# Stochastic Approximations

Mauro Camara Escudero

School of Mathematics, University of Bristol

# Table of contents

# Robbins-Monro Procedure

## Robbins-Monro

- Want to find unique root $\boldsymbol{\theta}_*$ of a function $h(\boldsymbol{\theta})$ so that $h(\boldsymbol{\theta}_*) = 0$.

# Robbins-Monro

- Want to find unique root $\boldsymbol{\theta}_*$ of a function $h(\boldsymbol{\theta})$ so that $h(\boldsymbol{\theta}_*) = 0$.
- Can only observe noisy measurements $H(\boldsymbol{\theta}, X)$ with $X$ RV and

$$h(\boldsymbol{\theta}) = \mathbb{E}_X[H(\boldsymbol{\theta}, X)] = \int H(\boldsymbol{\theta}, x)p(x; \boldsymbol{\theta})dx$$

# Robbins-Monro

- Want to find unique root $\boldsymbol{\theta}_*$ of a function $h(\boldsymbol{\theta})$ so that $h(\boldsymbol{\theta}_*) = 0$.
- Can only observe noisy measurements $H(\boldsymbol{\theta}, X)$ with $X$ RV and

$$h(\boldsymbol{\theta}) = \mathbb{E}_X[H(\boldsymbol{\theta}, X)] = \int H(\boldsymbol{\theta}, x)p(x; \boldsymbol{\theta})dx$$

- Under regularity and stability conditions the sequence

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \gamma_{k+1} H(\boldsymbol{\theta}_k, X_{k+1})$$

converges a.s. to $\boldsymbol{\theta}_*$. Here $\gamma_{k+1}$ are *small* step-sizes.

# Robbins-Monro Assumptions

- $(\gamma_k)_k$ big enough to explore, small enough for $\boldsymbol{\theta}_k$ to converge.

$$\sum \gamma_k = \infty \qquad \sum \gamma_k^2 < \infty$$

# Robbins-Monro Assumptions

- $(\gamma_k)_k$ big enough to explore, small enough for $\boldsymbol{\theta}_k$ to converge.

$$\sum \gamma_k = \infty \qquad \sum \gamma_k^2 < \infty$$

- Stability: Recursion $\boldsymbol{\theta}_k$ doesn't blow up. Related to convexity.

$$\sup_{\epsilon \leq |\boldsymbol{\theta} - \boldsymbol{\theta}_*| \leq \frac{1}{\epsilon}} (\boldsymbol{\theta} - \boldsymbol{\theta}_*)^\top h(\boldsymbol{\theta}) < 0 \qquad \forall \epsilon > 0$$

# Robbins-Monro Assumptions

- $(\gamma_k)_k$ big enough to explore, small enough for $\boldsymbol{\theta}_k$ to converge.

$$\sum \gamma_k = \infty \qquad \sum \gamma_k^2 < \infty$$

- Stability: Recursion $\boldsymbol{\theta}_k$ doesn't blow up. Related to convexity.

$$\sup_{\epsilon \leq |\boldsymbol{\theta} - \boldsymbol{\theta}_*| \leq \frac{1}{\epsilon}} (\boldsymbol{\theta} - \boldsymbol{\theta}_*)^\top h(\boldsymbol{\theta}) < 0 \qquad \forall \epsilon > 0$$

- Existence: $h(\boldsymbol{\theta})$ exists as $\exists$ constant $C$ bounding variance

$$\sigma^2(\boldsymbol{\theta}) = \int |H(\boldsymbol{\theta}, x)|^2 p(x; \boldsymbol{\theta}) dx \leq C(1 + |\boldsymbol{\theta}|^2)$$

# Robbins-Monro Assumptions

- $(\gamma_k)_k$ big enough to explore, small enough for $\boldsymbol{\theta}_k$ to converge.

$$\sum \gamma_k = \infty \qquad \sum \gamma_k^2 < \infty$$

- Stability: Recursion $\boldsymbol{\theta}_k$ doesn't blow up. Related to convexity.

$$\sup_{\epsilon \leq |\boldsymbol{\theta}-\boldsymbol{\theta}_*| \leq \frac{1}{\epsilon}} (\boldsymbol{\theta}-\boldsymbol{\theta}_*)^\top h(\boldsymbol{\theta}) < 0 \qquad \forall \epsilon > 0$$

- Existence: $h(\boldsymbol{\theta})$ exists as $\exists$ constant $C$ bounding variance

$$\sigma^2(\boldsymbol{\theta}) = \int |H(\boldsymbol{\theta}, x)|^2 p(x; \boldsymbol{\theta}) dx \leq C(1 + |\boldsymbol{\theta}|^2)$$

- Memory: Conditional expectation of $X_{k+1}$ given past $\mathcal{F}_k$ depends only on $\boldsymbol{\theta}_k$.

$$\mathbb{E}_X[g(\boldsymbol{\theta}_k, X_{k+1}) \mid \mathcal{F}_k] = \int g(\boldsymbol{\theta}_k, x) p(x; \boldsymbol{\theta}_k) dx$$

# Proof of Convergence

# Robbins-Siegmund Lemma

If $Z_k, B_k, C_k, D_k$ finite, non-negative RVs known given the past $\mathcal{F}_k$ and satisfying

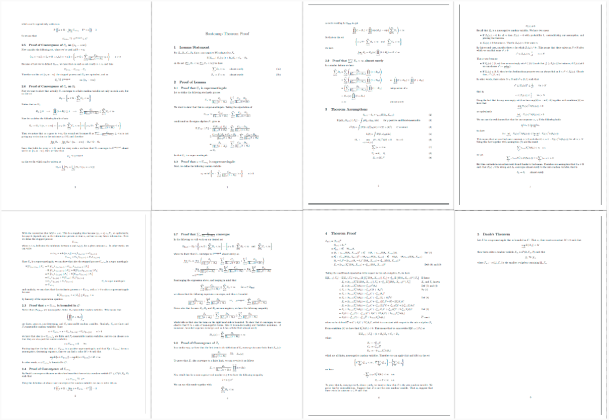
$$\mathbb{E}\left[Z_{k+1} \mid \mathcal{F}_k\right] \leq (1 + B_k) Z_k + C_k - D_k$$

on the set $\{\sum_k B_k < \infty, \sum_k C_k < \infty\}$ then

$$\sum_k D_k < \infty \qquad \text{almost surely} \tag{1a}$$

$$Z_k \to Z < \infty \qquad \text{almost surely} \tag{1b}$$

# Proof Outline



# Robbins-Monro Convergence Proof (I)

- **Find expression for distance** $Z_{k+1} := |T_{k+1}|^2 := |\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_*|^2$
  1. *Expand the square and plug-in recursion formula*

$$Z_{k+1} = \left(\boldsymbol{\theta}_k + \gamma_{k+1} H(\boldsymbol{\theta}_k, X_{k+1})\right)^2 + \boldsymbol{\theta}_*^2 - 2\left(\boldsymbol{\theta}_k + \gamma_{k+1} H(\boldsymbol{\theta}_k, X_{k+1})\right) \boldsymbol{\theta}_*$$

# Robbins-Monro Convergence Proof (I)

- **Find expression for distance** $Z_{k+1} := |T_{k+1}|^2 := |\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_*|^2$
    1. *Expand the square and plug-in recursion formula*

    $$Z_{k+1} = (\boldsymbol{\theta}_k + \gamma_{k+1} H(\boldsymbol{\theta}_k, X_{k+1}))^2 + \boldsymbol{\theta}_*^2 - 2\left(\boldsymbol{\theta}_k + \gamma_{k+1} H(\boldsymbol{\theta}_k, X_{k+1})\right) \boldsymbol{\theta}_*$$

    2. *Rearrange and rewrite in terms of* $T_k$ *and* $Z_k$

    $$Z_{k+1} = Z_k + 2\gamma_{k+1} T_k^\top H(\boldsymbol{\theta}_k, X_{k+1}) + \gamma_{k+1}^2 |H(\boldsymbol{\theta}_k, X_{k+1})|^2$$

- **Bound the expected distance** $\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k]$
    1. *Take conditional expectation given past* $\mathcal{F}_k$*. Use def of* $\sigma^2(\boldsymbol{\theta})$ *and* $h(\boldsymbol{\theta})$.

    $$\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] = Z_k + 2\gamma_{k+1} T_k^\top h(\boldsymbol{\theta}_k) + \gamma_{k+1}^2 \sigma^2(\boldsymbol{\theta}_k)$$

    2. *Use bound on* $\sigma^2(\boldsymbol{\theta})$ and rearrange all constants.

    $$\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] \leq Z_k(1 + \gamma_{k+1}^2 C) + \gamma_{k+1}^2 \overline{C} + 2\gamma_{k+1} T_k^\top h(\boldsymbol{\theta}_k)$$

    where $\overline{C} := C + \boldsymbol{\theta}_* C + 2T_k \boldsymbol{\theta}_* C$ is constant given $\mathcal{F}_k$.

# Robbins-Monro Convergence Proof (II)

- **Prove convergence of $Z_k$ to a finite RV Z**
    1. *Use stability condition $T_k^\top h(\boldsymbol{\theta}_k) < 0$ to write it as Lemma, with*

$$B_k := \gamma_{k+1}^2 C \quad C_k := \gamma_{k+1}^2 \overline{C} \quad D_k := -2\gamma_{k+1} T_k^\top h(\boldsymbol{\theta}_k)$$

# Robbins-Monro Convergence Proof (II)

- **Prove convergence of $Z_k$ to a finite RV Z**
    1. *Use stability condition $T_k^\top h(\boldsymbol{\theta}_k) < 0$ to write it as Lemma, with*

$$B_k := \gamma_{k+1}^2 C \quad C_k := \gamma_{k+1}^2 \overline{C} \quad D_k := -2\gamma_{k+1} T_k^\top h(\boldsymbol{\theta}_k)$$

    2. *Lemma gives convergence on $\left\{\sum \gamma_{k+1}^2 < \infty\right\}$*

$$Z_k \xrightarrow{\text{a.s.}} Z < \infty \qquad \text{and} \qquad -\sum \gamma_{k+1} T_k^\top h(\boldsymbol{\theta}_k) < \infty \text{ a.s.}$$

- **Prove $Z$ is the zero random variable**
    1. *By contradiction assume $\exists \omega \in \Omega$ such that $Z(\omega) \neq 0$.*
    2. *Using some analysis and the stability condition we finally find*

$$\sum_k -\gamma_{k+1} T_k^\top(\omega) h(\boldsymbol{\theta}_k(\omega)) \geq \alpha \sum_k \gamma_{k+1} = +\infty \quad \text{for some } \alpha > 0$$

    which is a contradiction.

# Application to EM Algorithm

# Maximum Likelihood Estimation with Missing Data

- $\mathbf{y}$ observed data, $\mathbf{x}$ missing data

# Maximum Likelihood Estimation with Missing Data

- $\mathbf{y}$ observed data, $\mathbf{x}$ missing data
- $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})$ complete-data likelihood

# Maximum Likelihood Estimation with Missing Data

- $\mathbf{y}$ observed data, $\mathbf{x}$ missing data
- $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})$ complete-data likelihood
- Find $\boldsymbol{\psi}^*$ maximizing $p(\mathbf{y} \mid \boldsymbol{\psi})$ incomplete-data likelihood

# Maximum Likelihood Estimation with Missing Data

- $\mathbf{y}$ observed data, $\mathbf{x}$ missing data
- $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})$ complete-data likelihood
- Find $\boldsymbol{\psi}^*$ maximizing $p(\mathbf{y} \mid \boldsymbol{\psi})$ incomplete-data likelihood
- Conditional distribution

$$p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi}) = \begin{cases} \frac{p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})}{p(\mathbf{y} \mid \boldsymbol{\psi})} & \text{if } p(\mathbf{y} \mid \boldsymbol{\psi}) \neq 0 \\ 0 & \text{if } p(\mathbf{y} \mid \boldsymbol{\psi}) = 0 \end{cases}$$

# EM Algorithm

- E-step

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

- M-step

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

*Intuition*: Want to maximize $\log p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})$ but it's unknown, so maximize current expectation given data and current estimate of $\boldsymbol{\psi}_k$. EM useful when maximizing $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_k)$ easier than maximizing $p(\mathbf{y} \mid \boldsymbol{\psi})$. $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_k)$ can be intractable!

# Monte Carlo EM

Replace Q($\boldsymbol{\psi}$| $\boldsymbol{\psi}_k$) with Monte Carlo estimate

$$Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_k) \approx \frac{1}{m_k} \sum_{i=1}^{m_k} \log p(\mathbf{x}_k^{(i)} \mid \boldsymbol{\psi})$$

where $\mathbf{x}_k^{(1)}, \ldots, \mathbf{x}_k^{(m_k)}$ are drawn from $p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi}_k)$.
*Problem*: At every new iteration, we discard previous samples.

# SAEM Setting

- $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})$ follows a distribution from curved exponential family

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

# SAEM Setting

- $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})$ follows a distribution from curved exponential family

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

- Define $\overline{\theta}(\boldsymbol{\psi}) := \mathbb{E}_{p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi})}\left[\widetilde{\theta}(\mathbf{X})\right]$. The E-step becomes

$$Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_k) = L(\overline{\theta}(\boldsymbol{\psi}_k), \boldsymbol{\psi})$$

## SAEM Setting

- $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})$ follows a distribution from curved exponential family

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

- Define $\overline{\theta}(\boldsymbol{\psi}) := \mathbb{E}_{p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi})}\left[\widetilde{\theta}(\mathbf{X})\right]$. The E-step becomes

$$Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_k) = L(\overline{\theta}(\boldsymbol{\psi}_k), \boldsymbol{\psi})$$

- Assume there exists function $\widehat{\psi}$ that maximizes the E-step

$$L(\theta, \widehat{\psi}(\theta)) \geq L(\theta, \boldsymbol{\psi}) \qquad \forall\, \theta \quad \forall \boldsymbol{\psi}$$

# SAEM Setting

- $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})$ follows a distribution from curved exponential family

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

- Define $\overline{\theta}(\boldsymbol{\psi}) := \mathbb{E}_{p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi})}\left[\widetilde{\theta}(\mathbf{X})\right]$. The E-step becomes

$$Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_k) = L(\overline{\theta}(\boldsymbol{\psi}_k), \boldsymbol{\psi})$$

- Assume there exists function $\widehat{\psi}$ that maximizes the E-step

$$L(\theta, \widehat{\psi}(\theta)) \geq L(\theta, \boldsymbol{\psi}) \qquad \forall\, \theta \quad \forall \boldsymbol{\psi}$$

- Basically $\boldsymbol{\psi}_{k+1} = \widehat{\psi}(\overline{\theta}(\boldsymbol{\psi}_k)$

# SAEM Setting

- $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\psi})$ follows a distribution from curved exponential family

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

- Define $\overline{\theta}(\boldsymbol{\psi}) := \mathbb{E}_{p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi})}\left[\widetilde{\theta}(\mathbf{X})\right]$. The E-step becomes

$$Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_k) = L(\overline{\theta}(\boldsymbol{\psi}_k), \boldsymbol{\psi})$$

- Assume there exists function $\widehat{\psi}$ that maximizes the E-step

$$L(\theta, \widehat{\psi}(\theta)) \geq L(\theta, \boldsymbol{\psi}) \qquad \forall \, \theta \quad \forall \boldsymbol{\psi}$$

- Basically $\boldsymbol{\psi}_{k+1} = \widehat{\psi}(\overline{\theta}(\boldsymbol{\psi}_k)$
- Define $\boldsymbol{\theta}_{k+1} = \overline{\theta}(\boldsymbol{\psi}_k)$

# Diagram of Function Compositions

$$\begin{tikzcd}
\boldsymbol{\theta}_k \arrow[r, "\widehat{\psi}"] \arrow[d, dotted, "G"] & \boldsymbol{\psi}_k \arrow[dl, "\overline{\theta}"] \arrow[d, dotted, "T"] \\
\boldsymbol{\theta}_{k+1} \arrow[r, "\widehat{\psi}"] & \boldsymbol{\psi}_{k+1}
\end{tikzcd}$$

# Diagram of Function Compositions

$$\begin{tikzcd}[column sep=huge, row sep=huge]
\boldsymbol{\theta}_k \arrow[r, "\widehat{\psi}"] \arrow[d, dotted, "G"] & \boldsymbol{\psi}_k \arrow[dl, "\text{E-step}"] \arrow[d, dotted, "\text{EM-Algorithm}"] \\
\boldsymbol{\theta}_{k+1} \arrow[r, "\text{M-step}"] & \boldsymbol{\psi}_{k+1}
\end{tikzcd}$$

Can consider EM algorithm on conditional expectations $\boldsymbol{\theta}_k$'s instead of $\boldsymbol{\psi}_k$'s.

$$\boldsymbol{\theta}_{k+1} = G(\boldsymbol{\theta}_k)$$

Aim: Find $\boldsymbol{\theta}_*$ fixed point of $G$ satisfying $G(\boldsymbol{\theta}_*) = \boldsymbol{\theta}_*$

# Going back to Robbins-Monro

- Fixed point $\boldsymbol{\theta}_*$ of $G$ is unique root of $h(\boldsymbol{\theta}) := G(\boldsymbol{\theta}) - \boldsymbol{\theta}$

# Going back to Robbins-Monro

- Fixed point $\boldsymbol{\theta}_*$ of $G$ is unique root of $h(\boldsymbol{\theta}) := G(\boldsymbol{\theta}) - \boldsymbol{\theta}$
- Noisy measurement of

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

is a Monte Carlo estimate with $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ drawn from $p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi})$

$$H(\boldsymbol{\theta}, \mathbf{X}) = \frac{1}{m}\sum_{j=1}^{m} \widetilde{\theta}(\mathbf{x}^{(j)}) - \boldsymbol{\theta}$$

# Going back to Robbins-Monro

- Fixed point $\boldsymbol{\theta}_*$ of $G$ is unique root of $h(\boldsymbol{\theta}) := G(\boldsymbol{\theta}) - \boldsymbol{\theta}$
- Noisy measurement of

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

  is a Monte Carlo estimate with $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ drawn from $p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi})$

$$H(\boldsymbol{\theta}, \mathbf{X}) = \frac{1}{m}\sum_{j=1}^{m} \widetilde{\theta}(\mathbf{x}^{(j)}) - \boldsymbol{\theta}$$

- RM procedure converges to $\boldsymbol{\theta}_*$ under regularity conditions.

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

  where $\mathbf{x}_k^{(j)}$ are drawn from $p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi}_{k-1})$

# Going back to Robbins-Monro

- Fixed point $\boldsymbol{\theta}_*$ of $G$ is unique root of $h(\boldsymbol{\theta}) := G(\boldsymbol{\theta}) - \boldsymbol{\theta}$
- Noisy measurement of

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

  is a Monte Carlo estimate with $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ drawn from $p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi})$

$$H(\boldsymbol{\theta}, \mathbf{X}) = \frac{1}{m}\sum_{j=1}^{m} \widetilde{\theta}(\mathbf{x}^{(j)}) - \boldsymbol{\theta}$$

- RM procedure converges to $\boldsymbol{\theta}_*$ under regularity conditions.

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

  where $\mathbf{x}_k^{(j)}$ are drawn from $p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\psi}_{k-1})$
- Using linearity $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}_{k-1}) = -\xi(\boldsymbol{\psi}) + \langle \boldsymbol{\theta}_k \,, \phi(\boldsymbol{\psi})\rangle$ we get

$$\widehat{Q}_k(\boldsymbol{\psi}) = \widehat{Q}_{k-1}(\boldsymbol{\psi}) + \gamma_k \left(\frac{1}{m_k}\sum_{j=1}^{m_k} \log p(\mathbf{x}_k^{(j)}, \mathbf{y} \mid \boldsymbol{\psi}) - \widehat{Q}_{k-1}(\boldsymbol{\psi})\right)$$

# Thank you

# References i

A. Benveniste, P. Priouret, and M. Métivier.
***Adaptive Algorithms and Stochastic Approximations*.**
Springer-Verlag, Berlin, Heidelberg, 1990.

B. Delyon, M. Lavielle, and E. Moulines.
**Convergence of a stochastic approximation version of the em algorithm.**
*Ann. Statist.*, 27(1):94–128, 03 1999.

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.

M. Duflo.
***Random Iterative Models*.**
Springer-Verlag, Berlin, Heidelberg, 1st edition, 1997.

M. G. Gu and S. Li.
**A stochastic approximation algorithm for maximum-likelihood estimation with incomplete data.**
*Canadian Journal of Statistics*, 26(4):567–582, 1998.

H. Kushner and G. Yin.
***Stochastic Approximation and Recursive Algorithms and Applications*.**
Stochastic Modelling and Applied Probability. Springer New York, 2003.

# References iii


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

H. Robbins and S. Monro.
**A stochastic approximation method.**
*Ann. Math. Statist.*, 22(3):400–407, 09 1951.

H. Robbins and S. Monro.
**A stochastic approximation method.**
*Ann. Math. Statist.*, 22(3):400–407, 09 1951.

C. F. J. Wu.
**On the convergence properties of the em algorithm.**
*Ann. Statist.*, 11(1):95–103, 03 1983.