Introduction à l'algorithme EM

Philippe Langlais
felipe@iro.umontreal.ca

October 19, 2012

Plan

EM par l'exemple

Légitimité d'EM

Applications de EM

- Retour à notre exemple
- Jets de pièces
- Estimation des coefficients d'une combinaison linéaire de modèles
- Retour sur Jelinek & Mercer

Crédits

- Exposition à EM par l'exemple

  Cette partie reprend la présentation faite à EMNLP'2001 par Ted Pedersen lors d'un pannel sur l'algorithme EM (Pedersen 2001a)

- Fondements de EM (Dempster, Laird, and Rubin 1977; Baum 1972) (d'après Jelinek 1998)
- Application à l'estimation de coefficients de pondération dans une mixture de modèles (d'après Berger 2000)

- (Pedersen 2001b) propose une bonne liste de pointeurs sur EM.

EM par l'exemple

Légitimité d'EM

Applications de EM

- Retour à notre exemple
- Jets de pièces
- Estimation des coefficients d'une combinaison linéaire de modèles
- Retour sur Jelinek & Mercer

EM par l'exemple

Soit:

- n entités classées dans 4 catégories $Y = (y_1, y_2, y_3, y_4)$
- $\theta = (\theta_1 = \frac{1}{2} + \frac{1}{4}\pi, \theta_2 = \frac{1}{4}(1-\pi), \theta_3 = \frac{1}{4}(1-\pi), \theta_4 = \frac{1}{4}\pi)$ les probabilités associées à chaque catégorie, définies par rapport à un paramètre π que l'on souhaite apprendre.

Alors:

- la probabilité d'observer sur n tirages la classification en 4 classes selon des comptes y_1, y_2, y_3 et y_4 est donnée par:

$$\mathcal{L}(\pi) = p(y_1, y_2, y_3, y_4) = \underbrace{\frac{n!}{y_1! y_2! y_3! y_4!}}_{\alpha} \theta_1^{y_1} \theta_2^{y_2} \theta_3^{y_3} \theta_4^{y_4}$$

Et $\log \mathcal{L}(\pi) =$
$\log \alpha + y_1 \log(\frac{1}{2} + \frac{1}{4}\pi) + (y_2 + y_3) \log(\frac{1}{4}(1-\pi)) + y_4 \log(\frac{1}{4}\pi)$

Estimateur à maximum de vraisemblance

Rappel

$$\frac{\delta}{\delta\pi}\log\mathscr{L}(\pi) = y_1\frac{1/4}{(2+\pi)/4} + (y_2+y_3)\frac{-1}{1-\pi} + y_4\frac{1}{\pi}$$

Maximum si:

$$\frac{y_1}{2+\pi} - \frac{y_2+y_3}{1-\pi} + \frac{y_4}{\pi} = 0$$

- Si on observe pour $n = 197$ les comptes: $(125, 18, 20, 34)$
- Alors: $\pi = 0.627$

La connaissance des y_i est suffisante pour le calcul analytique du maximum de vraisemblance. On parle de statistique suffisante.

Changement des données du problème

- En regardant mieux le problème, on réalise:
  - qu'il y a 5 classes
  - que y_1 est en fait la composition de deux classes: $x_1 + x_2$
  - que $p(x_1) = \frac{1}{2}$ et que $p(x_2) = \frac{1}{4}\pi$
- **Problème:** les expérimentations sont terminées et on ne peut plus faire de mesures des comptes de x_1 et x_2
  - On ne peut pas calculer le maximum de vraisemblance. On dit qu'on est en présence d'une statistique (ou de données) incomplète(s)
- EM nous permet de contourner le problème.
  - Principe: prendre les espérances des comptes manquants afin d'obtenir une statistique suffisante pour calculer le maximum de vraisemblance.
  - Boucler tant qu'on améliore la vraisemblance des données incomplètes.

La recette EM sur notre exemple

Données du nouveau problème

- $X = (x_1, x_2, y_2, y_3, y_4) = (x_1, x_2, 18, 20, 34)$
- $x_1 + x_2 = y_1 = 125$
- $\theta = (p_1 = \frac{1}{2}, p_2 = \frac{1}{4}\pi, \theta_2 = \frac{1}{4}(1-\pi), \theta_3 = \frac{1}{4}(1-\pi), \theta_4 = \frac{1}{4}\pi)$

- **M-STEP** (Maximization)
  - $\log \mathscr{L}(\pi) =$
    $-\log \alpha' + x_1 \log(\frac{1}{2}) + (x_2 + y_4)\log(\frac{\pi}{4}) + (y_2 + y_3)\log(\frac{1-\pi}{4})$
  - $\frac{\delta}{\delta\pi}\log \mathscr{L}(\pi) = \frac{x_2+y_4}{\pi} - \frac{y_2+y_3}{1-\pi} = 0$
  - $\pi = \frac{x_2+y_4}{x_2+y_4+y_3+y_2}$

  mais x_2 n'est pas connu !
- **E-STEP** (Expectation).

  $X = (y_1 - \hat{x_2}, \hat{x_2}, y_2, y_3, y_4)$ constitue alors une statistique complète à partir de laquelle on peut calculer le maximum de vraisemblance

Mise en place

- On a 2 catégories, X_1 et X_2 avec pour comptes respectifs (inconnus) x_1 et x_2, sachant qu'il y a eu $y_1 = 125$ tirages indépendants avec:
  - p_1, la chance d'avoir un élément classé X_1
  - $p_2 = (1 - p_1)$ la probabilité d'avoir un élément classé X_2.
- Soit $p(x_1)$ la probabilité d'obtenir x_1 occurrences de la classe X_1:
  - $p(x_1) = \binom{y_1}{x_1} {p_1}^{x_1} {p_2}^{y_1 - x_1}$
  - et on sait que $E[x_1 | y_1] = y_1 p_1$
- Comme $p_1 = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{4}\pi}$ alors $\begin{cases} \hat{x_1} & = 125 / \left(2 \times \left(\frac{1}{2} + \frac{1}{4}\pi\right)\right) \\ \hat{x_2} & = 125 - \hat{x_1} \end{cases}$

La recette EM sur notre exemple

2 lancements de l'algorithme

INIT	$\pi \longleftarrow 0.5$	$\pi \longleftarrow 0.1$
E-STEP 1	$\hat{x}_2 = 25$	$\hat{x}_2 = 5.95$
M-STEP 1	$\pi \longleftarrow 0.6082$	$\pi \longleftarrow 0.5125$
E-STEP 2	$\hat{x}_2 = 29.15$	$\hat{x}_2 = 25.49$
M-STEP 2	$\pi \longleftarrow 0.6243$	$\pi \longleftarrow 0.6102$
E-STEP 3	$\hat{x}_2 = 29.74$	$\hat{x}_2 = 29.22$
M-STEP 3	$\pi \longleftarrow 0.6264$	$\pi \longleftarrow 0.6245$
E-STEP 4	$\hat{x}_2 = 29.82$	$\hat{x}_2 = 29.81$
M-STEP 4	$\pi \longleftarrow 0.6267$	$\pi \longleftarrow 0.6267$
E-STEP 5	$\hat{x}_2 = 29.82$	$\hat{x}_2 = 29.81$
M-STEP 5	$\pi \longleftarrow 0.6268$	$\pi \longleftarrow 0.6267$
E-STEP 6	$\hat{x}_2 = 29.82$	$\hat{x}_2 = 29.82$
M-STEP 6	$\pi \longleftarrow 0.6268$	$\pi \longleftarrow 0.6268$

EM par l'exemple

Légitimité d'EM

Applications de EM

- Retour à notre exemple
- Jets de pièces
- Estimation des coefficients d'une combinaison linéaire de modèles
- Retour sur Jelinek & Mercer

EM: la recette

- On observe des réalisations y d'une distribution que l'on essaye de modéliser à l'aide d'un modèle paramétrique de paramètres λ.
- On fait l'hypothèse que y est une statistique incomplète et qu'il existe une variable cachée h dont la connaissance nous donne une statistique complète (y, h).

  Note: Cela revient à dire que l'on fait une hypothèse d'une distribution jointe qu'il est plus facile de modéliser que celle de y: $p(y, h|\lambda)$
- L'idée de EM est de faire une estimée de la vraisemblance de la donnée complète $p(y, h)$ à partir de notre estimée courante λ' que nous utilisons pour estimer les nouveaux paramètres λ.

EM: la recette

- **INIT** définir des valeurs initiales pour les paramètres λ'
- **E-STEP** (Expectation): Calcul de la fonction auxiliaire:
  - $Q(\lambda, \lambda') = E_h[\log p_\lambda(y, h)|\lambda', y]$
- **M-STEP** (Maximization):
  - $\hat{\lambda} = \text{argmax}_\lambda \, Q(\lambda, \lambda')$
- Boucler sur **E-STEP** avec $\lambda' \longleftarrow \hat{\lambda}$ tant que la convergence n'est pas déjà atteinte

Cette recette, nous amène (lorsqu'elle est applicable) à un maximum (souvent local) de la vraisemblance des données incomplètes.

Conditions d'application

- Être capable d'identifier la statistique suffisante et disposer d'un moyen de calculer les espérances des données manquantes
- Pouvoir résoudre le problème de maximisation sur la statistique complète
- Le choix des valeurs initiales des paramètres peut conditionner le résultat de l'optimisation
  - dans le cas général, la vraisemblance a de nombreux maxima locaux, et EM n'est garanti de trouver qu'un de ces maxima

Légitimité de EM

gain de vraisemblance: $\log p(y|\lambda) - \log p(y|\lambda')$

$$
\begin{aligned}
&= \overbrace{\sum_h p(h|y,\lambda')}^{1} \log p(y|\lambda) - \overbrace{\sum_h p(h|y,\lambda')}^{1} \log p(y|\lambda') \\
&= \textstyle\sum_h p(h|y,\lambda') \log p(y|\lambda) \frac{p(h,y|\lambda)}{p(h,y|\lambda)} - \sum_h p(h|y,\lambda') \log p(y|\lambda') \frac{p(h,y|\lambda')}{p(h,y|\lambda')} \\
&= \textstyle\sum_h p(h|y,\lambda') \log \frac{p(h,y|\lambda)}{p(h|y,\lambda)} - \sum_h p(h|y,\lambda') \log \frac{p(h,y|\lambda')}{p(h|y,\lambda')} \\
&= \textstyle\sum_h p(h|y,\lambda') \log p(h,y|\lambda) - \sum_h p(h|y,\lambda') \log p(h,y|\lambda') \\
&+ \textstyle\sum_h p(h|y,\lambda') \log p(h|y,\lambda') - \sum_h p(h|y,\lambda') \log p(h|y,\lambda) \\
&\\
&\geq \textstyle\sum_h p(h|y,\lambda') \log p(h,y|\lambda) - \sum_h p(h|y,\lambda') \log p(h,y|\lambda') \\
&= \textstyle\sum_h p(h|y,\lambda') \log \frac{p(h,y|\lambda)}{p(h,y|\lambda')}
\end{aligned}
$$

L'inégalité étant la résultante de l'application de l'inégalité de Jensen:

- ici: $\sum_x p(x) \log \frac{p(x)}{q(x)} \geq 0$, égalité ssi $p = q$

Légitimité de EM

$$
\begin{array}{lll}
\text{si} & \sum_h p(h|y,\lambda')\log p(h,y|\lambda) & > \quad \sum_h p(h|y,\lambda')\log p(h,y|\lambda') \\
\text{alors} & \log p(y|\lambda) & > \quad \log p(y|\lambda') \\
\text{cad} & p(y|\lambda) & > \quad p(y|\lambda')
\end{array}
$$

- Si on arrive à trouver λ tel que:
  - $\sum_h p(h|y,\lambda')\log p(h,y|\lambda) > \sum_h p(h|y,\lambda')\log p(h,y|\lambda')$
  - alors le modèle sous le régime λ ne peut que s'améliorer (sur les données d'entraînement).
- Il suffit de maximiser le terme de gauche:

$$
\begin{aligned}
& \operatorname{argmax}_\lambda \sum_h p(h|y,\lambda')\log p(h,y|\lambda) \\
= \quad & \operatorname{argmax}_\lambda \sum_h \underbrace{p(h|y,\lambda')\times p(y|\lambda')}\log p(h,y|\lambda) \\
= \quad & \operatorname{argmax}_\lambda \sum_h p(h,y|\lambda')\log p(h,y|\lambda) \\
= \quad & \operatorname{argmax}_\lambda Q(\lambda,\lambda')
\end{aligned}
$$

Application de EM à notre exemple

Rappel du problème

- $Y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$ est une donnée incomplète
- $X = (x_1, x_2, \underbrace{x_3}_{y_2}, \underbrace{x_4}_{y_3}, \underbrace{x_5}_{y_4})$ est la donnée complète
- mais x_1 et x_2 sont inconnus et contraints par $x_1 + x_2 = y_1$
- la multinomiale régissant X est telle que:
  $\theta = (p_1, p_2, p_3, p_4, p_5) = (\frac{1}{2}, \frac{\pi}{4}, \frac{1-\pi}{4}, \frac{1-\pi}{4}, \frac{\pi}{4})$
  qui ne dépendent que d'un paramètre (π) que nous souhaitons estimer.

Application de EM à notre exemple

Application de la recette

$$\frac{\delta}{\delta\pi}\sum_{x_2} P_{\theta'}(x_2|Y)\log P(X|\theta) = 0$$

$$\frac{\delta}{\delta\pi}\sum_{x_2} P_{\theta'}(x_2|Y)\left[\log\alpha + \sum_{i=1}^{5} x_i \log p_i\right] = 0$$

$$\underbrace{\frac{\delta}{\delta\pi}\sum_{x_2} P_{\theta'}(x_2|Y)\log\alpha}_{0} + \frac{\delta}{\delta\pi}\sum_{x_2} P_{\theta'}(x_2|Y)\left[\sum_{i=1}^{5} x_i \log p_i\right] = 0$$

$$\sum_{x_2} P_{\theta'}(x_2|Y)\sum_{i=1}^{5} x_i \frac{\delta}{\delta\pi}\log p_i = 0$$

Application de EM à notre exemple

Application de la recette

$$\sum_{x_2} P_{\theta'}(x_2|Y)\left[x_1 \times 0 + (x_2+x_5)\frac{\delta}{\delta\pi}\left(\log\frac{\pi}{4}\right) + (x_3+x_4)\frac{\delta}{\delta\pi}\left(\log\frac{1-\pi}{4}\right)\right]$$

$$\sum_{x_2} P_{\theta'}(x_2|Y)\left[\frac{x_2+x_5}{\pi} - \frac{x_3+x_4}{1-\pi}\right] = 0$$

$$\sum_{x_2} P_{\theta'}(x_2|Y)\left[(1-\pi)(x_2+x_5) - \pi(x_3+x_4)\right] = 0$$

$$\sum_{x_2} P_{\theta'}(x_2|Y)\left[(1-\pi)(x_2+x_5)\right] - \pi(x_3+x_4) = 0$$

$$\sum_{x_2} P_{\theta'}(x_2|Y)\left[(x_2+x_5) - \pi(x_2+x_5)\right] - \pi(x_3+x_4) = 0$$

$$\underbrace{\sum_{x_2} P_{\theta'}(x_2|Y)x_2}_{\bar{x}_2} + x_5 - \sum_{x_2} P_{\theta'}(x_2|Y)\pi x_2 - \pi x_5 - \pi(x_3+x_4) = 0$$

Application de EM à notre exemple

Application de la recette

$$\bar{x}_2 + x_5 - \pi\bar{x}_2 - \pi x_5 - \pi(x_3 + x_4) = 0$$

d'où:

$$\pi = \frac{\bar{x}_2 + x_5}{\bar{x}_2 + x_3 + x_4 + x_5}$$

- EM consiste donc bien ici à:
  - **E-STEP**: calculer ici l'espérance du compte manquant x_2 (sous le régime π') de manière à
  - **M-STEP**: pouvoir réestimer π

EM par l'exemple

Légitimité d'EM

Applications de EM

- Retour à notre exemple
- Jets de pièces
- Estimation des coefficients d'une combinaison linéaire de modèles
- Retour sur Jelinek & Mercer

Jets de pièces

- Soit N observations
  $X = \{X_1 = \langle x_{11}, x_{12}, \ldots, x_{1K} \rangle, \ldots, X_N = \langle x_{N1}, x_{N2}, \ldots, x_{NK} \rangle\}$
  issues du processus (génératif) suivant:
  Pour chaque $i \in [1, N]$:
  - choisir une pièce c_i parmi P ($c_i \in [1, P]$)
  - générer K tirages indépendants à l'aide de cette pièce (X_i)
- Les paramètres (θ inconnus) du problème sont:
  - λ_c, $c \in [1, P]$, la probabilité à priori de choisir la pièce c
  - p_c, $c \in [1, P]$ la probabilité de tirer Face (F) avec la pièce c.
- Cherchons à estimer $\theta = \{(p_c, \lambda_c)\}_{c \in [1,P]}$ à partir de X, sous la contrainte $\sum_{c=1}^{P} \lambda_c = 1$.

Jets de pièces

- Donnée manquante: $Y = \{Y_1, \ldots, Y_N\}$,
  $Y_i \in [1, P]$, la pièce choisie pour le ie tirage.
- EM nous dit de calculer:

$$\begin{aligned} \hat{\theta} &= \text{argmax}_\theta \, E\left[\log P(X, Y|\theta)|X, \theta'\right] \\ &= \text{argmax}_\theta \sum_{i=1}^{N} E\left[\log P(X_i, Y_i|\theta)|X_i, \theta'\right] \end{aligned}$$

- Or:

$$\begin{aligned} P(X_i, Y_i|\theta) &= P(Y_i|\theta) \times P(X_i|\theta, Y_i) \\ &= \lambda_c \times P(F_i = \sum_{k=1}^{K} \delta(X_{ik}, F)|\theta, Y_i) \\ &= \lambda_c \times \binom{K}{F_i} p_c^{F_i} (1 - p_c)^{K - F_i} \end{aligned}$$

- Donc:

$$E = \sum_i \sum_c p(c|X_i, \theta') \left[\log \lambda_c + \log cte + F_i \log p_c + (K - F_i) \log(1 - p_c\right.$$

Jets de pièces

L' apriori sur la pièce c, λ_c, vérifie (β un coefficient de Lagrange):

$$\frac{\delta}{\delta \lambda_c} E - \beta = 0$$

Soit:

$$\lambda_c = \frac{\sum_i p(c|X_i, \theta')}{\beta}$$

Et donc:

$$\lambda_c = \frac{\sum_i p(c|X_i, \theta')}{\sum_c \sum_i p(c|X_i, \theta') = N}$$

car:

$$\sum_c \sum_i p(c|X_i, \theta') = \sum_i \sum_c p(c|X_i, \theta') = \sum_i 1 = N$$

Jets de pièces

La probabilité p_c d'un tirage F étant donnée la pièce c vérifie:

$$\frac{\delta E}{\delta p_c} = \sum_i p(c|X_i, \theta') \left[\frac{F_i}{p_c} - \frac{K - F_i}{1 - p_c} \right] = 0$$

cad:

$$\begin{aligned} \sum_i p(c|X_i, \theta') \left[F_i(1 - p_c) - (K - F_i) p_c \right] &= 0 \\ \sum_i p(c|X_i, \theta') F_i - \sum_i p(c|X_i, \theta') K p_c &= 0 \end{aligned}$$

D'où:

$$p_c = \frac{\sum_i F_i p(c|X_i, \theta')}{K \sum_i p(c|X_i, \theta')} = \frac{\sum_i \frac{F_i}{K} \times p(c|X_i, \theta')}{\sum_i p(c|X_i, \theta')}$$

Note: $p(c|X_i, \theta')$ l'à postériori (sous l'ancien régime) d'avoir choisi la pièce c étant donnée l'observation du ie tirage X_i:

$$p(Y_i = c|X_i, \theta') = \frac{p(Y_i = c, X_i|\theta')}{p(X_i|\theta')} = \frac{p(Y_i = c, X_i|\theta')}{\sum_{c=1}^{P} p(Y_i = c, X_i|\theta')}$$

EM par l'exemple

Légitimité d'EM

Applications de EM

- Retour à notre exemple
- Jets de pièces
- Estimation des coefficients d'une combinaison linéaire de modèles
- Retour sur Jelinek & Mercer

Estimation des coefficients d'une combinaison linéaire de modèles

Problème

- on dispose de N modèles (connus): $p_1(y), p_2(y), \ldots, p_N(y)$
- et d'un corpus d'entraînement $O = y_1, \ldots, y_T$, et:

$$p_\lambda(y) = \sum_{i=1}^{N} \lambda_i p_i(y) \text{ avec } \lambda_i \in [0,1] \text{ et } \sum_{i=1}^{N} \lambda_i = 1$$

- On cherche $\hat{\lambda} = \text{argmax}_\lambda \log p_\lambda(y)$
- On peut appliquer EM:
  - La variable cachée est l'état i ($i \in [1, N]$) dans lequel se trouve le modèle p_λ au moment de la prédiction.
    Pensez: dans l'état i, le modèle combiné s'appuie sur le i-ème modèle.

Estimation des coefficients d'une combinaison linéaire de modèles

Notre fonction auxiliaire (λ' le jeu de paramètres courant) est:

$$Q(\lambda, \lambda') = \sum_y \tilde{p}(y) \sum_{i=1}^{N} p_{\lambda'}(s = i|y) \log p_\lambda(y, s = i)$$

où $\tilde{p}$ est la distribution empirique, et:

- λ_i est l'*apriori* d'être dans l'état i
- $p_\lambda(s = i, y) = \lambda_i \times p_i(y)$ est la probabilité d'être dans l'état i et de générer y
- et $p_{\lambda'}(s = i|y) = \frac{\lambda'_i p_i(y)}{\sum_i \lambda'_i p_i(y)}$ est la probabilité d'être dans l'état i sachant que y est l'observation courante

Estimation des coefficients d'une combinaison linéaire de modèles

- EM nous dit de maximiser (sur λ) la fonction auxiliaire sous la contrainte que les coefficients somment à 1.
- On introduit pour cela un multiplicateur de Lagrange α:

$$
\begin{aligned}
\frac{\delta}{\delta\lambda_i}\left[Q(\lambda,\lambda') - \alpha\left(\sum_i \lambda_i - 1\right)\right] &= 0 \\
\sum_y \tilde{p}(y) p_{\lambda'}(s=i|y) \frac{\partial}{\partial\lambda_i}\left[\log \lambda_i p_i(y)\right] - \alpha &= 0 \\
\sum_y \tilde{p}(y) p_{\lambda'}(s=i|y) \frac{1}{\lambda_i} - \alpha &= 0 \\
\underbrace{\frac{1}{\lambda_i}\sum_y \tilde{p}(y) p_{\lambda'}(s=i|y)}_{C_i} - \alpha &= 0
\end{aligned}
$$

- Finalement: $\lambda_i = \frac{C_i}{\sum_i C_i}$

Estimation des coefficients d'une combinaison linéaire de modèles

Notre algorithme EM revient donc à:

- Initialiser λ' (n'importe quel choix tel que $\sum_i \lambda'_i = 1$)
- Repéter jusqu'à convergence:

  E-step calcul des comptes C_i selon la formule précédente (c'est une fonction de λ')

  M-step pour tout i, $\lambda_i \leftarrow \frac{C_i}{\sum_i C_i}$

Note: si l'on voit C_i comme le nombre espéré de fois où le modèle i sera utilisé pour générer l'observation (étant donné un jeu de paramètre λ'), alors cet algorithme est assez intuitif.

EM par l'exemple

Légitimité d'EM

Applications de EM

- Retour à notre exemple
- Jets de pièces
- Estimation des coefficients d'une combinaison linéaire de modèles
- Retour sur Jelinek & Mercer

Retour au modèle Jelinek-Mercer

Rappel du problème

- on a N modèles n-grammes (en pratique pour un trigramme, $N = 3$) que l'on combine avec des coefficients λ_i qui dépendent du contexte:

$$p(w_t|\overbrace{w_{t-2}w_{t-1}}^{h_t}) = \sum_{i=1}^{N} \lambda_i(\theta(h_t))p_i(w_t|h_{t,i})$$

  - où $\theta(h_t)$ est n'importe quelle fonction appliquée à l'historique (ex: $\theta(h) \rightarrow \lfloor \log|h| \rfloor$)
  - et $h_{t,i}$ désigne les $N-i$ derniers mots du contexte h_t.
- On souhaite estimer les coefficients λ à maximum de vraisemblance
- c'est une instance du modèle précédent à ceci près que les comptes sont conditionnés par les historiques.

Retour au modèle Jelinek-Mercer

- Statistique manquante: la contribution $\hat{c}_i(\theta(w''w'))$ de chaque modèle p_i lors d'une prédiction dans le contexte $w''w'$.
- On applique la recette EM précédente:
  - E-STEP:

$$\hat{c}_i(\theta(w''w')) = \sum_{t=1:\theta(w_{t-2},w_{t-1})=\theta(w'',w')}^{T} \frac{\lambda_i(\theta(w''w'))p_i(w_t|h_{t,i})}{\sum_{i=1}^{N}\lambda_i(\theta(w''w'))p_i(w_t|h_{t,i})}$$

  - M-STEP:

$$\lambda_i(\theta(w''w')) \longleftarrow \frac{\hat{c}_i(\theta(w''w'))}{\sum_{i=1}^{N}\hat{c}_i(\theta(w''w'))}$$

- **Note:** En général, les estimées obtenues dépendent grandement des valeurs initiales des λ (optimum local).
- Estimées sur held-out !!!

Retour au modèle Jelinek-Mercer

Un codage possible

- Soit S le domaine de θ
- $\lambda = (\lambda_1 \dots \lambda_N)$ et $c = (c_1 \dots c_N)$, deux matrices N$\times|S|$.
- Soit M un tableau de N réels

loop
  {E-STEP}
  for all $s \in S$, $i = 1 \rightarrow N$ **do**
    $c_i[s] \leftarrow 0$
  for all $t = 1 \rightarrow |\mathscr{T}|$ **do**
    $somme \leftarrow 0$
    for all $i = 1 \rightarrow N$ **do**
      $M[i] = \lambda_i[\theta(h_t)] \times p_i(w_t|h_{t,i})$
      $somme+ = M[i]$
    for all $i = 1 \rightarrow N$ **do**
      $c_i[\theta(h_t)]+ = M[i]/somme$

{M-STEP}
for all $s \in S$ **do**
  $sum \leftarrow 0$
  for all $i = 1 \rightarrow N$ **do**
    $sum+ = c_i[\theta(h)]$
  for all $i = 1 \rightarrow N$ **do**
    $\lambda_i[\theta(h)] \leftarrow c_i[\theta(h)]/sum$

Technique de minimisation générique

La descente de gradient (version de base)

- minimiser une fonction $f(\theta)$ d'un vecteur de paramètre θ:

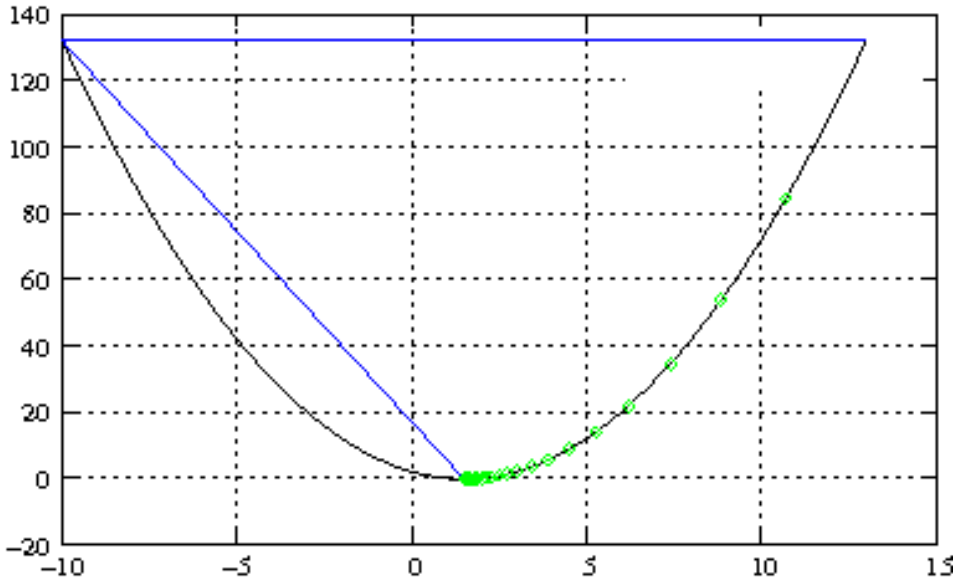
  Initialise θ, fixe ε
  $t \leftarrow 0$
  repeat
    $\delta \leftarrow \eta(t)\nabla f(\theta)$
    $\theta \leftarrow \theta - \delta$
    $t \leftarrow t + 1$
  until $\delta < \varepsilon$
  return θ

- ε est un méta-paramètre (contrôle le test d'arrêt)
- $\eta(t)$ est appelée le learning rate:
  - ralentit la descente si trop petit
  - une valeur trop grande peut diverger
  - souvent une fonction décroissante du temp: ex. $\eta(t) = \frac{1}{1+t}$

Technique de minimisation générique

Minimisation de $x^2 - 3x + 2$

Init: $x = 13$, $\varepsilon = 0.000000001$



t	x	$f(x)$	$\nabla f(x)$	$\eta(t)$	δ
			$\eta(t) = 1/(1+t)$		
0	-10	132	23	1	23
1	1.5	-0.25	-23	0.5	-11.5
2	1.5	-0.25	0	0.3	0
			$\eta(t) = 0.1$		
0	10.7	84.39	23	0.1	2.3
1	8.86	53.92	18.4	0.1	1.84
2	7.38	34.42	14.72	0.1	1.47
			...		
95	1.5	-0.25	1.4e-08	0.1	1.4e-09
96	1.5	-0.25	1.1e-08	0.1	1.1e-09
97	1.5	-0.25	9.1e-09	0.1	9.1e-10

Un peu lourd pour trouver $x = \frac{3}{2}$ ...

Références I


Baum, L.E. (1972). “An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process”. In: *Inequalities* 3, pp. 1–8.


Berger, Adam (2000). *Convexity, Maximum Likelihood and All That*. School of Computer Sceince, Carnegie Mellon University.


Dempster, A.P., N.M. Laird, and D.B. Rubin (1977). “Maximum Likelihood from Incomplete Data via the EM algorithm”. In: *Journal of the Royal Statistical Society* 39(1), pp. 1–38.


Jelinek, Frederick (1998). *Statistical Methods for Speech Recognition*. Cambridge, Massachusetts: The MIT Press.


Pedersen, Ted (2001a). *A Gentle Introduction to the EM algorithm*. Transparents présentés au panel sur l’algorithme EM qui s’est tenu à la conférence Empirical Methods in Natural Language Processing (EMNLP).


— (2001b). “The EM algorithm, Selected readings”. Unpublished notes to accompany the panel discussion on the EM algorithm at EMNLP 2001.