Restricted ML, EM-algorithm and Numerical integration

We start with some motivation: `https://xiuming.info/docs/tutorials/reml.pdf` has some good explanations. Restricted Maximum Likelihood (ReML) exists as Maximum likelihood estimates are often suboptimal in obtaining good properties such as unbiasedness.

Motivation – fixed effect linear model: Let us consider the basic linear regression model,

$$y_i = \mathbf{x}_i\beta + \epsilon_i, \quad \epsilon_i \sim \text{Normal}(0, \sigma^2),$$

where $\mathbf{x}_i$'s are p-dimensional data points. The likelihood of the above model is $\prod_i \phi\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right)$, where $\phi(\cdot)$ stands for the normal density. What are the Maximum Likelihood estimates (MLE) of $\boldsymbol{\beta}$ and σ? These are $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}^T\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}^T\hat{\boldsymbol{\beta}}) = \frac{1}{n}\mathbf{y}^T(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$, where $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{X}$ is the $n \times p$ design matrix. So, $\mathbf{x}_i$ is the *i-th* row of $\mathbf{X}$.

Using the results of χ^2 distribution, we know $\mathbf{y}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y} \sim \chi^2(n-p)$ as the rank of $(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$ is $n-p$. Thus $\mathbb{E}(\hat{\sigma}^2) = \frac{n-p}{n}\sigma$. Thus with increasing p, the MLE estimate of σ contains more bias.

Hence, instead of calculating the MLE of β and σ simultaneously from the likelihood $\prod_i \phi\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right)$, we could calculate MLE of β only from $\prod_i \phi\left(\frac{y_i - \mathbf{x}_i^T\beta}{\sigma}\right)$ and for σ, we re-write the model removing $\boldsymbol{\beta}$ term completely and use the likelihood of that to calculate the MLE of σ. This is the guiding principle in REML.

Motivation – linear mixed model: Say we are collecting longitudinal data on blood pressure (BP) of patients, and we are interested to study association between BP and BMI. Every patient will be visiting the clinic after every 3 months for a year. Hence, there are a total of 5 visits for each patient, visit0, visit3, visit6, visit9 and visit12. Let $y_{i,j}$ be the BP rating for *i-th* patient at *j-th* visit where $j = 0, 3, 6, 9, 12$ and $x_{i,j}$ stands for the BMI rating for *i-th* patient at *j-th* visit.

Using a simple regression model, we can write $\mathbb{E}(y_{i,j}) = \beta x_{i,j}$. Hence, an appropriate model may be $y_{i,j} = \beta x_{i,j} + \epsilon_{i,j}$ for $\epsilon_{i,j} \sim \text{Normal}(0, \sigma^2)$. [Note that I write $y_{i,j} = \beta x_{i,j} + \epsilon_{i,j}, \quad \epsilon_{i,j} \sim \text{Normal}(0, \sigma^2)$ and $y_{i,j} \sim \text{Normal}(\beta x_{i,j}, \sigma^2)$ alternatively as they mean the same thing.] Is this an adequate model?

Well under this model $\text{cov}(y_{i,0}, y_{i,3}) = 0$ or more generally $\text{cov}(y_{i,j}, y_{i,j'}) = 0$ for any $j \neq j'$. This is unreasonable, as $y_{i,0}$ and $y_{i,3}$ both correspond to the same subject. Thus $y_{i,0}$ and $y_{i,3}$ are expected to be correlated. To introduce this correlation, we bring in another source of noise via random effect and finally write the complete model as,

$$\begin{aligned} \text{Model: } & y_{i,j} \sim \text{Normal}(\beta x_{i,j} + b_i, \sigma^2). \\ \text{Random effect: } & b_i \sim \text{Normal}(0, \sigma_1^2) \end{aligned}$$

Now, $\text{cov}(y_{i,0}, y_{i,3}) = \sigma_1^2$. One may argue that the covariance is not general enough, as it introduces the same level of correlation for all the pairs of repeated measures. However, the generalization is easy. For simplicity, we will not discuss the generalization. For multivariate analysis working with repeated measures, the likelihood is given by the multivariate Gaussian distribution:

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}_y) = \prod_i \int \prod_j \phi\left(\frac{y_{i,j} - \beta x_{i,j} - b_i}{\sigma}\right) \phi\left(\frac{b_i}{\sigma_1}\right) \mathrm{d}b_i,$$

where ϕ stands for the normal density.

Restricted maximum likelihood estimation (REML)

The idea of Restricted Maximum Likelihood (REML) comes from realization that the variance estimator given by the Maximum Likelihood (ML) for linear mixed model is biased. What is an estimator and in which way it is biased? An estimator is simply an approximation / estimate of model parameters. Assuming that statistical observations follow Normal distribution, there are two parameters: μ (mean) and σ^2 (variance) to estimate if one wants to summarize the observations. It turns out that the variance estimator given by Maximum Likelihood (ML) is biased, i.e. the value we obtain from the ML model over- or under-estimates the true variance, see the figure below.

In practice, when we e.g. solve a Linear Regression model using ML, we rarely think about the bias in the variance estimator, since we are usually interested in the coefficients of the linear model, which is the mean, and often do not even realize that in parallel we estimate one more fitting parameter, which is the variance. In this case, the variance is considered to be a so-called nuisance parameter that is not of our primary interest.

When $y_i \sim \text{Normal}(\mu, \sigma^2)$ for $i = 1, \ldots, n$, ML estimates of μ and σ^2 are $\bar{y}$ and $\text{Var}(y) = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$ respectively. However, we know that $\mathbb{E}(= \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$, thus it is a biased estimate. The problem continues even for likelihood based estimates of

linear mixed model.

The problem with the biased variance estimator by ML appears to be due to the fact that we used an unknown estimator for the mean for computing the variance estimator. Instead, if we make sure that the log-likelihood function does not contain any information about the mean, we can optimize it with respect to the variance components and get an unbiased variance estimator. This is essentially what Restricted Maximum Likelihood (REML) does. In this case, the mean (not the variance like for ML) is considered to be a nuisance parameter that should be somehow removed from the equation. A way to get rid of the information about the mean from the log-likelihood function is to compute a marginal probability, i.e. integrate the log-likelihood over the mean.

ReML for fixed effect linear regression
Our original model can be written as $\mathbf{y} \sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.

Notice that $\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}\mathbf{X} = \mathbf{X} - \mathbf{X} = 0$. Let $\mathbf{P}_X = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Thus, our model can be made independent of $\boldsymbol{\beta}$ by pre-multiplying our linear regression model with $\{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}$ and get $\{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}\mathbf{y} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}\boldsymbol{\epsilon} = \boldsymbol{\eta}$ (say). Since $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \sim \text{Normal}(0, \sigma^2\mathbf{I}_n)$, we have $\boldsymbol{\eta} \sim \text{Normal}(0, \sigma^2\{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\})$ as $\{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}^2 = \{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}$ (It is called an idempotent matrix when $\mathbf{A}^2 = \mathbf{A}$).

Let $\mathbf{y}' = \{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}\mathbf{y}$. This reduces to $\mathbf{y}' \sim \text{Normal}(0, \sigma^2\{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\})$. MLE of σ^2 from this 'restricted' model will be $\frac{1}{n-p}\mathbf{y}^T\{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}\mathbf{y}$ (the unbiased estimate). Verify that using the expectation formula of quadratic form (`https://en.wikipedia.org/wiki/Quadratic_form_(statistics)`).

[Above results uses the fact of Gaussian distribution that: if $\mathbf{x} \sim \text{MVN}(\mu, \boldsymbol{\Sigma})$, then $\mathbf{A}\mathbf{x} \sim \text{MVN}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^2)$. So, $\boldsymbol{\epsilon} \sim \text{Normal}(0, \sigma^2\mathbf{I})$ implies $\beta \sim \text{MVN}(0, \sigma^2\mathbf{P}_X\mathbf{P}_X^T)$. It is easy to check that $\mathbf{P}_X\mathbf{P}_X^T = \mathbf{P}_X$.]

ReML for mixed effect linear regression For mixed effect case, to get a good estimate of the parameters we need to consider the integrated likelihood. We need to consider the marginal model to get the estimates. For example, b_i is a nuisance parameter. We first marginalize it. This will give us one layer model $\mathbf{y}_i \sim \text{Normal}(\beta\mathbf{x}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 & \cdots & \sigma_1^2 \\ \sigma_1^2 & \sigma^2 + \sigma_1^2 & \cdots & \sigma_1^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1^2 & \sigma_1^2 & \cdots & \sigma^2 + \sigma_1^2 \end{pmatrix}$.

Here $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,n_i})$ and $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,n_i})$ where n_i is the number of longitudinal visits by the *i-th* subject.

R implementation:`lme4` (https://cran.r-project.org/web/packages/lme4/index.html) is one of the most popular packages for mixed effect regression using restricted ML approach.

EM algorithm

Illustrative model Let there be n_i many observations in $\mathbf{y}_i$ for $i = 1, \ldots, N$.

$$\begin{aligned} \text{Model: } & \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{A}_i\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \\ & \boldsymbol{\epsilon}_i = \{\epsilon_{i,1}, \ldots, \epsilon_{i,n_i}\}, \quad \epsilon_{i,j} \sim \text{Normal}(0, \sigma^2), \\ \text{Random effect: } & \boldsymbol{\eta}_i \sim \text{Normal}(0, \boldsymbol{\Omega}) \end{aligned}$$

In statistics, an expectation–maximization (EM) algorithm is an iterative method to find (local) maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

Specifically, Let the statistical model has observed data $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, a set of unobserved latent data or missing values $\boldsymbol{\eta} = \{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N\}$, and a set of unknown parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma, \boldsymbol{\Omega}\}$. The likelihood function is given by $L(\boldsymbol{\theta}; \mathbf{Y}, \boldsymbol{\eta}) = P(\mathbf{Y}, \boldsymbol{\eta} \mid \boldsymbol{\theta})$, the maximum likelihood estimate (MLE) of the unknown parameters is determined by maximizing the marginal likelihood of the observed data

$L(\boldsymbol{\theta}; \mathbf{Y}) = p(\mathbf{Y} \mid \boldsymbol{\theta}) = \int p(\mathbf{Y}, \boldsymbol{\eta} \mid \boldsymbol{\theta})\, d\boldsymbol{\eta} = \int p(\mathbf{Y} \mid \boldsymbol{\eta}, \boldsymbol{\theta}) p(\boldsymbol{\eta} \mid \boldsymbol{\theta})\, d\boldsymbol{\eta}$ However, this quantity is often intractable since $\boldsymbol{\eta}$ is unobserved and the distribution of $\boldsymbol{\eta}$ is unknown before attaining $\boldsymbol{\theta}$ (The distribution of $\boldsymbol{\eta}$ may be specified using some parameters that are part of $\boldsymbol{\theta}$).

The EM algorithm [1] seeks to find the MLE of the marginal likelihood by iteratively applying these two steps:

- Expectation step (E step): Define $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ as the expected value of the log likelihood function of $\boldsymbol{\theta}$, with respect to the current conditional distribution of $\boldsymbol{\eta}$ given $\mathbf{Y}$ and the current estimates of the parameters $\boldsymbol{\theta}^{(t)}$: $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\boldsymbol{\eta}|\mathbf{Y},\boldsymbol{\theta}^{(t)}}\left[\log L(\boldsymbol{\theta}; \mathbf{Y}, \boldsymbol{\eta})\right] = \mathrm{E}_{\boldsymbol{\eta}|\mathbf{Y},\boldsymbol{\theta}^{(t)}}[\log\{p(\mathbf{Y}, \boldsymbol{\eta} \mid \boldsymbol{\theta})\}]$, i.e. expectation of the full likelihood under the conditional

distribution of $\boldsymbol{\eta} \mid (\mathbf{Y}, \boldsymbol{\theta}^{(t)})$

- Maximization step (M step): Find the parameters that maximize this quantity: $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$

Explanation:
We have $p(\mathbf{Y}, \boldsymbol{\eta} \mid \boldsymbol{\theta}) = p(\boldsymbol{\eta} \mid \mathbf{Y}, \boldsymbol{\theta}) p(\mathbf{Y} \mid \boldsymbol{\theta})$

Thus $\log\{p(\mathbf{Y}, \boldsymbol{\eta} \mid \boldsymbol{\theta})\} = \log\{p(\boldsymbol{\eta} \mid \mathbf{Y}, \boldsymbol{\theta})\} + \log\{p(\mathbf{Y} \mid \boldsymbol{\theta})\}$

Start from the log-likelihood: $\log\{p(\mathbf{Y}|\boldsymbol{\theta})\} = \log\{\int p(\mathbf{Y}, \boldsymbol{\eta} \mid \boldsymbol{\theta}) d\boldsymbol{\eta}\} = \log\{\int \frac{p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})}{q(\boldsymbol{\eta})} q(\boldsymbol{\eta}) d\boldsymbol{\eta}\}$, $q(\boldsymbol{\eta})$ is some distribution of $\boldsymbol{\eta}$ (which will be characterized below). It is also called the variational distribution.

Applying Jensen's inequality, $\log\{\int \frac{p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})}{q(\boldsymbol{\eta})} q(\boldsymbol{\eta}) d\boldsymbol{\eta}\} = \log\{\mathbb{E}_{\boldsymbol{\eta} \sim q(\boldsymbol{\eta})} \frac{p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})}{q(\boldsymbol{\eta})}\} \geq \mathbb{E}_{\boldsymbol{\eta} \sim q(\boldsymbol{\eta})}[\log\{\frac{p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})}{q(\boldsymbol{\eta})}\}] = \int [\log\{\frac{p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})}{q(\boldsymbol{\eta})}\}] q(\boldsymbol{\eta}) d\boldsymbol{\eta}$.

Let's quantify the difference, $\log\{p(\mathbf{Y}|\boldsymbol{\theta})\} - \int [\log\{\frac{p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})}{q(\boldsymbol{\eta})}\}] q(\boldsymbol{\eta}) d\boldsymbol{\eta} = \int [\log\{p(\mathbf{Y}|\boldsymbol{\theta})\}] q(\boldsymbol{\eta}) d\boldsymbol{\eta} - \int [\log\{\frac{p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})}{q(\boldsymbol{\eta})}\}] q(\boldsymbol{\eta}) d\boldsymbol{\eta}$, as $\int q(\boldsymbol{\eta}) d\boldsymbol{\eta} = 1$ as it is valid probability density.

Using $\log\{p(\mathbf{Y}, \boldsymbol{\eta} \mid \boldsymbol{\theta})\} = \log\{p(\boldsymbol{\eta} \mid \mathbf{Y}, \boldsymbol{\theta})\} + \log\{p(\mathbf{Y} \mid \boldsymbol{\theta})\}$, we have
$\int [\log\{p(\mathbf{Y}|\boldsymbol{\theta})\}] q(\boldsymbol{\eta}) d\boldsymbol{\eta} - \int [\log\{\frac{p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})}{q(\boldsymbol{\eta})}\}] q(\boldsymbol{\eta}) d\boldsymbol{\eta} = \int \log\{\frac{q(\boldsymbol{\eta})}{p(\boldsymbol{\eta}|\mathbf{Y}, \boldsymbol{\theta})}\} q(\boldsymbol{\eta}) d\boldsymbol{\eta} = \mathbb{E}_{\boldsymbol{\eta} \sim q(\boldsymbol{\eta})}[\log\{\frac{q(\boldsymbol{\eta})}{p(\boldsymbol{\eta}|\mathbf{Y}, \boldsymbol{\theta})}\}]$

Thus, $\log p(\mathbf{Y} \mid \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\eta} \sim q(\boldsymbol{\eta})}[\log\{\frac{p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})}{q(\boldsymbol{\eta})}\}] + \mathbb{E}_{\boldsymbol{\eta} \sim q(\boldsymbol{\eta})}[\log\{\frac{q(\boldsymbol{\eta})}{p(\boldsymbol{\eta}|\mathbf{Y}, \boldsymbol{\theta})}\}] = \mathbb{E}_{\boldsymbol{\eta} \sim q(\boldsymbol{\eta})}[\log\{\frac{p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})}{q(\boldsymbol{\eta})}\}]$, setting the choice of variational distribution as $q(\boldsymbol{\eta}) = p(\boldsymbol{\eta} \mid \mathbf{Y}, \boldsymbol{\theta})$.

Specifically, we set $q(\boldsymbol{\eta}) = p(\boldsymbol{\eta} \mid \mathbf{Y}, \boldsymbol{\theta}^{(t)})$ Maximization of $\log p(\mathbf{Y} \mid \boldsymbol{\theta})$ is equivalent to maximization of $\mathbb{E}_{\boldsymbol{\eta} \sim q(\boldsymbol{\eta})}[\log\{\frac{p(\mathbf{Y}, \boldsymbol{\eta}|\boldsymbol{\theta})}{p(\boldsymbol{\eta}|\mathbf{Y}, \boldsymbol{\theta}^{(t)})}\}] = \mathbb{E}_{\boldsymbol{\eta} \sim p(\boldsymbol{\eta}|\mathbf{Y}, \boldsymbol{\theta}^{(t)})}[\log\{p(\mathbf{Y}, \boldsymbol{\eta} \mid \boldsymbol{\theta})\}] - \mathbb{E}_{\boldsymbol{\eta} \sim p(\boldsymbol{\eta}|\mathbf{Y}, \boldsymbol{\theta}^{(t)})}[\log\{p(\boldsymbol{\eta} \mid \mathbf{Y}, \boldsymbol{\theta}^{(t)})\}] = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) - \mathbb{E}_{\boldsymbol{\eta} \sim p(\boldsymbol{\eta}|\mathbf{Y}, \boldsymbol{\theta}^{(t)})}[\log\{p(\boldsymbol{\eta} \mid \mathbf{Y}, \boldsymbol{\theta}^{(t)})\}]$

$\mathbb{E}_{\boldsymbol{\eta} \sim p(\boldsymbol{\eta}|\mathbf{Y}, \boldsymbol{\theta}^{(t)})}[\log\{p(\boldsymbol{\eta} \mid \mathbf{Y}, \boldsymbol{\theta}^{(t)})\}]$ is the negative entropy, which only relies on $\boldsymbol{\theta}^{(t)}$. Thus, this term will not affect maximization with respect to $\boldsymbol{\theta}$. Hence, maximization of the log-likelihood $\log\{p(\mathbf{Y} \mid \boldsymbol{\theta})\}$ can be done by maximizing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ which is the logic behind the EM algorithm to work.

`http://sia.webpopix.org/EMlme.html` has a great illustration of EM-algorithm for linear mixed model. (which is also attached with this note separately)

It is a special case of minorization-maximization (MM) algorithm [2], which is discussed below. `https://www.stat.berkeley.edu/~aldous/Colloq/lange-talk.pdf` is a good tutorial for interested students.

For maximization of $f(\boldsymbol{\theta})$, minorization-maximization (MM) algorithm is:

- Minorization step: Construct a surrogate function $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ such that $f(\boldsymbol{\theta}) \geq g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ (dominance condition) and $f(\boldsymbol{\theta}^{(t)}) = g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ (tangent condition):
- Maximization step:
$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$
- Ascent property of minorization-maximization algorithm $f(\boldsymbol{\theta}^{(t)}) = g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \leq g(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \leq f(\boldsymbol{\theta}^{(t+1)})$

Figure 1 illustrates this algorithm. EM is a special case of minorization-maximization (MM) algorithm where g-function is set to Q.

For minimization of $f(\boldsymbol{\theta})$, majorization-minimization (MM) algorithm is:

- Majorization step: Construct a surrogate function $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ such that $f(\boldsymbol{\theta}) \leq g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ (dominance condition) and $f(\boldsymbol{\theta}^{(t)}) = g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ (tangent condition):
- Minimization step:
$$\boldsymbol{\theta}^{(t+1)} = \arg\min_{\boldsymbol{\theta}} g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$
- Similarly we have the *descent* property of the majorization-minimization algorithm.

Numerical integration: Our focus is to integrate $f(x)$ with in $[a, b]$ i.e. we want to approximately compute $F(x) = \int_a^b f(x)$. The basic architecture of all the numerical integration methods is to first break the interval $[a, b]$ into several small intervals (usually of equal size). Let $a_1, a_2, \ldots, a_{M+1}$ are $M + 1$ break points such that $a_1 = a$ and $a_{M+1} = b$ with $a_{i+1} - a_i = \delta$ for all $i = 1, \ldots, M$. Then $\delta M = b - a$. Thus, $\delta = \frac{b-a}{M}$.

Firstly, $F(x) = \int_a^b f(x) = \sum_{i=1}^M \int_{a_i}^{a_{i+1}} f(x)$. We approximate each component $\int_{a_i}^{a_{i+1}} f(x)$ of the sum. Different methods replace $f(x)$ with different approximations. Different methods have different approximation errors.

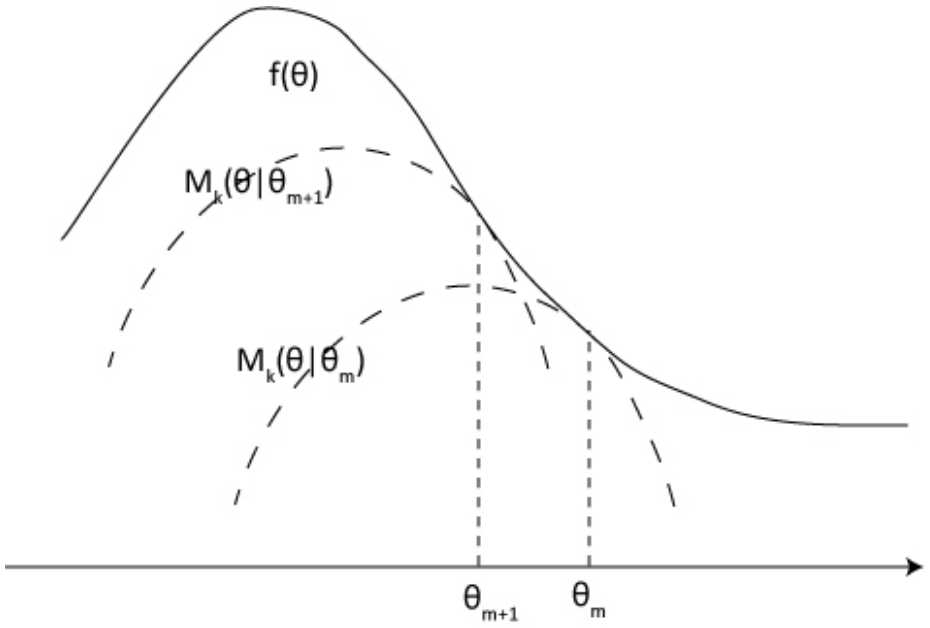

Figure 1: Here to maximize $f(\boldsymbol{\theta})$, the surrogate functions are $M_k(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ in m-th iteration. Figure courtesy: `https://en.wikipedia.org/wiki/MM_algorithm`.

Piece-wise constant or Riemann sum: In this case, the functional value is kept constant at $f(a_i)$ or $f(a_{i+1})$ in each of the intervals $[a_i, a_{i+1}]$. If it is the first one then $F(x) \approx \sum_{i=1}^{M} \delta f(a_i)$.

Trapezoidal rule: In this case, the functional value is kept constant at $\{f(a_i)+f(a_{i+1})\}/2$ in each of the intervals $[a_i, a_{i+1}]$. If it is the first one, then $F(x) \approx \sum_{i=1}^{M} \delta \frac{f(a_i)+f(a_{i+1})}{2}$.

Simpson's 1/3-rd rule: In this case, the function is replaced by a polynomial function $P_i(x)$ separately for each interval, such that $P_i(a_i) = f(a_i)$, $P_i(a_{i+1}) = f(a_{i+1})$ and $P_i(\{a_i + a_{i+1}\}/2) = f(\{a_i + a_{i+1}\}/2)$. Using Lagrange polynomial interpolation, a possible solution is

$$P_i(x) = f(a_i)\frac{(x-m)(x-a_{i+1})}{(a_i-m)(a_i-a_{i+1})} + f(m)\frac{(x-a_i)(x-a_{i+1})}{(m-a_i)(m-a_{i+1})} + f(a_{i+1})\frac{(x-a_i)(x-m)}{(a_{i+1}-a_i)(a_{i+1}-m)},$$

where $m = \{a_i + a_{i+1}\}/2$. By integration by substitution,

$$\int_{a_i}^{a_{i+1}} P_i(x)\,dx = \frac{1}{6}\delta\left[f(a_i) + 4f\left(\frac{a_i+a_{i+1}}{2}\right) + f(a_{i+1})\right],$$

when $a_{i+1} - a_i = \delta$. And subsequently, $F(x) \approx \sum_{i=1}^{M} \int_{a_i}^{a_{i+1}} P_i(x)\,dx = \sum_{i=1}^{M} \delta \frac{f(a_i)+4f(\frac{a_i+a_{i+1}}{2})+f(a_{i+1})}{6}$.

Laplace approximation: This method is used to approximate integrals of the form $\int_a^b e^{Mf(x)}dx$. It's approximation using Laplace's method is given by, $\sqrt{\frac{2\pi}{M|f''(x_0)|}}e^{Mf(x_0)}$ where x_0 is the global maxima of $f(x)$ between (a, b).

Proof for the simpler case is very straightforward. Since, x_0 is the global maxima of $f(x)$ between (a, b), we must have $f'(x_0) = 0$. By Taylor series expansion of $f(x)$ around x_0 is $f(x) \approx f(x_0) + f'(x_0)(x - x_0) - \frac{1}{2}|f''(x_0)|(x - x_0)^2 = f(x_0) - \frac{1}{2}|f''(x_0)|(x - x_0)^2$.

We can integrate $\int \exp(-\frac{1}{2}|f''(x_0)|(x-x_0)^2)dx$ by identifying its similarities with Gaussian density (specifically mean x_0 and variance $1/|f"(x_0)|$). This completes the proof.

Integration by sampling: To motivate integration by sampling, we can go back to our method of moment estimate. We know sample mean is an unbiased estimate of the population mean. Population mean is actually represented by an integration $\int xf(x)dx$, where x is a random variable with probability density $f(x)$. An estimate of that can be obtained by sample mean. Let $x_1, \ldots, x_n$ are n samples of x. Then we can approximate $\int xf(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} x_i$.

In case of normal distribution, we know $\int x \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x-\mu)^2/(2\sigma^2)) = \mu \approx \frac{1}{n}\sum_{i=1}^{n} x_i$, where $x_1, \ldots, x_n$ are samples from Normal(μ, σ^2).

The above approximation holds due to the Strong Law of Large Numbers. We can even compute the approximation errors using Central limit theorem (CLT), which is called confidence interval in the context of parameter estimation. So, the confidence interval can be interpreted as the amount of error in the approximation.

Another type of numerical integration is called 'Monte Carlo integration' which is based on 'Monte Carlo' (read: random) samples. The formulation depends on the specific problem. One of the most popular approaches for this purpose is using the importance sampling scheme.

The main idea: To approximate $\int t(x)P(x \mid \theta)dx$ for some function of x, we can draw samples of x from posterior $P(x \mid \theta)$. Let $x_1, \ldots, x_n$ are n samples of x from posterior $P(x \mid \theta)$. Then for n large enough, $\int t(x)P(x \mid \boldsymbol{\theta})dx \approx \frac{1}{n}\sum_{i=1}^{n} t(x_i)$. This is only applicable for random variables with a pre-specified probability distribution.

References

[1] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[2] Kenneth Lange, David R Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000.