Preliminaries

Expectation-Maximization

Let $\mathbf{Y}$ be the random vector of observed data, whose realizations $\mathbf{y}$ we call **observed data**. Similarly, let $\mathbf{X}$ be the random vector of **missing** data whose realizations $\mathbf{x}$ we do not observe. Our aim is to find the maximum likelihood estimator $\boldsymbol{\psi}_{\text{MLE}}$ of the *complete-data* likelihood $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})$ but since realizations $\mathbf{x}$ are missing we aim instead to maximize the incomplete-data likelihood $p(\mathbf{y} \mid \boldsymbol{\psi})$.

Algorithm 1: Expectation-Maximization

Initialize $\boldsymbol{\psi}_0$ randomly.
until convergence:
Expectation Step: Calculate expectation of **complete-data** log-likelihood.

$$Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_k) = \mathbb{E}_{p(\mathbf{x}|\mathbf{y},\boldsymbol{\psi}_k)} \left[\log p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})\right]$$

Maximization Step: Maximize the expectation with respect to $\boldsymbol{\psi}$

$$\boldsymbol{\psi}_{k+1} := \underset{\boldsymbol{\psi} \in \boldsymbol{\Psi}}{\arg\max}\, Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_k)$$

Check convergence of the sequence $(\boldsymbol{\psi}_k)_k$.

The EM algorithm has the powerful property that an increase in the $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_k)$ function will force an increase at least as big in the incomplete-data likelihood $p(\mathbf{y} \mid \boldsymbol{\psi})$. The method was introduced more rigorously by Dempster et al. [1977] although the correct proof of convergence for cases when $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})$ does not follow a distribution in the exponential family was provided later by Wu [1983].

Robbins-Monro Procedure

Suppose we want to find the root $\boldsymbol{\theta}_*$ of a function $h(\boldsymbol{\theta})$ of which we can only obtain *noisy measurements* $H(\boldsymbol{\theta}, \mathbf{X})$ where $\mathbf{X}$ is a random vector and $\mathbb{E}[H(\boldsymbol{\theta}, \mathbf{X})] = h(\boldsymbol{\theta})$. Then under some regularity conditions the following iterative scheme converges to one of the roots

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + \gamma_n H(\boldsymbol{\theta}_{n-1}, \mathbf{X}_n)$$

where $(\gamma_n)_n$ is a non-negative sequence of step-sizes converging to 0. This procedure was first introduced by Robbins and Monro [1951] in the context of finding the value $\boldsymbol{\theta}_*$ of a *regression function* $h(\boldsymbol{\theta})$ giving $h(\boldsymbol{\theta}_*) = \alpha$.

General Form of Stochastic Approximations

Benveniste et al. [1990] provides an extensive study of the general form of stochastic approximation algorithms

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \gamma_{n+1} H(\boldsymbol{\theta}_n, \mathbf{X}_{n+1}) + \gamma_{n+1}^2 \rho_{n+1}(\boldsymbol{\theta}_n, \mathbf{X}_{n+1})$$

where $\boldsymbol{\theta}_n \in \mathbb{R}^d$ is the parameter of interest, $\mathbf{X}_n \in \mathbb{R}^k$ is considered to be a random vector representing the **state** of the system, $(\gamma_n)_n$ are step sizes, $H : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$ now represents how $\boldsymbol{\theta}$ changes due to new observations of $\mathbf{X}_n$ and $\rho_{n+1} : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$ represents a perturbation. The Robbins-Monro procedure can then be seen as the special case with no perturbation, additional restrictions on the evolution of the state vector $\mathbf{X}_n$ and additional conditions regarding the step sizes γ_n.

Curved Exponential Family

If a probability density function can be expressed as follows we say it belongs to the *exponential family.*

$$p(\mathbf{z} \mid \boldsymbol{\psi}) \propto \exp\left(-\xi(\boldsymbol{\psi}) + \langle T(\mathbf{z}) \,,\, \phi(\boldsymbol{\psi})\rangle\right)$$

If the dimension of $\boldsymbol{\psi}$ is less than the dimension of $\phi(\boldsymbol{\psi})$ then we say the family is *curved.*

Context of the Stochastic Approximation EM Algorithm

Setting and Pseudo-Algorithm

Suppose the **complete-data** likelihood has a distribution in the **curved exponential** family

$$p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi}) \propto \exp\left(-\xi(\boldsymbol{\psi}) + \langle \widetilde{\theta}(\mathbf{x}) \, , \, \phi(\boldsymbol{\psi}) \rangle\right)$$

where the dependence on the observed data implicit. If computation of the following integral is **intractable**

$$\mathbb{E}_{p(\mathbf{x}|\mathbf{y},\boldsymbol{\psi})}\left[\widetilde{\theta}(\mathbf{X})\right] = \int \widetilde{\theta}(\mathbf{x}) p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi}) d\mathbf{x}$$

then the Maximum Likelihood Estimate $\boldsymbol{\psi}_{\text{MLE}}$ can be found with the following algorithm.

Algorithm 2: SAEM

Initialize $\boldsymbol{\psi}_0$ randomly, and choose *large* step size γ_1. Set $\gamma_0 = 1$ and $\boldsymbol{\theta}_{-1} = \mathbf{0}$

until convergence:

Sample $\mathbf{x}_k^{(1)}, \ldots, \mathbf{x}_k^{(m_k)}$ from $p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi}_k)$ and calculate

$$\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_{k-1} + \gamma_k \left(\frac{1}{m_k} \sum_{j=1}^{m_k} \widetilde{\theta}(\mathbf{x}_k^{(j)}) - \boldsymbol{\theta}_{k-1} \right)$$

Maximize stochastic approximation of conditional expectation.

$$\boldsymbol{\psi}_{k+1} := \arg\max_{\boldsymbol{\psi} \in \boldsymbol{\Psi}} -\xi(\boldsymbol{\psi}) + \langle \boldsymbol{\theta}_k \, , \, \phi(\boldsymbol{\psi}) \rangle$$

Compute *averaged sequence* of parameters

$$\overline{\boldsymbol{\psi}}_{k+1} \leftarrow \overline{\boldsymbol{\psi}}_k + \frac{1}{n}\left(\boldsymbol{\psi}_{k+1} - \overline{\boldsymbol{\psi}}_k\right)$$

Check convergence of the averaged sequence.

Additional Details and Literature Review

- **Step Size and Averaged Sequence**: Polyak [1990] showed that using a step size γ_k that goes to zero slower than $1/k$ (but not too slow) will guarantee that the *averaged sequence* $(\overline{\boldsymbol{\psi}}_k)_k$ will reach its limit at an *optimum* rate. It is suggested to use $\gamma_k = k^{-2/3}$. In other words, we can use *larger* step sizes that might make $(\boldsymbol{\psi}_k)_k$ diverge, but might still allow $(\overline{\boldsymbol{\psi}}_k)_k$ to converge.

- **Averaged Sequence Equivalence**: Delyon et al. [1999], when presenting the SAEM algorithm, noticed that while the original paper by Polyak [1990] suggested to consider $(\overline{\boldsymbol{\theta}}_k)_k$ one can equivalently (and more appropriately) use averaging on the iterates of the original parameter $\boldsymbol{\psi}_k$.

- **Convergence of Stochastic Approximations**: This topic goes beyond the scope of the lecture, however, a good resource to learn the tools and techniques to prove convergence for such algorithms are covered in Benveniste et al. [1990]. A simpler stochastic approximation version of the EM algorithm has been proposed by Gu and Li [1998] and exploits the fact that maximizing an expectation is equivalent to finding the stationary points of the gradient, which is a root-finding problem so under regularity conditions Robbins-Monro procedure can be applied.

References

Albert Benveniste, Pierre Priouret, and Michel Métivier. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, Berlin, Heidelberg, 1990. ISBN 0-387-52894-6.

Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, 27(1):94–128, 03 1999. doi: 10.1214/aos/1018031103. URL `https://doi.org/10.1214/aos/1018031103`.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL `http://www.jstor.org/stable/2984875`.

Ming Gao Gu and Shaolin Li. A stochastic approximation algorithm for maximum-likelihood estimation with incomplete data. *Canadian Journal of Statistics*, 26(4):567–582, 1998. doi: 10.2307/3315718. URL `https://onlinelibrary.wiley.com/doi/abs/10.2307/3315718`.

Boris Polyak. New stochastic approximation type procedures. *Avtomatica i Telemekhanika*, 7:98–107, 01 1990.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951. doi: 10.1214/aoms/1177729586. URL `https://doi.org/10.1214/aoms/1177729586`.

C. F. Jeff Wu. On the convergence properties of the em algorithm. *Ann. Statist.*, 11(1):95–103, 03 1983. doi: 10.1214/aos/1176346060. URL `https://doi.org/10.1214/aos/1176346060`.