Quantitative Analysis of Stochastic Approximation Methods

Oberwolfach Workshop
Mathematical Logic: Proof Theory, Constructive Mathematics

Paulo Oliva
(joint work with Rob Arthan)

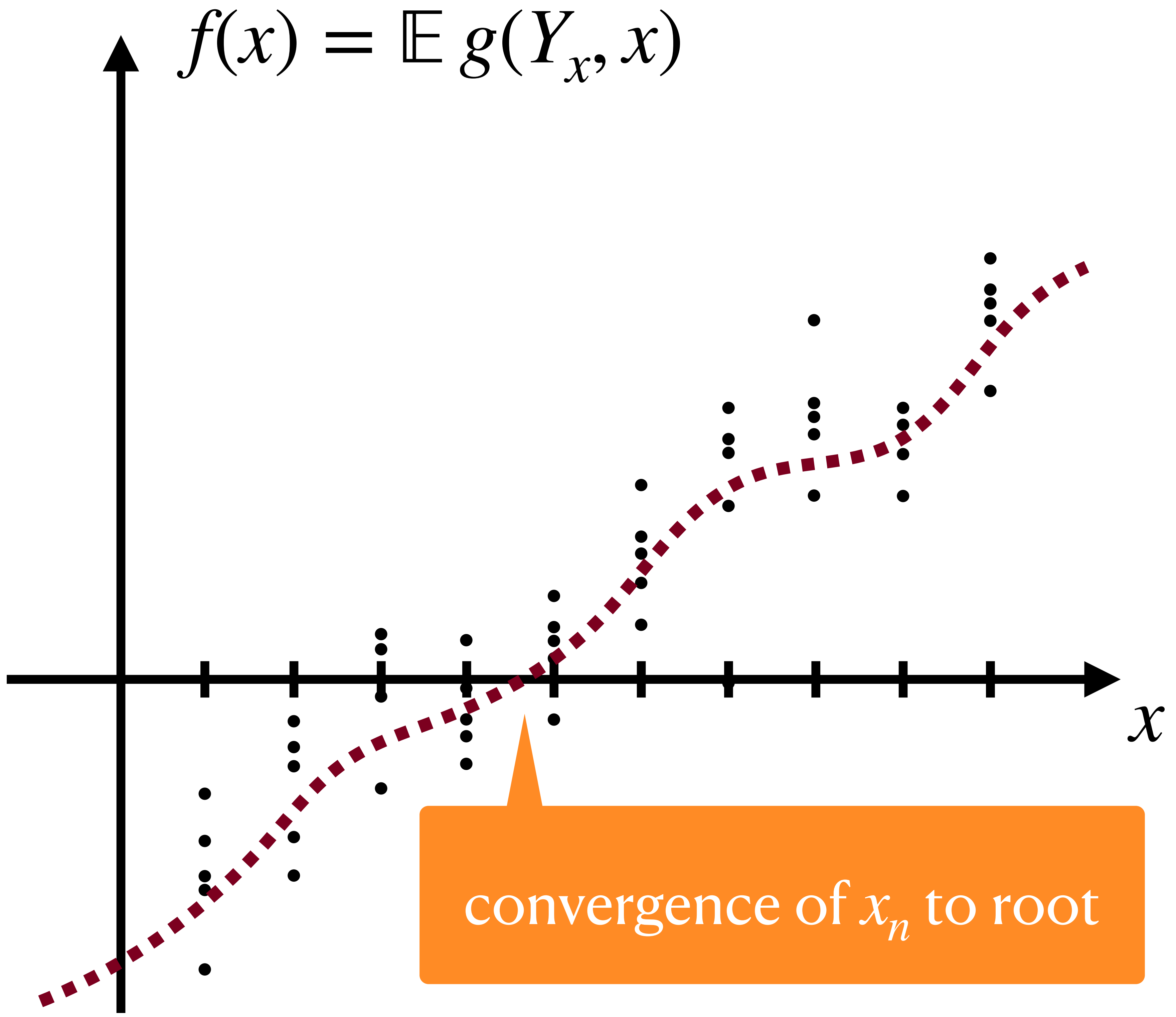
Thu 16 Nov 2023

Approximation Methods

Stochastic Approximation Methods

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 $\mathbb{E}\, g(Y_x, x) = 0$

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

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: $y_1, y_2, \ldots$

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

(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 $\mathbb{E}\, g(Y_x, x) = 0$

<u>Stochastic Approximation Algorithm</u>

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

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

Wolfowitz (1952)

- 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

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

Wolfowitz (1952):

Convergence in prob. when $f(y, x) = b - y$

Assumes:

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

Blum (1954)

- 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

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

Blum (1952):

A. s. convergence when $f(y, x) = b - y$

Assumes:

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

Dvoretzky Theorem (Derman-Sachs Proof)

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 $X_n \to 0$ w.p.1.

$^{*}X_n$ indep.

Derman-Sacks Proof (1959)

- 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

Derman-Sacks Proof (1959)

LEMMA 1. *Let* $\{a_n\}$, $\{b_n\}$, $\{c_n\}$, $\{\delta_n\}$, and $\{\xi_n\}$ *be sequences of real numbers satisfying*

(i) $\{a_n\}$, $\{b_n\}$, $\{c_n\}$, $\{\xi_n\}$ *are non-negative,*

(ii) $\lim_{n\to\infty} a_n = 0$, $\sum b_n < \infty$, $\sum c_n = \infty$, $\sum \delta_n$ *converges,*

and, for all n *larger than some* N_0,

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

Then, $\lim_{n\to\infty} \xi_n = 0$.

Transfer & Dialectica

- 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