Quantitative Analysis of Stochastic Approximation Methods

MathFound Seminar
University of Bath

Paulo Oliva
(joint work with Rob Arthan)

Tue 20 Feb 2024

Mining Proofs for Quantitative Data…

sequence $(a_n)_{n\in\mathbb{N}}$ **converges**

$$\forall \varepsilon > 0\ \boxed{\exists N}\ \forall i,j \geq N(|a_i - a_j| \leq \varepsilon)$$

$$\exists \varphi\ \forall \varepsilon > 0\ \forall i,j \geq \varphi(\varepsilon)(|a_i - a_j| \leq \varepsilon)$$

$(a_n)_{n\in\mathbb{N}}$ converges with **modulus of convergence** φ

series $\Sigma_n a_n$ **converges**

$$\forall \varepsilon > \boxed{0 \ \exists N} \ \forall i, j \geq N (|\Sigma_{k=i}^{j} a_k| \leq \varepsilon)$$

$$\exists \psi \ \forall \varepsilon > 0 \ \forall i, j \geq \psi(\varepsilon)(|\Sigma_{k=i}^{j} a_k| \leq \varepsilon)$$

$\Sigma_n a_n$ converges with **modulus of convergence** ψ

$$(a_n)_{n\in\mathbb{N}} \textbf{ converges} \quad \Rightarrow \quad (b_n)_{n\in\mathbb{N}} \textbf{ converges}$$

$$\exists \Phi \forall \varphi (\forall \varepsilon > 0\ \forall i, j \geq \varphi(\varepsilon)(|a_i - a_j| \leq \varepsilon)$$
$$\Rightarrow \quad \forall \varepsilon > 0\ \forall i, j \geq \Phi(\varphi, \varepsilon)(|b_i - b_j| \leq \varepsilon))$$

Φ turns a m.o.c for $(a_n)_{n\in\mathbb{N}}$ into a m.o.c for $(b_n)_{n\in\mathbb{N}}$

sequence of r.v. $(X_n)_{n\in\mathbb{N}}$ **converges with probability 1**

$$\mathbb{P}[\forall \varepsilon > 0\ \exists N\ \forall i,j \geq N(|X_i - X_j| \leq \varepsilon)] = 1$$

$$\mathbb{P}[\{\omega :\ \forall \varepsilon > 0\ \boxed{\exists N}\ \forall i,j \geq N(|X_i(\omega) - X_j(\omega)| \leq \varepsilon)\}] = 1$$

$$\exists \varphi\ \forall \varepsilon, \delta > 0\,(\mathbb{P}[\forall i,j \geq \varphi(\varepsilon,\delta)(|X_i - X_j| \leq \varepsilon)] \geq 1 - \delta)$$

φ a **uniform modulus of a.s. convergence** for $(X_n)_{n\in\mathbb{N}}$

Avigad-Dean-Rute (2011)

$$(a_n)_{n \in \mathbb{N}} \textbf{ converges} \quad \Rightarrow \quad (b_n)_{n \in \mathbb{N}} \textbf{ converges}$$

$$\exists \Phi \forall \varphi (\boxed{\forall \delta > 0} \forall i, j \geq \varphi(\delta)(|a_i - a_j| \leq \delta)$$

$$\Rightarrow \quad \forall \varepsilon > 0 \, \forall i, j \geq \Phi(\varphi, \varepsilon)(|b_i - b_j| \leq \varepsilon))$$

$$\exists \Phi, \Delta \forall \varphi, \varepsilon > 0$$

$$(\forall i, j \geq \varphi(\Delta(\varphi, \varepsilon))(|a_i - a_j| \leq \Delta(\varphi, \varepsilon))$$

$$\Rightarrow \quad \forall i, j \geq \Phi(\varphi, \varepsilon)(|b_i - b_j| \leq \varepsilon))$$

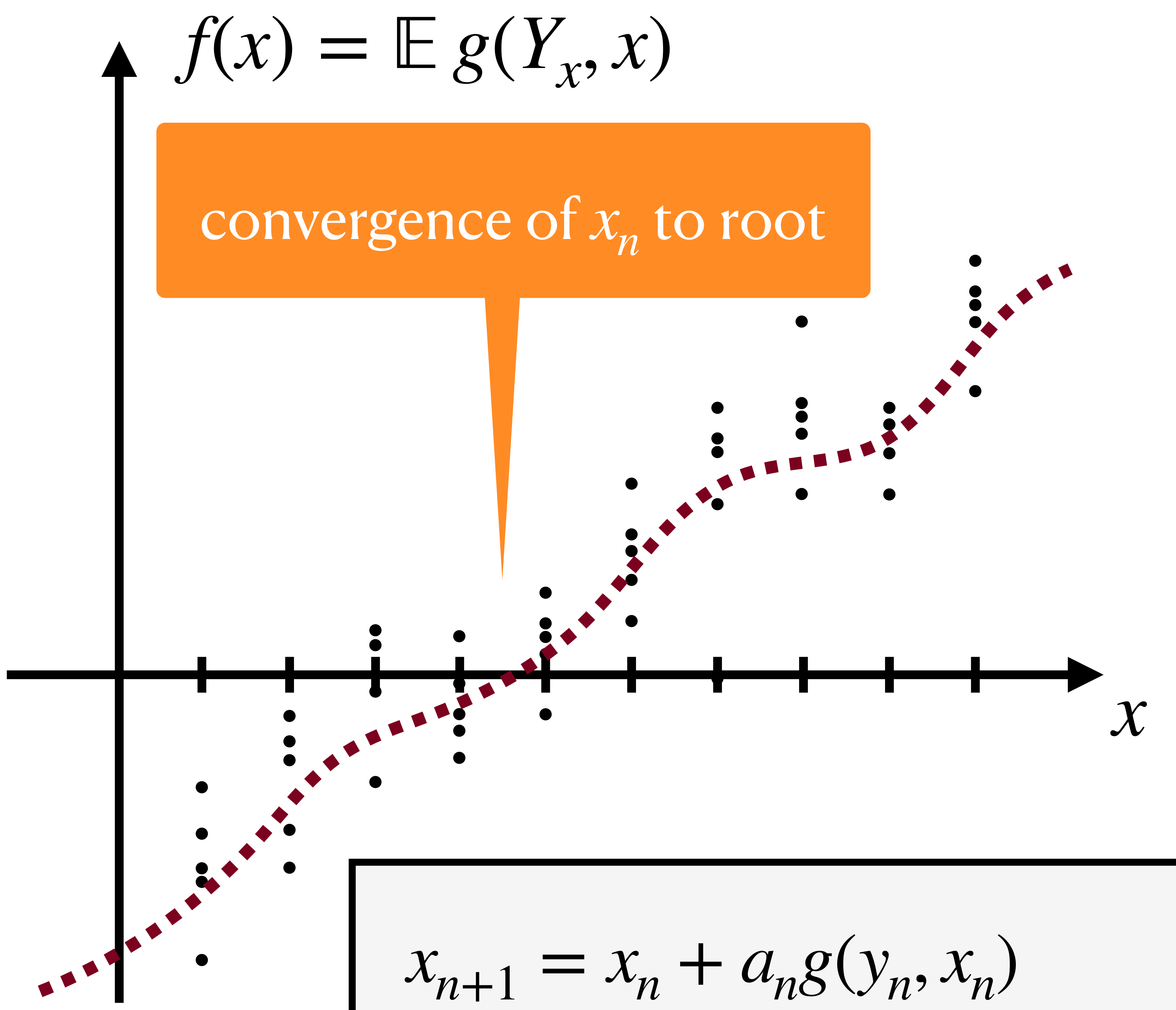
Stochastic Approximation Methods

Stochastic Approximation Methods

- **Iterated procedures** used to **approximate a target** value when target is unknown and observations are corrupted by **noise**
- Introduced by **Robbins-Monro** (1951)
- Robbins-Monro algorithms generalised by **Dvoretzky** (1956)
- Generalisation of (deterministic) approximation methods
- **Stochastic gradient descent** (in Machine Learning) is based on Stochastic Approximation Theory

Stochastic Approximation Methods

- Y_x real-valued random variable parametrised by x
- $P[Y_x]$ the probability of Y_x for given x
- Assume we can only observe $P[Y_x]$ via sampling
- $g(y, x)$ a given function
- **Problem**: Find x such that $\mathbb{E}\, g(Y_x, x) = 0$



$$x_{n+1} = x_n + a_n g(y_n, x_n)$$

given samples $y_0, y_1, \ldots$ of $Y_{x_1}, Y_{x_2}, \ldots$

Stochastic Approximation Methods

- Y_x real-valued random variable parametrised by x
- $P[Y_x]$ the probability of Y_x for given x
- Assume we can only observe $P[Y_x]$ via sampling
- $g(y, x)$ a given function
- **Problem**: Find x such that $\mathbb{E}\, g(Y_x, x) = 0$

<u>Stochastic Gradient Descent</u>

$L(y, x)$ a loss function, x model param.

Problem: Find x s.t. $\mathbb{E}\, L(Y_x, x)$ is minimal

$g(y, x) = -\nabla_x L(y, x)$

Training set: $y_1, y_2, \ldots$

$x_{n+1} = x_n + a_n g(y_n, x_n)$

(a_n learning rate)

SGD: (x_n) converges to critical point of loss function

Robbin-Monro Stochastic Approximation Method

Robbins-Monro (1951)

- Y_x real-valued random variable parametrised by x
- $P[Y_x]$ the probability of Y_x for given x
- Assume we can only observe $P[Y_x]$ via sampling
- $g(y, x)$ a given function
- **Problem**: Find x such that $\mathbb{E}\, g(Y_x, x) = 0$

Stochastic Approximation Algorithm

$$x_{n+1} = x_n + a_n g(y_n, x_n)$$

Robbins-Monro (1951):

L_2 convergence of x_n when $g(y, x) = b - y$

Assumes:

- $a_n \to 0,\ \Sigma\, a_n = \infty,\ \Sigma\, a_n^2 < \infty$
- Y_x bounded w. p. 1
- function $f(x) = \mathbb{E}[Y_x]$
    - non-decreasing
    - solution for $f(x) = b$ exists
    - derivative at solution is positive

Dvoretzky (1956)

- Vast generalisation of Robbins-Monro result
- Iteration has two parts:
  - Deterministic part which is assume to converge to solution
  - Stochastic part with expected value of zero and bounded variance

Assumptions:

$T_n\colon \mathbb{R}^n \to \mathbb{R}$

$Y_1, Y_2, \ldots$ (dependent on history of $x_1, x_2, \ldots$)

- $\mathbb{E}[Y_n \,|\, X_1, \ldots, X_n] = 0$, w. p. 1
- $\Sigma\, \mathbb{E}[Y_n^2] < \infty$

There exists x_* such that

$$|\,T_n(\vec{x}) - x_*\,| \leq \max(\alpha_n, (1+\beta_n)\,|\,x_n - x_*\,| - \gamma_n)$$

$\alpha_n, \beta_n, \gamma_n \in \mathbb{R}^+$ such that

$$\alpha_n \to 0, \Sigma\,\beta_n < \infty, \Sigma\,\gamma_n = \infty$$

Result:

The iteration

$$x_{n+1} = \underbrace{\boxed{T_n(x_1, \ldots, x_n)}}_{\textit{deterministic}} + \overbrace{\boxed{Y_n}}^{\textit{noise terms}}$$

converges a.s. to x_*

THEOREM 1. (Dvoretzky). *Let* $\{X_n\}$, $\{T_n(X_1, \cdots, X_n)\}$, $\{Y_n(X_1, \cdots, X_n)\}$ *be sequences of real random variables with* X_1 *arbitrary and*

$$X_{n+1} = T_n(X_1, \cdots, X_n) + Y_n(X_1, \cdots, X_n). \tag{6}$$

Assume

$$E\{Y_n \mid X_1, \cdots, X_n\} = 0 \qquad \text{w.p.1}, \tag{7}$$

$$\sum EY_n^2 < \infty, \tag{8}$$

and

$$|T_n| \leqq \max(\alpha_n, (1 + \beta_n)|X_n| - \gamma_n) \tag{9}$$

where α_n, β_n, γ_n *are positive numbers such that*

$$\alpha_n \to 0, \sum \beta_n < \infty, \qquad \sum \gamma_n = \infty. \tag{10}$$

Then $\boxed{X_n \to 0 \text{ w.p.1.}}$

Dvoretzky Theorem (Derman-Sachs Proof)

Derman-Sacks Proof (1959)

$^{*}X_n$ indep.

- Borel-Cantelli lemma (1st) $\Sigma P[X_n] < \infty \Rightarrow P[X_n \text{ i.o.}] = 0$
- Chebyshev inequality $P[\,|X-\mu| \geq k\sigma] \leq 1/k^2$
- Abel's test Σa_n conv $\wedge$ b_n mon. and bounded $\Rightarrow \Sigma a_n b_n$ conv
- Slowdown lemma Σa_n conv $\Rightarrow \exists b_n(b_n \to 0 \wedge \Sigma a_n/b_n$ conv$)$
- Kolmogorov inequality* $P[\max_{1\leq k\leq n} |X_1 + \ldots + X_k| \geq \lambda) \leq 1/\lambda^2 \mathrm{Var}[X_1 + \ldots + X_n]$
- Variance lemma* $\Sigma \mathbb{E} X_n^2 < \infty \Rightarrow \Sigma X_n$ a.s.
- "Lemma 1" about $\mathbb{R}$ converging and diverging sequences and series

Abel’s Test

Abel's Test

Lemma (Abel's Test). If $\Sigma_n a_n$ converges and $(b_n)_{n\in\mathbb{N}}$ is monotone and converges then $\Sigma_n a_n b_n$ converges.

Proof: Summation by parts.

<u>Quantitative data</u>: Functional converting moduli of convergence for $\Sigma_n a_n$ and $(b_n)_{n\in\mathbb{N}}$ into a modulus of convergence for $\Sigma_n a_n b_n$.

Lemma 1 (Quantitative Abel's Test) *Let $\{\delta_n\}$ and $\{\nu_n\}$ be sequences of real numbers, with $\{\nu_n\}$ monotone. Suppose that for some D and V we have*

(i) $|\sum_{n=i}^{j} \delta_n| \leq D$, *for all* $j \geq i \geq 1$,

(ii) $|\nu_k| \leq V$, *for all* $k \geq 1$,

and $\Phi_{\{\nu_n\}}$ is a modulus of convergence for $\{\nu_n\}$, i.e.

(iii) $\forall \varepsilon > 0 \forall i, j \geq \Phi_{\{\nu_n\}}(\varepsilon)(|\nu_i - \nu_j| \leq \varepsilon)$.

Then, for any Ψ and $\varepsilon > 0$ if

$$\forall i, j \geq \Psi(\varepsilon/4V) \left(\left| \sum_{n=i}^{j} \delta_n \right| \leq \frac{\varepsilon}{4V} \right) \to \forall i, j \geq \Psi^*(\Psi, \varepsilon) \left(\left| \sum_{n=i}^{j} \delta_n \nu_n \right| \leq \varepsilon \right)$$

where

$$\Psi^*(\Psi, \varepsilon) = \max \left(\Phi_{\{\nu_n\}} \left(\frac{\varepsilon}{4D} \right), \Psi \left(\frac{\varepsilon}{4V} \right) \right).$$

Slowdown Lemma

Slowdown Lemma

Slowdown lemma. If $\Sigma_n a_n$ converges (with $a_n \geq 0$) then there exists a sequence $(b_n)_{n\in\mathbb{N}}$ which converges to 0 such that $\Sigma_n a_n / b_n$ also converges.

Proof: Take $b_n = \sqrt{\Sigma_{i\geq n} a_n}$.

Quantitative data: Finite approximation to infinite sum in b_n, modulus of convergence for $(b_n)_{n\in\mathbb{N}}$ and modulus of convergence for $\Sigma_n a_n / b_n$ given one for $\Sigma_n a_n$.

Lemma 5 *Let (t_n) be a sequence of non-negative real numbers. Assume the series $\sum_n t_n$ converges with Cauchy modulus $\Psi_{\sum t_n}$, i.e.,*

$$\forall \varepsilon > 0 \left(\sum_{k \geq \Psi_{\sum t_n}(\varepsilon)} t_k < \varepsilon \right) \tag{3}$$

For $\psi \colon \mathbb{R} \to \mathbb{R}$, let $F(\psi, 0) = 0$ and $F(\psi, n) = \max(F(\psi, n-1) + 1, \psi(\sqrt{t_n}/n^2))$ and let $r_n = \sum_{i=n}^{F(\Psi_{\sum t_n}, n)} t_i$. Define $\Psi_{\sum \frac{t_n}{\sqrt{r_n}}}$ by

$$\Psi_{\sum \frac{t_n}{\sqrt{r_n}}}(\varepsilon) = \max\left(\Psi_{\sum t_n}\left(\frac{\varepsilon^2}{4} \right), \left\lceil \frac{\pi^2}{2\varepsilon} + \frac{1}{2} \right\rceil \right) \tag{4}$$

Then the series $\sum_n \frac{t_n}{\sqrt{r_n}}$ converges with Cauchy modulus $\Psi_{\sum \frac{t_n}{\sqrt{r_n}}}$, i.e.,

$$\forall \varepsilon > 0 \left(\sum_{k \geq \Psi_{\sum \frac{t_n}{\sqrt{r_n}}}(\varepsilon)}^{\infty} \frac{t_k}{\sqrt{r_k}} < \varepsilon \right) \tag{5}$$

Derman-Sachs Lemma 1

Lemma 1

Lemma 1. $a_n, b_n, c_n, \delta_n, \xi_n$ sequences in $\mathbb{R}$

(i) a_n, b_n, c_n, ξ_n non-negative (δ_n can be negative)

(ii) $\lim a_n = 0, \Sigma\, b_n < \infty, \Sigma\, c_n = \infty, \Sigma\, \delta_n$ converges

(iii) $\xi_{n+1} \leq \max(a_n, (1 + b_n)\xi_n + \delta_n - c_n)$, for all large enough n

Then $\xi_n \to 0$

- Quantitative data: Constructive, modulus of convergence for ξ_n given moduli for assumptions and N from which point (iii) holds.

Lemma 4 (Derman-Sacks Lemma 1) *Let $(a_n), (b_n), (c_n), (\delta_n)$, and (ξ_n) be sequences of reals such that:*

(i) $(a_n), (b_n), (c_n)$, *and* (ξ_n) *are non-negative*

(ii) $\xi_n \le E$ *and* $|\sum_{n=i}^{j} \delta_n| \le D$, *for all* $j \ge i \ge 1$

(iii) for some $\Phi_{\{a_n\}}, \Psi_{\sum b_n}$ *and* $\Delta_{\sum c_n}$ *we have*

- $\forall \varepsilon > 0 \forall i \ge \Phi^0_{\{a_n\}}(\varepsilon)(a_i \le \varepsilon)$
- $\forall \varepsilon > 0 (\sum_{i \ge \Psi_{\sum b_n}(\varepsilon)} b_i \le \varepsilon)$
- $\forall K, k (\sum_{i=k}^{\Delta_{\sum c_n}(K,k)} c_i \ge K)$

(iv) and N_0 *be a natural number such that for all* $n \ge N_0$ *we have*

$$\xi_{n+1} \le \max(a_n, (1 + b_n)\xi_n + \delta_n - c_n).$$

Moreover, let

$$\Phi^*(\Psi, \varepsilon) = \max(\Delta_{\sum c_n / B_n}(E/B_{N(\Psi,\varepsilon)-1} + 2D, N(\Psi, \varepsilon)), N(\Psi, \varepsilon$$

where

$$\Delta_{\sum c_n / B_n}(K, k) = \Delta_{\sum c_n}(K \cdot B, k),$$

$$N(\Psi, \varepsilon) = \max\left(\Phi^0_{\{a_n\}}\left(\frac{\varepsilon}{2B}\right), \Psi^*\left(\Psi, \frac{\varepsilon}{2B}\right), N_0\right),$$

$$\Psi^*(\Psi, \varepsilon) = \max\left(\Phi_{\{\frac{1}{B_n}\}}\left(\frac{\varepsilon}{4D}\right), \Psi\left(\frac{\varepsilon}{4}\right)\right),$$

$$\Phi_{\{\frac{1}{B_n}\}}(\varepsilon) = \Phi_{\{B_n\}}\left(\frac{\min\{\varepsilon, 1/2\}}{1 - \min\{\varepsilon, 1/2\}}\right),$$

$$\Phi_{\{B_n\}}(\varepsilon) = \Psi_{\sum b_n}\left(\ln\left(\frac{\varepsilon}{B} + 1\right)\right),$$

$b = 1 + \sum_{i=1}^{\Psi_{\sum b_n}(1/2)-1} b_i$, $B = e^b$ *and* $B_n = \prod_{i=1}^{n}(1 + b_i)$. *Then, for all* Ψ *and* $\varepsilon > 0$ *we have that*

$$\forall i, j \ge \Psi\left(\frac{\varepsilon}{8B}\right)\left(\left|\sum_{n=i}^{j} \delta_n\right| \le \frac{\varepsilon}{8B}\right) \to \forall i \ge \Phi^*(\Psi, \varepsilon)(\xi_i \le \varepsilon).$$

Variance Lemma

Variance Lemma

Variance lemma. If

(i) $\mathbb{E}[Z_n \,|\, X_1, \ldots, X_n] = 0$, w. p. 1

(ii) $\Sigma_n \, \mathbb{E}[Z_n^2] < \infty$

Then $\Sigma_n \, Z_n$ converges almost surely.

Proof: Apply Kolmogorov inequality to $S_N = \Sigma_{n=0}^{N} Z_n$.

- Quantitative data: Modulus of almost sure convergence for $\Sigma_n Z_n$ given modulus of convergence for $\Sigma_n \mathbb{E}[Z_n^2]$.

Lemma 7 *Let $\{Z_n\}$ be a sequence of random variables such that $\mathbb{E}[Z_i Z_j] = 0$ for $i \neq j$, and suppose the series $\sum_n \mathbb{E}[Z_n^2]$ is such that*

(i) $\sum_{n=1}^{N} \mathbb{E}[Z_n^2] \leq C$, *for all* $N \geq 1$, *and*

(ii) $\sum_n \mathbb{E}[Z_n^2]$ *converges with a Cauchy modulus* $\Psi_{\sum \mathbb{E}[Z_n^2]}$, *i.e.,*

$$\forall \varepsilon > 0 \left(\sum_{k \geq \Psi_{\sum \mathbb{E}[Z_n^2]}(\varepsilon)} \mathbb{E}[Z_k^2] < \varepsilon \right). \tag{13}$$

Then the series $\sum Z_n$ converges with probability 1, with modulus $\Psi_{\sum \mathbb{E}[Z_n^2]}(\delta\varepsilon^2/4)$,

$$\forall \varepsilon, \delta > 0 \left(\mathbb{P}\left[\bigcap_{\substack{m \geq \Psi_{\sum \mathbb{E}[Z_n^2]}(\delta\varepsilon^2/4) \\ k \geq 1}} \left(\left| \sum_{l=m}^{m+k} Z_l \right| < \varepsilon \right) \right] \geq 1 - \delta \right) \tag{14}$$

Main Proof

Proof

Let $Z_n = Y_n \operatorname{sgn}(T_n)$. By the **Variance Lemma**

ΣZ_n converges a. s.

By **Chebyshev inequality** and **Borel-Cantelli lemma**

$|Z_n| \leq \alpha_n$, a. s., for large enough n

It follows that

$|X_{n+1}| \leq \max(\ldots)$, a. s., for large enough n

Then, apply **Lemma 1** pointwise on this event of prob. 1

- H. Robbins and S. Monro, **A Stochastic Approximation Method**, The Annals of Mathematical Statistics, 22:3, 1951
- A. Dvoretzky, **On Stochastic Approximation**, Berkeley Symposium on Mathematical Statistics and Probability, 3:1, 39–55 1956
- C. Derman and J. Sacks, **On Dvoretzky's Stochastic Approximation Theorem**, Ann. Math. Statist. 30(2): 601–606, 1959
- J. Avigad, E. Dean and J. Rute, **A metastable dominated convergence theorem**, Journal of Logic and Analysis, 4:3, 1–19, 2012
- R. Arthan and P. Oliva, **On the Borel-Cantelli Lemmas, the Erdós-Rényi Theorem, and the Kochen-Stone Theorem**, Journal of Logic and Analysis, 13:6, 1–23, 2021

The End

Stochastic Approximation Methods

- Y_x real-valued random variable parametrised by x
- $P[Y_x]$ the probability of Y_x for given x
- Assume we can only observe $P[Y_x]$ via sampling
- $g(y, x)$ a given function
- **Problem**: Find x such that $\mathbb{E}\, g(Y_x, x) = 0$

Kolmogorov Strong Law of Large Numbers

Y a r. v. (independent of x)

Problem: Find x s.t. $\mathbb{E}\, Y = x$

Instance:

$g(y, x) = y - x$

$x_{n+1} = x_n + (y_n - x_n)/(n+1)$

(given samples $y_0, y_1, \ldots$)

SLLN: (x_n) converges to $\mathbb{E}\, Y$ a.s.

Stochastic Approximation Methods

- Y_x real-valued random variable parametrised by x
- $P[Y_x]$ the probability of Y_x for given x
- Assume we can only observe $P[Y_x]$ via sampling
- $g(y, x)$ a given function
- **Problem**: Find x such that $\mathbb{E}\, g(Y_x, x) = 0$

Banach Fixed-Point Theorem

Contraction mapping $\phi : \mathbb{R} \to \mathbb{R}$

Problem: Find x s.t. $\phi(x) = x$

Instance:

$Y_x = \phi(x)$

$g(y, x) = y - x$

$x_{n+1} = x_n + (\phi(x_n) - x_n)/(n+1)$

BFT: (x_n) converges to f.p. of ϕ

Dvoretzky $\Rightarrow$ Robbins-Monro

(Dvoretzky) $$x_{n+1} = T_n(x_1, \ldots, x_n) + Y_n(x_1, \ldots, x_n)$$

$$\Downarrow$$

$$f(u) = \mathbb{E}(Z_u)$$

$$T_n(x_1, \ldots, x_n) = x_n + a_n f(x_n)$$

$$Y_n(x_1, \ldots, x_n) = a_n(Z_{x_n} - f(x_n))$$

(Robbins-Monro) $$x_{n+1} = x_n + a_n Z_{x_n}$$

Convergence of Sequences of Random Variables

Almost Sure Convergence

A sequence of random variables $X_1, X_2, X_3, \cdots$ converges **almost surely** to a random variable X, shown by $X_n \xrightarrow{a.s.} X$, if

$$P\left(\left\{s \in S : \lim_{n\to\infty} X_n(s) = X(s)\right\}\right) = 1.$$

Convergence in Mean

Let $r \geq 1$ be a fixed number. A sequence of random variables $X_1, X_2, X_3, \cdots$ converges **in the** r**th mean** or **in the** L^r **norm** to a random variable X, shown by $X_n \xrightarrow{L^r} X$, if

$$\lim_{n\to\infty} E\left(|X_n - X|^r\right) = 0.$$

If $r = 2$, it is called the **mean-square convergence**, and it is shown by $X_n \xrightarrow{m.s.} X$.

Convergence in Probability

A sequence of random variables $X_1, X_2, X_3, \cdots$ converges **in probability** to a random variable X, shown by $X_n \xrightarrow{p} X$, if

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \epsilon\right) = 0, \qquad \text{for all } \epsilon > 0.$$

Almost Sure Convergence

A sequence $(x_n) \in \mathbb{R}$ is **converging** if

$$\forall \varepsilon > 0 \, \exists N \, \forall m, n \geq N(|x_n - x_m| < \varepsilon)$$

A sequence of r. v. $(X_n) \in \Omega \to \mathbb{R}$ is **converging almost surely** if

$$P[\{\omega \colon \forall \varepsilon > 0 \, \exists N \, \forall m, n \geq N(|X_n(\omega) - X_m(\omega)| < \varepsilon)\}] = 1$$

Egorov Theorem

A sequence of r. v. $(X_n) \in \Omega \to \mathbb{R}$ **converges almost uniformly** if

$$\forall \varepsilon, \delta > 0 \, \exists N \left(P[\{\omega \colon \forall m, n \geq N(|X_n(\omega) - X_m(\omega)| < \varepsilon)\}] \geq 1 - \delta \right)$$

Avigad-Dean-Rute (2011)