Introduction to Computational Statistics

First concepts in Stochastic Optimization – juliette.chevallier@insa-toulouse.fr

Spring School on Statistics and Machine Learning

plmlab.math.cnrs.fr/chevallier-teaching/hcmus-springschool-computationalstatistics

All materials for the course are available at

plmlab.math.cnrs.fr/chevallier-teaching/hcmus-springschool-computationalstatistics

Motivation: (Bayesian) inference

Given a parametric model and observations "from" this model,

Estimate the parameters that best fit the model to the data

Definition (Estimator)

Given a measurable space $\mathcal{Y}$, and the set of *admissible* parameters Θ

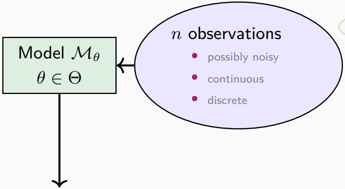
An estimator $\hat{\theta} \colon \mathcal{Y} \to \Theta$ is a function of the data

Example: *Estimation of the mean and variance of* $\mathcal{N}(\mu, \sigma^2)$

- ✓ $\bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i$ estimator of μ
- ✓ $\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$ estimator of σ^2
- × $\frac{1}{n} \sum_{i=1}^{n} (y_i - \mu)^2$ *NOT* an estimator since it depends on μ

Frequentist *vs.* Bayesian inference

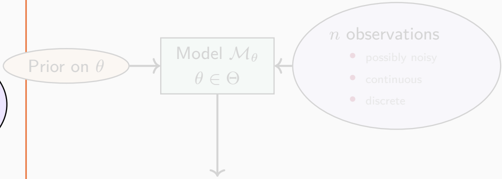
Frequentist



- Estimation of $\hat{\theta}_n$
  $\rightsquigarrow$ Maximum Likelihood Estimation
- Test on θ

When? No idea of possible prior
$\rightsquigarrow$ Better no prior than a wrong one

Bayesian



- Estimation of $\hat{\theta}_n$
  $\rightsquigarrow$ Maximum A Posteriori
- Estimation of $\mathbb{P}(\theta|\mathrm{obs})$

Combine both information, increase knowledge

When?

- When additive information is available, even weakly informative
- When $\dim \theta \geqslant \#\mathrm{observations}$
  $\rightsquigarrow$ Regularization

Frequentist *vs.* Bayesian inference

Frequentist

Model $\mathcal{M}_\theta$

$\theta \in \Theta$

n observations

- possibly noisy
- continuous
- discrete

- Estimation of $\hat{\theta}_n$

  $\rightsquigarrow$ Maximum Likelihood Estimation
- Test on θ

When? No idea of possible prior

$\rightsquigarrow$ Better no prior than a wrong one

Bayesian

Prior on θ

Model $\mathcal{M}_\theta$

$\theta \in \Theta$

n observations

- possibly noisy
- continuous
- discrete

- Estimation of $\hat{\theta}_n$

  $\rightsquigarrow$ Maximum A Posteriori
- Estimation of $\mathbb{P}(\theta|\mathrm{obs})$

Combine both information, increase knowledge

When?

- When additive information is available, even weakly informative
- When $\dim\theta \geqslant \#\mathrm{observations}$

  $\rightsquigarrow$ Regularization

Frequentist *vs.* Bayesian inference

Frequentist

Model $\mathcal{M}_\theta$
$\theta \in \Theta$

n observations
- possibly noisy
- continuous
- discrete

- Estimation of $\hat{\theta}_n$
  $\rightsquigarrow$ Maximum Likelihood Estimation
- Test on θ

When? No idea of possible prior
$\rightsquigarrow$ Better no prior than a wrong one

Bayesian

Prior on θ

Model $\mathcal{M}_\theta$
$\theta \in \Theta$

n observations
- possibly noisy
- continuous
- discrete

- Estimation of $\hat{\theta}_n$
  $\rightsquigarrow$ Maximum A Posteriori
- Estimation of $\mathbb{P}(\theta|\text{obs})$

Combine both information, increase knowledge

When?

- When additive information is available, even weakly informative
- When $\dim \theta \geqslant \#\text{observations}$
  $\rightsquigarrow$ Regularization

Classical Estimators

Some estimators:

1. Method of moments
2. Mean square error $\rightsquigarrow$ *Cf. linear model*
3. Maximum likelihood or Maximum a posteriori (if Bayesian)

- Let $(f_\theta)_{\theta \in \Theta}$ be a family of pdf
- Let $(y_1, y_2, \ldots, y_n)$ be an i.i.d sample with respect to unknown $f_{\theta^*} \in (f_\theta)_{\theta \in \Theta}$

Definition (Likelihood)

- Likelihood (of the observations) $\mathcal{L}_n^{\text{MLE}} = \mathcal{L}_n \quad \longleftrightarrow \quad$ pdf of $y_1, y_2, \ldots, y_n$

$$\mathcal{L}_n : \theta \in \Theta \mapsto \mathcal{L}_n(\theta; y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} f_\theta(y_i)$$

- Posterior likelihood $\mathcal{L}_n^{\text{MAP}} \quad \longleftrightarrow \quad$ pdf of $\theta \,|\, y_1, y_2, \ldots, y_n$

$$\boxed{\hat{\theta}^{\text{MLE}} \in \underset{\theta \in \Theta}{\text{argmax}}\, \mathcal{L}_n^{\text{MLE}}(\theta)} \qquad \boxed{\hat{\theta}^{\text{MAP}} \in \underset{\theta \in \Theta}{\text{argmax}}\, \mathcal{L}_n^{\text{MAP}}(\theta)}$$

Remark: MLE/MAP may not exist, may not be unique

Classical Estimators

Some estimators:

1. Method of moments
2. Mean square error $\rightsquigarrow$ *Cf. linear model*
3. Maximum likelihood or Maximum a posteriori (if Bayesian)

- Let $(f_\theta)_{\theta\in\Theta}$ be a family of pdf
- Let $(y_1, y_2, \ldots, y_n)$ be an i.i.d sample with respect to unknown $f_{\theta^*} \in (f_\theta)_{\theta\in\Theta}$

Definition (Likelihood)

- Likelihood (of the observations) $\mathcal{L}_n^{\mathsf{MLE}} = \mathcal{L}_n \quad\longleftrightarrow\quad$ pdf of $y_1, y_2, \ldots, y_n$

$$\mathcal{L}_n : \; \theta \in \Theta \;\mapsto\; \mathcal{L}_n(\theta; y_1, y_2, \ldots, y_n) \;=\; \prod_{i=1}^{n} f_\theta(y_i)$$

- Posterior likelihood $\mathcal{L}_n^{\mathsf{MAP}} \quad\longleftrightarrow\quad$ pdf of $\theta \,|\, y_1, y_2, \ldots, y_n$

$$\boxed{\hat{\theta}^{\mathsf{MLE}} \in \operatorname*{argmax}_{\theta\in\Theta} \mathcal{L}_n^{\mathsf{MLE}}(\theta)} \qquad \boxed{\hat{\theta}^{\mathsf{MAP}} \in \operatorname*{argmax}_{\theta\in\Theta} \mathcal{L}_n^{\mathsf{MAP}}(\theta)}$$

Remark: MLE/MAP may not exist, may not be unique

Decision Theory

- Several choice for the estimation of θ given a n-sample

? *How can we choose the best one?*

$\rightsquigarrow$ Quantify the quality/goodness-of-fit of the estimators and compare them

$\rightsquigarrow$ Loss function

Definition (Loss function)

Loss $\ell(\theta^*, \hat{\theta}) \equiv$ function which quantifies the difference between $\theta^\star$ and $\hat{\theta}$

- $\ell(a, b) \geqslant 0$
- $\ell(a, b) = \ell(b, a)$
- $\ell(a, a) = 0$
- $\ell(a, b) = 0 \implies a = b$
- $\mathcal{C}^0$, differentiable, triangular inequality

Remark: $\ell(\theta, \hat{\theta}(Y))$ is a random variable

$\rightsquigarrow$ Even if the estimation is excellent, $\ell(\theta, \hat{\theta}(Y))$ may be large

$\rightsquigarrow$ Consider the average loss

Definition (Risk function)

Risk $R(\theta, \hat{\theta}) \equiv$ Average loss: $\boxed{R(\theta, \hat{\theta}) = \mathbb{E}_{Y|\theta}\left[\ell\left(\theta, \hat{\theta}(Y)\right)\right]}$

- Several choice for the estimation of θ given a n-sample

? *How can we choose the best one?*

$\rightsquigarrow$ Quantify the quality/goodness-of-fit of the estimators and compare them

$\rightsquigarrow$ Loss function

Definition (Loss function)

Loss $\ell(\theta^*, \hat{\theta}) \equiv$ function which quantifies the difference between $\theta^\star$ and $\hat{\theta}$

- $\ell(a, b) \geqslant 0$
- $\ell(a, b) = \ell(b, a)$
- $\ell(a, a) = 0$
- $\ell(a, b) = 0 \implies a = b$
- $\mathcal{C}^0$, differentiable, triangular inequality

Remark: $\ell(\theta, \hat{\theta}(Y))$ is a random variable

$\rightsquigarrow$ Even if the estimation is excellent, $\ell(\theta, \hat{\theta}(Y))$ may be large

$\rightsquigarrow$ Consider the average loss

Definition (Risk function)

Risk $R(\theta, \hat{\theta}) \equiv$ Average loss: $\boxed{R(\theta, \hat{\theta}) = \mathbb{E}_{Y|\theta}\left[\ell\left(\theta, \hat{\theta}(Y)\right)\right]}$

Decision Theory

- Several choice for the estimation of θ given a n-sample

? *How can we choose the best one?*

$\rightsquigarrow$ Quantify the quality/goodness-of-fit of the estimators and compare them

$\rightsquigarrow$ Loss function

Definition (Loss function)

Loss $\ell(\theta^*, \hat{\theta}) \equiv$ function which quantifies the difference between $\theta^\star$ and $\hat{\theta}$

- $\ell(a,b) \geqslant 0$
- $\ell(a,b) = \ell(b,a)$
- $\ell(a,a) = 0$
- $\ell(a,b) = 0 \implies a = b$
- $\mathcal{C}^0$, differentiable, triangular inequality

Remark: $\ell(\theta, \hat{\theta}(Y))$ is a random variable

$\rightsquigarrow$ Even if the estimation is excellent, $\ell(\theta, \hat{\theta}(Y))$ may be large

$\rightsquigarrow$ Consider the average loss

Definition (Risk function)

Risk $R(\theta, \hat{\theta}) \equiv$ Average loss: $\boxed{R(\theta, \hat{\theta}) = \mathbb{E}_{Y|\theta}\left[\ell\left(\theta, \hat{\theta}(Y)\right)\right]}$

Let $Y \sim \mathcal{B}in(100, \theta) \quad \theta \in [0, 1]$

Estimate quality measured by quadratic loss

- Naiv estimator: $\hat{\theta}_1 = \dfrac{Y}{100}$

  $\rightsquigarrow$Quadratic loss $\ell_1(\theta, \hat{\theta}_1) = \mathbb{E}\left[\left(\theta - \dfrac{Y}{100}\right)^2\right] = \dfrac{\theta(1-\theta)}{100}$

- Other choice: $\hat{\theta}_2 = \dfrac{Y+3}{100}$

  $\rightsquigarrow$ Quadratic loss $\ell_2(\theta, \hat{\theta}_2) = \mathbb{E}\left[\left(\theta - \dfrac{Y+3}{100}\right)^2\right] = \dfrac{9}{100^2} + \dfrac{\theta(1-\theta)}{100}$

Question: *Which estimator to choose?*

Problem: The choice depends on the *unknown* θ

$\rightsquigarrow$ Consider the area under the curve

Note: It is equivalent to a Bayes uniform prior

$\rightsquigarrow$ If possible, choose a suitable prior!

Let $Y \sim \mathcal{B}in(100, \theta) \quad \theta \in [0,1]$

Estimate quality measured by quadratic loss

- Naiv estimator: $\hat{\theta}_1 = \dfrac{Y}{100}$

  $\rightsquigarrow$Quadratic loss $\ell_1(\theta, \hat{\theta}_1) = \mathbb{E}\left[\left(\theta - \dfrac{Y}{100}\right)^2\right] = \dfrac{\theta(1-\theta)}{100}$

- Other choice: $\hat{\theta}_2 = \dfrac{Y+3}{100}$

  $\rightsquigarrow$ Quadratic loss $\ell_2(\theta, \hat{\theta}_2) = \mathbb{E}\left[\left(\theta - \dfrac{Y+3}{100}\right)^2\right] = \dfrac{9}{100^2} + \dfrac{\theta(1-\theta)}{100}$

Question: *Which estimator to choose?*

Problem: The choice depends on the *unknown* θ

$\rightsquigarrow$ Consider the area under the curve

Note: It is equivalent to a Bayes uniform prior

$\rightsquigarrow$ If possible, choose a suitable prior!

Decision Theory in Practice

Let $Y \sim \mathcal{B}in(100, \theta) \quad \theta \in [0,1]$

Estimate quality measured by quadratic loss

- Naiv estimator: $\hat{\theta}_1 = \frac{Y}{100}$

  $\rightsquigarrow$Quadratic loss $\ell_1(\theta, \hat{\theta}_1) = \mathbb{E}\left[\left(\theta - \frac{Y}{100}\right)^2\right] = \frac{\theta(1-\theta)}{100}$

- Other choice: $\hat{\theta}_2 = \frac{Y+3}{100}$

  $\rightsquigarrow$ Quadratic loss $\ell_2(\theta, \hat{\theta}_2) = \mathbb{E}\left[\left(\theta - \frac{Y+3}{100}\right)^2\right] = \frac{9}{100^2} + \frac{\theta(1-\theta)}{100}$

Question: *Which estimator to choose?*

Problem: The choice depends on the *unknown* θ

$\rightsquigarrow$ Consider the area under the curve

Note: It is equivalent to a Bayes uniform prior

$\rightsquigarrow$ If possible, choose a suitable prior!

Decision Theory in Practice

Let $Y \sim \mathcal{B}in(100, \theta) \quad \theta \in [0,1]$

Estimate quality measured by quadratic loss

- Naiv estimator: $\hat{\theta}_1 = \frac{Y}{100}$

  $\rightsquigarrow$Quadratic loss $\ell_1(\theta, \hat{\theta}_1) = \mathbb{E}\left[\left(\theta - \frac{Y}{100}\right)^2\right] = \frac{\theta(1-\theta)}{100}$

- Other choice: $\hat{\theta}_2 = \frac{Y+3}{100}$

  $\rightsquigarrow$ Quadratic loss $\ell_2(\theta, \hat{\theta}_2) = \mathbb{E}\left[\left(\theta - \frac{Y+3}{100}\right)^2\right] = \frac{9}{100^2} + \frac{\theta(1-\theta)}{100}$

Question: *Which estimator to choose?*

Problem: The choice depends on the *unknown* θ

$\rightsquigarrow$ Consider the area under the curve

Note: It is equivalent to a Bayes uniform prior

$\rightsquigarrow$ If possible, choose a suitable prior!

Let $Y \sim \mathcal{B}in(100, \theta)$ $\theta \in [0, 1]$

Estimate quality measured by quadratic loss

- Naiv estimator: $\hat{\theta}_1 = \dfrac{Y}{100}$

  $\rightsquigarrow$Quadratic loss $\ell_1(\theta, \hat{\theta}_1) = \mathbb{E}\left[\left(\theta - \dfrac{Y}{100}\right)^2\right] = \dfrac{\theta(1-\theta)}{100}$

- Other choice: $\hat{\theta}_2 = \dfrac{Y+3}{100}$

  $\rightsquigarrow$ Quadratic loss $\ell_2(\theta, \hat{\theta}_2) = \mathbb{E}\left[\left(\theta - \dfrac{Y+3}{100}\right)^2\right] = \dfrac{9}{100^2} + \dfrac{\theta(1-\theta)}{100}$

Question: *Which estimator to choose?*

Problem: The choice depends on the *unknown* θ

$\rightsquigarrow$ Consider the area under the curve

Note: It is equivalent to a Bayes uniform prior

$\rightsquigarrow$ If possible, choose a suitable prior!

Note that [computations left to run]

$$\boxed{\begin{aligned}\mathbb{E}_{Y|\theta}\left[\|\hat{\theta}(Y)-\theta\|^2\right] &= \mathbb{E}_{Y|\theta}\left[\left\|\hat{\theta}(Y)-\mathbb{E}_{Y|\theta}\left[\hat{\theta}(Y)\right]\right\|^2\right] + \left\|\mathbb{E}_{Y|\theta}\left[\hat{\theta}(Y)\right]-\theta\right\|^2 \\ &= \mathcal{V}ar\left(\hat{\theta}(Y)\right) + \mathrm{Bias}\left(\hat{\theta}(Y)\right)^2\end{aligned}}$$

Definition (Bayesian risk)

Given a prior π on $\theta \in \Theta$: $\boxed{\hat{R}(\hat{\theta}) = \mathbb{E}_\theta\left[\mathbb{E}_{Y|\theta}\left[\ell\left(\theta, \hat{\theta}(Y)\right)\right]\right]}$

Remark:

- $R(\theta, \hat{\theta}) = \int \ell\left(\theta, \hat{\theta}(Y)\right) \mathbb{P}_\theta(Y)\, \mathrm{d}Y$
- $\hat{R}(\hat{\theta}) = \int\int \ell\left(\theta, \hat{\theta}(Y)\right) \mathbb{P}_\theta(Y)\pi(\theta)\, \mathrm{d}Y\, \mathrm{d}\theta$
- $\hat{R}$ does not depend on θ

Contribution of the Bayesian Framework to Decision Theory

Note that [computations left to run]

$$\boxed{\begin{aligned}\mathbb{E}_{Y|\theta}\left[\|\hat{\theta}(Y)-\theta\|^2\right] &= \mathbb{E}_{Y|\theta}\left[\left\|\hat{\theta}(Y)-\mathbb{E}_{Y|\theta}\left[\hat{\theta}(Y)\right]\right\|^2\right]+\left\|\mathbb{E}_{Y|\theta}\left[\hat{\theta}(Y)\right]-\theta\right\|^2\\ &= \mathcal{V}ar\left(\hat{\theta}(Y)\right)+\mathrm{Bias}\left(\hat{\theta}(Y)\right)^2\end{aligned}}$$

Definition (Bayesian risk)

Given a prior π on $\theta \in \Theta$: $\boxed{\hat{R}(\hat{\theta}) = \mathbb{E}_{\theta}\left[\mathbb{E}_{Y|\theta}\left[\ell\left(\theta, \hat{\theta}(Y)\right)\right]\right]}$

Remark:

- $R(\theta,\hat{\theta}) = \displaystyle\int \ell\left(\theta,\hat{\theta}(Y)\right)\mathbb{P}_{\theta}(Y)\,\mathrm{d}Y$
- $\hat{R}(\hat{\theta}) = \displaystyle\int\int \ell\left(\theta,\hat{\theta}(Y)\right)\mathbb{P}_{\theta}(Y)\pi(\theta)\,\mathrm{d}Y\,\mathrm{d}\theta$
- $\hat{R}$ does not depend on θ

Quadratic Loss and Bayesian Prior

- Consider a discrete distribution, with discrete loss $\ell(\theta, \hat{\theta}) = \mathbb{1}_{\theta \neq \hat{\theta}}$

  Hence: *Bayesian risk*

$$\hat{R}(\hat{\theta}) = \sum_{\theta \in \Theta} \sum_{Y \in \mathcal{Y}} \ell\left(\theta, \hat{\theta}(Y)\right) \mathbb{P}(Y|\theta)\pi(\theta) = \sum_{\theta \in \Theta} \sum_{Y \in \mathcal{Y}} \ell\left(\theta, \hat{\theta}(Y)\right) \mathbb{P}(Y, \theta)$$

  and [computations left to run]

$$\hat{\theta}(Y) \in \operatorname{argmin} \hat{R}(\hat{\theta}) \iff \hat{\theta}(Y) \in \operatorname{argmax} \mathbb{P}(\theta|Y) \rightsquigarrow \text{MAP}$$

- Likewise with L^2 quadratic loss: $\hat{\theta}(Y) = \mathbb{E}[\theta|Y]$ $\rightsquigarrow$ Mean A Posteriori

---

- Previous result generalize well with improper prior
- For estimate, one need to calculate either expectations or maxima, or sample to approximate these quantities (MLE, MAP, Mean a posteriori)

Quadratic Loss and Bayesian Prior

- Consider a discrete distribution, with discrete loss $\ell(\theta, \hat{\theta}) = \mathbb{1}_{\theta \neq \hat{\theta}}$

  Hence: *Bayesian risk*

$$\hat{R}(\hat{\theta}) = \sum_{\theta \in \Theta} \sum_{Y \in \mathcal{Y}} \ell\left(\theta, \hat{\theta}(Y)\right) \mathbb{P}(Y|\theta)\pi(\theta) = \sum_{\theta \in \Theta} \sum_{Y \in \mathcal{Y}} \ell\left(\theta, \hat{\theta}(Y)\right) \mathbb{P}(Y, \theta)$$

  and [computations left to run]

$$\hat{\theta}(Y) \in \operatorname{argmin} \hat{R}(\hat{\theta}) \iff \hat{\theta}(Y) \in \operatorname{argmax} \mathbb{P}(\theta|Y) \rightsquigarrow \text{MAP}$$

- Likewise with L^2 quadratic loss: $\hat{\theta}(Y) = \mathbb{E}[\theta|Y] \rightsquigarrow$ Mean A Posteriori

---

- Previous result generalize well with improper prior
- For estimate, one need to calculate either expectations or maxima, or sample to approximate these quantities (MLE, MAP, Mean a posteriori)

Quadratic Loss and Bayesian Prior

- Consider a discrete distribution, with discrete loss $\ell(\theta, \hat{\theta}) = \mathbb{1}_{\theta \neq \hat{\theta}}$

  Hence: *Bayesian risk*

$$\hat{R}(\hat{\theta}) = \sum_{\theta \in \Theta} \sum_{Y \in \mathcal{Y}} \ell\left(\theta, \hat{\theta}(Y)\right) \mathbb{P}(Y|\theta)\pi(\theta) = \sum_{\theta \in \Theta} \sum_{Y \in \mathcal{Y}} \ell\left(\theta, \hat{\theta}(Y)\right) \mathbb{P}(Y, \theta)$$

  and [computations left to run]

$$\hat{\theta}(Y) \in \operatorname{argmin} \hat{R}(\hat{\theta}) \iff \hat{\theta}(Y) \in \operatorname{argmax} \mathbb{P}(\theta|Y) \rightsquigarrow \text{MAP}$$

- Likewise with L^2 quadratic loss: $\hat{\theta}(Y) = \mathbb{E}[\theta|Y]$ $\rightsquigarrow$ Mean A Posteriori

---

- Previous result generalize well with improper prior
- For estimate, one need to calculate either expectations or maxima, or sample to approximate these quantities (MLE, MAP, Mean a posteriori)

Stochastic Gradient Descent

Reminder of Gradient Descent

Consider the problem $\min_{\theta \in \Theta} J(\theta)$ with

- $\exists\, Y : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathcal{Y}$ random variable
- $\exists\, j : \Theta \times \mathcal{Y} \to \overline{\mathbb{R}}$

and $J(\theta) = \mathbb{E}_Y [j(\theta, Y)]$

$\Longrightarrow$ $\boxed{\min_{\theta \in \Theta} \mathbb{E}\, [j(\theta, Y)]}$

Algorithm 1: Gradient Descent (GD)

Data: $\theta_0 = \theta_{\text{init}} \in \Theta$

Result: Local minimizer $\theta_{\text{end}} \in \Theta$

```
Set k = 0. Compute ∇J(θ_0)
while ||∇J(θ_k)|| ⩾ ε do
    θ_{k+1} = proj_Θ { θ_k − γ ∇J(θ_k) }
    Compute ∇J(θ_{k+1})
return θ_{k+1}
```

- Typically, $j(\theta, Y) := \ell(\theta, \hat{\theta}(Y))$
- Under regularity constraints for J, GD algorithm converges toward local minimum of J
- Other deterministic minimization methods apply (*e.g.* Newton)

  But: All require the calculation ∇J, involving an integral

$\rightsquigarrow$ Very long computation times, especially for high-dimensional Y

$\rightsquigarrow$ **Solution**: Monte Carlo approach to computing $\mathbb{E}$

Consider the problem $\min_{\theta \in \Theta} J(\theta)$ with

- $\exists\, Y : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathcal{Y}$ random variable
- $\exists\, j : \Theta \times \mathcal{Y} \to \overline{\mathbb{R}}$

and $J(\theta) = \mathbb{E}_Y[j(\theta, Y)]$

$\Longrightarrow$ $\boxed{\min_{\theta \in \Theta} \mathbb{E}\left[j(\theta, Y)\right]}$

Algorithm 2: Gradient Descent (GD)

Data: $\theta_0 = \theta_{\text{init}} \in \Theta$

Result: Local minimizer $\theta_{\text{end}} \in \Theta$

Set $k = 0$. Compute $\nabla J(\theta_0)$

while $\underline{\|\nabla J(\theta_k)\| \geqslant \varepsilon}$ **do**

  $\theta_{k+1} = \mathrm{proj}_\Theta \left\{ \theta_k - \gamma \nabla J(\theta_k) \right\}$

  Compute $\nabla J(\theta_{k+1})$

return $\underline{\theta_{k+1}}$

- Typically, $j(\theta, Y) := \ell(\theta, \hat{\theta}(Y))$
- Under regularity constraints for J, GD algorithm converges toward local minimum of J
- Other deterministic minimization methods apply (*e.g.* Newton)

  But: All require the calculation ∇J, involving an integral

$\rightsquigarrow$ Very long computation times, especially for high-dimensional Y

$\rightsquigarrow$ **Solution**: Monte Carlo approach to computing $\mathbb{E}$

Reminder of Gradient Descent

Consider the problem $\min_{\theta \in \Theta} J(\theta)$ with

- $\exists\, Y : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathcal{Y}$ random variable
- $\exists\, j : \Theta \times \mathcal{Y} \to \overline{\mathbb{R}}$

and $J(\theta) = \mathbb{E}_Y[j(\theta, Y)]$

$$\Longrightarrow \quad \boxed{\min_{\theta \in \Theta} \mathbb{E}\left[j(\theta, Y)\right]}$$

Algorithm 3: Gradient Descent (GD)

Data: $\theta_0 = \theta_{\text{init}} \in \Theta$

Result: Local minimizer $\theta_{\text{end}} \in \Theta$

```
Set k = 0. Compute ∇J(θ_0)
while ‖∇J(θ_k)‖ ⩾ ε do
    θ_{k+1} = proj_Θ{ θ_k − γ ∇J(θ_k) }
    Compute ∇J(θ_{k+1})
return θ_{k+1}
```

- Typically, $j(\theta, Y) := \ell(\theta, \hat{\theta}(Y))$
- Under regularity constraints for J, GD algorithm converges toward local minimum of J
- Other deterministic minimization methods apply (*e.g.* Newton)

  But: All require the calculation ∇J, involving an integral

$\rightsquigarrow$ Very long computation times, especially for high-dimensional Y

$\rightsquigarrow$ **Solution**: Monte Carlo approach to computing $\mathbb{E}$

Reminder of Gradient Descent

Consider the problem $\min_{\theta \in \Theta} J(\theta)$ with

- $\exists\, Y : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathcal{Y}$ random variable
- $\exists\, j : \Theta \times \mathcal{Y} \to \overline{\mathbb{R}}$

and $J(\theta) = \mathbb{E}_Y[j(\theta, Y)]$

$\Longrightarrow$ $\boxed{\min_{\theta \in \Theta} \mathbb{E}\,[j(\theta, Y)]}$

Algorithm 4: Gradient Descent (GD)

Data: $\theta_0 = \theta_{\text{init}} \in \Theta$

Result: Local minimizer $\theta_{\text{end}} \in \Theta$

```
Set k = 0. Compute ∇J(θ_0)
while ‖∇J(θ_k)‖ ⩾ ε do
    θ_{k+1} = proj_Θ{ θ_k − γ ∇J(θ_k) }
    Compute ∇J(θ_{k+1})
return θ_{k+1}
```

- Typically, $j(\theta, Y) := \ell(\theta, \hat{\theta}(Y))$
- Under regularity constraints for J, GD algorithm converges toward local minimum of J
- Other deterministic minimization methods apply (*e.g.* Newton)

  But: All require the calculation ∇J, involving an integral

$\rightsquigarrow$ Very long computation times, especially for high-dimensional Y

$\rightsquigarrow$ **Solution**: Monte Carlo approach to computing $\mathbb{E}$

Algorithm 5: Gradient Descent (GD)

Data: $\theta_0 = \theta_{\text{init}} \in \Theta$, sequence $(\gamma_k)_k$

Result: Local minimizer $\theta_{\text{end}} \in \Theta$

```
for k = 0 → maxIter do
    θ_{k+1} = proj_Θ{ θ_k − γ_k ∇J(θ_k) }
return θ_{k+1}
```

$$\theta_{k+1} = \operatorname{proj}_\Theta\left\{\, \theta_k - \gamma_k \, \nabla J(\theta_k) \,\right\}$$

Algorithm 6: Stochastic Gradient Descent (SGD)

Data: $\theta_0 = \theta_{\text{init}} \in \Theta$, `sampler` $\mathbb{P}_Y$, sequence $(\gamma_k)_k$

Result: Local minimizer $\theta_{\text{end}} \in \Theta$

```
for k = 0 → maxIter do
    Sample y_k ~ P_Y
    θ_{k+1} = proj_Θ{ θ_k − γ_k ∂j/∂θ(θ_k, y_{k+1}) }
return θ_{k+1}
```

$$\theta_{k+1} = \operatorname{proj}_\Theta\left\{\, \theta_k - \gamma_k \, \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \,\right\}$$

Stochastic Gradient Descent: Monte Carlo approach to computing $\mathbb{E}$

Algorithm 7: Gradient Descent (GD)

Data: $\theta_0 = \theta_{\text{init}} \in \Theta$, sequence $(\gamma_k)_k$

Result: Local minimizer $\theta_{\text{end}} \in \Theta$

for $k = 0 \to \texttt{maxIter}$ **do**

$\quad \theta_{k+1} = \mathrm{proj}_\Theta \left\{ \theta_k - \gamma_k \nabla J(\theta_k) \right\}$

return θ_{k+1}

Algorithm 8: Stochastic Gradient Descent (SGD)

Data: $\theta_0 = \theta_{\text{init}} \in \Theta$, sampler $\mathbb{P}_Y$, sequence $(\gamma_k)_k$

Result: Local minimizer $\theta_{\text{end}} \in \Theta$

for $k = 0 \to \texttt{maxIter}$ **do**

$\quad$ Sample $y_k \sim \mathbb{P}_Y$

$\quad \theta_{k+1} = \mathrm{proj}_\Theta \left\{ \theta_k - \gamma_k \dfrac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \right\}$

return θ_{k+1}

Intuition behind Stochastic Gradient Decent

1. Consider $k_0 \in \mathbb{N}$
2. Sum the SGD formula k times from k_0

$$\theta_{k+1} = \theta_k - \gamma_k \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \implies \theta_{k_0+k} = \theta_{k_0} - \sum_{\ell=0}^{k-1} \gamma_{k_0+\ell} \frac{\partial j}{\partial \theta}(\theta_{k_0+\ell}, y_{k_0+\ell+1})$$

Assume:

- $\theta \mapsto \dfrac{\partial j}{\partial \theta}(\theta, y)$ sufficiently regular
- $|\theta_{k_0+\ell} - \theta_{k_0+\ell'}|$ small (to be rigorously defined !)

Hence, according to Cesàro mean lemma (2nd line)

$$\theta_{k_0+k} \simeq \theta_{k_0} - \sum_{\ell=0}^{k-1} \gamma_{k_0+\ell} \frac{\partial j}{\partial \theta}(\theta_{k_0}, y_{k_0+\ell+1})$$

$$\simeq \theta_{k_0} - \left(\sum_{\ell=0}^{k-1} \gamma_{k_0+\ell}\right) \nabla J(\theta_{k_0})$$

$\rightsquigarrow$ *One step of GD algorithm!*

Cesàro lemma

Let x_k be a sequence in an Hibert space such that $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x_k = \mu$

Let $(\rho_k)_k$ a sequence of positive numbers $\searrow 0$

Assume that $\varepsilon_k = (\rho_k - \rho_{k+1}) > 0$ and $\sum_k \varepsilon_k$ diverge

Then

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} \rho_k x_k}{\sum_{k=1}^{n} \rho_k} = \mu$$

Intuition behind Stochastic Gradient Decent

1. Consider $k_0 \in \mathbb{N}$
2. Sum the SGD formula k times from k_0

$$\theta_{k+1} = \theta_k - \gamma_k \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \implies \theta_{k_0+k} = \theta_{k_0} - \sum_{\ell=0}^{k-1} \gamma_{k_0+\ell} \frac{\partial j}{\partial \theta}(\theta_{k_0+\ell}, y_{k_0+\ell+1})$$

Assume:

- $\theta \mapsto \dfrac{\partial j}{\partial \theta}(\theta, y)$ sufficiently regular
- $|\theta_{k_0+\ell} - \theta_{k_0+\ell'}|$ small (to be rigorously defined !)

Hence, according to Cesàro mean lemma (2nd line)

$$\theta_{k_0+k} \simeq \theta_{k_0} - \sum_{\ell=0}^{k-1} \gamma_{k_0+\ell} \frac{\partial j}{\partial \theta}(\theta_{k_0}, y_{k_0+\ell+1})$$

$$\simeq \theta_{k_0} - \left(\sum_{\ell=0}^{k-1} \gamma_{k_0+\ell} \right) \nabla J(\theta_{k_0})$$

$\rightsquigarrow$ *One step of GD algorithm!*

Cesàro lemma

Let x_k be a sequence in an Hibert space such that $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x_k = \mu$

Let $(\rho_k)_k$ a sequence of positive numbers $\searrow 0$

Assume that $\varepsilon_k = (\rho_k - \rho_{k+1}) > 0$ and $\sum_k \varepsilon_k$ diverge

Then

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} \rho_k x_k}{\sum_{k=1}^{n} \rho_k} = \mu$$

Intuition behind Stochastic Gradient Decent

1. Consider $k_0 \in \mathbb{N}$
2. Sum the SGD formula k times from k_0

$$\theta_{k+1} = \theta_k - \gamma_k \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \implies \theta_{k_0+k} = \theta_{k_0} - \sum_{\ell=0}^{k-1} \gamma_{k_0+\ell} \frac{\partial j}{\partial \theta}(\theta_{k_0+\ell}, y_{k_0+\ell+1})$$

Assume:

- $\theta \mapsto \frac{\partial j}{\partial \theta}(\theta, y)$ sufficiently regular
- $|\theta_{k_0+\ell} - \theta_{k_0+\ell'}|$ small (to be rigorously defined !)

Hence, according to Cesàro mean lemma (2nd line)

$$\begin{aligned}\theta_{k_0+k} &\simeq \theta_{k_0} - \sum_{\ell=0}^{k-1} \gamma_{k_0+\ell} \frac{\partial j}{\partial \theta}(\theta_{k_0}, y_{k_0+\ell+1}) \\ &\simeq \theta_{k_0} - \left(\sum_{\ell=0}^{k-1} \gamma_{k_0+\ell}\right) \nabla J(\theta_{k_0})\end{aligned}$$

$\rightsquigarrow$ *One step of GD algorithm!*

Cesàro lemma

Let x_k be a sequence in an Hibert space such that $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x_k = \mu$

Let $(\rho_k)_k$ a sequence of positive numbers $\searrow 0$

Assume that $\varepsilon_k = (\rho_k - \rho_{k+1}) > 0$ and $\sum_k \varepsilon_k$ diverge

Then

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} \rho_k x_k}{\sum_{k=1}^{n} \rho_k} = \mu$$

Monte Carlo Sum and Stochastic Gradient

Aim: Compute the expectation $\mathbb{E}[Y]$ of a random variable $Y : \Omega \to \mathbb{R}$, $Y \sim \mu$

- Approximation of $\mathbb{E}[Y]$ by Monte Carlo sum:
  - Let $y_1, \dots, y_k \overset{i.i.d}{\sim} \mu$. Denote $\theta_k = \frac{1}{k}\sum_{\ell=1}^{k} y_\ell$
  - According to law of large numbers, $\theta_k \xrightarrow[k\to\infty]{p.s.} \mathbb{E}[Y]$
- Note that $\frac{1}{k+1} = \frac{1}{k} - \frac{1}{k(k+1)}$. Hence:

$$\theta_{k+1} = \frac{1}{k+1}\sum_{\ell=1}^{k} y_\ell + \frac{y_{k+1}}{k+1} = \frac{1}{k}\sum_{\ell=1}^{k} y_\ell - \frac{1}{k+1}\left(\frac{1}{k}\sum_{\ell=1}^{k} y_\ell - y_{k+1}\right)$$

$$= \theta_k - \frac{1}{k+1}(\theta_k - y_{k+1})$$

$$= \theta_k - \gamma_k \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \quad \text{where} \quad \gamma_k = \frac{1}{k+1} \quad \text{and} \quad j(\theta, y) = \frac{1}{2}(\theta - y)^2$$

- However: Expectation of a random variable $\equiv$ minimum value of the dispersion criterion of a point cloud $\rightsquigarrow \mathbb{E}[Y] = \operatorname{argmin}_{\theta\in\mathbb{R}} \frac{1}{2}(\theta - Y)^2$

$\Longrightarrow$ *Monte Carlo approximation is a stochastic gradient on the dispersion function*

Monte Carlo Sum and Stochastic Gradient

Aim: Compute the expectation $\mathbb{E}[Y]$ of a random variable $Y : \Omega \to \mathbb{R}$, $Y \sim \mu$

- Approximation of $\mathbb{E}[Y]$ by Monte Carlo sum:
  - Let $y_1, \dots, y_k \overset{i.i.d}{\sim} \mu$. Denote $\theta_k = \frac{1}{k}\sum_{\ell=1}^{k} y_\ell$
  - According to law of large numbers, $\theta_k \xrightarrow[k\to\infty]{p.s.} \mathbb{E}[Y]$
- Note that $\frac{1}{k+1} = \frac{1}{k} - \frac{1}{k(k+1)}$. Hence:

$$\begin{aligned}\theta_{k+1} &= \frac{1}{k+1}\sum_{\ell=1}^{k} y_\ell + \frac{y_{k+1}}{k+1} = \frac{1}{k}\sum_{\ell=1}^{k} y_\ell - \frac{1}{k+1}\left(\frac{1}{k}\sum_{\ell=1}^{k} y_\ell - y_{k+1}\right)\\ &= \theta_k - \frac{1}{k+1}(\theta_k - y_{k+1})\\ &= \theta_k - \gamma_k \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \quad \text{where} \quad \gamma_k = \frac{1}{k+1} \quad \text{and} \quad j(\theta, y) = \frac{1}{2}(\theta - y)^2\end{aligned}$$

- However: Expectation of a random variable $\equiv$ minimum value of the dispersion criterion of a point cloud $\rightsquigarrow \mathbb{E}[Y] = \operatorname{argmin}_{\theta\in\mathbb{R}} \frac{1}{2}(\theta - Y)^2$

$\Longrightarrow$ *Monte Carlo approximation is a stochastic gradient on the dispersion function*

Monte Carlo Sum and Stochastic Gradient

Aim: Compute the expectation $\mathbb{E}[Y]$ of a random variable $Y : \Omega \to \mathbb{R}$, $Y \sim \mu$

- Approximation of $\mathbb{E}[Y]$ by Monte Carlo sum:
  - Let $y_1, \ldots, y_k \overset{i.i.d}{\sim} \mu$. Denote $\theta_k = \frac{1}{k}\sum_{\ell=1}^{k} y_\ell$
  - According to law of large numbers, $\theta_k \xrightarrow[k\to\infty]{p.s.} \mathbb{E}[Y]$
- Note that $\frac{1}{k+1} = \frac{1}{k} - \frac{1}{k(k+1)}$. Hence:

$$\begin{aligned}
\theta_{k+1} &= \frac{1}{k+1}\sum_{\ell=1}^{k} y_\ell + \frac{y_{k+1}}{k+1} = \frac{1}{k}\sum_{\ell=1}^{k} y_\ell - \frac{1}{k+1}\left(\frac{1}{k}\sum_{\ell=1}^{k} y_\ell - y_{k+1}\right)\\
&= \theta_k - \frac{1}{k+1}(\theta_k - y_{k+1})\\
&= \theta_k - \gamma_k \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \quad \text{where} \quad \gamma_k = \frac{1}{k+1} \quad \text{and} \quad j(\theta, y) = \frac{1}{2}(\theta - y)^2
\end{aligned}$$

- However: Expectation of a random variable $\equiv$ minimum value of the dispersion criterion of a point cloud $\rightsquigarrow \mathbb{E}[Y] = \operatorname{argmin}_{\theta\in\mathbb{R}} \frac{1}{2}(\theta - Y)^2$

$\Longrightarrow$ *Monte Carlo approximation is a stochastic gradient on the dispersion function*

Monte Carlo Sum and Stochastic Gradient

Aim: Compute the expectation $\mathbb{E}[Y]$ of a random variable $Y : \Omega \to \mathbb{R}$, $Y \sim \mu$

- Approximation of $\mathbb{E}[Y]$ by Monte Carlo sum:
  - Let $y_1, \ldots, y_k \overset{i.i.d}{\sim} \mu$. Denote $\theta_k = \frac{1}{k} \sum_{\ell=1}^{k} y_\ell$
  - According to law of large numbers, $\theta_k \xrightarrow[k \to \infty]{p.s.} \mathbb{E}[Y]$
- Note that $\frac{1}{k+1} = \frac{1}{k} - \frac{1}{k(k+1)}$. Hence:

$$
\begin{aligned}
\theta_{k+1} &= \frac{1}{k+1} \sum_{\ell=1}^{k} y_\ell + \frac{y_{k+1}}{k+1} = \frac{1}{k} \sum_{\ell=1}^{k} y_\ell - \frac{1}{k+1} \left(\frac{1}{k} \sum_{\ell=1}^{k} y_\ell - y_{k+1} \right) \\
&= \theta_k - \frac{1}{k+1} (\theta_k - y_{k+1}) \\
&= \theta_k - \gamma_k \frac{\partial j}{\partial \theta} (\theta_k, y_{k+1}) \quad \text{where} \quad \gamma_k = \frac{1}{k+1} \quad \text{and} \quad j(\theta, y) = \frac{1}{2} (\theta - y)^2
\end{aligned}
$$

- However: Expectation of a random variable $\equiv$ minimum value of the dispersion criterion of a point cloud $\rightsquigarrow \mathbb{E}[Y] = \operatorname{argmin}_{\theta \in \mathbb{R}} \frac{1}{2} (\theta - Y)^2$

$\Longrightarrow$ *Monte Carlo approximation is a stochastic gradient on the dispersion function*

A Few Comments

Remark:

- γ_k converges to 0, but not too fast: $\lim_{k \to +\infty} \gamma_k = 0$, but $\sum_k \gamma_k$ diverges
- Monte Carlo method converges p.s.
  Can we get the equivalent result for the stochastic gradient? Yes !
- There are central limit theorems (CLT) for Monte Carlo sums.
  What about stochastic gradient? Yes !

$\rightsquigarrow$ *Cf. Stochastic Approximation theory*

A Few Comments

Remark:

- γ_k converges to 0, but not too fast: $\lim_{k\to+\infty}\gamma_k = 0$, but $\sum_k \gamma_k$ diverges
- Monte Carlo method converges p.s.
  Can we get the equivalent result for the stochastic gradient? Yes !
- There are central limit theorems (CLT) for Monte Carlo sums.
  What about stochastic gradient? Yes !

$\rightsquigarrow$ *Cf. Stochastic Approximation theory*

Stochastic Gradient Descent

Batch Size

- In all the above, we present a gradient step with one simulation
  But we could also run a Monte Carlo sum on several simulations
  - $\rightsquigarrow$ We refer to batch size

Batch GD	Mini-Batch GD	Stochastic GD
At each iteration, gradient obtained by sampling n realizations $y_{k+1,i}$	*At each iteration,* gradient obtained by sampling $1 \leqslant m \leqslant n$ realizations $y_{k+1,j}$	*At each iteration,* gradient obtained by sampling one realization y_{k+1}
$\theta_{k+1} = \theta_k - \frac{\gamma}{n} \sum_{i=1}^{n} \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1,i})$	$\theta_{k+1} = \theta_k - \frac{\gamma}{m} \sum_{i=1}^{m} \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1,i})$	$\theta_{k+1} = \theta_k - \gamma \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1})$

An epoch corresponds to the simulation of n observations

- Batch gradient descent: 1 iteration per epoch
- Mini-batch gradient descent: m iterations per epoch
- Stochastic gradient descent: n iterations per epoch

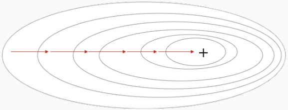
Batch Size

- In all the above, we present a gradient step with one simulation
  But we could also run a Monte Carlo sum on several simulations
  - $\rightsquigarrow$ We refer to batch size

Batch GD

At each iteration, gradient obtained by sampling n realizations $y_{k+1,i}$

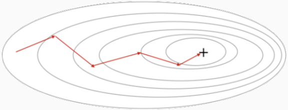
$$\theta_{k+1} = \theta_k - \frac{\gamma}{n} \sum_{i=1}^{n} \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1,i})$$



Mini-Batch GD

At each iteration, gradient obtained by sampling $1 \leqslant m \leqslant n$ realizations $y_{k+1,j}$

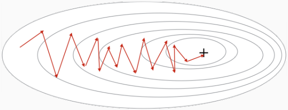
$$\theta_{k+1} = \theta_k - \frac{\gamma}{m} \sum_{i=1}^{m} \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1,i})$$



Stochastic GD

At each iteration, gradient obtained by sampling one realization y_{k+1}

$$\theta_{k+1} = \theta_k - \gamma \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1})$$



An epoch corresponds to the simulation of n observations

- Batch gradient descent: 1 iteration per epoch
- Mini-batch gradient descent: m iterations per epoch
- Stochastic gradient descent: n iterations per epoch

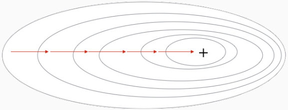
Batch Size

- In all the above, we present a gradient step with one simulation
  But we could also run a Monte Carlo sum on several simulations
  - $\rightsquigarrow$ We refer to batch size

Batch GD

At each iteration, gradient obtained by sampling n realizations $y_{k+1,i}$

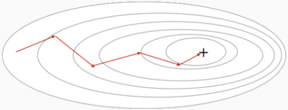
$$\theta_{k+1} = \theta_k - \frac{\gamma}{n} \sum_{i=1}^{n} \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1,i})$$



Mini-Batch GD

At each iteration, gradient obtained by sampling $1 \leqslant m \leqslant n$ realizations $y_{k+1,j}$

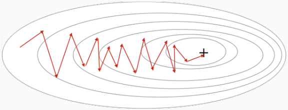
$$\theta_{k+1} = \theta_k - \frac{\gamma}{m} \sum_{i=1}^{m} \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1,i})$$



Stochastic GD

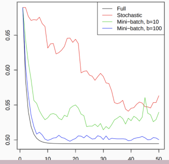
At each iteration, gradient obtained by sampling one realization y_{k+1}

$$\theta_{k+1} = \theta_k - \gamma \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1})$$



An epoch corresponds to the simulation of n observations

- Batch gradient descent: 1 iteration per epoch
- Mini-batch gradient descent: m iterations per epoch
- Stochastic gradient descent: n iterations per epoch



Stop criterion:

- Cannot, as with gradient descent, rely on $\|\theta_{k+1} - \theta_k\|$
  As $\gamma_k \to 0$ and $|\dfrac{\partial j}{\partial \theta}| \leqslant n$, $\|\theta_{k+1} - \theta_k\| \to 0$ by construction
- Nor on $\frac{\partial j}{\partial \theta}$, which is not informative about ∇J
- An approximation of $\mathbb{E}\left[\frac{\partial j}{\partial \theta}\right]$ can be used
- Most often a number is set by the user $\rightsquigarrow$ `maxIter`

Step size:

- In theory, the highest assymptotic speed requires $\gamma_k = \frac{1}{k}$
- But, in practice, $\dfrac{1}{k^\alpha}$, $\alpha \in \left]\frac{1}{2}, 1\right[$ is sometimes better
  $\rightsquigarrow$ It allows $\frac{\partial j}{\partial \theta}$ to have a fairly high weight compared to the previous value.

Finally, we may also require a burn-in time, which consists in removing the first values (considered aberrant).

Stop Criterion & Step Size

Stop criterion:

- Cannot, as with gradient descent, rely on $\|\theta_{k+1} - \theta_k\|$

  As $\gamma_k \to 0$ and $|\dfrac{\partial j}{\partial \theta}| \leqslant n$, $\|\theta_{k+1} - \theta_k\| \to 0$ by construction
- Nor on $\frac{\partial j}{\partial \theta}$, which is not informative about ∇J
- An approximation of $\mathbb{E}\left[\frac{\partial j}{\partial \theta}\right]$ can be used
- Most often a number is set by the user $\rightsquigarrow$ `maxIter`

Step size:

- In theory, the highest assymptotic speed requires $\gamma_k = \frac{1}{k}$
- But, in practice, $\dfrac{1}{k^\alpha}$, $\alpha \in \left]\frac{1}{2}, 1\right[$ is sometimes better

  $\rightsquigarrow$ It allows $\frac{\partial j}{\partial \theta}$ to have a fairly high weight compared to the previous value.

Finally, we may also require a burn-in time, which consists in removing the first values (considered aberrant).

Stop Criterion & Step Size

Stop criterion:

- Cannot, as with gradient descent, rely on $\|\theta_{k+1} - \theta_k\|$

  As $\gamma_k \to 0$ and $|\dfrac{\partial j}{\partial \theta}| \leqslant n$, $\|\theta_{k+1} - \theta_k\| \to 0$ by construction
- Nor on $\frac{\partial j}{\partial \theta}$, which is not informative about ∇J
- An approximation of $\mathbb{E}\left[\frac{\partial j}{\partial \theta}\right]$ can be used
- Most often a number is set by the user $\rightsquigarrow$ `maxIter`

Step size:

- In theory, the highest assymptotic speed requires $\gamma_k = \frac{1}{k}$
- But, in practice, $\dfrac{1}{k^\alpha}$, $\alpha \in \left]\frac{1}{2}, 1\right[$ is sometimes better

  $\rightsquigarrow$ It allows $\frac{\partial j}{\partial \theta}$ to have a fairly high weight compared to the previous value.

Finally, we may also require a burn-in time, which consists in removing the first values (considered aberrant).

Stochastic Gradient Descent

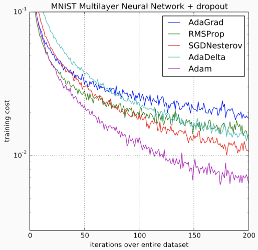
- Due to its low numerical cost, stochastic gradient is widely used in (very) large parametric models, such as neural networks

  $\rightsquigarrow$ Few possible/necessary adjustments to make SGD more practical in this context

- In high-dimensional parameter spaces, the topology of the objective function makes gradient descent difficult or even inefficient
  - Many local minima
  - Each "small" gradient $\frac{\partial j}{\partial \theta}$ is uninformative compared to the "large" gradient ∇J
  - ...

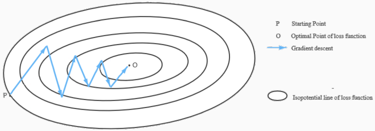
- Some suitable optimizers:

  NAG, AdaGrad, Adadelta, RMSProp, Adam, AdaMax, Nadam, AMSGrad, AdamW, QHAdam, YellowFin, AggMo, QHM, Demon Adam, *etc.*



Momentum

- The SGD updates the parameters after viewing only a subset of the training set
  ⇝ Add an inertia or momentum term:
  - Reduces variance
  - Limits oscillations along the convergence path
  - Avoids getting stuck too easily in a local minimum



In practice: we adapt the SGD algorithm to take account of previous gradients and smooth the update

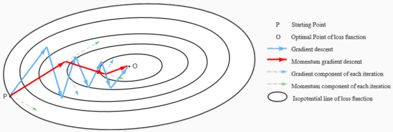
$$\theta_{k+1} = \theta_k - \gamma \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \quad \rightsquigarrow \quad \begin{cases} v_{k+1} = \beta v_k - \gamma \dfrac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \\ \theta_{k+1} = \theta_k + v_k \end{cases}$$

- Velocity v: Direction in which parameters will be modified
- $\beta \in]0,1[$ quantifies the relative importance of previous gradients compared to the current one

Note: $\beta \simeq 0.9$ in general

Momentum

- The SGD updates the parameters after viewing only a subset of the training set
  $\rightsquigarrow$ Add an inertia or momentum term:
  - Reduces variance
  - Limits oscillations along the convergence path
  - Avoids getting stuck too easily in a local minimum



In practice: we adapt the SGD algorithm to take account of previous gradients and smooth the update

$$\theta_{k+1} = \theta_k - \gamma \frac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \quad \rightsquigarrow \quad \begin{cases} v_{k+1} = \beta v_k - \gamma \dfrac{\partial j}{\partial \theta}(\theta_k, y_{k+1}) \\ \theta_{k+1} = \theta_k + v_k \end{cases}$$

- Velocity v: Direction in which parameters will be modified
- $\beta \in]0, 1[$ quantifies the relative importance of previous gradients compared to the current one

Note: $\beta \simeq 0.9$ in general

SGD Enhancement Example: `AdaGrad`

- `AdaGrad` introduces a form of learning rate adaptation by accumulating the squares of the previous gradients

- Balance the power of gradients:
  - Smaller updates, *i.e.* low learning rates, for parameters associated with high gradients
  - And larger updates, *i.e.* high learning rates, for parameters associated with low gradients

Implementation:

1. Compute the gradient $g_{k+1} = \dfrac{\partial j}{\partial \theta}(\theta_k, y_{k+1})$
2. Gradient accumulation $r_{k+1} = r_k + \|g_{k+1}\|$
3. Parameter update $\theta_{k+1} = \theta_k - \dfrac{1}{\sqrt{r_{k+1}}} g_{k+1}$

- `RMSProp` is almost identical to `AdaGrad`, but the impact of older gradients is altered by a multiplicative coefficient $\rho \in]0, 1[$ (weight decay)

Implementation:

1. Compute the gradient $g_{k+1} = \dfrac{\partial j}{\partial \theta}(\theta_k, y_{k+1})$
2. Gradient accumulation $r_{k+1} = \rho\, r_k + (1-\rho)\|g_{k+1}\|$
3. Parameter update $\theta_{k+1} = \theta_k - \dfrac{1}{\sqrt{r_{k+1}}}\, g_{k+1}$

- `Adam` is similar to `RMSProp`, but also adapts momentum

EM Algorithm and Variants

Framework: Latent Variables Model

- **Aim**: Given a latent variables model, find the MLE (or MAP if Bayesian)

$$\theta^* \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, q(y\,;\,\theta) \;=\; \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathbb{E}_Z\left[q(y, z\,;\,\theta)\right]$$

- **Model**: Latent or Hierarchical model

Observations: $Y \leftrightarrow y = (y_i)_{i \in [\![1,n]\!]} \in \mathcal{Y}$

$Y_i \,|\, Z_i\,;\,\theta \sim q(y_i|z_i\,;\,\theta)$

Latent variables: $Z \leftrightarrow z = (z_i)_{i \in [\![1,n]\!]} \in \mathcal{Z}$

$Z_i\,;\,\theta \sim q(z_i\,;\,\theta)$

$$\begin{cases} Y_i \,|\, Z_i\,;\,\theta \sim q(y_i|z_i\,;\,\theta) \\ Z_i\,;\,\theta \sim q(z_i\,;\,\theta) \\ \theta \sim q_{\text{prior}}(\theta) \end{cases}$$

Parameter: $\theta \in \Theta$, θ set of admissible parameters

If Bayesian framework, $\theta \sim q_{\text{prior}}(\theta)$

- **Remarks**
  - We observe only y
  - The law of Z may depend on θ
  - Why EM instead of GDS?
    To avoid the computation of the expectation of the gradient *(See in few minutes)*

- **Aim**: Given a latent variables model, find the MLE (or MAP if Bayesian)

$$\theta^* \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, q(y\,;\,\theta) \;=\; \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathbb{E}_Z\left[q(y, z\,;\,\theta)\right]$$

- **Model**: Latent or Hierarchical model

Observations: $Y \leftrightarrow y = (y_i)_{i \in [\![1,n]\!]} \in \mathcal{Y}$

$Y_i \mid Z_i\,;\,\theta \;\sim\; q(y_i|z_i\,;\,\theta)$

Latent variables: $Z \leftrightarrow z = (z_i)_{i \in [\![1,n]\!]} \in \mathcal{Z}$

$Z_i\,;\,\theta \;\sim\; q(z_i\,;\,\theta)$

$$\begin{cases} Y_i \mid Z_i\,;\,\theta \;\sim\; q(y_i|z_i\,;\,\theta) \\ Z_i\,;\,\theta \;\sim\; q(z_i\,;\,\theta) \\ \theta \sim q_{\text{prior}}(\theta) \end{cases}$$

Parameter: $\theta \in \Theta$, θ set of admissible parameters

If Bayesian framework, $\theta \sim q_{\text{prior}}(\theta)$

- **Remarks**
  - We observe only y
  - The law of Z may depend on θ
  - Why EM instead of GDS?
    To avoid the computation of the expectation of the gradient *(See in few minutes)*

Framework: Latent Variables Model

- **Aim**: Given a latent variables model, find the MLE (or MAP if Bayesian)

$$\boxed{\theta^* \in \operatorname*{argmax}_{\theta \in \Theta} q(y\,;\,\theta) \;=\; \operatorname*{argmax}_{\theta \in \Theta} \mathbb{E}_Z\left[q(y, z\,;\,\theta)\right]}$$

- **Model**: Latent or Hierarchical model

Observations: $Y \leftrightarrow y = (y_i)_{i \in [\![1,n]\!]} \in \mathcal{Y}$

$Y_i \mid Z_i\,;\,\theta \;\sim\; q(y_i|z_i\,;\,\theta)$

Latent variables: $Z \leftrightarrow z = (z_i)_{i \in [\![1,n]\!]} \in \mathcal{Z}$

$Z_i\,;\,\theta \;\sim\; q(z_i\,;\,\theta)$

$$\boxed{\begin{cases} Y_i \mid Z_i\,;\,\theta \;\sim\; q(y_i|z_i\,;\,\theta) \\ Z_i\,;\,\theta \;\sim\; q(z_i\,;\,\theta) \\ \theta \sim q_{\text{prior}}(\theta) \end{cases}}$$

Parameter: $\theta \in \Theta$, θ set of admissible parameters

If Bayesian framework, $\theta \sim q_{\text{prior}}(\theta)$

- **Remarks**
  - We observe only y
  - The law of Z may depend on θ
  - Why EM instead of GDS?
    To avoid the computation of the expectation of the gradient *(See in few minutes)*

Intuitions behind the EM Algorithm

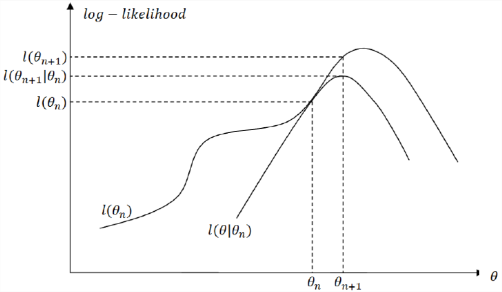
We seek $\theta^* \in \mathrm{argmax}_{\theta\in\Theta}\, q(y\,;\,\theta)$, equivalent to $\theta^* \in \mathrm{argmax}_{\theta\in\Theta} \log q(y\,;\,\theta)$

But: Hard to derive due to the $\mathbb{E}_Z \equiv \int_{\mathcal{Z}}$

An idea: Construct a sequence $(\theta_k)_k$ from an initial point θ_0 and

1. Find a function B such that $B(\theta|\theta_k) \leqslant \log q(y\,;\,\theta)$ for all θ
2. Maximize $B(\cdot|\theta_k)$ instead of $\log q(y\,;\,\theta)$

$\rightsquigarrow$ As soon as B is “optimal”, repeating step 1 & 2, we may hope a convergence toward local minima of q



Intuitions behind the EM Algorithm

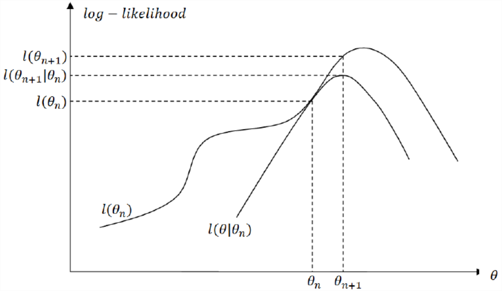
We seek $\theta^* \in \operatorname{argmax}_{\theta \in \Theta} q(y\,;\,\theta)$, equivalent to $\theta^* \in \operatorname{argmax}_{\theta \in \Theta} \log q(y\,;\,\theta)$

But: Hard to derive due to the $\mathbb{E}_Z \equiv \int_{\mathcal{Z}}$

An idea: Construct a sequence $(\theta_k)_k$ from an initial point θ_0 and

1. Find a function B such that $B(\theta|\theta_k) \leqslant \log q(y\,;\,\theta)$ for all θ
2. Maximize $B(\cdot|\theta_k)$ instead of $\log q(y\,;\,\theta)$

$\rightsquigarrow$ As soon as B is “optimal”, repeating step 1 & 2, we may hope a convergence toward local minima of q



Construction of the Minimizer B (1/2)

- Let f_k be a density function, which may depend on the current value of θ

$$\log q(y\,;\theta) = \log \int_{\mathcal{Z}} q(y, z\,;\theta)\,\mathrm{d}z = \log \int_{\mathcal{Z}} \frac{q(y, z\,;\theta)}{f_k(z)} f_k(z)\,\mathrm{d}z$$

- According to the Jensen inequality,

$$\log \int_{\mathcal{Z}} \frac{q(y, z\,;\theta)}{f_k(z)} f_k(z)\,\mathrm{d}z \leqslant \int_{\mathcal{Z}} \log\left(\frac{q(y, z\,;\theta)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z$$

- Since the Jensen convex inequality is optimal (we can reach the equality), we set

$$\boxed{B(\theta|\theta_k) = \int_{\mathcal{Z}} \log\left(\frac{q(y, z\,;\theta)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z}$$

- **Aim**: Find the optimal density f_k, *i.e.* the one that "best" minimize $\log q$
  - We seek f_k such that $B(\theta_k|\theta_k)$ "touch" $\log q(y\,;\theta_k)$
  - *But*: $B(\cdot|\theta_k)$ minimizes $\log q(y\,;\cdot)$, so "touching" in θ_k is equivalent for $B(\cdot|\theta_k)$ to be maximal in θ_k $\rightsquigarrow$ Find f_k such that $B(\theta_k|\theta_k)$ is maximal

$$\Longrightarrow\ f_k^* \in \underset{f_k}{\operatorname{argmax}} \int_{\mathcal{Z}} \log\left(\frac{q(y, z\,;\theta_k)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z \quad \text{s.t.} \quad f_k \geqslant 0 \ \ \& \ \int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z = 1$$

Construction of the Minimizer B (1/2)

- Let f_k be a density function, which may depend on the current value of θ

$$\log q(y\,;\theta) \;=\; \log \int_{\mathcal{Z}} q(y,z\,;\theta)\,\mathrm{d}z \;=\; \log \int_{\mathcal{Z}} \frac{q(y,z\,;\theta)}{f_k(z)} f_k(z)\,\mathrm{d}z$$

- According to the Jensen inequality,

$$\log \int_{\mathcal{Z}} \frac{q(y,z\,;\theta)}{f_k(z)} f_k(z)\,\mathrm{d}z \leqslant \int_{\mathcal{Z}} \log\left(\frac{q(y,z\,;\theta)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z$$

- Since the Jensen convex inequality is optimal (we can reach the equality), we set

$$\boxed{B(\theta|\theta_k) = \int_{\mathcal{Z}} \log\left(\frac{q(y,z\,;\theta)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z}$$

- **Aim**: Find the optimal density f_k, *i.e.* the one that "best" minimize $\log q$
  - We seek f_k such that $B(\theta_k|\theta_k)$ "touch" $\log q(y\,;\theta_k)$
  - *But*: $B(\cdot|\theta_k)$ minimizes $\log q(y\,;\cdot)$, so "touching" in θ_k is equivalent for $B(\cdot|\theta_k)$ to be maximal in θ_k $\rightsquigarrow$ Find f_k such that $B(\theta_k|\theta_k)$ is maximal

$$\Longrightarrow\; f_k^* \in \underset{f_k}{\operatorname{argmax}} \int_{\mathcal{Z}} \log\left(\frac{q(y,z\,;\theta_k)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z \;\text{ s.t. }\; f_k \geqslant 0 \;\&\; \int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z = 1$$

- Let f_k be a density function, which may depend on the current value of θ

$$\log q(y\,;\theta) \;=\; \log \int_{\mathcal{Z}} q(y, z\,;\theta)\,\mathrm{d}z \;=\; \log \int_{\mathcal{Z}} \frac{q(y, z\,;\theta)}{f_k(z)} f_k(z)\,\mathrm{d}z$$

- According to the Jensen inequality,

$$\log \int_{\mathcal{Z}} \frac{q(y, z\,;\theta)}{f_k(z)} f_k(z)\,\mathrm{d}z \leqslant \int_{\mathcal{Z}} \log\left(\frac{q(y, z\,;\theta)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z$$

- Since the Jensen convex inequality is optimal (we can reach the equality), we set

$$\boxed{B(\theta|\theta_k) = \int_{\mathcal{Z}} \log\left(\frac{q(y, z\,;\theta)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z}$$

- **Aim**: Find the optimal density f_k, *i.e.* the one that "best" minimize $\log q$
  - We seek f_k such that $B(\theta_k|\theta_k)$ "touch" $\log q(y\,;\theta_k)$
  - *But*: $B(\cdot|\theta_k)$ minimizes $\log q(y\,;\cdot)$, so "touching" in θ_k is equivalent for $B(\cdot|\theta_k)$ to be maximal in θ_k $\rightsquigarrow$ Find f_k such that $B(\theta_k|\theta_k)$ is maximal

$$\Longrightarrow\; f_k^* \in \operatorname*{argmax}_{f_k} \int_{\mathcal{Z}} \log\left(\frac{q(y, z\,;\theta_k)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z \;\text{ s.t. }\; f_k \geqslant 0 \;\;\&\; \int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z = 1$$

Construction of the Minimizer B (1/2)

- Let f_k be a density function, which may depend on the current value of θ

$$\log q(y\,;\theta) \;=\; \log \int_{\mathcal{Z}} q(y,z\,;\theta)\,\mathrm{d}z \;=\; \log \int_{\mathcal{Z}} \frac{q(y,z\,;\theta)}{f_k(z)} f_k(z)\,\mathrm{d}z$$

- According to the Jensen inequality,

$$\log \int_{\mathcal{Z}} \frac{q(y,z\,;\theta)}{f_k(z)} f_k(z)\,\mathrm{d}z \leqslant \int_{\mathcal{Z}} \log\left(\frac{q(y,z\,;\theta)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z$$

- Since the Jensen convex inequality is optimal (we can reach the equality), we set

$$\boxed{B(\theta|\theta_k) = \int_{\mathcal{Z}} \log\left(\frac{q(y,z\,;\theta)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z}$$

- **Aim**: Find the optimal density f_k, *i.e.* the one that "best" minimize $\log q$
  - We seek f_k such that $B(\theta_k|\theta_k)$ "touch" $\log q(y\,;\theta_k)$
  - *But*: $B(\cdot|\theta_k)$ minimizes $\log q(y\,;\cdot)$, so "touching" in θ_k is equivalent for $B(\cdot|\theta_k)$ to be maximal in θ_k $\rightsquigarrow$ Find f_k such that $B(\theta_k|\theta_k)$ is maximal

$$\Longrightarrow\; f_k^* \in \underset{f_k}{\operatorname{argmax}} \int_{\mathcal{Z}} \log\left(\frac{q(y,z\,;\theta_k)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z \;\; \text{s.t.} \;\; f_k \geqslant 0 \;\;\&\;\; \int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z = 1$$

Construction of the Minimizer B (1/2)

- Let f_k be a density function, which may depend on the current value of θ

$$\log q(y\,;\theta) \;=\; \log \int_{\mathcal{Z}} q(y,z\,;\theta)\,\mathrm{d}z \;=\; \log \int_{\mathcal{Z}} \frac{q(y,z\,;\theta)}{f_k(z)} f_k(z)\,\mathrm{d}z$$

- According to the Jensen inequality,

$$\log \int_{\mathcal{Z}} \frac{q(y,z\,;\theta)}{f_k(z)} f_k(z)\,\mathrm{d}z \leqslant \int_{\mathcal{Z}} \log\left(\frac{q(y,z\,;\theta)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z$$

- Since the Jensen convex inequality is optimal (we can reach the equality), we set

$$\boxed{B(\theta|\theta_k) = \int_{\mathcal{Z}} \log\left(\frac{q(y,z\,;\theta)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z}$$

- **Aim**: Find the optimal density f_k, *i.e.* the one that "best" minimize $\log q$
  - We seek f_k such that $B(\theta_k|\theta_k)$ "touch" $\log q(y\,;\theta_k)$
  - *But*: $B(\cdot|\theta_k)$ minimizes $\log q(y\,;\cdot)$, so "touching" in θ_k is equivalent for $B(\cdot|\theta_k)$ to be maximal in θ_k $\rightsquigarrow$ Find f_k such that $B(\theta_k|\theta_k)$ is maximal

$$\Longrightarrow\; f_k^* \in \underset{f_k}{\operatorname{argmax}} \int_{\mathcal{Z}} \log\left(\frac{q(y,z\,;\theta_k)}{f_k(z)}\right) f_k(z)\,\mathrm{d}z \;\text{ s.t. }\; f_k \geqslant 0 \;\;\&\;\; \int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z = 1$$

"Reminder" on Functional differential

- **Lagrangian multiplicator** $G(f_k, \lambda) = \lambda\Big(1 - G_1(f_k)\Big) + G_2(f_k) - G_3(f_k)$, with
  - $G_1(f_k) = \displaystyle\int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z$
  - $G_2(f_k) = \displaystyle\int_{\mathcal{Z}} \log q(y, z\,; \theta_k) f_k(z)\,\mathrm{d}z$
  - $G_3(f_k) = \displaystyle\int_{\mathcal{Z}} \log f_x(z) f_k(z)\,\mathrm{d}z$

[Reminder ?] Functional differential/Directional derivative

- Let $F(f)$ be a functional defined from a function f
- Let ϕ be a test function sufficiently regular

⇝ *Differential of F in f in the direction ϕ:*

$$dF[f, \phi] = \int_{\mathcal{X}} \frac{\partial F}{\partial f}(x)\phi(x)\,\mathrm{d}x$$

$$dF[f, \phi] = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\left(F(f + \varepsilon\phi) - F(f)\right) = \left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} F(f + \varepsilon\phi)\right]$$

"Reminder" on Functional differential

- **Lagrangian multiplicator** $G(f_k, \lambda) = \lambda\big(1 - G_1(f_k)\big) + G_2(f_k) - G_3(f_k)$, with
  - $G_1(f_k) = \displaystyle\int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z$
  - $G_2(f_k) = \displaystyle\int_{\mathcal{Z}} \log q(y, z\,; \theta_k) f_k(z)\,\mathrm{d}z$
  - $G_3(f_k) = \displaystyle\int_{\mathcal{Z}} \log f_x(z) f_k(z)\,\mathrm{d}z$

[Reminder ?] Functional differential/Directional derivative

- Let $F(f)$ be a functional defined from a function f
- Let ϕ be a test function sufficiently regular

$\rightsquigarrow$ *Differential of F in f in the direction ϕ:*

$$\boxed{\begin{aligned} dF[f,\phi] &= \int_{\mathcal{X}} \frac{\partial F}{\partial f}(x)\phi(x)\,\mathrm{d}x \\ dF[f,\phi] &= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\left(F(f+\varepsilon\phi) - F(f)\right) = \left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} F(f+\varepsilon\phi)\right] \end{aligned}}$$

Construction of the Minimizer B (2/2)

- **Lagrangian multiplicator** $G(f_k, \lambda) = \lambda(1 - G_1(f_k)) + G_2(f_k) - G_3(f_k)$, with

$$G_1(f_k) = \int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z \qquad G_2(f_k) = \int_{\mathcal{Z}} \log q(y, z\,; \theta_k) f_k(z)\,\mathrm{d}z \qquad G_3(f_k) = \int_{\mathcal{Z}} \log f_x(z) f_k(z)\,\mathrm{d}z$$

$$\begin{cases} \dfrac{\partial G}{\partial f_t}(z) \;=\; -\lambda + \log q(y, z\,; \theta_k) - \log f_x(z) - 1 & (1) \\ \dfrac{\partial G}{\partial \lambda}(z) \;=\; 1 - \displaystyle\int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z & (2) \end{cases}$$

$$(1) \iff \log f_k(z) = \log q(y, z\,; \theta_k) - \lambda - 1 \iff f_k(z) = \mathrm{e}^{-\lambda-1} q(y, z\,; \theta_k)$$

$$(2) \iff 1 = \int_{\mathcal{Z}} \mathrm{e}^{-\lambda-1} q(y, z\,; \theta_k)\,\mathrm{d}z = \mathrm{e}^{-\lambda-1} q(y\,; \theta_k) \iff \mathrm{e}^{-\lambda-1} = \frac{1}{q(y\,; \theta_k)}$$

$$\implies f_k(z) = \frac{1}{q(y\,; \theta_k)} \times q(y, z\,; \theta_k) = q(z|y\,; \theta_k)$$

- We **check** that $B(\theta_k|\theta_k) = q(y\,; \theta_k)$

$$B(\theta_k|\theta_k) = \int_{\mathcal{Z}} \log\left(\frac{q(y, z\,; \theta_k)}{f_k(z)}\right) q(z|y\,; \theta_k)\,\mathrm{d}z = [\ldots] = q(y\,; \theta_k)$$

Construction of the Minimizer B (2/2)

- **Lagrangian multiplicator** $G(f_k, \lambda) = \lambda\,(1 - G_1(f_k)) + G_2(f_k) - G_3(f_k)$, with

$$G_1(f_k) = \int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z \quad G_2(f_k) = \int_{\mathcal{Z}} \log q(y, z\,;\theta_k) f_k(z)\,\mathrm{d}z \quad G_3(f_k) = \int_{\mathcal{Z}} \log f_x(z) f_k(z)\,\mathrm{d}z$$

$$\begin{cases} \dfrac{\partial G}{\partial f_t}(z) \;=\; -\lambda + \log q(y, z\,;\theta_k) - \log f_x(z) - 1 & (1) \\ \dfrac{\partial G}{\partial \lambda}(z) \;=\; 1 - \displaystyle\int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z & (2) \end{cases}$$

$$(1) \iff \log f_k(z) = \log q(y, z\,;\theta_k) - \lambda - 1 \iff f_k(z) = \mathrm{e}^{-\lambda-1}\, q(y, z\,;\theta_k)$$

$$(2) \iff 1 = \int_{\mathcal{Z}} \mathrm{e}^{-\lambda-1}\, q(y, z\,;\theta_k)\,\mathrm{d}z = \mathrm{e}^{-\lambda-1}\, q(y\,;\theta_k) \iff \mathrm{e}^{-\lambda-1} = \frac{1}{q(y\,;\theta_k)}$$

$$\implies f_k(z) = \frac{1}{q(y\,;\theta_k)} \times q(y, z\,;\theta_k) = q(z|y\,;\theta_k)$$

- We **check** that $B(\theta_k|\theta_k) = q(y\,;\theta_k)$

$$B(\theta_k|\theta_k) = \int_{\mathcal{Z}} \log\left(\frac{q(y, z\,;\theta_k)}{f_k(z)}\right) q(z|y\,;\theta_k)\,\mathrm{d}z \;=\; [\ldots] \;=\; q(y\,;\theta_k)$$

- **Lagrangian multiplicator** $G(f_k,\lambda) = \lambda\,(1 - G_1(f_k)) + G_2(f_k) - G_3(f_k)$, with

$$G_1(f_k) = \int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z \qquad G_2(f_k) = \int_{\mathcal{Z}} \log q(y,z\,;\theta_k) f_k(z)\,\mathrm{d}z \qquad G_3(f_k) = \int_{\mathcal{Z}} \log f_x(z) f_k(z)\,\mathrm{d}z$$

$$\begin{cases} \dfrac{\partial G}{\partial f_t}(z) \;=\; -\lambda + \log q(y,z\,;\theta_k) - \log f_x(z) - 1 & (1) \\[2ex] \dfrac{\partial G}{\partial \lambda}(z) \;=\; 1 - \displaystyle\int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z & (2) \end{cases}$$

$$(1) \iff \log f_k(z) = \log q(y,z\,;\theta_k) - \lambda - 1 \iff f_k(z) = \mathrm{e}^{-\lambda-1}\, q(y,z\,;\theta_k)$$

$$(2) \iff 1 = \int_{\mathcal{Z}} \mathrm{e}^{-\lambda-1}\, q(y,z\,;\theta_k)\,\mathrm{d}z = \mathrm{e}^{-\lambda-1}\, q(y\,;\theta_k) \iff \mathrm{e}^{-\lambda-1} = \frac{1}{q(y\,;\theta_k)}$$

$$\implies f_k(z) = \frac{1}{q(y\,;\theta_k)} \times q(y,z\,;\theta_k) = q(z|y\,;\theta_k)$$

- We **check** that $B(\theta_k|\theta_k) = q(y\,;\theta_k)$

$$B(\theta_k|\theta_k) = \int_{\mathcal{Z}} \log\left(\frac{q(y,z\,;\theta_k)}{f_k(z)}\right) q(z|y\,;\theta_k)\,\mathrm{d}z = [\ldots] = q(y\,;\theta_k)$$

Construction of the Minimizer B (2/2)

- **Lagrangian multiplicator** $G(f_k, \lambda) = \lambda\,(1 - G_1(f_k)) + G_2(f_k) - G_3(f_k)$, with

$$G_1(f_k) = \int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z \qquad G_2(f_k) = \int_{\mathcal{Z}} \log q(y, z\,; \theta_k) f_k(z)\,\mathrm{d}z \qquad G_3(f_k) = \int_{\mathcal{Z}} \log f_x(z) f_k(z)\,\mathrm{d}z$$

$$\begin{cases} \dfrac{\partial G}{\partial f_t}(z) \; = \; -\lambda + \log q(y, z\,; \theta_k) - \log f_x(z) - 1 & (1) \\[2ex] \dfrac{\partial G}{\partial \lambda}(z) \; = \; 1 - \displaystyle\int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z & (2) \end{cases}$$

$$(1) \iff \log f_k(z) = \log q(y, z\,; \theta_k) - \lambda - 1 \iff f_k(z) = \mathrm{e}^{-\lambda-1}\, q(y, z\,; \theta_k)$$

$$(2) \iff 1 = \int_{\mathcal{Z}} \mathrm{e}^{-\lambda-1}\, q(y, z\,; \theta_k)\,\mathrm{d}z = \mathrm{e}^{-\lambda-1}\, q(y\,; \theta_k) \iff \mathrm{e}^{-\lambda-1} = \frac{1}{q(y\,; \theta_k)}$$

$$\implies f_k(z) = \frac{1}{q(y\,; \theta_k)} \times q(y, z\,; \theta_k) = q(z|y\,; \theta_k)$$

- We **check** that $B(\theta_k|\theta_k) = q(y\,; \theta_k)$

$$B(\theta_k|\theta_k) = \int_{\mathcal{Z}} \log\left(\frac{q(y, z\,; \theta_k)}{f_k(z)}\right) q(z|y\,; \theta_k)\,\mathrm{d}z \; = \; [\ldots] \; = \; q(y\,; \theta_k)$$

Construction of the Minimizer B (2/2)

- **Lagrangian multiplicator** $G(f_k, \lambda) = \lambda\,(1 - G_1(f_k)) + G_2(f_k) - G_3(f_k)$, with

$$G_1(f_k) = \int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z \qquad G_2(f_k) = \int_{\mathcal{Z}} \log q(y, z\,; \theta_k) f_k(z)\,\mathrm{d}z \qquad G_3(f_k) = \int_{\mathcal{Z}} \log f_x(z) f_k(z)\,\mathrm{d}z$$

$$\begin{cases} \dfrac{\partial G}{\partial f_t}(z) \;=\; -\lambda + \log q(y, z\,; \theta_k) - \log f_x(z) - 1 & (1) \\[2ex] \dfrac{\partial G}{\partial \lambda}(z) \;=\; 1 - \displaystyle\int_{\mathcal{Z}} f_k(z)\,\mathrm{d}z & (2) \end{cases}$$

$$(1) \iff \log f_k(z) = \log q(y, z\,; \theta_k) - \lambda - 1 \iff f_k(z) = \mathrm{e}^{-\lambda-1}\, q(y, z\,; \theta_k)$$

$$(2) \iff 1 = \int_{\mathcal{Z}} \mathrm{e}^{-\lambda-1}\, q(y, z\,; \theta_k)\,\mathrm{d}z = \mathrm{e}^{-\lambda-1}\, q(y\,; \theta_k) \iff \mathrm{e}^{-\lambda-1} = \frac{1}{q(y\,; \theta_k)}$$

$$\implies f_k(z) = \frac{1}{q(y\,; \theta_k)} \times q(y, z\,; \theta_k) = q(z|y\,; \theta_k)$$

- We **check** that $B(\theta_k|\theta_k) = q(y\,; \theta_k)$

$$B(\theta_k|\theta_k) = \int_{\mathcal{Z}} \log\left(\frac{q(y, z\,; \theta_k)}{f_k(z)}\right) q(z|y\,; \theta_k)\,\mathrm{d}z \;=\; [\ldots] \;=\; q(y\,; \theta_k)$$

- **Maximization of B:** $\implies$ $\theta_{k+1} \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k) + \log q(\theta)$

$$\begin{aligned} B(\theta|\theta_k) &= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,;\theta) - \log f_k(Z)\right] \\ &= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\Big[\log q(y,Z\,|\,\theta) + \underbrace{\log q(\theta)}_{\text{If Bayesian}}\Big] + \underbrace{\mathcal{H}(f_k)}_{\text{Entropy of } f_k} \\ &= Q(\theta|\theta_k) + \log q(\theta) + \mathcal{H}(f_k) \end{aligned}$$

Expectation-Maximization (EM) Algorithm

E-step Compute the conditional expectation

$$\boxed{Q(\theta|\theta_k) = \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,;\theta)\right]}$$

M-step Maximize $Q(\theta|\theta_k)$ in the feasible set: $\boxed{\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k)}$

Remarks:

✓ We only need to compute one integral, no derivative (like in SGD)

× We have to compute an integral...

- **Maximization of B:** $\Longrightarrow \quad \theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k) + \log q(\theta)$

$$\begin{aligned} B(\theta|\theta_k) &= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,;\theta) - \log f_k(Z)\right] \\ &= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,|\,\theta) + \underbrace{\log q(\theta)}_{\text{If Bayesian}}\right] + \underbrace{\mathcal{H}(f_k)}_{\text{Entropy of } f_k} \\ &= Q(\theta|\theta_k) + \log q(\theta) + \mathcal{H}(f_k) \end{aligned}$$

Expectation-Maximization (EM) Algorithm

E-step Compute the conditional expectation

$$\boxed{Q(\theta|\theta_k) = \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,;\theta)\right]}$$

M-step Maximize $Q(\theta|\theta_k)$ in the feasible set: $\boxed{\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k)}$

Remarks:

✓ We only need to compute one integral, no derivative (like in SGD)

× We have to compute an integral...

- **Maximization of B:** $\Longrightarrow \quad \theta_{k+1} \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k) + \log q(\theta)$

$$
\begin{aligned}
B(\theta|\theta_k) &= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y, Z\,;\theta) - \log f_k(Z)\right] \\
&= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\Big[\log q(y, Z \mid \theta) + \underbrace{\log q(\theta)}_{\text{If Bayesian}}\Big] + \underbrace{\mathcal{H}(f_k)}_{\text{Entropy of } f_k} \\
&= Q(\theta|\theta_k) + \log q(\theta) + \mathcal{H}(f_k)
\end{aligned}
$$

Expectation-Maximization (EM) Algorithm

E-step Compute the conditional expectation

$$
\boxed{Q(\theta|\theta_k) = \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y, Z\,;\theta)\right]}
$$

M-step Maximize $Q(\theta|\theta_k)$ in the feasible set: $\boxed{\theta_{k+1} \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k)}$

Remarks:

✓ We only need to compute one integral, no derivative (like in SGD)

× We have to compute an integral...

- **Maximization of B:** $\implies \theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k) + \log q(\theta)$

$$
\begin{aligned}
B(\theta|\theta_k) &= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,;\theta) - \log f_k(Z)\right] \\
&= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,|\,\theta) + \underbrace{\log q(\theta)}_{\text{If Bayesian}}\right] + \underbrace{\mathcal{H}(f_k)}_{\text{Entropy of } f_k} \\
&= Q(\theta|\theta_k) + \log q(\theta) + \mathcal{H}(f_k)
\end{aligned}
$$

Expectation-Maximization (EM) Algorithm

E-step Compute the conditional expectation

$$\boxed{Q(\theta|\theta_k) = \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,;\theta)\right]}$$

M-step Maximize $Q(\theta|\theta_k)$ in the feasible set: $\boxed{\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k)}$

Remarks:

- ✓ We only need to compute one integral, no derivative (like in SGD)
- × We have to compute an integral...

Maximization of $B \rightsquigarrow$ EM Algorithm

- **Maximization of B:** $\implies \theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k) + \log q(\theta)$

$$\begin{aligned} B(\theta|\theta_k) &= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,;\theta) - \log f_k(Z)\right] \\ &= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\Big[\log q(y,Z\,|\,\theta) + \underbrace{\log q(\theta)}_{\text{If Bayesian}}\Big] + \underbrace{\mathcal{H}(f_k)}_{\text{Entropy of } f_k} \\ &= Q(\theta|\theta_k) + \log q(\theta) + \mathcal{H}(f_k) \end{aligned}$$

Expectation-Maximization (EM) Algorithm

E-step Compute the conditional expectation

$$\boxed{Q(\theta|\theta_k) = \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,;\theta)\right]}$$

M-step Maximize $Q(\theta|\theta_k)$ in the feasible set: $\boxed{\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k)}$

Remarks:

- ✓ We only need to compute one integral, no derivative (like in SGD)
- × We have to compute an integral. . .

Maximization of $B \rightsquigarrow$ EM Algorithm

- **Maximization of** B: $\implies \quad \theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k) + \log q(\theta)$

$$
\begin{aligned}
B(\theta|\theta_k) &= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,;\theta) - \log f_k(Z)\right] \\
&= \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,|\,\theta) + \underbrace{\log q(\theta)}_{\text{If Bayesian}}\right] + \underbrace{\mathcal{H}(f_k)}_{\text{Entropy of } f_k} \\
&= Q(\theta|\theta_k) + \log q(\theta) + \mathcal{H}(f_k)
\end{aligned}
$$

Expectation-Maximization (EM) Algorithm

E-step Compute the conditional expectation

$$\boxed{Q(\theta|\theta_k) = \mathbb{E}_{Z\sim q(\cdot|y\,;\theta_k)}\left[\log q(y,Z\,;\theta)\right]}$$

M-step Maximize $Q(\theta|\theta_k)$ in the feasible set: $\boxed{\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k)}$

Remarks:

✓ We only need to compute one integral, no derivative (like in SGD)

× We have to compute an integral...

EM Algorithm and Variants

Mixture model of m components: *Given:*

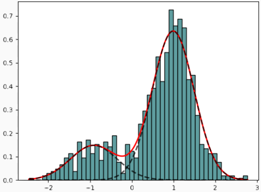
- $(\alpha_j)_{j \in [\![1,m]\!]}$ the *proportions* of the mixture

$$\forall j \in [\![1,m]\!]\,, \quad \alpha_j \in [0,1] \qquad \text{and} \qquad \sum_{j=1}^{m} \alpha_j = 1\,;$$

- For all j, $f_j\big(\,\cdot\,;\omega_j\big)$ the *density* of the j-th sub-population, which (possibly) depends on a parameter ω_j; *and*

- $\theta = \big(\alpha_j, \omega_j\big)_{j \in [\![1,m]\!]}$ the whole *parameters* of the mixture model.

We define: $$\boxed{q\big(\,\cdot\,;\theta\big) = \sum_{j=1}^{m} \alpha_j\, f_j\big(\,\cdot\,;\omega_j\big)}$$



Mixture model of m components: *Given:*

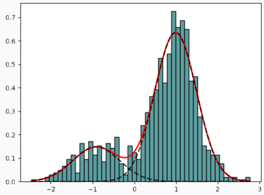
- $(\alpha_j)_{j\in[\![1,m]\!]}$ the *proportions* of the mixture

$$\forall j \in [\![1,m]\!]\,,\quad \alpha_j \in [0,1] \qquad \text{and} \qquad \sum_{j=1}^{m} \alpha_j = 1\,;$$

- For all j, $f_j\big(\cdot\,;\,\omega_j\big)$ the *density* of the j-th sub-population, which (possibly) depends on a parameter ω_j; *and*

- $\theta = \big(\alpha_j, \omega_j\big)_{j\in[\![1,m]\!]}$ the whole *parameters* of the mixture model.

We define: $$\boxed{q\big(\cdot\,;\theta\big) \;=\; \sum_{j=1}^{m} \alpha_j\, f_j\big(\cdot\,;\,\omega_j\big)}$$



Mixture model: $q\big(\cdot\,;\theta\big) = \sum_{j=1}^{m} \alpha_j\, f_j\big(\cdot\,;\omega_j\big) \quad \longleftrightarrow \quad \begin{cases} \text{Proportions } \alpha_j \\ \text{Densities } f_j\big(\cdot\,;\omega_j\big) \end{cases}$

Question: *How to generate data according to such a model?* $(y_i)_{i\in[\![1,n]\!]}$

For all individual $i \in [\![1,n]\!]$,

(i) Let $\mathcal{P}_m = \{\mathcal{C}_1, \dots, \mathcal{C}_m\}$ be a partition of $[\![1,n]\!]$ into m classes, such that the individual i belongs to $\mathcal{C}_j$ with probability α_j: $\mathbb{P}\,(i \in \mathcal{C}_j;\, \alpha_j) = \alpha_j$;

(ii) Then, y_i is generated according to the density $f_j\big(\cdot\,;\omega_j\big)$ if $i \in \mathcal{C}_j$.

$\rightsquigarrow$ Latent variables $(z_i)_{i\in[\![1,n]\!]}$ to encode the classes.

For all individual $i \in [\![1,n]\!]$,

$$\boxed{\begin{cases} z_i \,|\, (\alpha_j)_{j\in[\![1,m]\!]} \sim \sum_{k=j}^{m} \alpha_j \delta_j \\ y_i \,|\, z_i,\, \theta = (\alpha_j, \omega_j)_{j\in[\![1,m]\!]} \sim f_{z_i}\big(\cdot\,;\omega_{z_i}\big) \end{cases}}$$

Mixture model: $q\big(\cdot\,;\theta\big) = \sum_{j=1}^{m} \alpha_j\, f_j\big(\cdot\,;\,\omega_j\big) \quad \longleftrightarrow \quad \begin{cases} \text{Proportions } \alpha_j \\ \text{Densities } f_j\big(\cdot\,;\,\omega_j\big) \end{cases}$

Question: *How to generate data according to such a model?* $(y_i)_{i\in[\![1,n]\!]}$

For all individual $i \in [\![1,n]\!]$,

(i) Let $\mathcal{P}_m = \{\mathcal{C}_1, \dots, \mathcal{C}_m\}$ be a partition of $[\![1,n]\!]$ into m classes, such that the individual i belongs to $\mathcal{C}_j$ with probability α_j: $\mathbb{P}\left(i \in \mathcal{C}_j;\, \alpha_j\right) = \alpha_j$;

(ii) Then, y_i is generated according to the density $f_j\big(\cdot\,;\,\omega_j\big)$ if $i \in \mathcal{C}_j$.

$\rightsquigarrow$ Latent variables $(z_i)_{i\in[\![1,n]\!]}$ to encode the classes.

For all individual $i \in [\![1,n]\!]$, $\begin{cases} z_i \,|\, (\alpha_j)_{j\in[\![1,m]\!]} \sim \sum_{k=j}^{m} \alpha_j \delta_j \\ y_i \,|\, z_i,\, \theta = (\alpha_j, \omega_j)_{j\in[\![1,m]\!]} \sim f_{z_i}\big(\cdot\,;\,\omega_{z_i}\big) \end{cases}$

Mixture Models *vs.* Latent Variables

Mixture model: $q\big(\cdot\,;\theta\big) = \sum_{j=1}^{m} \alpha_j\, f_j\big(\cdot\,; \omega_j\big) \quad \longleftrightarrow \quad \begin{cases} \text{Proportions } \alpha_j \\ \text{Densities } f_j\big(\cdot\,; \omega_j\big) \end{cases}$

Question: *How to generate data according to such a model?* $(y_i)_{i \in [\![1,n]\!]}$

For all individual $i \in [\![1,n]\!]$,

(i) Let $\mathcal{P}_m = \{\mathcal{C}_1, \dots, \mathcal{C}_m\}$ be a partition of $[\![1,n]\!]$ into m classes, such that the individual i belongs to $\mathcal{C}_j$ with probability α_j: $\mathbb{P}\left(i \in \mathcal{C}_j;\, \alpha_j\right) = \alpha_j$;

(ii) Then, y_i is generated according to the density $f_j\big(\cdot\,; \omega_j\big)$ if $i \in \mathcal{C}_j$.

$\rightsquigarrow$ Latent variables $(z_i)_{i \in [\![1,n]\!]}$ to encode the classes.

For all individual $i \in [\![1,n]\!]$, $\boxed{\begin{cases} z_i \,|\, (\alpha_j)_{j \in [\![1,m]\!]} \sim \sum_{k=j}^{m} \alpha_j \delta_j \\ y_i \,|\, z_i,\, \theta = (\alpha_j, \omega_j)_{j \in [\![1,m]\!]} \sim f_{z_i}\big(\cdot\,; \omega_{z_i}\big) \end{cases}}$

Hierarchical Writing of Mixture Models

- **Mixture Models**: For all $i \in [\![1, n]\!]$, $\begin{cases} z_i \,|\, (\alpha_j)_{j \in [\![1,m]\!]} \sim \sum_{j=1}^{m} \alpha_j \delta_j \,, \\ y_i \,|\, z_i,\, (\alpha_j, \omega_j)_{j \in [\![1,m]\!]} \sim f_{z_i}\big(\cdot\,;\, \omega_{z_i}\big). \end{cases}$

- **Complete likelihood**: For all $y = (y_i)_{i \in [\![1,n]\!]}$ and $z = (z_i)_{i \in [\![1,n]\!]}$,

$$q(y, z\,;\,\theta) = \prod_{i=1}^{n} q(y_i, z_i\,;\,\theta) = \prod_{i=1}^{n} q(y_i | z_i\,;\,\theta)\, q(z_i\,;\,\theta) = \prod_{i=1}^{n} \alpha_{z_i}\, f_{z_i}\big(y_i\,;\,\omega_{z_i}\big)\,.$$

- **Conditional likelihood**: For all $i \in [\![1, n]\!]$,

$$\begin{aligned} q(y_i\,;\,\theta) &= \sum_{j=1}^{m} q(y_i, \{z_i = j\}\,;\,\theta) \\ &= \sum_{j=1}^{m} q(y_i \,|\, \{z_i = j\}\,;\,\theta)\, q(\{z_i = j\}\,;\,\theta) = \sum_{j=1}^{m} \alpha_j\, f_j\big(y_i\,;\omega_j\big) \end{aligned}$$

Hierarchical Writing of Mixture Models

- **Mixture Models**: For all $i \in [\![1, n]\!]$, $\begin{cases} z_i \,|\, (\alpha_j)_{j\in[\![1,m]\!]} \sim \sum_{j=1}^{m} \alpha_j \delta_j \,, \\ y_i \,|\, z_i,\, (\alpha_j, \omega_j)_{j\in[\![1,m]\!]} \sim f_{z_i}\big(\,\cdot\,;\, \omega_{z_i}\big). \end{cases}$

- **Complete likelihood**: For all $y = (y_i)_{i\in[\![1,n]\!]}$ and $z = (z_i)_{i\in[\![1,n]\!]}$,

$$q(y, z\,;\,\theta) = \prod_{i=1}^{n} q(y_i, z_i\,;\,\theta) = \prod_{i=1}^{n} q(y_i|z_i\,;\,\theta)\, q(z_i\,;\,\theta) = \prod_{i=1}^{n} \alpha_{z_i}\, f_{z_i}\big(y_i\,;\,\omega_{z_i}\big)\,.$$

- **Conditional likelihood**: For all $i \in [\![1, n]\!]$,

$$\begin{aligned} q(y_i\,;\,\theta) &= \sum_{j=1}^{m} q(y_i, \{z_i = j\}\,;\,\theta) \\ &= \sum_{j=1}^{m} q(y_i\,|\,\{z_i = j\}\,;\,\theta)\, q(\{z_i = j\}\,;\,\theta) = \sum_{j=1}^{m} \alpha_j\, f_j\big(y_i\,;\omega_j\big) \end{aligned}$$

Hierarchical Writing of Mixture Models

- **Mixture Models**: For all $i \in [\![1, n]\!]$, $\begin{cases} z_i \,|\, (\alpha_j)_{j \in [\![1,m]\!]} \sim \displaystyle\sum_{j=1}^{m} \alpha_j \delta_j \,, \\ y_i \,|\, z_i,\, (\alpha_j, \omega_j)_{j \in [\![1,m]\!]} \sim f_{z_i}\big(\,\cdot\,;\, \omega_{z_i}\big). \end{cases}$

- **Complete likelihood**: For all $y = (y_i)_{i \in [\![1,n]\!]}$ and $z = (z_i)_{i \in [\![1,n]\!]}$,

$$q(y, z\,;\,\theta) = \prod_{i=1}^{n} q(y_i, z_i\,;\,\theta) = \prod_{i=1}^{n} q(y_i|z_i\,;\,\theta)\, q(z_i\,;\,\theta) = \prod_{i=1}^{n} \alpha_{z_i}\, f_{z_i}\big(y_i\,;\,\omega_{z_i}\big)\,.$$

- **Conditional likelihood**: For all $i \in [\![1, n]\!]$,

$$\begin{aligned} q(y_i\,;\,\theta) &= \sum_{j=1}^{m} q(y_i, \{z_i = j\}\,;\,\theta) \\ &= \sum_{j=1}^{m} q(y_i \,|\, \{z_i = j\}\,;\,\theta)\, q(\{z_i = j\}\,;\,\theta) = \sum_{j=1}^{m} \alpha_j\, f_j\big(y_i\,;\omega_j\big) \end{aligned}$$

Parameters Estimation through the EM Algorithm

E-step: Compute the conditional expected log-likelihood

$$\begin{aligned} Q(\theta|\theta_k) &= \int_{\mathcal{Z}} \log q\big(y, z\,;\,\theta\big)\, q\big(z|y\,;\,\theta_k\big)\,\mathrm{d}z \\ &= \mathbb{E}_{Z\sim q(\,\cdot\,|y\,;\,\theta_k)}\Big[\log q\big(y, Z\,;\,\theta\big)\Big] \end{aligned}$$

Here: $\boxed{Q(\theta|\theta_k) = \sum_{i=1}^{n}\sum_{j=1}^{m}\Big[\log(\alpha_j) + \log\left(f_j(y_i\,;\,\omega_j)\right)\Big]\tau_{ij}^{(k)}}$

where $\tau_{ij}^{(k)} = \mathbb{P}\left(Z_i = j \,|\, y_i\,;\,\theta_k\right) = \dfrac{\alpha_j^{(k)} f_j(y_i\,;\,\omega_j^{(k)})}{\sum_{\ell=1}^{m}\alpha_\ell^{(k)} f_\ell(y_i\,;\,\omega_\ell^{(k)})}$.

M-step: Maximize $Q(\,\cdot\,|\theta_k)$ in the feasible set θ: $\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k)$

Here: $\boxed{\begin{cases} \forall j \in [\![1,m]\!], \quad \alpha_j^{(k+1)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\tau_{ij}^{(k)} \\ \left(\omega_k^{(k+1)}\right)_{j\in[\![1,m]\!]} \in \underset{\omega=(\omega_j)\in\Omega}{\operatorname{argmax}} \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m}\tau_{ij}^{(k)}\log\left(f_j(y_i\,;\,\omega_j)\right) \end{cases}}$

Parameters Estimation through the EM Algorithm

E-step: Compute the conditional expected log-likelihood

$$\begin{aligned} Q(\theta|\theta_k) &= \int_{\mathcal{Z}} \log q\big(y, z\,;\,\theta\big)\, q\big(z|y\,;\,\theta_k\big)\, \mathrm{d}z \\ &= \mathbb{E}_{Z\sim q(\,\cdot\,|y\,;\,\theta_k)}\Big[\log q\big(y, Z\,;\,\theta\big)\Big] \end{aligned}$$

Here: $$\boxed{Q(\theta|\theta_k) = \sum_{i=1}^{n}\sum_{j=1}^{m}\Big[\log(\alpha_j) + \log\left(f_j(y_i\,;\,\omega_j)\right)\Big]\,\tau_{ij}^{(k)}}$$

where $\tau_{ij}^{(k)} = \mathbb{P}\left(Z_i = j \,|\, y_i\,;\,\theta_k\right) = \dfrac{\alpha_j^{(k)} f_j(y_i\,;\,\omega_j^{(k)})}{\sum_{\ell=1}^{m}\alpha_\ell^{(k)} f_\ell(y_i\,;\,\omega_\ell^{(k)})}$.

M-step: Maximize $Q(\,\cdot\,|\theta_k)$ in the feasible set θ: $\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k)$

Here: $$\boxed{\begin{cases} \forall j \in [\![1, m]\!], \quad \alpha_j^{(k+1)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\tau_{ij}^{(k)} \\ (\omega_k^{(k+1)})_{j\in[\![1,m]\!]} \in \underset{\omega=(\omega_j)\in\Omega}{\operatorname{argmax}} \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m}\tau_{ij}^{(k)}\log\left(f_j(y_i\,;\,\omega_j)\right) \end{cases}}$$

Parameters Estimation through the EM Algorithm

E-step: Compute the conditional expected log-likelihood

$$\begin{aligned} Q(\theta|\theta_k) &= \int_{\mathcal{Z}} \log q\big(y, z\,;\,\theta\big)\, q\big(z|y\,;\,\theta_k\big)\,\mathrm{d}z \\ &= \mathbb{E}_{Z\sim q(\,\cdot\,|y\,;\,\theta_k)}\Big[\log q\big(y, Z\,;\,\theta\big)\Big] \end{aligned}$$

Here: $$\boxed{Q(\theta|\theta_k) = \sum_{i=1}^{n}\sum_{j=1}^{m}\Big[\log(\alpha_j) + \log\left(f_j(y_i\,;\,\omega_j)\right)\Big]\,\tau_{ij}^{(k)}}$$

where $\tau_{ij}^{(k)} = \mathbb{P}\left(Z_i = j \mid y_i\,;\,\theta_k\right) = \dfrac{\alpha_j^{(k)} f_j(y_i\,;\,\omega_j^{(k)})}{\sum_{\ell=1}^{m}\alpha_\ell^{(k)} f_\ell(y_i\,;\,\omega_\ell^{(k)})}$.

M-step: Maximize $Q(\,\cdot\,|\theta_k)$ in the feasible set θ: $\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta_k)$

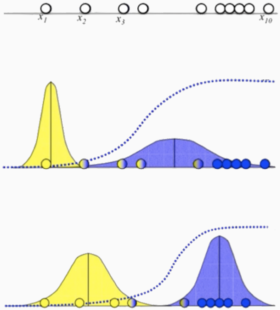
Here: $$\boxed{\begin{cases} \forall j \in [\![1, m]\!], \quad \alpha_j^{(k+1)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\tau_{ij}^{(k)} \\ (\omega_k^{(k+1)})_{j\in[\![1,m]\!]} \in \underset{\omega=(\omega_j)\in\Omega}{\operatorname{argmax}} \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m}\tau_{ij}^{(k)}\log\Big(f_j(y_i\,;\,\omega_j)\Big) \end{cases}}$$

Example 1: One-dimensional Gaussian Mixture Model

Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) \qquad m = 2$

Images from Victor Lavrenko.

$$\begin{cases} q(y\,;\,\theta) \;=\; \alpha\,\phi(y\,;\,\mu_1, \sigma_1^2) \;+\; (1-\alpha)\,\phi(y\,;\,\mu_2, \sigma_2^2) \\ \phi(y\,;\,\mu, \sigma^2) \;=\; \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(y-\mu)^2}{2\sigma^2}\right) \end{cases}$$



E-step: $Q(\theta|\theta_k) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(k)} \\ \tau_{i2}^{(k)} = 1 - \tau_{i1}^{(k)} \end{cases}$

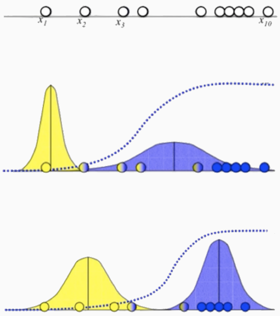
$$\tau_{i1}^{(k)} = \frac{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)})}{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)}) + (1-\alpha^{(k)})\,\phi(y; \mu_2^{(k)}, \sigma_2^{2(t)})}$$

M-step: $\alpha^{(k+1)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \tau_{i1}^{(k)}$

$$\forall j \in \{1, 2\}, \quad \mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} y_i}{\sum_{i=1}^{n} \tau_{ij}^{(k)}} \quad \& \quad \sigma_j^{2(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} \left(y_i - \mu_j^{(k+1)}\right)^2}{\sum_{i=1}^{n} \tau_{ij}^{(k)}}$$

Example 1: One-dimensional Gaussian Mixture Model

Images from Victor Lavrenko.



Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) \qquad m = 2$

$$\begin{cases} q(y\,;\,\theta) \;=\; \alpha\,\phi(y\,;\,\mu_1, \sigma_1^2) \;+\; (1-\alpha)\,\phi(y\,;\,\mu_2, \sigma_2^2) \\ \phi(y\,;\,\mu, \sigma^2) \;=\; \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(y-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta_k) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(k)} \\ \tau_{i2}^{(k)} = 1 - \tau_{i1}^{(k)} \end{cases}$

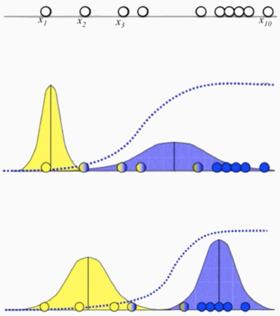
$$\tau_{i1}^{(k)} \;=\; \frac{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)})}{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)}) + (1-\alpha^{(k)})\,\phi(y; \mu_2^{(k)}, \sigma_2^{2(t)})}$$

M-step: $\alpha^{(k+1)} \;=\; \dfrac{1}{n}\sum_{i=1}^{n} \tau_{i1}^{(k)}$

$$\forall j \in \{1, 2\}, \quad \mu_j^{(k+1)} \;=\; \frac{\sum_{i=1}^n \tau_{ij}^{(k)} y_i}{\sum_{i=1}^n \tau_{ij}^{(k)}} \quad \& \quad \sigma_j^{2(t+1)} \;=\; \frac{\sum_{i=1}^n \tau_{ij}^{(k)} \left(y_i - \mu_j^{(k+1)}\right)^2}{\sum_{i=1}^n \tau_{ij}^{(k)}}$$

Example 1: One-dimensional Gaussian Mixture Model

Images from Victor Lavrenko.



Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) \qquad m = 2$

$$\begin{cases} q(y\,;\,\theta) = \alpha\,\phi(y\,;\,\mu_1, \sigma_1^2) + (1-\alpha)\,\phi(y\,;\,\mu_2, \sigma_2^2) \\ \phi(y\,;\,\mu, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(y-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta_k) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(k)} \\ \tau_{i2}^{(k)} = 1 - \tau_{i1}^{(k)} \end{cases}$

$$\tau_{i1}^{(k)} = \frac{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)})}{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)}) + (1-\alpha^{(k)})\,\phi(y; \mu_2^{(k)}, \sigma_2^{2(t)})}$$

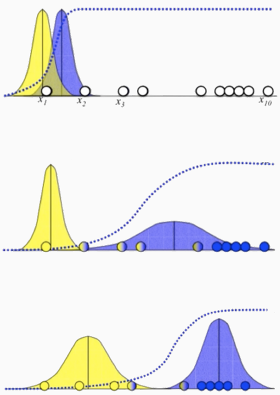
M-step: $\alpha^{(k+1)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \tau_{i1}^{(k)}$

$$\forall j \in \{1,2\}, \quad \mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} y_i}{\sum_{i=1}^{n} \tau_{ij}^{(k)}} \quad \& \quad \sigma_j^{2(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} \left(y_i - \mu_j^{(k+1)}\right)^2}{\sum_{i=1}^{n} \tau_{ij}^{(k)}}$$

Example 1: One-dimensional Gaussian Mixture Model

Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) \qquad m = 2$

Images from Victor Lavrenko.



$$\begin{cases} q(y\,;\,\theta) = \alpha\,\phi(y\,;\,\mu_1, \sigma_1^2) + (1-\alpha)\,\phi(y\,;\,\mu_2, \sigma_2^2) \\ \phi(y\,;\,\mu, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(y-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta_k) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(k)} \\ \tau_{i2}^{(k)} = 1 - \tau_{i1}^{(k)} \end{cases}$

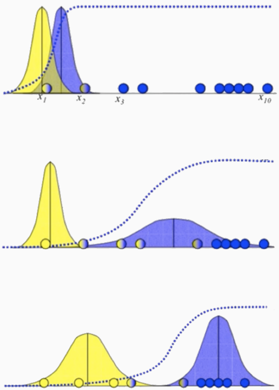
$$\tau_{i1}^{(k)} = \frac{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)})}{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)}) + (1-\alpha^{(k)})\,\phi(y; \mu_2^{(k)}, \sigma_2^{2(t)})}$$

M-step: $\alpha^{(k+1)} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \tau_{i1}^{(k)}$

$$\forall j \in \{1, 2\}\,, \quad \mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} y_i}{\sum_{i=1}^{n} \tau_{ij}^{(k)}} \quad \& \quad \sigma_j^{2(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} \left(y_i - \mu_j^{(k+1)}\right)^2}{\sum_{i=1}^{n} \tau_{ij}^{(k)}}$$

Example 1: One-dimensional Gaussian Mixture Model

Images from Victor Lavrenko.



Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) \qquad m = 2$

$$\begin{cases} q(y\,;\,\theta) = \alpha\,\phi(y\,;\,\mu_1, \sigma_1^2) + (1-\alpha)\,\phi(y\,;\,\mu_2, \sigma_2^2) \\ \phi(y\,;\,\mu, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(y-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta_k) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(k)} \\ \tau_{i2}^{(k)} = 1 - \tau_{i1}^{(k)} \end{cases}$

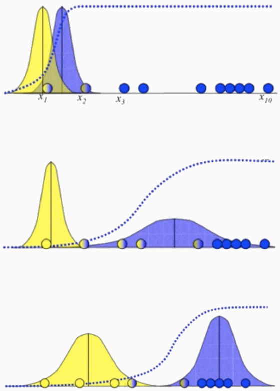
$$\tau_{i1}^{(k)} = \frac{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)})}{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)}) + (1-\alpha^{(k)})\,\phi(y; \mu_2^{(k)}, \sigma_2^{2(t)})}$$

M-step: $\alpha^{(k+1)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \tau_{i1}^{(k)}$

$$\forall j \in \{1, 2\}\,, \quad \mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} y_i}{\sum_{i=1}^{n} \tau_{ij}^{(k)}} \quad \& \quad \sigma_j^{2(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} \left(y_i - \mu_j^{(k+1)}\right)^2}{\sum_{i=1}^{n} \tau_{ij}^{(k)}}$$

Example 1: One-dimensional Gaussian Mixture Model

Images from Victor Lavrenko.



Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) \qquad m = 2$

$$\begin{cases} q(y\,;\,\theta) = \alpha\,\phi(y\,;\,\mu_1, \sigma_1^2) + (1-\alpha)\,\phi(y\,;\,\mu_2, \sigma_2^2) \\ \phi(y\,;\,\mu, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(y-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta_k) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(k)} \\ \tau_{i2}^{(k)} = 1 - \tau_{i1}^{(k)} \end{cases}$

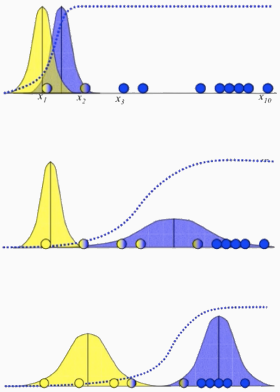
$$\tau_{i1}^{(k)} = \frac{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)})}{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)}) + (1-\alpha^{(k)})\,\phi(y; \mu_2^{(k)}, \sigma_2^{2(t)})}$$

M-step: $\alpha^{(k+1)} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \tau_{i1}^{(k)}$

$$\forall j \in \{1, 2\}, \quad \mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} y_i}{\sum_{i=1}^{n} \tau_{ij}^{(k)}} \quad \& \quad \sigma_j^{2(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} \left(y_i - \mu_j^{(k+1)}\right)^2}{\sum_{i=1}^{n} \tau_{ij}^{(k)}}$$

Example 1: One-dimensional Gaussian Mixture Model

Images from Victor Lavrenko.



Likelihood: $\theta = \left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) \qquad m = 2$

$$\begin{cases} q(y\,;\,\theta) \;=\; \alpha\,\phi(y\,;\,\mu_1, \sigma_1^2) \;+\; (1-\alpha)\,\phi(y\,;\,\mu_2, \sigma_2^2) \\ \phi(y\,;\,\mu, \sigma^2) \;=\; \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(y-\mu)^2}{2\sigma^2}\right) \end{cases}$$

E-step: $Q(\theta|\theta_k) \quad \longleftrightarrow \quad \begin{cases} \tau_{i1}^{(k)} \\ \tau_{i2}^{(k)} = 1 - \tau_{i1}^{(k)} \end{cases}$

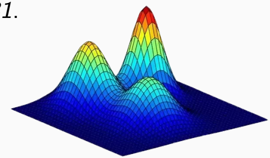
$$\tau_{i1}^{(k)} \;=\; \frac{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)})}{\alpha^{(k)}\,\phi(y; \mu_1^{(k)}, \sigma_1^{2(t)}) + (1-\alpha^{(k)})\,\phi(y; \mu_2^{(k)}, \sigma_2^{2(t)})}$$

M-step: $\alpha^{(k+1)} \;=\; \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \tau_{i1}^{(k)}$

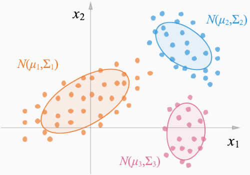
$$\forall j \in \{1, 2\}\,, \quad \mu_j^{(k+1)} \;=\; \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} y_i}{\sum_{i=1}^{n} \tau_{ij}^{(k)}} \quad \& \quad \sigma_j^{2(t+1)} \;=\; \frac{\sum_{i=1}^{n} \tau_{ij}^{(k)} \left(y_i - \mu_j^{(k+1)}\right)^2}{\sum_{i=1}^{n} \tau_{ij}^{(k)}}$$

Multivariate Gaussian Mixtures

- **Quantitative data**: $y_i \in \mathbb{R}^d$, *Generalization of slide 31.*



- **Likelihood**: $\theta = (\alpha_j, \mu_j, \Sigma_j)_{j \in [\![1,m]\!]}$



$$\begin{cases} q(y\,;\theta) = \displaystyle\sum_{j=1}^{m} \alpha_j \, \phi\,(y\,;\mu_j, \Sigma_j)\;. \\[2ex] \phi\,(y\,;\mu, \Sigma) = \dfrac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\dfrac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)\right)\;. \end{cases}$$

- **Gaussian Mixture Model**:

$$\begin{cases} \Theta = \left\{ (\alpha_j, \mu_j, \Sigma_j)_{j \in [\![1,m]\!]} \in \left([0,1] \times \mathbb{R}^d \times \mathcal{S}_d\mathbb{R}\right)^K \,\middle|\, \displaystyle\sum_{j=1}^{m} \alpha_j = 1 \right\} \\[2ex] \mathcal{M} = \left\{ \theta \in \Theta \,\middle|\, y \in \mathbb{R}^d \mapsto q_K(x\,;\theta) \right\} \end{cases}$$

- **Estimation** through the EM algorithm *(See tutorials $\rightsquigarrow$ Friday!)*

EM Algorithm and Variants

Convergence of the EM Algorithm

- Several demonstrations of convergence, with assumptions that are more or less complicated to ensure in practice
- We state the version of [Delyon, Lavielle, Moulines], whose assumptions are often meet

(M1)
- $\Theta \subset \mathbb{R}^{n_\theta}$ open subset of $\mathbb{R}^{n_\theta}$
- $\exists S : \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_\theta}$ a Borelian function such that
  - Convex envelope $\mathcal{C}onv\big(S(\mathbb{R}^{n_z})\big) \subset \mathcal{S}$
  - $\forall \theta \in \Theta\,,\ \int_{\mathbb{R}^{n_z}} |S(y,z)| q(z|y\,;\theta)\,\mathrm{d}z < +\infty$
  - $\boxed{q(y,z\,;\theta) = \mathcal{E}xp\left(-\psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta) \right\rangle\right)}$

$\rightsquigarrow$ *Exponential family,* $S \equiv$ *Exhaustive statistics*

(M2) $\phi\,,\ \psi \in \mathcal{C}^2(\Theta)$

(M3) $\bar{s} : \Theta \to \mathcal{S}$ s.t. $\bar{s}(\theta) = \int_{\mathbb{R}^{n_z}} S(y,z)\, q(z|y\,;\theta)\,\mathrm{d}z$ is $\mathcal{C}^1(\Theta)$

(M4) $\ell(\theta) = \log q(y\,;\theta)$ is $\mathcal{C}^1(\Theta)$. Moreover, $\partial\theta \int_{\mathbb{R}^{n_z}} \log q(y,z\,;\theta)\,\mathrm{d}z = \int_{\mathbb{R}^{n_z}} \partial\theta \log q(y,z\,;\theta)\,\mathrm{d}z$

(M5) Let $S : \mathcal{S} \times \Theta \to \mathbb{R}$ s.t. $L(s,\theta) = \psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta) \right\rangle$

$\exists \hat{\theta} : \mathcal{S} \to \Theta$, of class $\mathcal{C}^1(\mathcal{S})$ and $\forall s \in \mathcal{S}, \forall \theta \in \Theta$, $L(s, \hat{\theta}(s)) \geqslant L(s,\theta)$

EM Convergence

Assume (M1–5).
Then for all $\theta_0 \in \Theta$

- The sequence $\ell(\theta_k)_k$ produced by the EM algorithm is increasing
- $\boxed{\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0}$

$\boxed{\mathcal{L} = \{\theta \in \Theta s.t. \partial_\theta \ell(\theta) = 0\}}$

Convergence of the EM Algorithm

- Several demonstrations of convergence, with assumptions that are more or less complicated to ensure in practice
- We state the version of [Delyon, Lavielle, Moulines], whose assumptions are often meet

(M1)
- $\Theta \subset \mathbb{R}^{n_\theta}$ open subset of $\mathbb{R}^{n_\theta}$
- $\exists S \colon \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_\theta}$ a Borelian function such that
  - Convex envelope $\mathcal{C}onv\Big(S(\mathbb{R}^{n_z})\Big) \subset \mathcal{S}$
  - $\forall \theta \in \Theta\,,\ \int_{\mathbb{R}^{n_z}} |S(y,z)| q(z|y\,;\theta)\,\mathrm{d}z < +\infty$
  - $\boxed{q(y,z\,;\theta) = \mathcal{E}xp\left(-\psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta)\right\rangle\right)}$

$\rightsquigarrow$ *Exponential family, $S \equiv$ Exhaustive statistics*

(M2) $\phi\,,\ \psi \in \mathcal{C}^2(\Theta)$

(M3) $\bar{s} \colon \Theta \to \mathcal{S}$ s.t. $\bar{s}(\theta) = \int_{\mathbb{R}^{n_z}} S(y,z)\, q(z|y\,;\theta)\,\mathrm{d}z$ is $\mathcal{C}^1(\Theta)$

(M4) $\ell(\theta) = \log q(y\,;\theta)$ is $\mathcal{C}^1(\Theta)$. Moreover, $\partial\theta \int_{\mathbb{R}^{n_z}} \log q(y,z\,;\theta)\,\mathrm{d}z = \int_{\mathbb{R}^{n_z}} \partial\theta \log q(y,z\,;\theta)\,\mathrm{d}z$

(M5) Let $S \colon \mathcal{S} \times \Theta \to \mathbb{R}$ s.t. $L(s,\theta) = \psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta)\right\rangle$

$\exists \hat{\theta} \colon \mathcal{S} \to \Theta$, of class $\mathcal{C}^1(\mathcal{S})$ and $\forall s \in \mathcal{S}, \forall \theta \in \Theta$, $L(s, \hat{\theta}(s)) \geqslant L(s,\theta)$

EM Convergence

Assume (M1–5).

Then for all $\theta_0 \in \Theta$

- The sequence $\ell(\theta_k)_k$ produced by the EM algorithm is increasing
- $\boxed{\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0}$

$$\boxed{\mathcal{L} = \{\theta \in \Theta s.t. \partial_\theta \ell(\theta) = 0\}}$$

Convergence of the EM Algorithm

- Several demonstrations of convergence, with assumptions that are more or less complicated to ensure in practice
- We state the version of [Delyon, Lavielle, Moulines], whose assumptions are often meet

(M1)
- $\Theta \subset \mathbb{R}^{n_\theta}$ open subset of $\mathbb{R}^{n_\theta}$
- $\exists S \colon \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_\theta}$ a Borelian function such that
  - Convex envelope $\mathcal{C}onv\Big(S(\mathbb{R}^{n_z})\Big) \subset \mathcal{S}$
  - $\forall \theta \in \Theta\,,\ \int_{\mathbb{R}^{n_z}} |S(y,z)| q(z|y\,;\theta)\,\mathrm{d}z < +\infty$
  - $\boxed{q(y,z\,;\theta) = \mathcal{E}xp\left(-\psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta) \right\rangle\right)}$

$\rightsquigarrow$ *Exponential family, $S \equiv$ Exhaustive statistics*

(M2) $\phi\,,\ \psi \in \mathcal{C}^2(\Theta)$

(M3) $\bar{s} \colon \Theta \to \mathcal{S}$ s.t. $\bar{s}(\theta) = \int_{\mathbb{R}^{n_z}} S(y,z)\, q(z|y\,;\theta)\,\mathrm{d}z$ is $\mathcal{C}^1(\Theta)$

(M4) $\ell(\theta) = \log q(y\,;\theta)$ is $\mathcal{C}^1(\Theta)$. Moreover, $\partial\theta \int_{\mathbb{R}^{n_z}} \log q(y,z\,;\theta)\,\mathrm{d}z = \int_{\mathbb{R}^{n_z}} \partial\theta \log q(y,z\,;\theta)\,\mathrm{d}z$

(M5) Let $S \colon \mathcal{S} \times \Theta \to \mathbb{R}$ s.t. $L(s,\theta) = \psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta) \right\rangle$

$\exists \hat{\theta} \colon \mathcal{S} \to \Theta$, of class $\mathcal{C}^1(\mathcal{S})$ and $\forall s \in \mathcal{S}, \forall \theta \in \Theta$, $L(s, \hat{\theta}(s)) \geqslant L(s,\theta)$

EM Convergence

Assume (M1–5).

Then for all $\theta_0 \in \Theta$

- The sequence $\ell(\theta_k)_k$ produced by the EM algorithm is increasing
- $\boxed{\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0}$

$$\boxed{\mathcal{L} = \{\theta \in \Theta s.t. \partial_\theta \ell(\theta) = 0\}}$$

Convergence of the EM Algorithm

- Several demonstrations of convergence, with assumptions that are more or less complicated to ensure in practice
- We state the version of [Delyon, Lavielle, Moulines], whose assumptions are often meet

(M1)
- $\Theta \subset \mathbb{R}^{n_\theta}$ open subset of $\mathbb{R}^{n_\theta}$
- $\exists S \colon \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_\theta}$ a Borelian function such that
  - Convex envelope $\mathcal{C}onv\Big(S(\mathbb{R}^{n_z})\Big) \subset \mathcal{S}$
  - $\forall \theta \in \Theta\,,\ \int_{\mathbb{R}^{n_z}} |S(y,z)| q(z|y\,;\theta)\,\mathrm{d}z < +\infty$
  - $\boxed{q(y,z\,;\theta) = \mathcal{E}xp\left(-\psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta) \right\rangle\right)}$

$\rightsquigarrow$ *Exponential family, $S \equiv$ Exhaustive statistics*

(M2) $\phi\,,\ \psi \in \mathcal{C}^2(\Theta)$

(M3) $\bar{s} \colon \Theta \to \mathcal{S}$ s.t. $\bar{s}(\theta) = \int_{\mathbb{R}^{n_z}} S(y,z)\, q(z|y\,;\theta)\,\mathrm{d}z$ is $\mathcal{C}^1(\Theta)$

(M4) $\ell(\theta) = \log q(y\,;\theta)$ is $\mathcal{C}^1(\Theta)$. Moreover, $\partial\theta \int_{\mathbb{R}^{n_z}} \log q(y,z\,;\theta)\,\mathrm{d}z = \int_{\mathbb{R}^{n_z}} \partial\theta \log q(y,z\,;\theta)\,\mathrm{d}z$

(M5) Let $S \colon \mathcal{S} \times \Theta \to \mathbb{R}$ s.t. $L(s,\theta) = \psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta) \right\rangle$

$\exists \hat{\theta} \colon \mathcal{S} \to \Theta$, of class $\mathcal{C}^1(\mathcal{S})$ and $\forall s \in \mathcal{S}, \forall \theta \in \Theta$, $L(s, \hat{\theta}(s)) \geqslant L(s,\theta)$

EM Convergence

Assume (M1–5).
Then for all $\theta_0 \in \Theta$

- The sequence $\ell(\theta_k)_k$ produced by the EM algorithm is increasing
- $\boxed{\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0}$

$$\boxed{\mathcal{L} = \{\theta \in \Theta s.t. \partial_\theta \ell(\theta) = 0\}}$$

Convergence of the EM Algorithm

- Several demonstrations of convergence, with assumptions that are more or less complicated to ensure in practice
- We state the version of [Delyon, Lavielle, Moulines], whose assumptions are often meet

(M1)
- $\Theta \subset \mathbb{R}^{n_\theta}$ open subset of $\mathbb{R}^{n_\theta}$
- $\exists S \colon \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_\theta}$ a Borelian function such that
  - Convex envelope $\mathcal{C}onv\Big(S(\mathbb{R}^{n_z})\Big) \subset \mathcal{S}$
  - $\forall \theta \in \Theta\,,\ \int_{\mathbb{R}^{n_z}} |S(y,z)| q(z|y\,;\theta)\,\mathrm{d}z < +\infty$
  - $\boxed{q(y,z\,;\theta) = \mathcal{E}xp\left(-\psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta) \right\rangle\right)}$

$\rightsquigarrow$ *Exponential family,* $S \equiv$ *Exhaustive statistics*

(M2) $\phi\,,\ \psi \in \mathcal{C}^2(\Theta)$

(M3) $\bar{s} \colon \Theta \to \mathcal{S}$ s.t. $\bar{s}(\theta) = \int_{\mathbb{R}^{n_z}} S(y,z)\, q(z|y\,;\theta)\,\mathrm{d}z$ is $\mathcal{C}^1(\Theta)$

(M4) $\ell(\theta) = \log q(y\,;\theta)$ is $\mathcal{C}^1(\Theta)$. Moreover, $\partial\theta \int_{\mathbb{R}^{n_z}} \log q(y,z\,;\theta)\,\mathrm{d}z = \int_{\mathbb{R}^{n_z}} \partial\theta \log q(y,z\,;\theta)\,\mathrm{d}z$

(M5) Let $S \colon \mathcal{S} \times \Theta \to \mathbb{R}$ s.t. $L(s,\theta) = \psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta) \right\rangle$

$\exists \hat{\theta} \colon \mathcal{S} \to \Theta$, of class $\mathcal{C}^1(\mathcal{S})$ and $\forall s \in \mathcal{S}, \forall \theta \in \Theta$, $L(s,\hat{\theta}(s)) \geqslant L(s,\theta)$

EM Convergence

Assume (M1–5).
Then for all $\theta_0 \in \Theta$

- The sequence $\ell(\theta_k)_k$ produced by the EM algorithm is increasing
- $\boxed{\lim_{k \to \infty} d(\theta_k, \mathcal{L}) = 0}$

$\boxed{\mathcal{L} = \{\theta \in \Theta s.t. \partial_\theta \ell(\theta) = 0\}}$

Convergence of the EM Algorithm

- Several demonstrations of convergence, with assumptions that are more or less complicated to ensure in practice
- We state the version of [Delyon, Lavielle, Moulines], whose assumptions are often meet

(M1)
- $\Theta \subset \mathbb{R}^{n_\theta}$ open subset of $\mathbb{R}^{n_\theta}$
- $\exists S \colon \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_\theta}$ a Borelian function such that
  - Convex envelope $\mathcal{C}onv\Big(S(\mathbb{R}^{n_z})\Big) \subset \mathcal{S}$
  - $\forall \theta \in \Theta\,,\ \int_{\mathbb{R}^{n_z}} |S(y,z)| q(z|y\,;\theta)\,\mathrm{d}z < +\infty$
  - $\boxed{q(y,z\,;\theta) = \mathcal{E}xp\left(-\psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta) \right\rangle\right)}$

$\rightsquigarrow$ *Exponential family, $S \equiv$ Exhaustive statistics*

(M2) $\phi\,,\ \psi \in \mathcal{C}^2(\Theta)$

(M3) $\bar{s} \colon \Theta \to \mathcal{S}$ s.t. $\bar{s}(\theta) = \int_{\mathbb{R}^{n_z}} S(y,z)\, q(z|y\,;\theta)\,\mathrm{d}z$ is $\mathcal{C}^1(\Theta)$

(M4) $\ell(\theta) = \log q(y\,;\theta)$ is $\mathcal{C}^1(\Theta)$. Moreover, $\partial\theta \int_{\mathbb{R}^{n_z}} \log q(y,z\,;\theta)\,\mathrm{d}z = \int_{\mathbb{R}^{n_z}} \partial\theta \log q(y,z\,;\theta)\,\mathrm{d}z$

(M5) Let $S \colon \mathcal{S} \times \Theta \to \mathbb{R}$ s.t. $L(s,\theta) = \psi(\theta) + \left\langle S(y,z) \,\middle|\, \phi(\theta) \right\rangle$

$\exists \hat{\theta} \colon \mathcal{S} \to \Theta$, of class $\mathcal{C}^1(\mathcal{S})$ and $\forall s \in \mathcal{S}, \forall \theta \in \Theta$, $L(s, \hat{\theta}(s)) \geqslant L(s,\theta)$

EM Convergence

Assume (M1–5).
Then for all $\theta_0 \in \Theta$
- The sequence $\ell(\theta_k)_k$ produced by the EM algorithm is increasing
- $\boxed{\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0}$

$$\boxed{\mathcal{L} = \{\theta \in \Theta s.t. \partial_\theta \ell(\theta) = 0\}}$$

Convergence of the EM Algorithm

- Several demonstrations of convergence, with assumptions that are more or less complicated to ensure in practice
- We state the version of [Delyon, Lavielle, Moulines], whose assumptions are often meet

(M1)
- $\Theta \subset \mathbb{R}^{n_\theta}$ open subset of $\mathbb{R}^{n_\theta}$
- $\exists S \colon \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_\theta}$ a Borelian function such that
  - Convex envelope $\mathcal{C}onv\Big(S(\mathbb{R}^{n_z})\Big) \subset \mathcal{S}$
  - $\forall \theta \in \Theta\,,\ \int_{\mathbb{R}^{n_z}} |S(y,z)| q(z|y\,;\theta)\,\mathrm{d}z < +\infty$
  - $\boxed{q(y,z\,;\theta) = \mathcal{E}xp\left(-\psi(\theta) + \Big\langle\, S(y,z) \,\Big|\, \phi(\theta) \,\Big\rangle\right)}$

$\rightsquigarrow$ *Exponential family,* $S \equiv$ *Exhaustive statistics*

(M2) $\phi\,,\ \psi \in \mathcal{C}^2(\Theta)$

(M3) $\bar{s} \colon \Theta \to \mathcal{S}$ s.t. $\bar{s}(\theta) = \int_{\mathbb{R}^{n_z}} S(y,z)\, q(z|y\,;\theta)\,\mathrm{d}z$ is $\mathcal{C}^1(\Theta)$

(M4) $\ell(\theta) = \log q(y\,;\theta)$ is $\mathcal{C}^1(\Theta)$. Moreover, $\partial\theta \int_{\mathbb{R}^{n_z}} \log q(y,z\,;\theta)\,\mathrm{d}z = \int_{\mathbb{R}^{n_z}} \partial\theta \log q(y,z\,;\theta)\,\mathrm{d}z$

(M5) Let $S \colon \mathcal{S} \times \Theta \to \mathbb{R}$ s.t. $L(s,\theta) = \psi(\theta) + \Big\langle\, S(y,z) \,\Big|\, \phi(\theta) \,\Big\rangle$

$\exists \hat{\theta} \colon \mathcal{S} \to \Theta$, of class $\mathcal{C}^1(\mathcal{S})$ and $\forall s \in \mathcal{S}, \forall \theta \in \Theta$, $L(s,\hat{\theta}(s)) \geqslant L(s,\theta)$

EM Convergence

Assume (M1–5).
Then for all $\theta_0 \in \Theta$
- The sequence $\ell(\theta_k)_k$ produced by the EM algorithm is increasing
- $\boxed{\lim_{k\to\infty} d(\theta_k,\mathcal{L}) = 0}$

$$\boxed{\mathcal{L} = \{\theta \in \Theta s.t. \partial_\theta \ell(\theta) = 0\}}$$

Convergence Considerations

Remarks

- The sequence $(\theta_k)_k$ is deterministic, and therefore depends on θ_0
  However, for a θ_0, we may obtain a local minimum

  $\rightsquigarrow$ Test several θ_0 and choose the one for which $\ell(\theta_0)$ is maximum

- (M1) *Exponential family:* Many models fall into this category, including complex models

  $\rightsquigarrow$ *Not a constraint*

- (M4) ensures regularity of the log-likelihood

- (M5) allows easy updating of θ knowing S

- Convergence toward a stationary point of ℓ: A saddle point, for example, may be encountered

  $\rightsquigarrow$ Convexity constraints, at least local, are required to ensure that we reach a maximum

1. **Speeding-up** the EM Algorithm.

2. Limitations concerning the **M-step**.

   GEM: Generalized EM Algorithm [?, ?]

   Idea: Instead of maximize $Q(\cdot|\theta_k)$ at each step, only find a point θ_k "that makes it grow"

3. Limitations concerning the **E-step**.

   SEM: Stochastic EM Algorithm [?]
   MCEM: Monte-Carlo EM Algorithm [?]
   SAEM: Stochastic-Approximation EM Algorithm [?]

   Idea: Construct a stochastic approximation of $Q(\cdot|\theta_k)$, denoted Q_k

Variants of the EM Algorithm

1. **Speeding-up** the EM Algorithm.

2. Limitations concerning the **M-step**.

   GEM: Generalized EM Algorithm [?, ?]

   Idea: Instead of maximize $Q(\cdot|\theta_k)$ at each step, only find a point θ_k "that makes it grow"

3. Limitations concerning the **E-step**.

   SEM: Stochastic EM Algorithm [?]
   MCEM: Monte-Carlo EM Algorithm [?]
   SAEM: Stochastic-Approximation EM Algorithm [?]

   Idea: Construct a stochastic approximation of $Q(\cdot|\theta_k)$, denoted Q_k

1. **Speeding-up** the EM Algorithm.

2. Limitations concerning the **M-step**.

   GEM: Generalized EM Algorithm [?, ?]

   Idea: Instead of maximize $Q(\cdot|\theta_k)$ at each step, only find a point θ_k "that makes it grow"

3. Limitations concerning the **E-step**.

   SEM: Stochastic EM Algorithm [?]
   MCEM: Monte-Carlo EM Algorithm [?]
   SAEM: Stochastic-Approximation EM Algorithm [?]

   Idea: Construct a stochastic approximation of $Q(\cdot|\theta_k)$, denoted Q_k

Stochastic Variants of the EM Algorithm

SEM – Stochastic EM

S-step: Draw one observed sample

$$z_k \sim q\big(\cdot|y;\theta_k\big)$$

"E"-step: Estim. of $Q(\cdot|\theta_k)$

$$Q_k(\theta) = \log q(y, z_k; \theta)$$

M-step: Maximize Q_{k+1}:

$$\theta_{k+1} \in \operatorname*{argmax}_{\theta\in\Theta} Q_{k+1}(\theta)$$

- ✓ Very easy to install
- ✓ Randomness reduces the dependence on θ_0, more general exploration of the modes of $q(z|y;\theta)$
- × Convergence proved on average only

MCEM – Monte Carlo EM

S-step: Draw m samples

$$z_{k,j} \sim q\big(\cdot|y;\theta_k\big)$$

"E"-step: Monte-Carlo estim.

$$Q_k(\theta) = \frac{1}{m}\sum_{j=1}^{m} \log q(y, z_{k,j}; \theta)$$

M-step: Maximize Q_{k+1}:

$$\theta_{k+1} \in \operatorname*{argmax}_{\theta\in\Theta} Q_{k+1}(\theta)$$

- × Longer computation times
- × No theoretical convergence results known (except in very specific cases)

SAEM – Stochastic Approx.

S-step: Draw a sample

$$z_k \sim q\big(\cdot|y;\theta_k\big)$$

SA-step: Update $Q_k(\theta)$ as

$$Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k\left(\log q(y, z_k; \theta) - Q_k(\theta)\right)$$

M-step: Maximize Q_{k+1}:

$$\theta_{k+1} \in \operatorname*{argmax}_{\theta\in\Theta} Q_{k+1}(\theta)$$

- ✓ Convergence speed
- ✓ With a few more hypotheses (in particular about the γ_k sequence), theoretical convergence demonstrated

Stochastic Variants of the EM Algorithm

SEM – Stochastic EM

S-step: Draw one observed sample

$$z_k \sim q\big(\cdot\,|y;\theta_k\big)$$

"E"-step: Estim. of $Q(\cdot|\theta_k)$

$$Q_k(\theta) = \log q(y, z_k; \theta)$$

M-step: Maximize Q_{k+1}:

$$\theta_{k+1} \in \operatorname*{argmax}_{\theta\in\Theta} Q_{k+1}(\theta)$$

- ✓ Very easy to install
- ✓ Randomness reduces the dependence on θ_0, more general exploration of the modes of $q(z|y\,;\theta)$
- × Convergence proved on average only

MCEM – Monte Carlo EM

S-step: Draw m samples

$$z_{k,j} \sim q\big(\cdot\,|y;\theta_k\big)$$

"E"-step: Monte-Carlo estim.

$$Q_k(\theta) = \frac{1}{m}\sum_{j=1}^{m} \log q(y, z_{k,j}; \theta)$$

M-step: Maximize Q_{k+1}:

$$\theta_{k+1} \in \operatorname*{argmax}_{\theta\in\Theta} Q_{k+1}(\theta)$$

- × Longer computation times
- × No theoretical convergence results known (except in very specific cases)

SAEM – Stochastic Approx.

S-step: Draw a sample

$$z_k \sim q\big(\cdot\,|y;\theta_k\big)$$

SA-step: Update $Q_k(\theta)$ as

$$Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k\left(\log q(y, z_k; \theta) - Q_k(\theta)\right)$$

M-step: Maximize Q_{k+1}:

$$\theta_{k+1} \in \operatorname*{argmax}_{\theta\in\Theta} Q_{k+1}(\theta)$$

- ✓ Convergence speed
- ✓ With a few more hypotheses (in particular about the γ_k sequence), theoretical convergence demonstrated

Stochastic Variants of the EM Algorithm

SEM – Stochastic EM

S-step: Draw one observed sample

$$z_k \sim q\Big(\cdot |y; \theta_k\Big)$$

"E"-step: Estim. of $Q(\cdot|\theta_k)$

$$Q_k(\theta) = \log q(y, z_k; \theta)$$

M-step: Maximize Q_{k+1}:

$$\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q_{k+1}(\theta)$$

- ✓ Very easy to install
- ✓ Randomness reduces the dependence on θ_0, more general exploration of the modes of $q(z|y\,;\theta)$
- × Convergence proved on average only

MCEM – Monte Carlo EM

S-step: Draw m samples

$$z_{k,j} \sim q\Big(\cdot |y; \theta_k\Big)$$

"E"-step: Monte-Carlo estim.

$$Q_k(\theta) = \frac{1}{m}\sum_{j=1}^{m} \log q(y, z_{k,j}; \theta)$$

M-step: Maximize Q_{k+1}:

$$\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q_{k+1}(\theta)$$

- × Longer computation times
- × No theoretical convergence results known (except in very specific cases)

SAEM – Stochastic Approx.

S-step: Draw a sample

$$z_k \sim q\Big(\cdot |y; \theta_k\Big)$$

SA-step: Update $Q_k(\theta)$ as

$$Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k \left(\log q(y, z_k; \theta) - Q_k(\theta)\right)$$

M-step: Maximize Q_{k+1}:

$$\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q_{k+1}(\theta)$$

- ✓ Convergence speed
- ✓ With a few more hypotheses (in particular about the γ_k sequence), theoretical convergence demonstrated

A Detour through Stochastic Approximation Theory

Stochastic Approximation

- A much broader framework than previously seen

  Stochastic Gradient Descent, Stochastic EM

- General framework: We seek for θ^* such that $\boxed{h(\theta^*) = \mathbb{E}_Y\left[H(\theta^*, Y)\right] = 0}$
  - H is known
  - The distribution $\mathbb{P}_Y$ (which may depend on θ^*) is unknown

Vocabulary

- The h function is called mean field

- The idea behind stochastic approximation is to determine $\theta^\star$ iteratively via a scheme of the form

$$\theta_n = \theta_{n-1} + \gamma_n\, h(\theta_{n-1}) + \gamma_n\, \eta_n$$

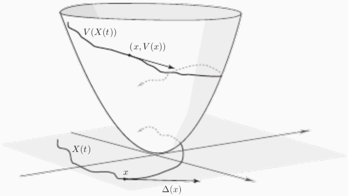
  - $h(\theta) = \mathbb{E}_Y\left[H(\theta, Y)\right]$ is the function we're trying to cancel out
  - $\eta_n = \mathbb{E}_Y\left[H(\theta_{n-1}, Y)\right] - h(\theta_{n-1})$ is a (random) perturbation

Remark: If $\eta_n = 0$, the schema is similar to that of the stochastic gradient

Convergence Considerations

Two steps to show the convergence of the sequence $(\theta_n)_n$

1. Find general conditions on the deterministic sequence of $(\eta_n)_n$ and on h that ensure the deterministic convergence of $(\theta_n)_n$
2. Show that these conditions are satisfied with probability 1, *i.e.* for almost any path of the noise process $(\eta_1(\omega), \ldots, \eta_n(\omega))$



$$\theta_n = \theta_{n-1} + \gamma_n\, h(\theta_{n-1}) + \gamma_n\, \eta_n$$

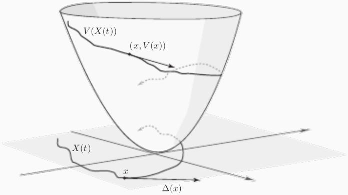
$$\longleftrightarrow \quad \frac{\theta_n - \theta_{n-1}}{\gamma_n} = h(\theta_{n-1}) + \eta_n$$

$\rightsquigarrow$ *Deterministic* convergence of scheme strongly linked to convergence of mean-field equation $\boxed{\frac{\mathrm{d}\theta_t}{\mathrm{d}t} = h(\theta_t)}$

Convergence Considerations

Two steps to show the convergence of the sequence $(\theta_n)_n$

1. Find general conditions on the deterministic sequence of $(\eta_n)_n$ and on h that ensure the deterministic convergence of $(\theta_n)_n$
2. Show that these conditions are satisfied with probability 1, *i.e.* for almost any path of the noise process $(\eta_1(\omega), \dots, \eta_n(\omega))$



$$\theta_n = \theta_{n-1} + \gamma_n \, h(\theta_{n-1}) + \gamma_n \, \eta_n$$

$$\longleftrightarrow \quad \frac{\theta_n - \theta_{n-1}}{\gamma_n} = h(\theta_{n-1}) + \eta_n$$

$\rightsquigarrow$ *Deterministic* convergence of scheme strongly linked to convergence of mean-field equation $\boxed{\dfrac{\mathrm{d}\theta_t}{\mathrm{d}t} = h(\theta_t)}$

A Detour through Stochastic Approximation Theory

Point-wise Convergence for Bounded θ: First Assumption

- Assume that $\eta_n = e_n + r_n \quad \rightsquigarrow \quad \boxed{\theta_n = \theta_{n-1} + \gamma_n\, h(\theta_n) + \gamma_n\, e_n + \gamma_n\, r_n}$

$$\gamma_n \geqslant 0 \qquad\qquad \lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i = \infty \qquad\qquad \lim_{n\to\infty} \gamma_n = 0$$

Assumption (SA1) : Existence of a Lyapunov function

Let $\mathcal{O} \subset \mathbb{R}^d$ an open set, with frontier denoted $\partial\mathcal{O}$.

- h is a continuous vector field over $\mathcal{O}$
- There exists a positive (or null) $\mathcal{C}^1$ function V such that
  (i) $\forall x \in \mathcal{O},\ \left\langle \nabla V(x) \,\middle|\, h(x) \right\rangle \leqslant 0$
  (ii) The set $\mathcal{S} = \left\{ x \middle| \left\langle \nabla V(x) \,\middle|\, h(x) \right\rangle = 0 \right\}$ has an empty interior: $\mathrm{int}\,(V(\mathcal{S})) = \varnothing$

Remark:

- We refer to the function V as a Lyapunov function
  - Not restrictive condition
  - For instance: If $h = -\nabla Q$ with Q of class $\mathcal{C}^1$, $V = Q$ is an admissible choice
- Condition (ii) can be proved using the Sard theorem (*differential geometry*)
  $\rightsquigarrow$ *For all V $\mathcal{C}^d$, $V(\{\nabla V = 0\})$ has an empty interior*

Point-wise Convergence for Bounded θ: First Assumption

- Assume that $\eta_n = e_n + r_n \rightsquigarrow$ $\boxed{\theta_n = \theta_{n-1} + \gamma_n h(\theta_n) + \gamma_n e_n + \gamma_n r_n}$

$$\gamma_n \geqslant 0 \qquad \lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i = \infty \qquad \lim_{n\to\infty} \gamma_n = 0$$

Assumption (SA1) : Existence of a Lyapunov function

Let $\mathcal{O} \subset \mathbb{R}^d$ an open set, with frontier denoted $\partial\mathcal{O}$.

- h is a continuous vector field over $\mathcal{O}$
- There exists a positive (or null) $\mathcal{C}^1$ function V such that
  - (i) $\forall x \in \mathcal{O}$, $\left\langle \nabla V(x) \,\middle|\, h(x) \right\rangle \leqslant 0$
  - (ii) The set $\mathcal{S} = \left\{ x \middle| \left\langle \nabla V(x) \,\middle|\, h(x) \right\rangle = 0 \right\}$ has an empty interior: $\mathrm{int}\,(V(\mathcal{S})) = \varnothing$

Remark:

- We refer to the function V as a Lyapunov function
  - Not restrictive condition
  - For instance: If $h = -\nabla Q$ with Q of class $\mathcal{C}^1$, $V = Q$ is an admissible choice
- Condition (ii) can be proved using the Sard theorem (*differential geometry*)

  $\rightsquigarrow$ *For all V $\mathcal{C}^d$, $V(\{\nabla V = 0\})$ has an empty interior*

Point-wise Convergence for Bounded θ: First Assumption

- Assume that $\eta_n = e_n + r_n \rightsquigarrow$ $\boxed{\theta_n = \theta_{n-1} + \gamma_n h(\theta_n) + \gamma_n e_n + \gamma_n r_n}$

$$\gamma_n \geqslant 0 \qquad \lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i = \infty \qquad \lim_{n\to\infty} \gamma_n = 0$$

Assumption (SA1) : Existence of a Lyapunov function

Let $\mathcal{O} \subset \mathbb{R}^d$ an open set, with frontier denoted $\partial\mathcal{O}$.

- h is a continuous vector field over $\mathcal{O}$
- There exists a positive (or null) $\mathcal{C}^1$ function V such that
  - (i) $\forall x \in \mathcal{O}$, $\left\langle \nabla V(x) \,\middle|\, h(x) \right\rangle \leqslant 0$
  - (ii) The set $\mathcal{S} = \left\{ x | \left\langle \nabla V(x) \,\middle|\, h(x) \right\rangle = 0 \right\}$ has an empty interior: $\mathrm{int}\,(V(\mathcal{S})) = \varnothing$

Remark:

- We refer to the function V as a Lyapunov function
  - Not restrictive condition
  - For instance: If $h = -\nabla Q$ with Q of class $\mathcal{C}^1$, $V = Q$ is an admissible choice
- Condition (ii) can be proved using the Sard theorem (*differential geometry*)

  $\rightsquigarrow$ *For all V $\mathcal{C}^d$, $V(\{\nabla V = 0\})$ has an empty interior*

Point-wise Convergence for Bounded θ

Definition (A-stable)

We say that algorithm $\theta_n = \theta_{n-1} + \gamma_n\, h(\theta_n) + \gamma_n\, e_n + \gamma_n\, r_n$ is A-stable if:

- θ_n stays in a compact subset of $\mathcal{O}$
- $\lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i e_i < \infty$ and $\lim_{n\to\infty} |r_n| = 0$

Theorem (Point-wise Convergence)

*Consider a A-stable algorithm and assume (SA1). **Then**:* $\boxed{\lim_{n\to\infty} d(\theta_n, \mathcal{S}) = 0}$

Especially, if $|\mathcal{S}| < \infty$, θ_n converges toward a point of $\mathcal{S}$

Remark

- The A-stable condition is the baseline assumption for convergence $\rightsquigarrow$ Ok !
- **But**: Assuming that θ stays in a compact is very restrictive !
  $\rightsquigarrow$ We can avoid this assumption by considering a sequence of growing compacts, and assuming that θ_n does not diverge to ∞

For more details: *Cf. the book/poly of Bernard Delyon*

Stochastic approximation with decreasing gain: Convergence and asymptotic theory (2000)

Definition (A-stable)

We say that algorithm $\theta_n = \theta_{n-1} + \gamma_n\, h(\theta_n) + \gamma_n\, e_n + \gamma_n\, r_n$ is A-stable if:

- θ_n stays in a compact subset of $\mathcal{O}$
- $\lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i e_i < \infty$ and $\lim_{n\to\infty} |r_n| = 0$

Theorem (Point-wise Convergence)

*Consider a A-stable algorithm and assume (SA1). **Then**:* $\boxed{\lim_{n\to\infty} d(\theta_n, \mathcal{S}) = 0}$

Especially, if $|\mathcal{S}| < \infty$, θ_n converges toward a point of $\mathcal{S}$

Remark

- The A-stable condition is the baseline assumption for convergence $\rightsquigarrow$ Ok !
- **But**: Assuming that θ stays in a compact is very restrictive !

  $\rightsquigarrow$ We can avoid this assumption by considering a sequence of growing compacts, and assuming that θ_n does not diverge to ∞

For more details: *Cf. the book/poly of Bernard Delyon*

Stochastic approximation with decreasing gain: Convergence and asymptotic theory (2000)

Point-wise Convergence for unBounded θ

Assumption (SA2) : Existence of a Lyapunov function

Let $\mathcal{O} \subset \mathbb{R}^d$ an open set, with frontier denoted $\partial\mathcal{O}$.

- h is a continuous vector field over $\mathcal{O}$
- There exists $K \subset \mathcal{O}$ compact
- There exists a positive (or null) $\mathcal{C}^1$ function V such that
  (i) $V(x) \to +\infty$ if $x \to \partial\mathcal{O}$ or $|x| \to +\infty$
  (ii) $\left\langle \nabla V(x) \,\middle|\, h(x) \right\rangle < 0$ if $x \notin K$

Theorem

Assume:

- *Hypothesis (SA2)*
- *There exists a compact $K_0 \subset \mathcal{O}$ such that $\theta_n \in K_0$ an infinity of times*
- *For all $N \in \mathbb{N}$,*

$$\lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i e_i \mathbb{1}_{\{V(\theta_{i-1}) \leqslant N\}} < \infty \quad \textit{and} \quad \lim_{n\to\infty} |r_n| \mathbb{1}_{\{V(\theta_{i-1}) \leqslant N\}} = 0$$

Then: *The algorithm is A-stable*

Point-wise Convergence for unBounded θ

Assumption (SA2) : Existence of a Lyapunov function

Let $\mathcal{O} \subset \mathbb{R}^d$ an open set, with frontier denoted $\partial\mathcal{O}$.

- h is a continuous vector field over $\mathcal{O}$
- There exists $K \subset \mathcal{O}$ compact
- There exists a positive (or null) $\mathcal{C}^1$ function V such that
  (i) $V(x) \to +\infty$ if $x \to \partial\mathcal{O}$ or $|x| \to +\infty$
  (ii) $\left\langle \nabla V(x) \,\middle|\, h(x) \right\rangle < 0$ if $x \notin K$

Theorem

Assume:

- *Hypothesis (SA2)*
- *There exists a compact $K_0 \subset \mathcal{O}$ such that $\theta_n \in K_0$ an infinity of times*
- *For all $N \in \mathbb{N}$,*

$$\lim_{n\to\infty} \sum_{i=1}^{n} \gamma_i e_i \mathbb{1}_{\{V(\theta_{i-1}) \leqslant N\}} < \infty \quad \textit{and} \quad \lim_{n\to\infty} |r_n| \mathbb{1}_{\{V(\theta_{i-1}) \leqslant N\}} = 0$$

Then: *The algorithm is A-stable*

A Detour through Stochastic Approximation Theory

Robins-Monroe Algorithms

Let's go toward a random noise!

Robins-Monroe Stochastic Approximation

Let the scheme $\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, Y_n)$

- $Y_n \sim \mathbb{P}_{\theta_{n-1}}$ with $\mathbb{P}(Y_n \in A | Y_{n-1}, Y_{n-2}, \ldots, \theta_0) = \mathbb{P}_{\theta_{n-1}}(Y_n \in A)$
- The algorithm seeks to solve $h(\theta) = \mathbb{E}_Y[H(\theta, Y)] = 0$

Rewrite the algorithm to make h explicit:
Set $e_n = H(\theta_{n-1}, Y_n) - h(\theta_{n-1})$,

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n e_n \qquad \text{(RM)}$$

Example: The SAEM and the SGD algorithms are of Robins-Monroe type

Almost-Sure Convergence of Robins-Monroe Schemes

Theorem

Assume:

(R1) $\sum \gamma_n = +\infty$ *and* $\sum \gamma_n^2 < \infty$

(R2)
- h *is continuous*
- $\mathcal{S} = \{\theta | h(\theta) = 0\}$ *is finite*
- *There exists* V *of class* $\mathcal{C}^1$ *such that*
  - (i) $\lim_{x \to \partial \mathcal{O}} V(x) = 0$ *OR* θ_n *remains in a compact*
  - (ii) $\left\langle \nabla V(\theta) \,\middle|\, h(\theta) \right\rangle \leqslant 0$
  - (iii) $\left\{ \left\langle \nabla V(\theta) \,\middle|\, h(\theta) \right\rangle = 0 \right\} = \mathcal{S}$

(R3) *For all compact* $K \subset \mathcal{O}$, $\sup_{\theta \in K} \mathbb{E}_Y \left[\|H(\theta, Y)\|^2 \right] < \infty$

Then: *The (RM) algorithm converge toward* θ^* *such that* $h(\theta^*) = 0$ *with proba 1*

Proof *(Compact case)*

- Compact case $\Longrightarrow$ For all $n \in \mathbb{N}$, $\theta_n \in K_0$

- **Aim**: Apply the previous theorem
  - (R1) and (R2) trivially induce the existence of a Lyapunov function (Made to that !)
  - $r_n = 0$ (and so $\lim_{n\to\infty} |r_n| = 0$)
  - What about $\sum \gamma_n \, e_n$?

- Set $X_n = \gamma_n \, e$ and $S_n = \displaystyle\sum_{i=1}^{n} X_i$

$$\begin{aligned}\mathbb{E}_Y\left[\,|X_n|^2 \,\middle|\, \mathcal{F}_{n-1}\right] &= \gamma_n^2 \, \mathbb{E}_Y\left[\,|e_n|^2 \,\middle|\, \mathcal{F}_{n-1}\right] \\ &\leqslant \sup_{\theta \in K_0} \gamma_n^2 \left[\,|H(\theta, Y) - h(\theta)|^2 \,\middle|\, \mathcal{F}_{n-1}\right]\end{aligned}$$

However, H and h continuous ans therefore bounded on K_0 $\oplus$ using (R3)

$$\mathbb{E}_Y\left[\,|X_n|^2 \,\middle|\, \mathcal{F}_{n-1}\right] \leqslant cste \times \gamma_n^2 \quad \Longrightarrow \quad \mathbb{E}_Y\left[\,|X_n|^2\right] \leqslant cste \times \gamma_n^2$$

Last, since $\sum \gamma_n^2 < \infty$, S_n converges almost surely $\square$

Proof *(Compact case)*

- Compact case $\Longrightarrow$ For all $n \in \mathbb{N}$, $\theta_n \in K_0$
- **Aim**: Apply the previous theorem
  - (R1) and (R2) trivially induce the existence of a Lyapunov function (Made to that !)
  - $r_n = 0$ (and so $\lim_{n\to\infty}|r_n| = 0$)
  - What about $\sum \gamma_n e_n$?
- Set $X_n = \gamma_n e$ and $S_n = \sum_{i=1}^{n} X_i$

$$\mathbb{E}_Y\left[|X_n|^2 \,\middle|\, \mathcal{F}_{n-1}\right] = \gamma_n^2\, \mathbb{E}_Y\left[|e_n|^2 \,\middle|\, \mathcal{F}_{n-1}\right]$$
$$\leqslant \sup_{\theta\in K_0} \gamma_n^2 \left[|H(\theta,Y) - h(\theta)|^2 \,\middle|\, \mathcal{F}_{n-1}\right]$$

However, H and h continuous ans therefore bounded on K_0 $\oplus$ using (R3)

$$\mathbb{E}_Y\left[|X_n|^2 \,\middle|\, \mathcal{F}_{n-1}\right] \leqslant cste \times \gamma_n^2 \quad \Longrightarrow \quad \mathbb{E}_Y\left[|X_n|^2\right] \leqslant cste \times \gamma_n^2$$

Last, since $\sum \gamma_n^2 < \infty$, S_n converges almost surely $\square$

Convergence of Robins-Monroe Schemes

Proof *(Compact case)*

- Compact case $\Longrightarrow$ For all $n \in \mathbb{N}$, $\theta_n \in K_0$
- **Aim**: Apply the previous theorem
  - (R1) and (R2) trivially induce the existence of a Lyapunov function (Made to that !)
  - $r_n = 0$ (and so $\lim_{n\to\infty}|r_n| = 0$)
  - What about $\sum \gamma_n \, e_n$?
- Set $X_n = \gamma_n \, e$ and $S_n = \sum_{i=1}^{n} X_i$

$$\mathbb{E}_Y \left[|X_n|^2 \,\middle|\, \mathcal{F}_{n-1} \right] = \gamma_n^2 \, \mathbb{E}_Y \left[|e_n|^2 \,\middle|\, \mathcal{F}_{n-1} \right]$$
$$\leqslant \sup_{\theta \in K_0} \gamma_n^2 \left[|H(\theta, Y) - h(\theta)|^2 \,\middle|\, \mathcal{F}_{n-1} \right]$$

However, H and h continuous ans therefore bounded on K_0 $\oplus$ using (R3)

$$\mathbb{E}_Y \left[|X_n|^2 \,\middle|\, \mathcal{F}_{n-1} \right] \leqslant cste \times \gamma_n^2 \quad \Longrightarrow \quad \mathbb{E}_Y \left[|X_n|^2 \right] \leqslant cste \times \gamma_n^2$$

Last, since $\sum \gamma_n^2 < \infty$, S_n converges almost surely $\square$

Proof *(Compact case)*

- Compact case $\Longrightarrow$ For all $n \in \mathbb{N}$, $\theta_n \in K_0$
- **Aim**: Apply the previous theorem
  - (R1) and (R2) trivially induce the existence of a Lyapunov function (Made to that !)
  - $r_n = 0$ (and so $\lim_{n\to\infty} |r_n| = 0$)
  - What about $\sum \gamma_n\, e_n$?
- Set $X_n = \gamma_n\, e$ and $S_n = \sum_{i=1}^{n} X_i$

$$\mathbb{E}_Y\left[|X_n|^2 \,\middle|\, \mathcal{F}_{n-1} \right] = \gamma_n^2\, \mathbb{E}_Y\left[|e_n|^2 \,\middle|\, \mathcal{F}_{n-1} \right]$$
$$\leqslant \sup_{\theta \in K_0} \gamma_n^2 \left[|H(\theta, Y) - h(\theta)|^2 \,\middle|\, \mathcal{F}_{n-1} \right]$$

However, H and h continuous ans therefore bounded on K_0 $\oplus$ using (R3)

$$\mathbb{E}_Y\left[|X_n|^2 \,\middle|\, \mathcal{F}_{n-1} \right] \leqslant cste \times \gamma_n^2 \quad \Longrightarrow \quad \mathbb{E}_Y\left[|X_n|^2 \right] \leqslant cste \times \gamma_n^2$$

Last, since $\sum \gamma_n^2 < \infty$, S_n converges almost surely $\square$

Proof *(Compact case)*

- Compact case $\implies$ For all $n \in \mathbb{N}$, $\theta_n \in K_0$
- **Aim**: Apply the previous theorem
  - (R1) and (R2) trivially induce the existence of a Lyapunov function (Made to that !)
  - $r_n = 0$ (and so $\lim_{n\to\infty}|r_n| = 0$)
  - What about $\sum \gamma_n \, e_n$?
- Set $X_n = \gamma_n \, e$ and $S_n = \sum_{i=1}^{n} X_i$

$$\mathbb{E}_Y\left[\,|X_n|^2 \,\middle|\, \mathcal{F}_{n-1}\right] = \gamma_n^2 \, \mathbb{E}_Y\left[\,|e_n|^2 \,\middle|\, \mathcal{F}_{n-1}\right]$$
$$\leqslant \sup_{\theta \in K_0} \gamma_n^2 \left[\,|H(\theta, Y) - h(\theta)|^2 \,\middle|\, \mathcal{F}_{n-1}\right]$$

However, H and h continuous ans therefore bounded on K_0 $\oplus$ using (R3)

$$\mathbb{E}_Y\left[\,|X_n|^2 \,\middle|\, \mathcal{F}_{n-1}\right] \leqslant cste \times \gamma_n^2 \quad \implies \quad \mathbb{E}_Y\left[\,|X_n|^2\right] \leqslant cste \times \gamma_n^2$$

Last, since $\sum \gamma_n^2 < \infty$, S_n converges almost surely $\square$