An introduction to stochastic approximation

Richard Combes (rcombes@kth.se)
Jie Lu
Alexandre Proutière

FEL 3310: Distributed optimization

A first example

First example of stochastic approximation (Robbins , 1951): a line search with noise.

- Parameter $x \in \mathbb{R}$
- System output $g(x) \in \mathbb{R}$, g smooth and increasing.
- Target value: $g^* = g(x^*)$.
- When x is used, we can observe $g(x) + M$, with $\mathbb{E}[M] = 0$ (noise)
- Goal: determine x^* sequentially

Proposed method , $\epsilon_n \sim 1/n$:

$$x_{n+1} = x_n + \epsilon_n(g^* - (g(x_n) + M_n))$$

A first example, some intuitions

$$x_m = x_n + \underbrace{\sum_{k=n}^{m-1} \epsilon_k (g^* - g(x_k))}_{\text{discretization term}} + \underbrace{\sum_{k=n}^{m-1} \epsilon_k M_k}_{\text{noise term}}$$

Error due to noise:

- Assume $\{M_n\}$ i.i.d Gaussian with unit variance.
- Noise term: $S_{n,m} = \sum_{k=n}^{m-1} M_k/k$,
- $\textbf{var}(S_{n,m}) \leq \sum_{k \geq n} 1/k^2 \to_{n \to +\infty} 0$
- Should be negligible using a law of large numbers type of result.

A first example, some intuitions

$$x_m = x_n + \underbrace{\sum_{k=n}^{m-1} \epsilon_k (g^* - g(x_k))}_{\text{discretization term}} + \underbrace{\sum_{k=n}^{m-1} \epsilon_k M_k}_{\text{noise term}}$$

Discretization error (assume no noise)

- Fundamental theorem of calculus: $(1/n)|g^* - g(x_n)| \leq (g'/n)|x^* - x_n|$.
- So for $n \geq g'$, we have either $x_n \leq x_{n+1} \leq x^*$ or $x_n \geq x_{n+1} \geq x^*$.
- $n \mapsto |g(x_n) - g^*|$ is decreasing for large n
- The discretization term is a Euler scheme for the o.d.e: $\dot{x} = g^* - g(x)$.

The associated o.d.e

General update equation:

$$x_{n+1} = x_n + \epsilon_n(h(x_n) + M_n),$$

with $h : \mathbb{R}^d \to \mathbb{R}^d$ and $x_n \in \mathbb{R}^d$, $M_n \in \mathbb{R}^d$, $\mathbb{E}[M_n] = 0$.
Associated o.d.e.:

$$\dot{x} = h(x).$$

- Main idea: The asymptotic behavior of $\{x_n\}$ can be derived from that of the o.d.e.
- With suitable assumptions, if the o.d.e. has a continosly differentiable Liapunov function V, then $V(x_n) \to_{n\to+\infty} 0$ a.s.

Why are stochastic approximation schemes so common ?

- Low memory requirements: Markovian updates, x_{n+1} is a function of x_n and the observation at time n. Implementation requires a small amount of memory.
- Influence of noise: replace a complicated, stochastic sequence by a deterministic o.d.e which does not depend on the noise statistics.
- Iterative updates: good models for agents updating their behavior through repeated interaction.

Example: stochastic gradient descent

- Goal: optimize a cost function with noise (Kiefer and Wolfowitz, 1952)
- Cost function $f : \mathbb{R} \to \mathbb{R}$ strongly convex, twice differentiable with a unique minimum x^*.
- Observation: $f(x_n) + M_n$
- Idea: approximate ∇f by finite differences, and use gradient descent:

$$x_{n+1} = x_n - \epsilon_n \frac{f(x_n + \delta_n) - f(x_n - \delta_n)}{2\delta_n},$$

- Provable convergence for (say): $\epsilon_n = n^{-1}$, $\delta_n = n^{-1/3}$.
- Useful for: on-line regression, training of neural networks, on-line optimization of MDPs etc.

Example: distributed updates

- Components of x_n are not updated simultaneously, agent k controls $x_{n,k}$.
- At time n , component $k(n)$ is updated, $k(n)$ uniformly distributed in $\{1, \ldots, d\}$.
- Update equation:

$$x_{n+1,k} = \begin{cases} x_{n,k} + \epsilon_n(h_k(x_n) + M_{n,k}) & , \ k = k(n) \\ x_{n,k} & , \ k \neq k(n) \end{cases} .$$

- The behavior of $\{x_n\}$ can be described by the ordinary differential equation (o.d.e.) $\dot{x} = h(x)$.
- Distributed and centralized updates have the same behavior.

Main theorem: assumptions

$\mathcal{F}_n$, σ-algebra generated by $(x_0, M_0, \ldots, x_n, M_n)$ (information available at time n).

(A1) (Lipshitz continuity of h) There exists $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$ $||h(x) - h(y)|| \leq L||x - y||$.

(A2) (Diminishing step sizes) $\sum_{n \geq 0} \epsilon_n = \infty$ and $\sum_{n \geq 0} \epsilon_n^2 < \infty$.

(A3) (Martingale difference noise) There exists $K \geq 0$ such that for all n we have that $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = 0$ and $\mathbb{E}[||M_{n+1}||^2|\mathcal{F}_n] \leq K(1 + ||x_n||)$.

(A4) (Boundedness of the iterates) $\sup_{n \geq 0} ||x_n|| < \infty$ a.s.

(A5) (Liapunov function) There exists a positive, radially unbounded, continuously differentiable function $V : \mathbb{R}^d \to \mathbb{R}$ such that for all $x \in \mathbb{R}^d$, $\langle \nabla V(x), h(x) \rangle \leq 0$ with strict inequality if $V(x) \neq 0$.

Main theorem: statement

Theorem

Assume that (A1) - (A5) hold, then we have that:

$$V(x_n) \to_{n \to \infty} 0, \; a.s.$$

Main theorem: lemma

Define $t(n) = \sum_{k=0}^{n-1} \epsilon_k$, and $\overline{x}$ linear by parts with $\overline{x}(t(n)) = x_n$.
Define x^n a solution of the o.d.e with $x^n(t(n)) = x_n$.

Lemma

For all $T > 0$, we have that:

$$\sup_{t \in [t(n), t(n)+T]} ||\overline{x}(t) - x^n(t)|| \to_{n\to\infty} 0 \text{ a.s.}$$

Appendix: Gronwall's lemma

Lemma (Gronwall's inequality)

Consider $T \geq 0$, $L \geq 0$ *and a function* $t \mapsto x(t)$ *such that* $\dot{x}(t) \leq L||x(t)||$, $t \in [0, T]$. *Then we have that* $\sup_{t \in [0,T]} ||x(t)|| \leq ||x(0)||e^{LT}$.

Lemma (Gronwall's inequality, discrete case)

Consider $K \geq 0$ *and positive sequences* $\{x_n\}$, $\{\epsilon_n\}$ *such that for all* $0 \leq n \leq N$*:*

$$x_{n+1} \leq K + \sum_{u=0}^{n} \epsilon_n x_n.$$

Then we have the upper bound: $\sup_{0 \leq n \leq N} x_n \leq Ke^{\sum_{n=0}^{N} \epsilon_n}$.

Appendix: Martingale convergence theorem

Theorem (Martingale convergence theorem)

Consider $\{M_n\}_{n\in\mathbb{N}}$ *a martingale in* $\mathbb{R}^d$ *with:*

$$\sum_{n\geq 0} \mathbb{E}[||M_{n+1} - M_n||^2|\mathcal{F}_n] < \infty,$$

then there exists a random variable $M_\infty \in \mathbb{R}^d$ *such that* $||M_\infty|| < \infty$ *a.s. and* $M_n \rightarrow_{n\rightarrow\infty} M_\infty$ *a.s.*