Advanced Section: Gaussian Mixture Models

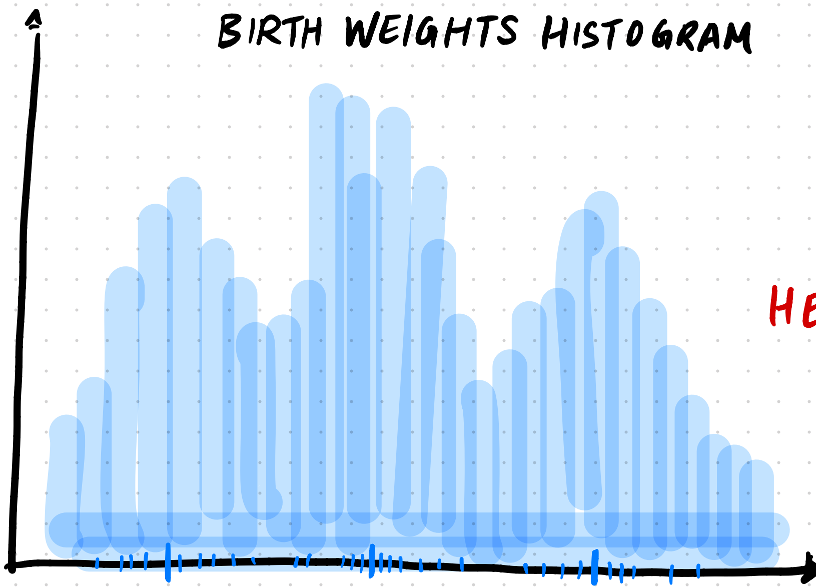
CS 109B

Spring, 2022

IACS

VE RI TAS

NON (PROBABILISTIC) MODEL BASED CLUSTERING



- HOW many clusters are there?
- How do we find them?

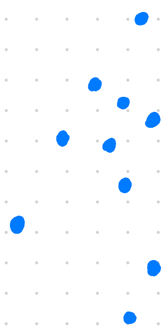
HEURISTIC: K-MEANS:

1. Assign points to clusters based on cluster means
2. Based on point assignments update cluster means

- How do we evaluate the quality of the clusters?

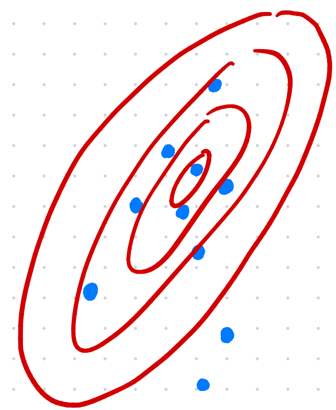
PROBABILISTIC VS NON-PROBABILISTIC APPROACHES

IS THIS ONE CLUSTER?



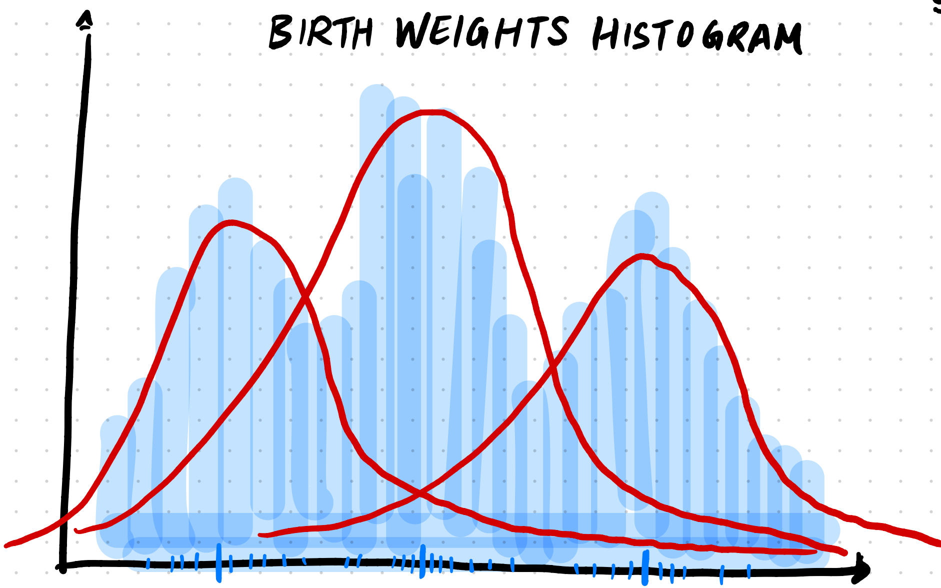
ASSUMPTIONS ABOUT CLUSTERS UNSTATED

DOES THE DATA COME FROM $N(\mu, \Sigma)$?



ASSUMPTIONS ABOUT CLUSTERS EXPLICITLY STATED

PROBABILISTIC MODELS FOR CLUSTERING



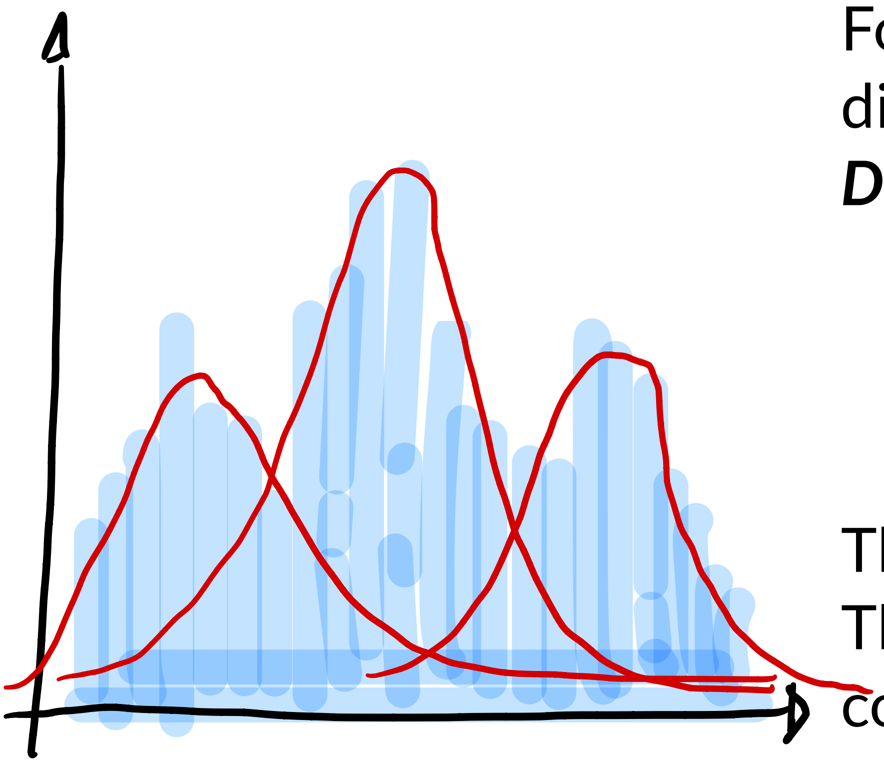
GAUSSIAN MIXTURE MODEL (GMM)

$$\ell = \pi_1 N(y_n; \mu_1, \sigma_1^2) + \pi_2 N(y_n; \mu_2, \sigma_2^2) + \pi_3 N(y_n; \mu_3, \sigma_3^2)$$

ASSUMPTIONS

- Each cluster is a Gaussian
- clusters are "mixed" as Gaussians overlap

A Similarity Measure for Distributions: Kullback–Leibler Divergence



Visually comparing models to the ***empirical distribution*** of the data is impractical. Fortunately, there are a large number of quantitative measures for comparing two distributions, these are called ***divergence measures***. For example, the ***Kullback–Leibler (KL) Divergence*** is defined for two distributions $p(\theta)$ and $q(\theta)$ supported on Θ as:

$$D_{\text{KL}}[q \parallel p] = \int_{\Theta} \log\left[\frac{q(\theta)}{p(\theta)}\right] q(\theta) d\theta$$

The KL-divergence $D_{\text{KL}}[q \parallel p]$ is bounded below by 0, which happens if and only if $q = p$. The KL-divergence has information theoretic interpretations that we will explore later in the course.

Note: The KL-divergence is defined in terms of the pdf's of p and q. If p is a distribution from which we only have samples and not the pdf (like the empirical distribution), we can nontheless estimate $D_{\text{KL}}[q \parallel p]$. Techniques that estimate the KL-divergence from samples are called ***non-parametric***. We will use them later in the course.

INFERENCE FOR GMM's: LIKELIHOOD MAXIMIZATION

$$\underbrace{\ell(\pi_k, \mu_k, \sigma_k^2)}_{\text{joint likelihood of the data set}} = \log \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \, N(y_n; \mu_k, \sigma_k^2)$$

$$= \sum_{n=1}^{N} \log \underbrace{\sum_{k=1}^{K} \pi_k \, N(y_n; \mu_k, \sigma_k^2)}_{\text{likelihood of each point}}$$

GOAL: find π_k, μ_k, σ_k^2 to maximize the likelihood of the data set.

HOW: Want to do gradient descent:

↳ $\nabla_{\pi_k, \mu_k, \sigma_k^2} \, \ell$

But:
- Gradient seem complicated
- This is secretly a constraine opt problem
  - ↳ $\sigma_k > 0$
  - ↳ $\sum_{k=1}^{K} \pi_k = 1$

Want:

1. Guess which Gaussians gets which points
2. Then it's easy to compute MLE of π_k, μ_k, σ_k^2

Ex: μ_k is empirical mean of pts in k-th Gaussian

Class Membership as a Latent Variable

We observe that there are three ***clusters*** in the data. We posit that there are three ***classes*** of infants in the study: infants with low birth weights, infants with normal birth weights and those with high birth weights. The numbers of infants in the classes are not equal.

For each observation Y_n, we model its class membership Z_n as a categorical variable,

$$Z_n \sim Cat(\pi),$$

where π_i in $\pi = [\pi_1, \pi_2, \pi_3]$ is the class proportion. Note that we don't have the class membership Z_n in the data! So Z_n is called a ***latent variable***.

Depending on the class, the n-th birth weight Y_n will have a different normal distribution,

$$Y_n | Z_n \sim \mathcal{N}\left(\mu_{Z_n}, \sigma^2_{Z_n}\right)$$

where μ_{Z_n} is one of the three class means $[\mu_1, \mu_2, \mu_3]$ and $\sigma^2_{Z_n}$ is one of the three class variances $[\sigma^2_1, \sigma^2_2, \sigma^2_3]$.

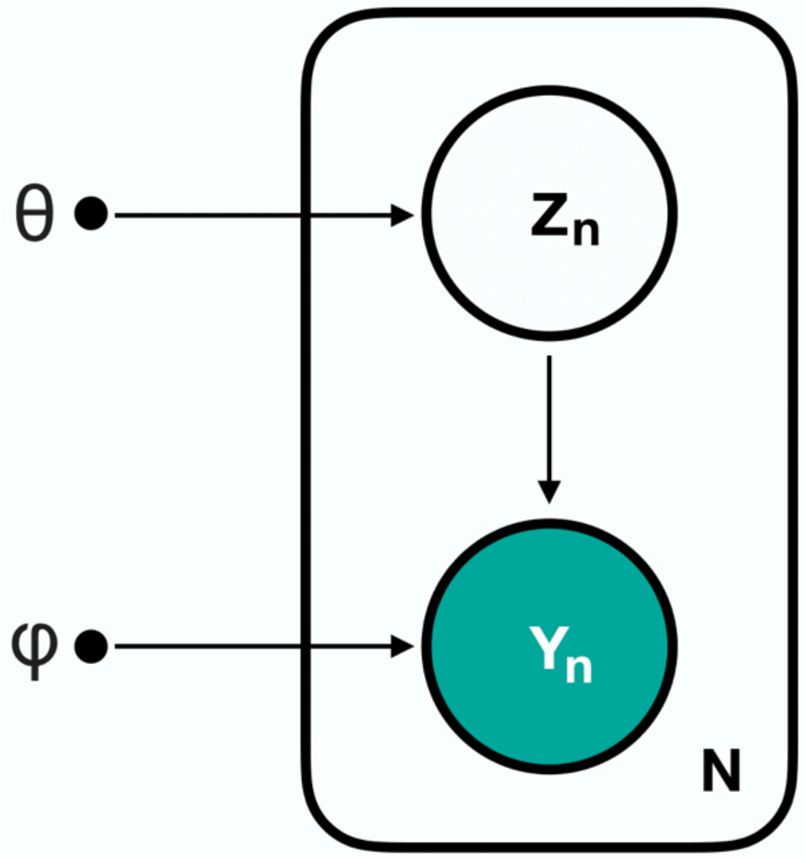
OBSERVED DATA LOG-LIKELIHOOD

$$P(y_n) = \int P(y_n|z_n)p(z_n)dz_n = \sum_{k=1}^{K} P(z_n=k)\, P(y_n|z_n) = \sum_{k=1}^{K} \pi_k P(y_n|z_n) = \sum_{k=1}^{K} \pi_k N(y_n; \mu_k, \sigma_k^2)$$

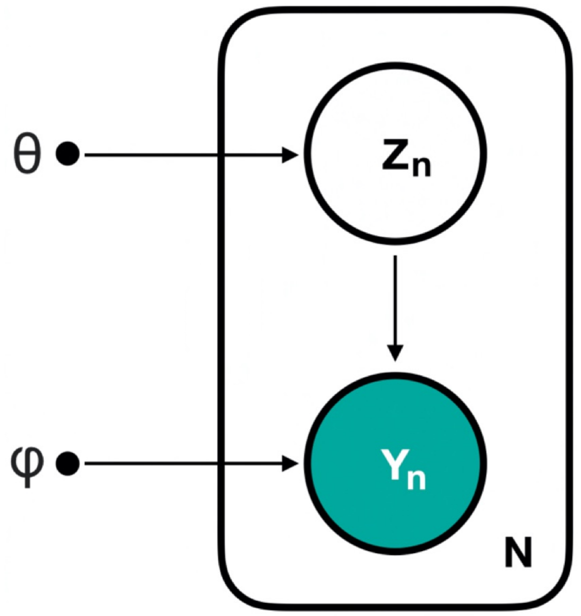
Common Latent Variable Models

Latent Variable Models

Models that include an observed variable Y and at least one unobserved variable Z are called ***latent variable models***. In general, our model can allow Y and Z to interact in many different ways. Today, we will study models with one type of interaction:



$$Z_n \sim p(Z|\theta)$$
$$Y_n|Z_n \sim p(Y|Z,\phi)$$
$$n = 1, \ldots, N$$



Item-Response Models

In ***item-response models***, we measure an real-valued unobserved trait Z of a subject by performing a series of experiments with binary observable outcomes, Y:

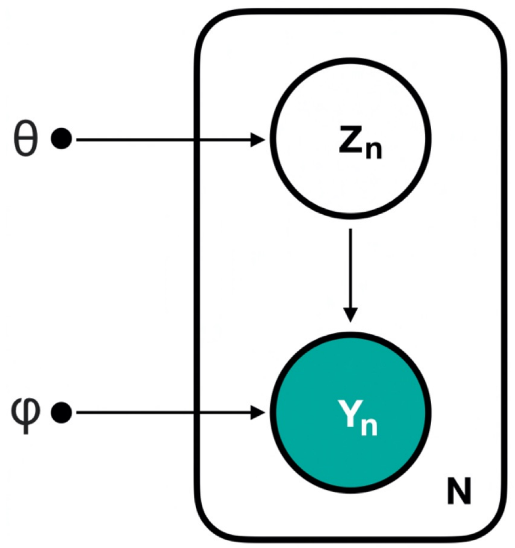
$$
\begin{aligned}
Z_n &\sim \mathcal{N}(\mu, \sigma^2), \\
\theta_n &= g(Z_n) \\
Y_n | Z_n &\sim Ber(\theta_n),
\end{aligned}
$$

where $n = 1, \ldots, N$ and g is some fixed function of Z_n.

Applications

Item response models are used to model the way "underlying intelligence" Z relates to scores Y on IQ tests.

Item response models can also be used to model the way "suicidality" Z relates to answers on mental health surveys. Building a good model may help to infer when a patient is at psychiatric risk based on in-take surveys at points of care through out the health-care system.



Factor Analysis Models

In ***factor analysis models***, we posit that the observed data Y with many measurements is generated by a small set of unobserved factors Z:

$$
\begin{aligned}
Z_n &\sim \mathcal{N}(0, I),\\
Y_n | Z_n &\sim \mathcal{N}(\mu + \Lambda Z_n, \Phi),
\end{aligned}
$$

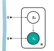
where $n = 1, \ldots, N$, $Z_n \in \mathbb{R}^{D'}$ and $Y_n \in \mathbb{R}^{D}$. We typically assume that D' is much smaller than D.

Applications

Factor analysis models are useful for biomedical data, where we typically measure a large number of characteristics of a patient (e.g. blood pressure, heart rate, etc), but these characteristics are all generated by a small list of health factors (e.g. diabetes, cancer, hypertension etc). Building a good model means we may be able to infer the list of health factors of a patient from their observed measurements.

Maximum Likelihood Estimation for Latent Variable Models: Expectation Maximization

MODEL



$z_n \sim p(z|\theta)$

$y_n \sim p(y_n|z_n, \phi)$

parameters: θ, ϕ

Calculus facts we need:

1. $\mathbb{E}_{x\sim p(x)}[f(x)] = \int f(x)\,p(x)\,dx$
2. properties of $\mathbb{E}$:
   - A. $\mathbb{E}_{x\sim p(x)}\left[a_i \sum_{i=1}^{I} f_i(x)\right] = a_i \sum_{i=1}^{I} \mathbb{E}_{x\sim p(x)}[f_i(x)]$
   - (Jensen's Inequality) → B. $\log \mathbb{E}_{x\sim p(x)}[f(x)] \geq \mathbb{E}_{x\sim p(x)}[\log f(x)]$
   - C. $\nabla_w \mathbb{E}_{x\sim p(x)}[f(x,w)] = \mathbb{E}_{x\sim p(x)}[\nabla_w f(x,w)]$
3. $D_{KL}[q(x)\|p(x)] = \int \log\left[\frac{q(x)}{p(x)}\right] q(x)\,dx$

   $= \mathbb{E}_{x\sim q(x)}\left[\log\left(\frac{q(x)}{p(x)}\right)\right]$

Likelihood over y & z:

$$\log \prod_{n=1}^{N} p(y_n|z_n,\phi)\,p(z_n|\theta) = \sum_{n=1}^{N} \log p(y_n|z_n,\phi) + \log p(z_n|\theta)$$

can we evaluate this?

Likelihood over observed data:

$$\log \prod_{n=1}^{N} p(y_n|\theta,\phi) = \sum_{n=1}^{N} \log p(y_n|\theta,\phi)$$

does the log help?

$$= \sum_{n=1}^{N} \log \int p(y_n|z_n,\phi)\,p(z_n|\theta)\,dz_n$$

$$= \sum_{n=1}^{N} \log \mathbb{E}_{z_n\sim p(z_n|\theta)}[p(y_n|z_n,\phi)]$$

$\ell_y(\theta,\phi)$

The maximum likelihood objective:

$$\phi_{MLE}, \theta_{MLE} = \underset{\theta,\phi}{\mathrm{argmax}}\ \ell_y(\theta,\phi) = \underset{\theta,\phi}{\mathrm{argmax}} \sum_{n=1}^{N} \log \mathbb{E}_{z_n\sim p(z_n|\theta)}[p(y_n|z_n,\phi)]$$

is this hard?

Trying out the optimization:

$$\nabla_{\theta,\phi}\,\ell_y(\theta,\phi) = \nabla_{\theta,\phi} \sum_{n=1}^{N} \log \mathbb{E}_{z_n\sim p(z_n|\theta)}[p(y_n|z_n,\phi)]$$

$$= \sum_{n=1}^{N} \nabla_{\theta,\phi} \log \mathbb{E}_{z_n\sim p(z_n|\theta)}[p(y_n|z_n,\phi)]$$

$$\text{"="} \sum_{n=1}^{N} \frac{\nabla_{\theta,\phi} \mathbb{E}_{z_n\sim p(z_n|\theta)}[p(y_n|z_n,\phi)]}{\mathbb{E}_{z_n\sim p(z_n|\theta)}[p(y_n|z_n,\phi)]}$$

can we MC estimate this?

An objective we can actually work with:

$$\max_{\theta,\phi} \ell_y(\theta,\phi) = \max_{\theta,\phi} \sum_{n=1}^{N} \log \mathbb{E}_{z_n\sim p(z_n|\theta)}[p(y_n|z_n,\phi)]$$

$$= \max_{\theta,\phi} \sum_{n=1}^{N} \log \int p(y_n|z_n,\phi)\,p(z_n|\theta)\,dz_n$$

$$= \max_{\theta,\phi} \sum_{n=1}^{N} \log \int \frac{p(y_n|z_n,\phi)\,p(z_n|\theta)}{q(z_n)}\,q(z_n)\,dz_n$$

Introduce auxillary variables $q(z_n)$! q is some distr. of your choice.

$$= \max_{\theta,\phi} \sum_{n=1}^{N} \log \mathbb{E}_{z\sim q(z_n)}\left[\frac{p(y_n|z_n,\phi)\,p(z_n|\theta)}{q(z_n)}\right]$$

is the log helping? does $\nabla_{\phi,\theta}$ commute with $\mathbb{E}_{z\sim p(z)}$?

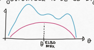
$$\geq \max_{\theta,\phi,q} \sum_{n=1}^{N} \mathbb{E}_{z\sim q(z_n)}\left[\log\left(\frac{p(y_n|z_n,\phi)\,p(z_n|\theta)}{q(z_n)}\right)\right]$$

The Evidence Lower Bound ELBO(θ,ϕ,q)

<u>Idea</u>: Instead of maximizing the log-likelihood, we maximize the lower bound ELBO.

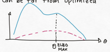
Why you should

When ELBO is maximized $\ell_y(\theta,\phi)$ is gauranteed to be as big.



Why you shouldn't

If ELBO is maximized then $\ell_y(\theta,\phi)$ can be far from optimized



How to maximize ELBO: coordinate ascent

I. maximize θ, ϕ, fixing q^*

$$\theta^*, \phi^* = \underset{\theta,\phi}{\mathrm{argmax}}\ \mathrm{ELBO}(\theta,\phi,q^*)$$

(M-step)

$$= \underset{\theta,\phi}{\mathrm{argmax}} \sum_{n=1}^{N} \mathbb{E}_{z\sim q^*(z_n)}\left[\log\left(\frac{p(y_n|z_n,\phi)\,p(z_n|\theta)}{q^*(z_n)}\right)\right]$$

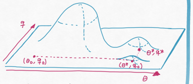
$$= \underset{\theta,\phi}{\mathrm{argmax}} \sum_{n=1}^{N} \left(\mathbb{E}_{z\sim q^*(z_n)}\left[\log[p(y_n|z_n,\phi)\,p(z_n|\theta)]\right] - \mathbb{E}_{z\sim q^*(z_n)}[q^*(z_n)]\right)$$

irrelevant for $\max_{\theta,\phi}$

$$\equiv \underset{\theta,\phi}{\mathrm{argmax}} \sum_{n=1}^{N} \mathbb{E}_{z\sim q^*(z_n)}\left[\log[p(y_n|z_n,\phi)\,p(z_n|\theta)]\right]$$

is this problem easier?

Are we done?



Maximizing each coordinate once is not sufficient! we need to iterate!

II. maximize q, fixing θ^*, ϕ^*

$$q^* = \underset{q}{\mathrm{argmax}}\ \mathrm{ELBO}(\theta^*,\phi^*,q)$$

<u>Note</u>: $\ell_y(\theta^*,\phi^*) - \mathrm{ELBO}(\theta^*,\phi^*,q) = \sum_{n=1}^{N}\left[\log p(y_n|\theta^*,\phi^*) - \mathbb{E}_{z_n\sim q(z_n)}\left[\log \frac{p(y_n,z_n|\theta^*,\phi^*)}{q(z_n)}\right]\right]$

$$= \sum_{n=1}^{N}\left[\mathbb{E}_{z_n\sim q(z_n)}[\log p(y_n|\theta^*,\phi^*)] - \mathbb{E}_{z_n\sim q(z_n)}\left[\log \frac{p(y_n,z_n|\theta^*,\phi^*)}{q(z_n)}\right]\right]$$

$$= \sum_{n=1}^{N}\left[\mathbb{E}_{z_n\sim q(z_n)}\left[\log p(y_n|\theta^*,\phi^*) - \log \frac{p(y_n,z_n|\theta^*,\phi^*)}{q(z_n)}\right]\right]$$

$$= \sum_{n=1}^{N}\left[\mathbb{E}_{z_n\sim q(z_n)} \log\left(\frac{p(y_n|\theta^*,\phi^*)\,q(z_n)}{p(y_n,z_n|\theta^*,\phi^*)}\right)\right]$$

$$= \sum_{n=1}^{N}\left[\mathbb{E}_{z_n\sim q(z_n)} \log\left(\frac{q(z_n)}{p(z_n|y_n,\theta^*,\phi^*)}\right)\right]$$

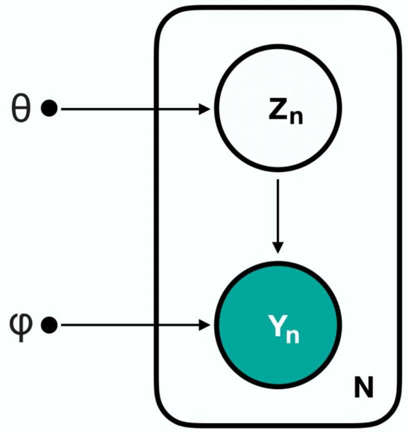
$$= \sum_{n=1}^{N} D_{KL}[q(z_n)\,\|\,p(z_n|y_n,\theta^*,\phi^*)]$$

$$q^* = \underset{q}{\mathrm{argmax}}\ \mathrm{ELBO}(\theta^*,\phi^*,q) = \underset{q}{\mathrm{argmin}}\ D_{KL}[q(z_n)\,\|\,p(z_n|y_n,\theta^*,\phi^*)]$$

$$q^* = p(z_n|y_n,\theta^*,\phi^*)$$

The Expectation Maximization Algorithm

The ***exepectation maximization (EM) algorithm*** maximize the ELBO of the model,



$$
\begin{aligned}
Z_n &\sim p(Z|\theta)\\
Y_n|Z_n &\sim p(Y|Z,\phi)\\
n &= 1,\dots,N
\end{aligned}
$$

1. **Initialization:** Pick θ_0, ϕ_0.

2. Repeat $i = 1, \dots, I$ times:

   E-Step:
   $$q_{\text{new}}(Z_n) = \underset{q}{\text{argmax}}\; ELBO(\theta_{\text{old}}, \phi_{\text{old}}, q) = p(Z_n|Y_n, \theta_{\text{old}}, \phi_{\text{old}})$$

   M-Step:
   $$
   \begin{aligned}
   \theta_{\text{new}}, \phi_{\text{new}} &= \underset{\theta,\phi}{\text{argmax}}\; ELBO(\theta, \phi, q_{\text{new}})\\
   &= \underset{\theta,\phi}{\text{argmax}}\; \sum_{n=1}^{N} \mathbb{E}_{Z_n \sim p(Z_n|Y_n, \theta_{\text{old}}, \phi_{\text{old}})}\left[\log(p(y_n, Z_n|\phi, \theta)\right].
   \end{aligned}
   $$

Example: EM for the Gaussian Mixture Model of Birth Weight

The Gaussian mixture model for the birth weight data has 3 Gaussians with meand $\mu = [\mu_1, \mu_2, \mu_3]$ and variances $\sigma^2 = [\sigma_1^2, \sigma_2^2, \sigma_3^2]$, and the model is defined as:

$$Z_n \sim Cat(\pi),$$
$$Y_n | Z_n \sim \mathcal{N}(\mu_{Z_n}, \sigma^2_{Z_n}),$$

where $n = 1, \ldots, N$ and $\sum_{k=1}^{3} \pi_k = 1$.

The E-Step

The E-step in EM computes the distribution:

$$q_{\text{new}}(Z_n) = \underset{q}{\text{argmax}}\ ELBO(\mu_{i-1}, \sigma^2_{i-1}, \pi_{i_1}, q) = p(Z_n | Y_n, \mu_{\text{old}}, \sigma^2_{\text{old}}, \pi_{\text{old}}).$$

Since Z_n is a label, $p(Z_n | Y_n, \ldots)$ is a categorical distribution, with the probability of $Z_n = k$ given by:

$$p(Z_n = k | Y_n, \mu_{\text{old}}, \sigma^2_{\text{old}}, \pi_{\text{old}}) = \frac{p(y_n | Z_n = k, \mu_{\text{old}}, \sigma^2_{\text{old}}) p(Z_n = k | \pi_{\text{old}})}{\sum_{k=1}^{K} p(y | Z_n = k, \mu_{\text{old}}, \sigma^2_{\text{old}}) p(Z_n = k | \pi_{\text{old}})} = \underbrace{\frac{\pi_{k,\text{old}}\ \mathcal{N}(y_n; \mu_{k,\text{old}}, \sigma^2_{k,\text{old}})}{\mathcal{Z}}}_{r_{n,k}},$$

where $\mathcal{Z} = \sum_{k=1}^{K} \pi_{k,\text{old}}\ \mathcal{N}(y_n; \mu_{k,\text{old}}, \sigma^2_{k,\text{old}})$.

Example: EM for the Gaussian Mixture Model of Birth Weight

Setting Up the M-Step

The M-step in EM maximize the following:

$$\underset{\mu,\sigma^2,\pi}{\mathrm{argmax}}\; ELBO(\mu, \sigma^2, \pi, q_{\text{new}}) = \underset{\mu,\sigma^2,\pi}{\mathrm{argmax}}\; \sum_{n=1}^{N} \mathbb{E}_{Z_n \sim p(Z_n|Y_n, \mu_{k,\text{old}}, \sigma^2_{k,\text{old}})} \left[\log\left(p(y_n, Z_n|\mu, \sigma^2, \pi)\right)\right].$$

If we expand the expectation a little, we get:

$$\begin{aligned}
\sum_{n=1}^{N} \mathbb{E}_{Z_n \sim p(Z_n|Y_n, \mu_{\text{old}}, \sigma^2_{\text{old}}, \pi_{\text{old}})} \left[\log\left(p(y_n, Z_n|\mu, \sigma^2, \pi)\right)\right] &= \underbrace{\sum_{n=1}^{N} \sum_{n=1}^{K} \log\left(\underbrace{p(y_n|Z_n = k, \mu, \sigma^2)p(Z_n = k|\pi)}_{\text{factoring the joint } p(y_n, Z_n|\ldots)}\right) p(Z_n = k|y_n, \theta_{\text{old}}, \phi_{\text{old}})}_{\text{expanding the expectation}} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \underbrace{r_{n,k}}_{p(Z_n=k|y_n, \theta_{\text{old}}, \phi_{\text{old}})} \left[\log \underbrace{\mathcal{N}(y_n; \mu_k, \sigma_k^2)}_{p(y_n|Z_n=k,\mu,\sigma^2)} + \log \underbrace{\pi_k}_{p(Z_n=k|\pi)}\right] \\
&= \underbrace{\sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \log \mathcal{N}(y_n; \mu_k, \sigma_k^2)}_{\text{Term \#1}} + \underbrace{\sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \pi_k}_{\text{Term \#2}}
\end{aligned}$$

We can maximize each Term #1 and Term #2 individually.

Example: EM for the Gaussian Mixture Model of Birth Weight

Solving the M-Step

We see that the optimization problem in the M-step:
$\mu_{\text{new}}, \sigma^2_{\text{new}}, \pi_{\text{new}} = \underset{\mu,\sigma^2,\pi}{\text{argmax}}\ ELBO(\mu, \sigma^2, \pi, q_{\text{new}})$ is equivalent to two problems

$$1. \quad \underset{\mu,\sigma^2}{\text{argmax}} \sum_{n=1}^{N}\sum_{k=1}^{K} r_{n,k} \log \mathcal{N}(y_n; \mu_k, \sigma_k^2)$$

$$2. \quad \underset{\pi}{\text{argmax}} \sum_{n=1}^{N}\sum_{k=1}^{K} r_{n,k} \pi_k$$

We can solve each optimization problem analytically by finding stationary points of the gradient (or the Lagrangian):

- $\mu_{\text{new}} = \frac{1}{\sum_{n=1}^{N} r_{n,k}} \sum_{n=1}^{N} r_{n,k} y_n$
- $\sigma^2_{\text{new}} = \frac{1}{\sum_{n=1}^{N} r_{n,k}} \sum_{n=1}^{N} r_{n,k} (y_n - \mu_{\text{new}})^2$
- $\pi_{\text{new}} = \frac{\sum_{n=1}^{N} r_{n,k}}{N}$

Example: EM for the Gaussian Mixture Model of Birth Weight

All Together

Initialization: Pick any π, μ, σ^2

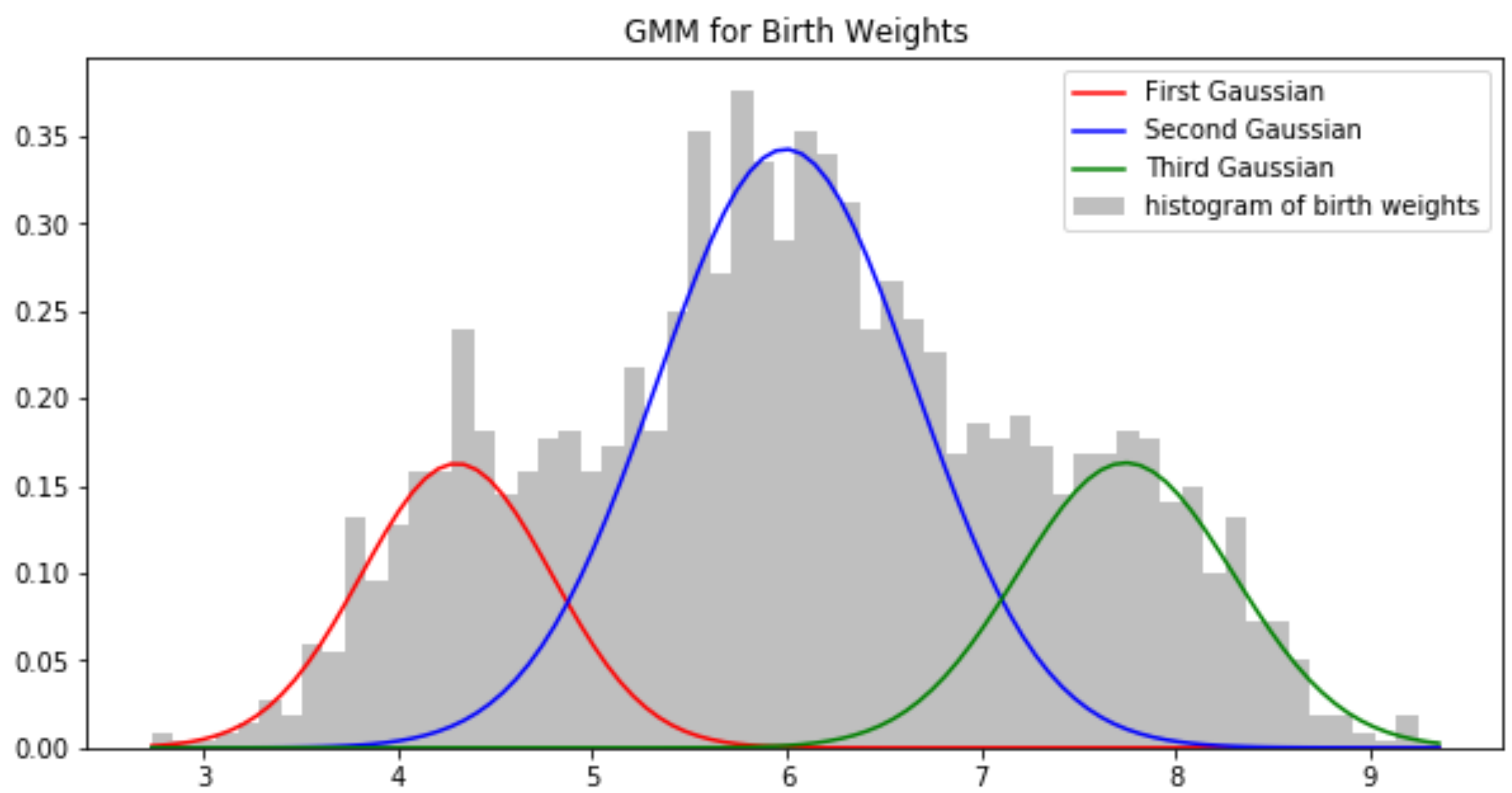
E-Step: Compute $r_{n,k} = \dfrac{\pi_{k,\text{old}}\ \mathcal{N}(y_n; \mu_{k,\text{old}}, \sigma^2_{k,\text{old}})}{\mathcal{Z}}$, where $\mathcal{Z} = \sum_{k=1}^{K} \pi_{k,\text{old}}\ \mathcal{N}(y_n; \mu_{k,\text{old}}, \sigma^2_{k,\text{old}})$.

M-Step: Compute model parameters:

- $\mu_{\text{new}} = \frac{1}{\sum_{n=1}^{N} r_{n,k}} \sum_{n=1}^{N} r_{n,k}\, y_n$
- $\sigma^2_{\text{new}} = \frac{1}{\sum_{n=1}^{N} r_{n,k}} \sum_{n=1}^{N} r_{n,k} (y_n - \mu_{\text{new}})^2$
- $\pi_{\text{new}} = \frac{\sum_{n=1}^{N} r_{n,k}}{N}$

Implementing EM for the Gaussian Mixture Model of Birth Weight

In [3]:
```
fig, ax = plt.subplots(1, 1, figsize=(10, 5))
ax.hist(y, bins=60, density=True, color='gray', alpha=0.5, label='histogram of b
irth weights')
ax.plot(x, pi_current[0] * sp.stats.norm(mu_current[0], sigma_current[0]**0.5).p
df(x), color='red', label='First Gaussian')
ax.plot(x, pi_current[1] * sp.stats.norm(mu_current[1], sigma_current[1]**0.5).p
df(x), color='blue', label='Second Gaussian')
ax.plot(x, pi_current[2] * sp.stats.norm(mu_current[2], sigma_current[2]**0.5).p
df(x), color='green', label='Third Gaussian')
ax.set_title('GMM for Birth Weights')
ax.legend(loc='best')
plt.show()
```



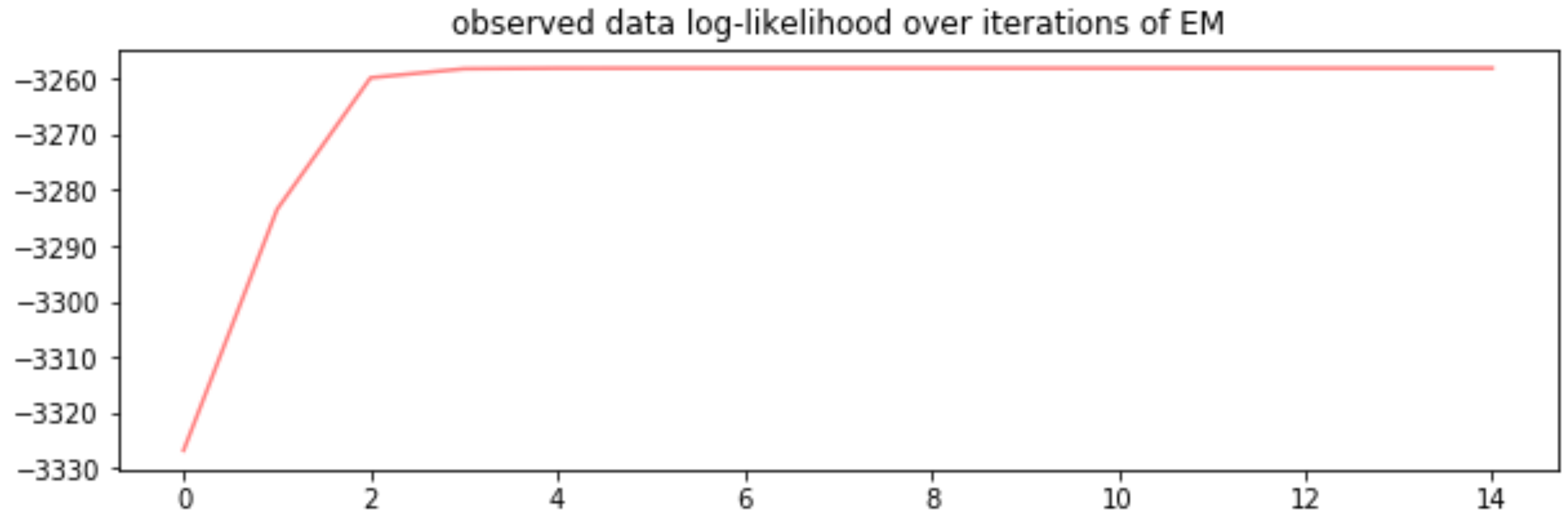
Sanity Check: Log-Likelihood During Training

Remember that ploting the MLE model against actual data is not always an option (e.g. high-dimensional data).

A sanity check for that your EM algorithm has been implemented correctly is to plot the observed data log-likelihood over the iterations of the algorithm:

$$\ell_y(\mu, \sigma^2, \pi) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \mathcal{N}(y_n; \mu_k, \sigma_k^2)\pi_k$$

In [4]:
```
fig, ax = plt.subplots(1, 1, figsize=(10, 3))
ax.plot(range(len(log_lkhd)), log_lkhd, color='red', alpha=0.5)
ax.set_title('observed data log-likelihood over iterations of EM')
plt.show()
```



Expectation Maximization versus Gradient-based Optimization

Pros of EM:

1. No learning rates to adjust
2. Don't need to worry about incorporating constraints (i.e. $p(Z_n|Y_n)$ is between 0 and 1)
3. Each iteration is guaranteed to increase or maintain observed data log-likelihood
4. Is guaranteed to converge to local optimum
5. Can be very fast to converge (when parameters are fewer)

Cons of EM:

1. Can get stuck in local optima
2. May not maximize observed data log-likelihood (the ELBO is just a lower bound)
3. Requires you to do math - you need analytic solutions for E-step and M-step
4. May be much slower than fancier gradient-based optimization