
\end{aligned}$$

where K is a normalization constant.

The solution takes the form of a Gaussian distribution probability density function (pdf). As we aim for $\mathbb{P}(x)$ to represent a probability distribution, we rely on the properties of the Gaussian distribution:

$$K = \frac{1}{\sqrt{2\pi\sigma^2}},\ \lambda = \sigma^2 \Rightarrow \boxed{\mathbb{P}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

3 Markov Chain

3.1 Uniform Stationary Distribution

Consider a finite, irreducible, and ergodic Markov chain with transition matrix P. If there are non-negative numbers $\pi = (\pi_0, \ldots, \pi_n)$ such that $\sum_{i=0}^{n} \pi_i = 1$ and if, for any pair of states i, j,

$$\pi_i P_{i,j} = \pi_j P_{j,i},$$

then π is the stationary distribution corresponding to P. This property is called time-reversibility. To prove this, consider the jth entry of πP. Using the assumption, we find that it equals

$$\sum_{i=0}^{n} \pi_i P_{i,j} = \sum_{i=0}^{n} \pi_j P_{j,i} = \pi_j.$$

Thus π satisfies $\pi = \pi P$. Since $\sum_{i=0}^{n} \pi_i = 1$, π must be the unique stationary distribution of the Markov chain.

Considering:

$$P_{x,y} = \begin{cases} \frac{1}{M} & \text{if } y \in N(x) \\ 0 & \text{if } y \neq x \text{ and } y \notin N(x) \\ 1 - \frac{|N(x)|}{M} & \text{if } y = x. \end{cases}$$

We want to prove that if the Markov chain constructed above is irreducible and aperiodic, then it has the uniform distribution as its unique stationary distribution.

For all $x \neq y$ we have either $P_{x,y} = P_{y,x} = 0$ or $P_{x,y} = P_{y,x} = \frac{1}{M}$, so the uniform distribution π satisfies

$$\pi_x P_{x,y} = \pi_y P_{y,x} \quad \text{for all } x, y.$$

It follows that the uniform distribution $\pi_x = \frac{1}{|\Omega|}$ is the stationary distribution.

3.2 What's Wrong?

We will argue that this approach will work poorly when each a_i is 1 and $b = \sqrt{n}$. Actually, the problem is that the solution set may not be **dense** enough in the sample space. Hence the uniform distribution over the whole sample space might not put enough **weight** on the solution set, and we may need to wait a long time before obtaining a good enough estimate of the number of solutions.

The expected number of 1's chosen this way is $n/2$. If $a_i = 1$ and $b = \sqrt{n}$, then by the Chernoff bound:

$$\Pr\left(\sum a_i X_i \leq \sqrt{n}\right) = \Pr\left(\sum a_i X_i \leq \left(1 - \frac{2}{\sqrt{n}}\right)\frac{n}{2}\right) \leq \exp\left(-\frac{n}{2}\left(\sqrt{n} - 2\sqrt{n}\right)^2/2\right) = O\left(e^{-n}\right)$$

3.3 Again, Markov Chain!

3.3.1

We first show that the chain is irreducible and aperiodic over the states of all valid solutions of the knapsack problem. The chain is therefore ergodic and has a unique stationary distribution. We then demonstrate that the uniform distribution satisfies the stationarity criterion.

Irreducibility: from any solution vector x, there is a positive probability of going back down to the all zero vector by zeroing out the non-zero x_is one by one. On the other hand, if x is a solution, then it is possible to reach x starting from the all zero vector. Hence the chain is irreducible.

Aperiodicity: Suppose $\sum_{i=1}^{n} a_i > b$, then there must exist $j \in [1, n]$ and a vector $x = (x_1, \ldots, x_j = 0, \ldots, x_n) \in \{0, 1\}^n$ such that $\sum_i a_i x_i < b$ but $\sum_i a_i x_i + a_j > b$. This means that in the Markov chain, the self-loop probability $P_{x,x} > 0$. Hence the chain is aperiodic.

Let M be the number of solutions and let x and y be two solutions that differ by one bit. Then $P_{x,y} = 1/n$ and $P_{y,x} = 1/n$. If $\pi_x = \pi_y = 1/M$, then $\pi_x P_{x,y} = \frac{1}{Mn} = \pi_y P_{y,x}$. This satisfies the time-reversibility condition and proves that the stationary distribution is uniform.

3.3.2

Here's a brief explanation of how this Markov chain can be used to estimate the number of valid solutions (M) for the knapsack problem:

1. Start with a random valid solution vector x.
2. Run the Markov chain for a large number of steps, allowing it to explore the state space of valid solutions.
3. At each step, check if the current state (solution vector) has been observed before.
4. Let N be the total number of distinct valid solution vectors observed over the course of the Markov chain simulation.
5. N can be used as an estimate for M, the total number of valid solutions to the knapsack problem.

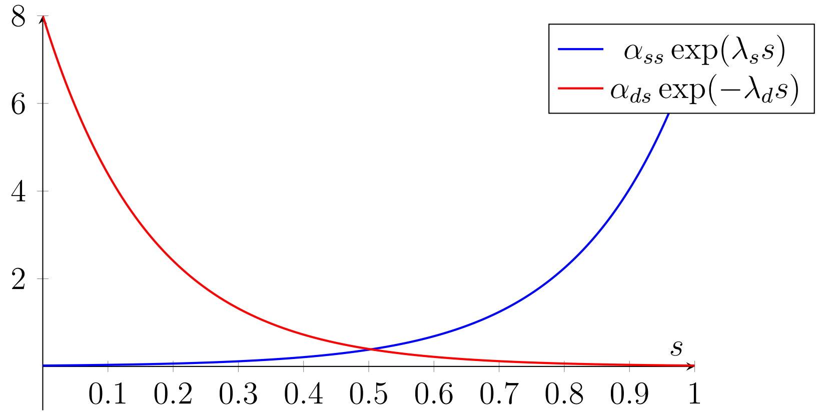
The key idea is that the Markov chain will eventually visit all valid solution vectors if run for a sufficiently long time, due to its irreducibility over the subspace of valid solutions (as shown in our proof). By counting the number of distinct valid solutions encountered, we can estimate the total number of valid solutions M.

However, this approach does not provide any theoretical guarantees on the accuracy of the estimate N, as it depends on the length of the simulation and the specific sequence of states explored by the Markov chain. Nevertheless, it can serve as a practical method to obtain an approximate estimate of M, particularly when an exact counting approach is computationally infeasible for large problem instances.

4 Bayesian Statistics

4.1 Everything Fine?

Let's draw the score distributions and provide an intuition for why these PDFs might be reasonable.



The PDFs are reasonable because if the persons are identical you get high probability densities for high scores and if the persons are identical you get high probability densities for low scores. $\alpha_{s,d}$ are normalization constants and compulsory to define the probability distributions. The normalization constant can be derived from the condition that the probability density function integrated over the whole range should be 1 ($\int_{-\infty}^{+\infty} f_s(s)ds = 1$).

4.2 Face Recongnition

This can be solved directly with Bayes' theorem.

$$\begin{aligned}&P\left(\text{same}_j, \text{different}_{i\neq j} \mid s_1, \ldots, s_N\right)\\ =&\frac{p\left(s_1, \ldots, s_N \mid \text{same}_j, \text{different}_{i\neq j}\right) P\left(\text{same}_j, \text{different}_{i\neq j}\right)}{p\left(s_1, \ldots, s_N\right)}\\ &\text{(since the scores are independent of each other and the}\\ &\text{a priory probability is the same for all points)} \rightarrow\\ =&\frac{p\left(s_j \mid \text{same}\right)\left(\prod_{i\neq j} p\left(s_i \mid \text{different}\right)\right)(1/N)}{p\left(s_1, \ldots, s_N\right)}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\alpha_{ss}\exp\left(\lambda_s s_j\right)\left(\prod_{i\neq j}\alpha_{ds}\exp\left(-\lambda_d s_i\right)\right)(1/N)}{p\left(s_1,\ldots,s_N\right)}\\
&\text{Consider } S := \sum s_i\\
&= \frac{\alpha_{ss}\alpha_{ds}^{N-1}\exp\left(\lambda_s s_j\right)\exp\left(\lambda_d\left(s_j - S\right)\right)(1/N)}{p\left(s_1,\ldots,s_N\right)}\\
&= \frac{\alpha_{ss}\alpha_{ds}^{N-1}\exp\left(\left(\lambda_s+\lambda_d\right)s_j\right)\exp\left(-\lambda_d S\right)(1/N)}{p\left(s_1,\ldots,s_N\right)}\\
&= \frac{\alpha_{ss}\alpha_{ds}^{N-1}\exp\left(\left(\lambda_s+\lambda_d\right)s_j\right)\exp\left(-\lambda_d S\right)(1/N)}{\sum_{j'}\alpha_{ss}\alpha_{ds}^{N-1}\exp\left(\left(\lambda_s+\lambda_d\right)s_{j'}\right)\exp\left(-\lambda_d S\right)(1/N)}\\
&= \frac{\exp\left(\left(\lambda_s+\lambda_d\right)s_j\right)}{\sum_{j'}\exp\left(\left(\lambda_s+\lambda_d\right)s_{j'}\right)}(\text{since } P \text{ must be normalized to 1}).
\end{aligned}
$$

This is a neat formula. It is instructive to calculate the ratio between the probability that j is the correct image and that k is the correct image, which is

$$
\begin{aligned}
\frac{P\left(\text{same}_j, \text{different}_{i\neq j} \mid s_1,\ldots,s_N\right)}{P\left(\text{same}_k, \text{different}_{i\neq k} \mid s_1,\ldots,s_N\right)} &\overset{(21)}{=} \frac{\exp\left(\left(\lambda_s+\lambda_d\right)s_j\right)}{\exp\left(\left(\lambda_s+\lambda_d\right)s_k\right)}\\
&= \exp\left(\left(\lambda_s+\lambda_d\right)\left(s_j - s_k\right)\right).
\end{aligned}
$$

We see that the ratio only depends on the difference between the score values but not on the values themselves.

4.3 Just the Highest!

The probability of correct recognition is the probability density that the correct picture gets a certain score s', i.e. $p\left(s' \mid \text{same}\right)$, times the probability that all the other gallery pictures get score below s', i.e. $P\left(s < s' \mid \text{different}\right)^{(N-1)}$, integrated over all possible scores s'. Note that the integration turns the probability density $p\left(s' \mid \text{same}\right)$ into a proper probability.

$$
\begin{aligned}
&P(\text{correct recognition})\\
&= \int_0^1 p\left(s' \mid \text{same}\right) P\left(s < s' \mid \text{different}\right)^{(N-1)} \mathrm{d}s'\\
&= \int_0^1 \alpha_{ss}\exp\left(\lambda_s s'\right)\left(\frac{\exp\left(-\lambda_d s'\right)-1}{\exp\left(-\lambda_d\right)-1}\right)^{(N-1)} \mathrm{d}s'\\
&= \int_0^1 \frac{\lambda_s\exp\left(\lambda_s s'\right)}{\exp\left(\lambda_s\right)-1}\left(\frac{\exp\left(-\lambda_d s'\right)-1}{\exp\left(-\lambda_d\right)-1}\right)^{(N-1)} \mathrm{d}s'
\end{aligned}
$$

$$= \underbrace{\left(\exp\left(\lambda_s\right) - 1\right)\left(\exp\left(-\lambda_d\right) - 1\right)^{(N-1)}}_{=:A} \int_0^1 \exp\left(\lambda_s s'\right)\left(\exp\left(-\lambda_d s'\right) - 1\right)^{(N-1)} \mathrm{d}s'$$

$= \ldots$ (one could simplify even further)

5 Estimation Theory

5.1 MLE for a Gaussian Distribution

$$\text{argmax}_\theta \mathbb{P}(X_1, X_2, \ldots, X_n|\theta) \overset{i.i.d}{=} \text{argmax}_\theta \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \text{argmax}_\theta \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}) = \mathcal{L}$$

$$\Rightarrow \log\mathcal{L} = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\Rightarrow \frac{\partial \log\mathcal{L}}{\partial\mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-\mu) = 0 \rightarrow \boxed{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i}$$

$$\Rightarrow \frac{\partial \log\mathcal{L}}{\partial\sigma^2} = -\frac{n}{2}\times\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \overset{\times 2\sigma^4}{\rightarrow} -n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\Rightarrow \boxed{\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$$

The obtained estimator is unbiased since:

$$\mathbb{E}\{\hat{\mu}\} = \mathbb{E}\{\frac{1}{n}\sum_{i=1}^{n}x_i\} \underset{i.i.d}{=} \frac{1}{n}\times n\times\mathbb{E}\{X_i\} = \mu$$

5.2 Function of Estimated Parameters

Using Induced Likelihood Function:

$$L^*(\hat{\eta} \mid \mathbf{x}) = \sup_\eta L^*(\eta \mid \mathbf{x}) = \sup_\eta \sup_{\theta\in\tau^{-1}(\eta)} L(\theta \mid \mathbf{x})$$

$$= \sup_\theta L(\theta \mid \mathbf{x}) = L(\hat{\theta} \mid \mathbf{x})$$

$$L(\hat{\theta} \mid \mathbf{x}) = \sup_{\theta\in\tau^{-1}(\tau(\hat{\theta}))} L(\theta \mid \mathbf{x}) = L^*[\tau(\hat{\theta}) \mid \mathbf{x}]$$

Hence, $L^*(\hat{\eta} \mid \mathbf{x}) = L^*[\tau(\hat{\theta}) \mid \mathbf{x}]$ and $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.

Using the mentioned theorem:

$$\tau(\theta) = \frac{1}{1+\mu^2} \Rightarrow \mu^*_{new} = \frac{1}{1+(\mu^*_{old})^2} = \frac{1}{1+(\frac{1}{n}\sum_{i=1}^{n}x_i)^2}$$

$$\Rightarrow \sigma^{2^*}_{new} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu^*_{new})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{1}{1+(\frac{1}{n}\sum_{i=1}^{n}x_i)^2})^2$$

5.3 Importance-Weighted Sampling

5.3.1

To show the IS estimator is an unbiased estimator, we shall prove that:

$$\mathbb{E}_q\{\hat{\mu}_{IS}\} = \mu$$

Therefore:

$$\Rightarrow \mathbb{E}_q\{\hat{\mu}_{IS}\} = \int q(x)\frac{f(x)p(x)}{q(x)}dx = \int f(x)p(x)dx = \mathbb{E}_p\{f(X)\} = \mu$$

5.3.2

$$\begin{aligned} Var(\hat{\mu}_{IS}) &= \mathbb{E}_q\{\hat{\mu}_{IS}^2\} - \mu^2 \\ &= \mathbb{E}_q\Big\{\int \frac{f^2(x)p^2(x)}{q^2(x)}dx\Big\} - \mu^2 = \int \frac{f^2(x)p^2(x)}{q(x)}dx - \mu^2 \end{aligned} \tag{3}$$

Remember that the minimum value of variance is zero. With careful consideration of the equation, it becomes evident that by selecting $q(x) = \frac{f(x)p(x)}{\mu}$, the variance would be reduced to zero.

$$q(x) = \frac{f(x)p(x)}{\mu} \Rightarrow Var(\hat{\mu}_{IS}) = \mu \underbrace{\int f(x)p(x)dx}_{=1} - \mu^2 = 0$$

Referring to Eq. 3, it is evident that $p(x)$ and $q(x)$ can be selected in a manner that drives the variance towards infinity. For instance, one can choose $q(x) = cf^2(x)p^2(x)$ (where c is a normalization constant, ensuring the integral of $q(x)$ is zero). With this choice, the integral in Eq. 3 will diverge, leading to infinite variance. Another example involves selecting $q(x) = cp^2(x)$ and $f(x)$ as an increasing function. Once again, the integral will diverge, resulting in infinite variance.

5.3.3

$$\hat{\mu}_{IS} = \frac{\sum_{i=1}^{n} w(x_i)f(x_i)}{\sum_{i=1}^{n} w(x_i)} = \frac{\frac{1}{n}\sum_{i=1}^{n} w(x_i)f(x_i)}{\frac{1}{n}\sum_{i=1}^{n} w(x_i)} = \frac{\bar{Z}}{\bar{W}}$$

Where $Z(x) = w(x)f(x)$.

Note that $\mathbb{E}_q\{w(X)\} = 1$, since:

$$\mathbb{E}_q\{w(X)\} = \int q(x)\frac{p(x)}{q(x)}dx = \int p(x)dx = 1$$

Using Taylor expansion around $\bar{W} = 1$:

$$\begin{aligned}\mathbb{E}_q\{\hat{\mu}_{IS}\} &= \mathbb{E}_q\{\bar{Z}[1 - (\bar{W} - 1) + (\bar{W} - 1)^2 + \cdots]\} \\ &\approx \mu + \frac{\mu \operatorname{Var}_q(w(X)) - \operatorname{Cov}_q(w(X), w(X) \cdot h(X))}{n}\end{aligned} \tag{4}$$

As observed in Eq.4, $\mathbb{E}_q\{\hat{\mu}_{IS}\}$ may not necessarily equal μ, indicating bias in the estimator. However, if n approaches infinity and $\mu \operatorname{Var}_q(w(X)) - \operatorname{Cov}_q(w(X), w(X) \cdot h(X)) < \infty$, then $\mathbb{E}_q\{\hat{\mu}_{IS}\}$ tends toward μ, rendering the estimator unbiased.

5.3.4

$$\hat{\mu}_{IS} = \frac{\sum_{i=1}^{n} w(x_i)f(x_i)}{\sum_{i=1}^{n} w(x_i)} = \sum_{i=1}^{n} \alpha_i f(x_i);\ \sum_{i=1}^{n} \alpha_i = 1,\ 0 \le \alpha_i \le 1$$

Since $\sum_{i=1}^{n} \alpha_i = 1,\ 0 \le \alpha_i \le 1$, it's obvious that:

$$\sup \hat{\mu}_{IS} \le \sup f(x) = M$$

$$\inf \hat{\mu}_{IS} \le \inf f(x) = m$$

Therefore $\Pr(m \le \hat{\mu}_{IS} \le M) = 1$. Then, since $m \le \hat{\mu}_{IS} \le M$,

$$0 \le \mathbb{E}[(M - \hat{\mu}_{IS})(\hat{\mu}_{IS} - m)] = -\mathbb{E}\left[\hat{\mu}_{IS}^2\right] - mM + (m + M)\mu.$$

Thus,

$$\sigma^2 = \mathbb{E}\left[\hat{\mu}_{IS}^2\right] - \mu^2 \le -mM + (m + M)\mu - \mu^2 = (M - \mu)(\mu - m).$$

Now, applying the Inequality of arithmetic and geometric means, $ab \le \left(\frac{a+b}{2}\right)^2$, with $a = M - \mu$ and $b = \mu - m$, yields the desired result:

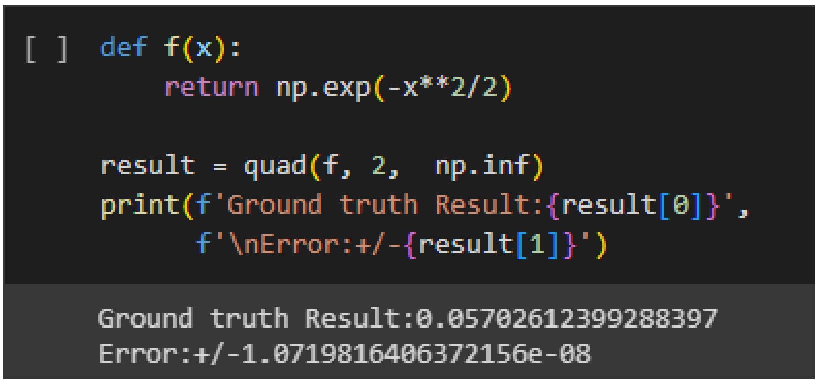
$$\sigma^2 \le (M - \mu)(\mu - m) \le \frac{(M - m)^2}{4}$$

5.3.5

In this part we aim to estimate the value of the following integral:

$$H = \int_{2}^{\infty} e^{-\frac{x^2}{2}}\, dx = \int_{2}^{\infty} f_X(x)\, dx$$

Figure 1 shows the true value of the integral calculated using Python's built-in numerical integration libraries, serving as the ground truth reference.



```
[ ]  def f(x):
         return np.exp(-x**2/2)

     result = quad(f, 2,  np.inf)
     print(f'Ground truth Result:{result[0]}',
           f'\nError:+/-{result[1]}')
```

```
     Ground truth Result:0.05702612399288397
     Error:+/-1.0719816406372156e-08
```

Figure 1: True value of the integral

If we define the function h as

$$h(x) = \begin{cases} 1 & \text{if } x > 2 \\ 0 & \text{if } x \leq 2 \end{cases}$$

we can rewrite this integral as

$$\int_{3}^{\infty} f_X(x)dx = \int_{-\infty}^{\infty} h(x) f_X(x)dx$$

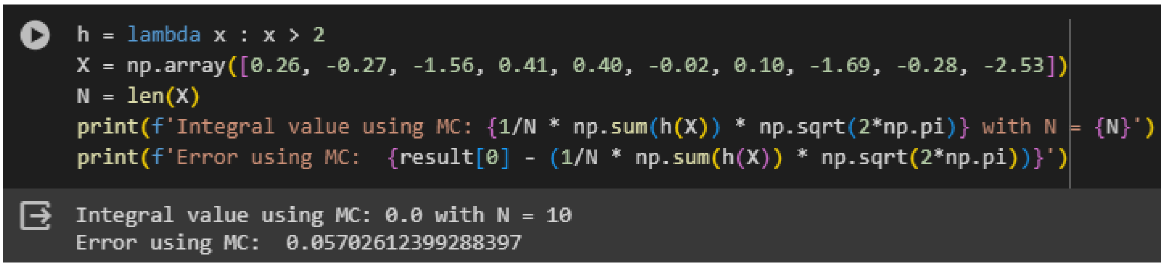
We can restate the integral above as

$$\int_{-\infty}^{\infty} h(x) f_X(x)dx = E[h(X)]$$

We can now easily estimate this expected value using **Monte Carlo** simulation. Given a random i.i.d. sample $x_1, x_2, \cdots, x_N$ generated by $f_X \sim \mathcal{N}(0,1)$, we can estimate $E[h(X)]$ using

$$\widehat{E}_{X \sim f_X}[h(X)] = \frac{1}{N} \sum_{i=1}^{N} h\left(x_i\right)$$

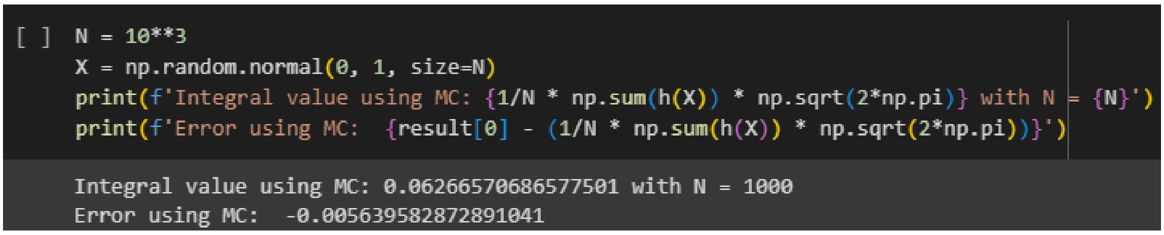
Figures 2-4 illustrate the value of the integral calculated using the Monte Carlo method with an increasing number of randomly sampled data points $10, 10^3$, and 10^7 samples respectively.



```
h = lambda x : x > 2
X = np.array([0.26, -0.27, -1.56, 0.41, 0.40, -0.02, 0.10, -1.69, -0.28, -2.53])
N = len(X)
print(f'Integral value using MC: {1/N * np.sum(h(X)) * np.sqrt(2*np.pi)} with N = {N}')
print(f'Error using MC:  {result[0] - (1/N * np.sum(h(X)) * np.sqrt(2*np.pi))}')

Integral value using MC: 0.0 with N = 10
Error using MC:  0.05702612399288397
```

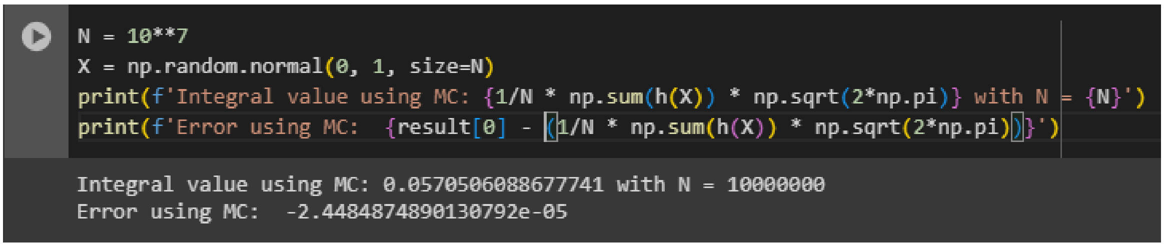
Figure 2: Monte Carlo, $N = 10$



```
[ ] N = 10**3
    X = np.random.normal(0, 1, size=N)
    print(f'Integral value using MC: {1/N * np.sum(h(X)) * np.sqrt(2*np.pi)} with N = {N}')
    print(f'Error using MC:  {result[0] - (1/N * np.sum(h(X)) * np.sqrt(2*np.pi))}')

    Integral value using MC: 0.06266570686577501 with N = 1000
    Error using MC:  -0.005639582872891041
```

Figure 3: Monte Carlo, $N = 10^3$



```
N = 10**7
X = np.random.normal(0, 1, size=N)
print(f'Integral value using MC: {1/N * np.sum(h(X)) * np.sqrt(2*np.pi)} with N = {N}')
print(f'Error using MC:  {result[0] - (1/N * np.sum(h(X)) * np.sqrt(2*np.pi))}')

Integral value using MC: 0.0570506088677741 with N = 10000000
Error using MC:  -2.4484874890130792e-05
```

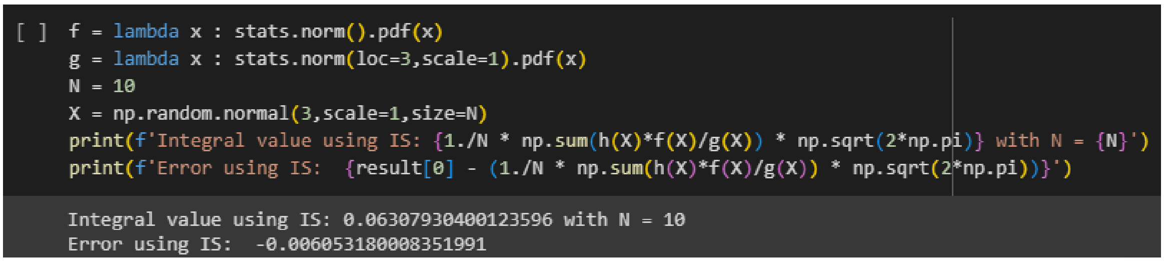
Figure 4: Monte Carlo, $N = 10^7$

Though this approach gets the job done, it turns out that this isn't very efficient. Since the probability of drawing a number greater than 2 from the standard normal distribution is so unlikely, it turns out we need many sample points to get a good approximation.

Importance sampling is one way to make Monte Carlo simulations converge much faster. We choose a different distribution to sample our points to generate more important points. With our example, we want to choose a distribution that would generate more numbers around 2 to get a more reliable estimate. The corresponding estimator is

$$\begin{aligned}\widehat{E}[h(X)] &= \widehat{E}\left[\frac{h(Y) f_X(Y)}{g_Y(Y)}\right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{h\left(y_i\right) f_X\left(y_i\right)}{g_Y\left(y_i\right)}\end{aligned} \tag{1.4}$$

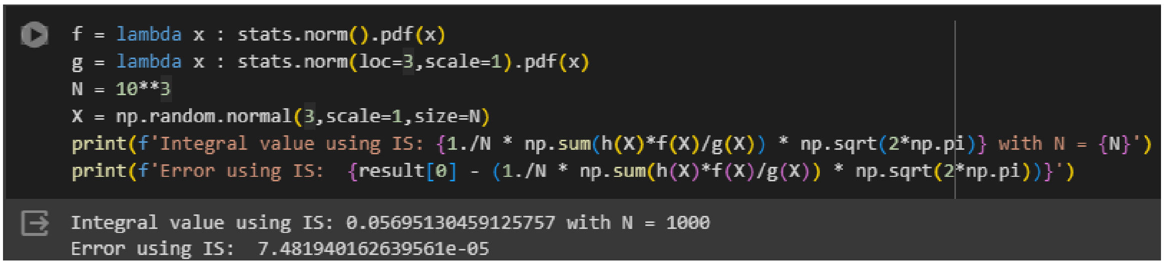
The function f_X is the pdf of the target distribution. The function g_Y is the pdf of the importance distribution. The fraction $\frac{f_X(X)}{g_Y(X)}$ is called the importance weight. This allows us to draw a sample from any distribution with pdf g_Y as long as we multiply $h(X)$ by the importance weight. Figures 5 and 6 show the calculated integral value using the Importance Sampling technique with just 100 and 1000 samples respectively.



```
[ ] f = lambda x : stats.norm().pdf(x)
    g = lambda x : stats.norm(loc=3,scale=1).pdf(x)
    N = 10
    X = np.random.normal(3,scale=1,size=N)
    print(f'Integral value using IS: {1./N * np.sum(h(X)*f(X)/g(X)) * np.sqrt(2*np.pi)} with N = {N}')
    print(f'Error using IS:  {result[0] - (1./N * np.sum(h(X)*f(X)/g(X)) * np.sqrt(2*np.pi))}')

    Integral value using IS: 0.06307930400123596 with N = 10
    Error using IS:  -0.006053180008351991
```

Figure 5: Importance Sampling, $N = 10$



```
f = lambda x : stats.norm().pdf(x)
g = lambda x : stats.norm(loc=3,scale=1).pdf(x)
N = 10**3
X = np.random.normal(3,scale=1,size=N)
print(f'Integral value using IS: {1./N * np.sum(h(X)*f(X)/g(X)) * np.sqrt(2*np.pi)} with N = {N}')
print(f'Error using IS:  {result[0] - (1./N * np.sum(h(X)*f(X)/g(X)) * np.sqrt(2*np.pi))}')

Integral value using IS: 0.05695130459125757 with N = 1000
Error using IS:  7.481940162639561e-05
```

Figure 6: Importance Sampling, $N = 10^3$

As expected, the Monte Carlo results in Figures 2-4 demonstrate that increasing the number of random samples improves the accuracy of the estimate, converging towards the true solution. However, an extremely high sample count of 10^7 is required to reach a reasonable approximation. In contrast, the Importance Sampling method in Figures 5 and 6 is able to achieve a value very close to the true integral with only 1000 carefully chosen samples. This significant performance advantage stems from Importance Sampling's ability to intentionally concentrate samples in the most "important" regions of the function that carry the highest weight towards the integral value.

By skewing the random samples towards areas under the curve that matter most, Importance Sampling dramatically reduces the number of evaluations required to numerically approximate the integral compared to traditional Monte Carlo methods that blindly sample the entire area. This variance reduction technique makes Importance Sampling a powerful tool for efficient numerical integration.

6 Variational Inference

6.1 Kullback-Leibler distance

Let $p(x), q(x), x \in \mathcal{X}$, be two probability mass functions. We will prove that $D(p\|q) \geq 0$ with equality if and only if $p(x) = q(x)$ for all x.

Let $A = \{x : p(x) > 0\}$ be the support set of $p(x)$. Then

$$\begin{aligned}
-D(p\|q) &= -\sum_{x\in A} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x\in A} p(x) \log \frac{q(x)}{p(x)} \\
&\leq \log \sum_{x\in A} p(x) \frac{q(x)}{p(x)} \quad \text{(By Jensen's Inequality)} \qquad (5) \\
&= \log \sum_{x\in A} q(x) \\
&\leq \log \sum_{x\in \mathcal{X}} q(x) \qquad (6) \\
&= \log 1 = 0
\end{aligned}$$

Since $\log t$ is a strictly concave function of t, we have equality in 5 if and only if $q(x)/p(x)$ is constant everywhere [i.e., $q(x) = cp(x) \quad \forall x$]. Thus, $\sum_{x\in A} q(x) = c \sum_{x\in A} p(x) = c$. We have equality in 6 only if $\sum_{x\in A} q(x) = \sum_{x\in \mathcal{X}} q(x) = 1$, which implies that $c = 1$. Hence, we have $D(p\|q) = 0$ if and only if $p(x) = q(x)$ for all x.

6.2 Use another distribution!

$$\begin{aligned}
D_{KL}(q(Z)||p(Z|X)) &= \int q(z) \log \frac{q(z)}{p(z|x)}\, dz = \int q(z) \log \frac{q(z)}{\frac{p(z,x)}{p(x)}}\, dz = \int q(z) \log \frac{q(z)p(x)}{p(z,x)}\, dz \\
&= \int q(z) \log p(x)\, dz + \int q(z) \log q(z)\, dz - \int q(z) \log p(x,z)\, dz \\
&= \log p(x) \underbrace{\int q(z)\, dz}_{=1} + \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z,x)] \\
&= \log p(x) - (\mathbb{E}_q[\log p(z,x)] - \mathbb{E}_q[\log q(z)]) = \boxed{\log p(x) - \mathbb{L}(q)}
\end{aligned}$$

6.3 Maximize the ELBO!

Using factorization property:

$$q(Z) = \prod_{i=1}^{n} q_i(z_i) \rightarrow \mathcal{L}(q) = \mathbb{E}_q[\log p(x,z)] - \mathbb{E}_q[\log q(z)]$$

Mean field approach:

$$\begin{aligned}
\mathcal{L}(q) &= \int q(z) \log p(x,z)\, dz - \int q(z) \log q(z)\, dz \\
&= \int \prod_{i=1}^{n} q_i(z_i) \log p(x,z)\, dz_1 \cdots dz_n - \int \prod_{i=1}^{n} q_i(z_i) \log \prod_{k=1}^{n} q_k(z_k)\, dz_1 \cdots dz_n \\
&= \int \prod_{i=1}^{n} q_i(z_i) \log p(x,z)\, dz_1 \cdots dz_n - \int \prod_{i=1}^{n} q_i(z_i) \sum_{k=1}^{n} \log q_k(z_k)\, dz_1 \cdots dz_n \\
&= \int \prod_{i=1}^{n} q_i(z_i) [\log p(x,z) - \sum_{k=1}^{n} \log q_k(z_k)]\, dz_1 \cdots dz_n \rightarrow \text{Select an arbitrary index} \Rightarrow \\
&= \int q_j(z_j) \int \prod_{\substack{i=1 \\ i\neq j}}^{n} q_i(z_i) [\log p(x,z)]\, dz_1 \cdots dz_n - \int q_j(z_j) \int \prod_{\substack{i=1 \\ i\neq j}}^{n} q_i(z_i) \sum_{k=1}^{n} \log q_k(z_k)\, dz_1 \cdots dz_n \\
&= \int q_j(z_j) \int \prod_{\substack{i=1 \\ i\neq j}}^{n} q_i(z_i) [\log p(x,z)]\, dz_1 \cdots dz_n - \int q_j(z_j) \log q_j(z_j)\, dz_j \underbrace{\int \prod_{\substack{i=1 \\ i\neq j}}^{n} q_i(z_i)\, dz_1 \cdots dz_n}_{=1} \\
&\quad - \underbrace{\int q_i(z_i)\, dz_i}_{=1} \int \prod_{\substack{i=1 \\ i\neq j}}^{n} q_i(z_i) \sum_{\substack{k=1 \\ k\neq j}}^{n} \log q_k(z_k)\, dz_1 \cdots dz_{j-1} dz_{j+1} \cdots dz_n \\
&= \int q_j(z_j) \int \prod_{\substack{i=1 \\ i\neq j}}^{n} q_i(z_i) [\log p(x,z)]\, dz_1 \cdots dz_n - \int q_j(z_j) \log q_j(z_j)\, dz_j - G(q_1, \cdots, q_{j-1}, q_{j+1}, \cdots, q_n)
\end{aligned}$$

$$\begin{aligned}
&\Rightarrow \text{argmax}_q \mathcal{L}(q) = H - G(q_1, \cdots, q_{j-1}, q_{j+1}, \cdots, q_n) = \\
&\quad \text{argmax}_q \int q_j(z_j) [\int \prod_{\substack{i=1 \\ i\neq j}}^{n} q_i(z_i) \log p(x,z) dz_1 \cdots dz_{j-1} dz_{j+1} \cdots dz_n - \log q_j(z_j)]\, dz_j - G \\
&\quad s.t. \int q_i(z_i) dz_i = 1
\end{aligned}$$

Using Lagrange multipliers:

$$\text{argmax}_q \mathcal{L}(q) = \text{argmax}_q H - G - \sum_{i=1}^{n} \lambda_i (\int q_i(z_i) dz_i - 1) \Rightarrow \frac{\partial}{\partial q_j(z_j)} \Rightarrow$$

$$\int \prod_{\substack{i=1\\i\neq j}}^{n} q_i(z_i)\log p(x,z)dz_1\cdots dz_{j-1}dz_{j+1}\cdots dz_n - \log q_j(z_j)\underbrace{-1-\lambda_j}_{\text{const.}} = 0$$

$$\Rightarrow \log q_j(z_j) = \int \prod_{\substack{i=1\\i\neq j}}^{n} q_i(z_i)\log p(x,z)dz_1\cdots dz_{j-1}dz_{j+1}\cdots dz_n + \text{const.}$$

$$= \mathbb{E}_{q_i, i\neq j}[\log p(X,Z)] + \text{const.}$$

6.4 Variational Mixture of Gaussians

we consider a variational distribution that factorizes between latent variables ($\boldsymbol{z}$) and parameters ($\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$):

$$q(\boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$$

In general, a factorized variational approximation to a full joint distribution of observed variables ($\boldsymbol{x}$) and hidden variables ($\boldsymbol{z}$) satisfies

$$\log q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i\neq j}\left[\log p(\boldsymbol{X}, \boldsymbol{Z})\right] + C \tag{7}$$

Which involves the following computation for our problem:

$$\log q^*(\boldsymbol{Z}) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] + C \tag{8}$$

$$\log q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathbb{E}_{\boldsymbol{z}}\left[\log p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] + C \tag{9}$$

In order to formulate a variational treatment of this model, we next write down the joint distribution of all of the random variables, which is given by

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})$$

We now make use of this decomposition. Note that we are only interested in the functional dependence of the right-hand side on the variable Z. Thus any terms that do not depend on Z can be absorbed into the additive normalization constant, giving

$$\log q^\star(\boldsymbol{Z}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\log p(\boldsymbol{Z}|\boldsymbol{\pi})\right] + \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\Lambda})\right] + \text{const.}$$

Some more things have been absorbed into that constant ($p(\boldsymbol{\Lambda})$, $p(\boldsymbol{\mu}|\boldsymbol{\Lambda})$ and $p(\boldsymbol{\pi})$) because they don't depend on $\boldsymbol{Z}$.

Then, since we know those conditional distributions, we can just substitute them in:

$$\log q^\star(\boldsymbol{Z}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\log \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}}\right] + \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}(x_n|\boldsymbol{\mu_k},\boldsymbol{\Lambda_k}^{-1})^{z_{nk}}\right] + \text{const.}$$

And then rearrange a bit:

$$\log q^\star(\boldsymbol{Z}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log \pi_k\right] + \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log \mathcal{N}(x_n|\boldsymbol{\mu_k},\boldsymbol{\Lambda_k}^{-1})\right] + \text{const.}$$

$$\log q^\star(\boldsymbol{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\mathbb{E}_{\boldsymbol{\pi}}\left[\log \pi_k\right] + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[\log \mathcal{N}(x_n|\boldsymbol{\mu_k},\boldsymbol{\Lambda_k}^{-1})\right] + \text{const.}$$

Then we can take everything inside the expectations and define a new variable ρ_{nk}:

$$\log \rho_{nk} = \mathbb{E}_{\boldsymbol{\pi}}\left[\log \pi_k\right] + \mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[\log \mathcal{N}(x_n|\boldsymbol{\mu_k},\boldsymbol{\Lambda_k}^{-1})\right]$$

$$\log \rho_{nk} = \mathbb{E}_{\boldsymbol{\pi}}\left[\log \pi_k\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}_k}\left[\log|\boldsymbol{\Lambda}_k|\right] - \frac{D}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right]$$

Then we can collapse the sums in a nice way and pretend the equation is much simpler than it really is:

$$\log q^\star(\boldsymbol{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log \rho_{nk} + \text{const.}$$

Now we can take the exponential of both sides:

$$q^\star(\boldsymbol{Z}) = \left(\prod_{n=1}^{N}\prod_{k=1}^{K} \rho_{nk}^{z_{nk}}\right) \times \text{const.} \quad \Rightarrow q^*(Z) \propto \prod_{n=1}^{N}\prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

Recall that we have defined

$$\log \rho_{nk} = \mathbb{E}[\log \pi_k] + \frac{1}{2}\mathbb{E}[\log|\Lambda_k|] - \frac{D}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)]$$

where D is the dimensionality of the data variable x. We see that this expression involves expectations with respect to the variational distributions of the parameters, and these are easily evaluated to give:

$$\mathbb{E}_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)] = D\beta_k^{-1} + \nu_k(x_n-m_k)^TW_k(x_n-m_k)$$

$$\log\tilde{\Lambda}_k \equiv \mathbb{E}[\log|\Lambda_k|] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right) + D\log 2 + \log|W_k|$$

$$\log\tilde{\pi}_k \equiv \mathbb{E}[\log \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha})$$

Where $\hat{\alpha} = \sum_k \alpha_k$ and $\psi(a) \equiv \frac{d}{da} \log \Gamma(a)$.

Requiring that this distribution be normalized, and noting that for each value of n the quantities z_{nk} are binary and sum to 1 over all values of k, we obtain

$$q^*(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}; \qquad r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}} \tag{10}$$

For the discrete distribution $q^*(Z)$ we have the standard result $\mathbb{E}[z_{nk}] = r_{nk}$. Determining values of r_{nk} is like the variational E-step. Now Let's prove the Equation 10: Requiring that this distribution be normalized, and nothing that for each value of n the quantities z_{nk} are binary and sum to 1 over all values of k...

We can go forward by name const. from above c:

$$q^{\star}(\boldsymbol{Z}) = c \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

For it to normalise:

$$1 = \sum_{Z} c \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

Then, since each vector of $\boldsymbol{z}_n$ is binary, with a single value being one, at each data point only a single ρ_{nj}; if j is where this one is located then the equation becomes:

$$1 = c \prod_{n=1}^{N} \sum_{j=1}^{K} \rho_{nj}$$

Solving for c:

$$c = \frac{1}{\prod_{n=1}^{N} \sum_{j=1}^{K} \rho_{nj}}$$

Then we can substitute that back into the original equation:

$$q^{\star}(\boldsymbol{Z}) = \frac{1}{\prod_{n=1}^{N} \sum_{j=1}^{K} \rho_{nj}} \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

And rearrange:

$$q^{\star}(\boldsymbol{Z}) = \prod_{n=1}^{N} \frac{1}{\sum_{j=1}^{K} \rho_{nj}} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

As z_{nk} must sum to one over all values of k ($\sum_{k}^{K} z_{nk} = 1$):

$$q^{\star}(\boldsymbol{Z}) = \prod_{n=1}^{N} \left(\frac{1}{\sum_{j=1}^{K} \rho_{nj}} \right)^{1} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

$$q^\star(\boldsymbol{Z}) = \prod_{n=1}^{N} \left(\frac{1}{\sum_{j=1}^{K} \rho_{nj}} \right)^{\sum_{k=1}^{K} z_{nk}} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

$$q^\star(\boldsymbol{Z}) = \prod_{n=1}^{N} \left(\prod_{k=1}^{K} \left(\frac{1}{\sum_{j=1}^{K} \rho_{nj}} \right)^{z_{nk}} \right) \left(\prod_{k=1}^{K} \rho_{nk}^{z_{nk}} \right)$$

$$q^\star(\boldsymbol{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}} \right)^{z_{nk}}$$

Then all we have to do is substitute in r_{nk} as defined above:

$$q^\star(\boldsymbol{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

And we're done.

Note that the optimal solution for the factor $q(\boldsymbol{Z})$ takes the same functional form as the prior $p(\boldsymbol{Z}|\pi)$.

Finding the expectation of this:

$$\mathbb{E}_{q^\star}[z_{nk}] = \sum_{Z} z_{nk} \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

The expectation selects a single element:

$$\mathbb{E}_{q^\star}[z_{nk}] = r_{nk}$$

Which is actually the expectation of any categorical distribution.

At this point, we shall define three statistics of the observed data set evaluated with respect to the responsibilities, given by

$$N_k = \sum_{n=1}^{N} r_{nk}$$

$$\bar{x}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n$$

$$\boldsymbol{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)^T$$

Note that these are analogous to quantities evaluated in the maximum likelihood EM algorithm for the Gaussian mixture model.

Now let us consider the factor $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ in the variational posterior distribution. Again using Equation 7 we have:

$$\log q^*(\pi, \mu, \Lambda) = \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k, \Lambda_k) + \mathbb{E}_Z[\log p(Z|\pi)]$$
$$+ \sum_{k=1}^{K} \sum_{n=1}^{N} \mathbb{E}[z_{nk}] \log \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1}) + \text{const}$$

We observe that the right-hand side of this expression decomposes into a sum of terms involving only π together with terms only involving μ and Λ, which implies that the variational posterior $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ factorizes to give $q(\boldsymbol{\pi})q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. Furthermore, the terms involving μ and Λ themselves comprise a sum over k of terms involving μ_k and Λ_k leading to the further factorization

$$q^\star(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$$

Identifying the terms on the right-hand side of that depend on π, we have

$$\log q^*(\pi) = (\alpha_0 - 1) \sum_{k=1}^{K} \log \pi_k + \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \log \pi_k + \text{const}$$

Taking the exponential of both sides, we recognize $q^*(\pi)$ as a Dirichlet distribution $q^*(\pi) = \text{Dir}(\pi|\alpha)$ where α has components α_k given by $\alpha_k = \alpha_0 + N_k$. To prove this:

$$\log q^\star(\boldsymbol{\pi}) = \sum_{k=1}^{K} (\alpha_0 - 1) \log \pi_k + \sum_{k=1}^{K} N_k \log \pi_k + \text{const.}$$

$$\log q^\star(\boldsymbol{\pi}) = \sum_{k=1}^{K} \log \pi_k (\alpha_0 - 1 + N_k) + \text{const.}$$

Then, taking the exponential:

$$q^\star(\boldsymbol{\pi}) = e^{\sum_{k=1}^{K} \log \pi_k (\alpha_0 - 1 + N_k) + \text{const.}}$$

$$q^\star(\boldsymbol{\pi}) = e^{\text{const.}} \prod_{k=1}^{K} e^{\log \pi_k (\alpha_0 - 1 + N_k)}$$

$$q^\star(\boldsymbol{\pi}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} e^{\log \pi_k (\alpha_0 - 1 + N_k)}$$

$$q^\star(\boldsymbol{\pi}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1 + N_k} = \mathbf{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

Where $\boldsymbol{\alpha}$ has components given by $\alpha_k = \alpha_0 - 1 + N_k$.

Finally we can write the variational posterior distribution as $q^\star(\boldsymbol{\mu}_k, \mathbf{\Lambda}_k) = q^\star(\boldsymbol{\mu}_k|\mathbf{\Lambda}_k)q^\star(\mathbf{\Lambda}_k)$ by the product rule. The result, as expected, is a Gaussian-Wishart distribution and is given:

$$q^*(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k|m_k, (\beta_k\Lambda_k)^{-1})W(\Lambda_k|W_k, \nu_k)$$

where we have defined

$$\begin{aligned}
\beta_k &= \beta_0 + N_k \\
m_k &= \frac{1}{\beta_k}(\beta_0 m_0 + N_k\bar{x}_k) \\
W_k^{-1} &= W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)^T \\
\nu_k &= \nu_0 + N_k
\end{aligned}$$

These update equations are analogous to the M-step equations of the EM algorithm for the maximum likelihood solution of the mixture of Gaussians.

If we substitute everything, we obtain the following result:

$$r_{nk} \propto \tilde{\pi}_k\tilde{\Lambda}_k^{1/2}\exp\left\{-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(x_n - m_k)^T W_k(x_n - m_k)\right\}$$

Notice the similarity to the corresponding result for the responsibilities in maximum likelihood EM

$$r_{nk} \propto \pi_k|\Lambda_k|^{1/2}\exp\left\{-\frac{1}{2}(x_n - \mu_k)^T\Lambda_k(x_n - \mu_k)\right\}$$

where we have used precision Λ_k instead of covariance Σ_k to highlight similarity. Thus, the optimization of the variational posterior distribution involves cycling between two stages analogous to the E and M steps of the maximum likelihood EM algorithm. For the variational Gaussian mixture model the expected values of the mixing coefficients in the posterior distribution are given by

$$\mathbb{E}[\pi_k] = \frac{\alpha_k + N_k}{K\alpha_0 + N}$$

As we have seen there is a close similarity between the variational solution for the Bayesian mixture of Gaussians and the EM algorithm for maximum likelihood. In fact if we consider the limit $N \to \infty$ then the Bayesian treatment converges to the maximum likelihood EM algorithm.

Variational lower bound For the variational mixture of Gaussians, the lower bound is given by:

$$\begin{aligned}\mathcal{L} &= \sum_Z \int\int\int q(Z,\pi,\mu,\Lambda)\log\{\frac{p(X,Z,\pi,\mu,\Lambda)}{q(Z,\pi,\mu,\Lambda)}\}dZd\mu d\Lambda \\ &= \mathbb{E}[\log p(X,Z,\pi,\mu,\Lambda)] - \mathbb{E}[\log q(Z,\pi,\mu,\Lambda)] \\ &= \mathbb{E}[\log p(X|Z,\mu,\Lambda)] + \mathbb{E}[\log p(Z|\pi)] + \mathbb{E}[\log p(\pi)] + \mathbb{E}[\log p(\mu,\Lambda)] \\ &\quad - \mathbb{E}[\log q(Z)] - \mathbb{E}[\log q(\pi)] - \mathbb{E}[\log q(\mu,\Lambda)]\end{aligned}$$

The various terms in the bound are easily evaluated to give the following results:

$$\begin{aligned}\mathbb{E}[\log p(X|Z,\mu,\Lambda)] &= \frac{1}{2}\sum_{k=1}^{K} N_k \Big\{\log\tilde{\Lambda}_k - D\beta_k^{-1} - \nu_k \mathrm{Tr}(S_k W_k) \\ &\qquad -\nu_k(\bar{x}_k - m_k)^T W_k(\bar{x}_k - m_k) - D\log(2\pi)\Big\} \\ \mathbb{E}[\log p(Z|\pi)] &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\log\tilde{\pi}_k \\ \mathbb{E}[\log p(\pi)] &= \log C(\alpha_0) + (\alpha_0 - 1)\sum_{k=1}^{K}\log\tilde{\pi}_k \\ \mathbb{E}[\log p(\mu,\Lambda)] &= \frac{1}{2}\sum_{k=1}^{K}\Big\{D\log(\frac{\beta_0}{2\pi}) + \log\tilde{\Lambda}_k - \frac{D\beta_0}{\beta_k} - \frac{\beta_0\nu_k}{\beta_k}(m_k - m_0)^T W_k(m_k - m_0)\Big\} \\ &\quad + K\log B(W_0,\nu_0) + \frac{\nu_0 - D - 1}{2}\sum_{k=1}^{K}\log\tilde{\Lambda}_k - \frac{1}{2}\sum_{k=1}^{K}\nu_k \mathrm{Tr}(W_0^{-1}W_k) \\ \mathbb{E}[\log q(Z)] &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\log r_{nk} \\ \mathbb{E}[\log q(\pi)] &= \sum_{k=1}^{K}(\alpha_k - 1)\log\tilde{\pi}_k + \log C(\alpha) \\ \mathbb{E}[\log q(\mu,\Lambda)] &= \sum_{k=1}^{K}\left\{\frac{1}{2}\log\tilde{\Lambda}_k + \frac{D}{2}\log\left(\frac{\beta_k}{2\pi}\right) - \frac{D}{2} - \mathbb{H}[q(\Lambda_k)]\right\}\end{aligned}$$

where D is the dimensionality of x, $\mathbb{H}[q(\Lambda_k)]$ is the entropy of the Wishart distribution given by $H[\Lambda] = -\log B(W,\Lambda) - \frac{\nu-D-1}{2}\mathbb{E}[\log|\Lambda|] + \frac{\nu D}{2}$, and the coefficients $C(\alpha)$ and $B(W,\nu)$ are defined by $C(\alpha) = \frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha_0)\ldots\Gamma(\alpha_K)}$ and $B(\mathbf{W},\nu) \equiv |\mathbf{W}|^{-\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}$, respectively ($\hat{\alpha} = \sum_k \alpha_k$).

Note that the terms involving expectations of the logs of the q distributions simply represent the negative entropies of those distributions. Some simplifications and combination of terms can be performed when these expressions are summed to give the lower bound. However, we have kept the expressions separate for ease of understanding.