Stochastic Approximation in Nonconvex Optimization and Reinforcement Learning

M. Vidyasagar FRS

SERB National Science Chair, IIT Hyderabad

Reinforcement Learning Workshop
IISc, 26 February 2024

Summary

1. Stochastic Approximation: Overview
2. Nonconvex Optimization
   - A Linear Recursion
   - Assumptions on the Objective Function
   - Convergence Theorems
   - Numerical Example
3. Block Asynchronous SA (BASA)
   - Convergence Analysis
   - Application to Q-Learning
4. Some Directions for Future Research

Outline

1. Stochastic Approximation: Overview
2. Nonconvex Optimization
   - A Linear Recursion
   - Assumptions on the Objective Function
   - Convergence Theorems
   - Numerical Example
3. Block Asynchronous SA (BASA)
   - Convergence Analysis
   - Application to Q-Learning
4. Some Directions for Future Research

Original Problem Formulation

Stochastic Approximation (SA) was proposed in 1951 by Robbins & Monro.

Objective: Given a function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$, find a solution to $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}$, when only noisy measurements of $\mathbf{f}(\cdot)$ are available.

Iterative method: Start with $\boldsymbol{\theta}_0$ and update via

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t [\mathbf{f}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}],$$

where α_t is the "step size" and $\boldsymbol{\xi}_{t+1}$ is the measurement error.

Question: When does $\boldsymbol{\theta}_t$ converge to a solution?

Various Types of Updating

- **Synchronous SA:** At each time t, *every component* of $\boldsymbol{\theta}_t$ gets updated. Traditional approach.
- **Asynchronous SA:** At each time t, *exactly one component* of $\boldsymbol{\theta}_t$ gets updated. Used in Reinforcement Learning (RL).
- **Block Asynchronous SA:** At each time t, *some but not necessarily all components* of $\boldsymbol{\theta}_t$ get updated. Used in large-scale optimization.

We will briefly discuss each of these.

Solving Fixed Point Problems

- Suppose $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$, and we wish to find a **fixed point** of $\mathbf{g}(\cdot)$, that is, a $\boldsymbol{\theta}^*$ such that $\mathbf{g}(\boldsymbol{\theta}^*) = \boldsymbol{\theta}^*$.
- This is the same as solving $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}$, with $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\theta}) - \boldsymbol{\theta}$.
- The updating formula is now

$$\boldsymbol{\theta}_{t+1} = (1 - \alpha_t)\boldsymbol{\theta}_t + \alpha_t[\mathbf{g}(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}],$$

where, as before, $\boldsymbol{\xi}_{t+1}$ is a measurement error.

Many problems in RL (e.g., Temporal Difference learning, Q-learning) involve solving a fixed-point problem.

Nonconvex Optimization

- Suppose $J : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^1$. We wish to find a **stationary point** $\boldsymbol{\theta}^*$ such that $\nabla J(\boldsymbol{\theta}^*) = \mathbf{0}$.
- This is similar to above discussion, with $\mathbf{f}(\boldsymbol{\theta}) = -\nabla J(\boldsymbol{\theta}_t)$. (Why the minus sign?)
- Suppose $\mathbf{h}_{t+1}$ is the **search direction** at step t (not necessarily equal to $\nabla J(\boldsymbol{\theta}_t)$), which is also corrupted by measurement error.
- Several ways to choose the search direction: momentum, or accelerated methods, ADAM, NADAM, RMSPROP etc.
- The updating formula is now

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \mathbf{h}_{t+1}.$$

- Several possible error models.
- Ideally, *not* restricted to convex $J(\cdot)$ alone!

Outline

Some Notation

Suppose $\{\mathcal{F}_t\}$ is a **filtration**, i.e., be an increasing sequence of σ-algebras. Then $E_t(X)$ denotes the **conditional expectation** $E(X|\mathcal{F}_t)$, and $CV_t(X)$ denotes the **conditional variance**

$$CV_t(X) = E_t(\|X - E_t(X)\|_2^2).$$

Definition

A function $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ is said to **belong to Class $\mathcal{B}$** if $\eta(0) = 0$, and in addition

$$\inf_{\epsilon \le r \le M} \eta(r) > 0, \ \forall 0 < \epsilon \le M < \infty.$$

Example of a Class $\mathcal{B}$ Function

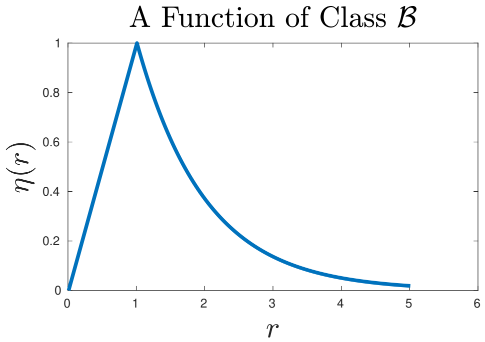
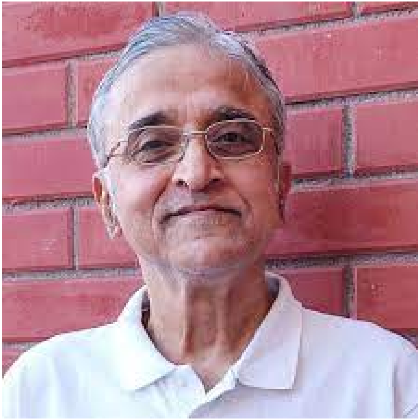


Figure: An illustration of a function in Class $\mathcal{B}$

Collaborators



Prof. R. L. Karandikar
Emeritus Prof., CMI

Tadipatri Uday Kiran Reddy
U. Penn (ex-IITH)

Outline

A Linear Recursion

Consider the linear stochastic recurrence relation

$$\boldsymbol{\theta}_{t+1} = (1 - \alpha_t)\boldsymbol{\theta}_t + \alpha_t \boldsymbol{\xi}_{t+1}, t \geq 0,$$

where $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, $\boldsymbol{\xi}_{t+1} \in \mathbb{R}^d$, and $\alpha_t \in (0,1)$, are all random variables for $t \geq 0$.

Despite its simple appearance, this equation is *all we need* to analyze all the problems studied here.

Assumption (N): Define $\mathcal{F}_t$ to be the σ-algebra generated by $\boldsymbol{\theta}_0, \alpha_0^t, \boldsymbol{\xi}_1^t$. Suppose there exist sequences of constants $\{\mu_t\}$, $\{M_t\}$ such that, for all $t \geq 0$ we have (almost surely)

$$\|E_t(\boldsymbol{\xi}_{t+1})\|_2 \leq \mu_t(1 + \|\boldsymbol{\theta}_t\|_2),$$

$$CV_t(\boldsymbol{\xi}_{t+1}) \leq M_t^2(1 + \|\boldsymbol{\theta}_t\|_2^2).$$

General Convergence Theorem

Theorem

(RLK-MV, 2024) Under assumptions (N), if (almost surely)

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty, \sum_{t=0}^{\infty} \mu_t \alpha_t < \infty, \sum_{t=0}^{\infty} M_t^2 \alpha_t^2 < \infty,$$

then $\{\boldsymbol{\theta}_t\}$ is bounded, and $\|\boldsymbol{\theta}_t\|_2$ converges to an $\mathbb{R}$-valued random variable. If in addition,

$$\sum_{t=0}^{\infty} \alpha_t = \infty,$$

then $\boldsymbol{\theta}_t \to \mathbf{0}$.

Assumption (N) is the *weakest assumption to date* on the error.

Outline

Reprise: Problem Formulation

Objective: Find a stationary point of a $\mathcal{C}^1$-function $J : \mathbb{R}^d \to \mathbb{R}$.

Approach: At each step t, choose a "search direction" $\mathbf{h}_{t+1}$, and set

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \mathbf{h}_{t+1},$$

where α_t is the step size.

Note: $\mathbf{h}_{t+1}$ need not equal $\nabla J(\boldsymbol{\theta}_t)$ plus noise: cf. momentum-based, accelerated, ADAM, NADAM, RMSPROP, etc.

Question: When does $\boldsymbol{\theta}_t$ converge to a stationary point of $J(\cdot)$, *even when $J(\cdot)$ is not convex?*

Known Bounds for Noise-Free Gradient Descent

Suppose $J(\cdot)$ is *convex* with a Lipschitz-continuous gradient. Assume that the unique global minimum of $J(\cdot)$ occurs at $\boldsymbol{\theta}^* = \mathbf{0}$ and equals zero.

- Choose $\mathbf{h}_{t+1} = \nabla J(\boldsymbol{\theta}_t)$ (gradient descent without noise). Then $J(\boldsymbol{\theta}_t) = O(t^{-1})$.[1]
- Nesterov's Accelerated Gradient (NAG) method achieves $J(\boldsymbol{\theta}_t) = O(t^{-2})$.
- No algorithm can achieve a faster rate.
- When a first-order approximation for $\nabla J(\boldsymbol{\theta}_t)$ is used, then $J(\boldsymbol{\theta}_t) = O(t^{-1/2})$.[2]

---

[1] Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer Scientific+Business Media (2004)

[2] Nesterov, Y., Spokoiny, V.: Random Gradient-Free Minimization of Convex Functions. Foundations of Computational Mathematics 17(2), 527–566 (2017)

Class of Nonconvex Functions Under Study

(TUKR-MV, 2023 and RLK-MV, 2024)

(J1.) $J : \mathbb{R}^d \rightarrow \mathbb{R}$ is $\mathcal{C}^1$, and $\nabla J(\cdot)$ is Lipschitz-continuous with constant L.

(J2.) There exists a constant H such that

$$\|\nabla J(\boldsymbol{\theta})\|_2^2 \leq H J(\boldsymbol{\theta}), \ \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

(J3.) There exists a function $\psi(\cdot)$ of Class $\mathcal{B}$ such that

$$\|\nabla J(\boldsymbol{\theta})\|_2^2 \geq \psi(J(\boldsymbol{\theta})), \ \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

(J3'.) There exists a constant K such that

$$\|\nabla J(\boldsymbol{\theta})\|_2^2 \geq K J(\boldsymbol{\theta}), \ \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

Discussions on Conditions

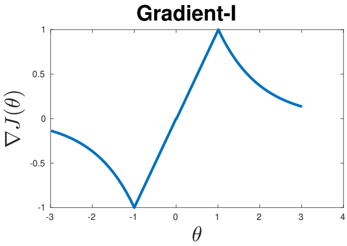
- (J2) holds for a convex function with Lipschitz-continuous gradient.
- (J3) is *weaker than* the Polyak-Lojawiesicz (PL) condition: There exists a $c > 0$ such that

$$\|\nabla J(\boldsymbol{\theta})\|_2^2 \geq cJ(\boldsymbol{\theta}), \ \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

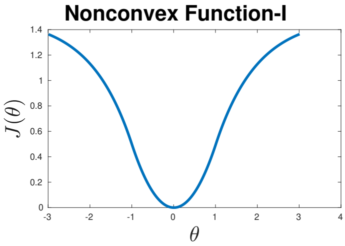
  In (J3), the linear term is replaced by a function of Class $\mathcal{B}$.
- A function satisfying (J1), (J2) and (J3) is "invex" – every local minimum is also a global minimum.
- (J3') is stronger than (J3), and is the PL condition.
- A *strongly* convex function satisfies (J3').

IIT Hyderabad

An Example of a Nonconvex Function that Satisfies (J3)



The function $\nabla J(\cdot)$



The function $J(\cdot)$

Figure: A nonconvex function that satisfies (J3) but not (J3')

$J(\cdot)$ satisfies (J3) and is not convex. It also *does not satisfy* the PL condition, because as $\theta \to \infty$, $J(\theta) \to \infty$ but $\nabla J(\boldsymbol{\theta}) \to 0$.

An Example of a Nonconvex Function that Satisfies (J3')

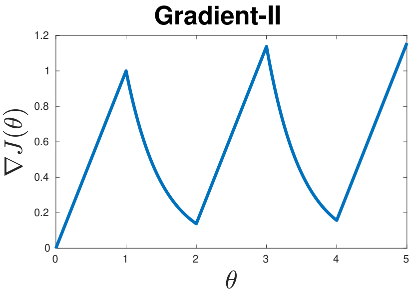


Figure: Gradient of a Function whose integral satisfies (J3')

Define $\nabla J(\cdot)$ to be the odd extension of the above, and $J(\cdot)$ to be its integral. Since $\nabla J(\cdot)$ is bounded both above and below by a linear function, $J(\cdot)$ satisfies (J3').

Outline

Convergence with Noisy Gradient: Set-Up

Suppose the search direction is a noise-corrupted gradient, i.e.,

$$\mathbf{h}_{t+1} = \nabla J(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1},$$

where the error satisfies

Assumption (N'): $E_t(\boldsymbol{\xi}_{t+1}) = \mathbf{0}$, and for some M, we have

$$CV_t(\boldsymbol{\xi}_{t+1}) \leq M^2(1 + \|\boldsymbol{\theta}_t\|_2^2), \ \forall t \geq 0.$$

Assumption (N') is more restrictive than Assumption (N):

$$\|E_t(\boldsymbol{\xi}_{t+1})\|_2 \leq \mu_t(1 + \|\boldsymbol{\theta}_t\|_2), CV_t(\boldsymbol{\xi}_{t+1}) \leq M_t^2(1 + \|\boldsymbol{\theta}_t\|_2^2).$$

Convergence Theorem

Theorem

1. *Suppose (J1) and (J2) hold, and*

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty.$$

Then $\{J(\boldsymbol{\theta}_t\}$ *and* $\{\nabla J(\boldsymbol{\theta}_t)\}$ *are bounded.*

2. *If in addition, (J3) holds and*

$$\sum_{t=0}^{\infty} \alpha_t = \infty,$$

then $J(\boldsymbol{\theta}_t) \to 0$ *and* $\|\nabla J(\boldsymbol{\theta}_t)\|_2 \to 0$ *as* $t \to \infty$.

Optimal Rate of Convergence with Noisy Gradient

Theorem

Suppose that $J(\cdot)$ satisfies (J1), (J2), (J3'), and that $\mathbf{h}_{t+1} = \nabla J(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}$ with Assumption (N') on $\boldsymbol{\xi}_{t+1}$. Suppose the step size sequence satisfies

$$\alpha_t = O(t^{-(1-\phi)}), \alpha_t = \Omega(t^{-(1-C)}), C \in (0, \phi]$$

for some $\phi \in (0, 0.5)$. Then $J(\boldsymbol{\theta}_t), \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda})$ for every $\lambda < 1 - 2\phi$.

In particular, we can make $J(\boldsymbol{\theta}_t), \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda})$ for any $\lambda < 1$ by choosing $\phi < (1 - \lambda)/2$.

We can achieve the same rate as Gradient Descent even *with* noisy measurements, provided (PL) holds.

Convergence with Approximate Gradient: Set-Up

Define the search direction $\mathbf{h}_{t+1} \in \mathbb{R}^d$ as follows:

$$h_{t+1,i} = \frac{[J(\boldsymbol{\theta}_t + c_t \boldsymbol{\Delta}_{t+1}) + \xi^+_{t+1,i}] - [J(\boldsymbol{\theta}_t - c_t \boldsymbol{\Delta}_{t+1}) - \xi^-_{t+1,i}]}{2 c_t \Delta_{t+1,i}},$$

where $\Delta_{t+1,i}, i \in [d]$ are d different and pairwise independent **Rademacher variables**. c_t is called the "increment."

Only 2 function evaluations, for every value of d. This is called SPSA (Simultaneous Perturbation SA) in Spall (1992).

Suppose the error $\boldsymbol{\xi}_{t+1}$ satisfies Assumption (N') (same as with noisy gradient):

$$E_t(\boldsymbol{\xi}_{t+1}) = \mathbf{0}, CV_t(\boldsymbol{\xi}_{t+1}) \leq M^2(1 + \|\nabla J(\boldsymbol{\theta}_t)\|_2^2), \ \forall t.$$

Convergence Theorem

Theorem

1. *Suppose Assumptions (J1), (J2) and (J3) hold. Then the iterations of the Stochastic Gradient Descent algorithm are bounded almost surely whenever*

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty, \sum_{t=0}^{\infty} \alpha_t c_t < \infty, \sum_{t=0}^{\infty} \alpha_t^2 / c_t^2 < \infty,$$

2. *If, in addition, we also have*

$$\sum_{t=0}^{\infty} \alpha_t = \infty,$$

then $J(\boldsymbol{\theta}_t) \to 0$ *and* $\nabla J(\boldsymbol{\theta}_t) \to \mathbf{0}$ *almost surely as* $t \to \infty$.

IIT Hyderabad

Convergence Theorem with Rates

Now *two* things to be adjusted: α_t and c_t.

Theorem

Suppose Assumption (J3) is strengthened to Assumption (J3'). Further, suppose that $\alpha_t = O(t^{-(1-\phi)})$ *and* $\alpha_t = \Omega(t^{-(1-C)})$*, where* $C \in (0, \phi]$*, and* $c_t = \Theta(t^{-s})$*. Suppose further that*

$$\phi < s, \phi + s < 0.5,$$

and define

$$\nu := \min\{1 - 2(\phi + s), s - \phi\}.$$

Then

$$J(\boldsymbol{\theta}_t), \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda}) \; \forall \lambda < \nu.$$

Optimal Choice of Parameters

In $\alpha_t = O(t^{-(1-\phi)})$, choose ϕ as small as possible and $C = \phi$ (large step sizes). Choose $c_t = O(t^{-1/3})$. Then

$$J(\boldsymbol{\theta}_t), \|\nabla J(\boldsymbol{\theta}_t)\|_2^2 = o(t^{-\lambda}) \ \forall \lambda < 1/3.$$

Compare with known bound of $t^{-1/2}$.

Multiple Measurement SPSA

In Bhatnagar and Prashanth (2022), they use $k+1$ measurements, not 2. Then $\mu_t = \Theta(c_t^k)$, not $\Theta(c_t)$.

The optimal convergence rate now is $o(t^{-\lambda})$ for $\lambda < k/(k+2)$, with the optimal increment being $c_t = t^{-s}$ with $s \approx 1/(k+2)$ (and ϕ being close to zero).

So we can achieve convergence arbitrarily close to $O(t^{-1})$ (the best bound for GD) by increasing k, *even with noisy measurements.*

Example: If $k = 2$ (3 function evaluations at each t), we match the rate of $O(t^{-1/2})$ of Nesterov-Spokiny (2017).

Block Updating

- Until now, *every* component of $\boldsymbol{\theta}_t$ is updated at each t (synchronous updating).
- In TUKR & MV, we study the case where *some but not* components are updated.
- If at time t, only components in some subset $S(t) \subseteq [d]$ are updated, then in principle, we need to compute $[\nabla J(\boldsymbol{\theta}_t)]_i$ only for $i \in S(t)$.
- However, if methods such as back-propagation are used, then it is just as easy to compute the full vector $\nabla J(\boldsymbol{\theta}_t)$.
- This approach is numerically less expensive than computing the full vector $\nabla J(\boldsymbol{\theta}_t)$ when *approximate gradients* are used.

Methods for Choosing Coordinates to be Updated

1. Full coordinate update.
2. Single coordinate update: Choose $i \in [d]$ at random and with equal probability at each time t, and update only the i-th component of $\boldsymbol{\theta}_t$.
3. Multiple coordinate update: Choose N different indices from $[d]$ *with replacement*, and update. If there are repetitions, update that coordinate twice (or more times).
4. Bernoulli update: At time $t+1$, pick a "rate" $\rho_{t+1} \in (0,1)$, and run d different Bernoulli processes with this rate. Update the i-th coordinate only if the i-th Bernoulli process equals 1.

IIT Hyderabad

Polyak's Heavy Ball Algorithm with Block Updating

In TUKR & MV, we study Polyak's Heavy Ball algorithm. In the full-coordinate update, we have

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_t[-\nabla J(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_{t+1}] + \mu(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}),$$

where $\boldsymbol{\xi}_{t+1}$ is the measurement error, and μ is the HB parameter. (Setting $\mu = 0$ gives Stochastic Gradient Descent.)

Suppose as before that (N) holds, i.e., there exist sequences of constants $\{\mu_t\}$, $\{M_t\}$ such that

$$\|E_t(\boldsymbol{\xi}_{t+1})\|_2 \leq \mu_t(1 + \|\boldsymbol{\theta}_t\|_2) \ \forall t,$$

$$CV_t(\boldsymbol{\xi}_{t+1}) \leq M_t^2(1 + \|\boldsymbol{\theta}_t\|_2^2) \ \forall t.$$

We can also apply each of the three other block-updating methods.

Convergence Theorem

Theorem

Suppose J satisfies Assumptions (J1) through (J3). Suppose any one of Options (1)–(4) is applied in the SHB algorithm.

1. *Suppose*

$$\sum_{t=0}^{\infty} \alpha_t^2 < \infty, \sum_{t=0}^{\infty} \alpha_t \mu_t < \infty, \sum_{t=0}^{\infty} \alpha_t^2 M_t^2 < \infty,$$

$$\sum_{t=0}^{\infty} \alpha_t = \infty.$$

   Then $\{J(\boldsymbol{\theta}_t)\}$ and $\{\boldsymbol{\theta}_t\}$ are bounded almost surely.

2. *If we add Assumption (J3'), then $\nabla J(\boldsymbol{\theta}_t) \to \mathbf{0}$ as $t \to \infty$, and $J(\boldsymbol{\theta}_t) \to 0$ as $t \to \infty$.*

Outline

1. Stochastic Approximation: Overview
2. Nonconvex Optimization
   - A Linear Recursion
   - Assumptions on the Objective Function
   - Convergence Theorems
   - Numerical Example
3. Block Asynchronous SA (BASA)
   - Convergence Analysis
   - Application to Q-Learning
4. Some Directions for Future Research

Numerical Example

We minimize the objective function

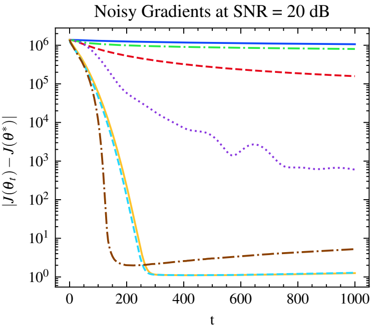
$$J(\boldsymbol{\theta}_t) = \boldsymbol{\theta}_t^\top A \boldsymbol{\theta}_t + \log \left(\sum_{i=0}^{d-1} e^{\theta_{t,i}} \right),$$

where $\boldsymbol{\theta}_t$ is a vector of 1 million parameters, and A is a block-diagonal matrix of size $(10^6 \times 10^6)$ consisting of 100 Hilbert matrices, each of dimension $10^4 \times 10^4$. The log-sum is convex, but the quadratic form is (*Very ill-conditioned.*)

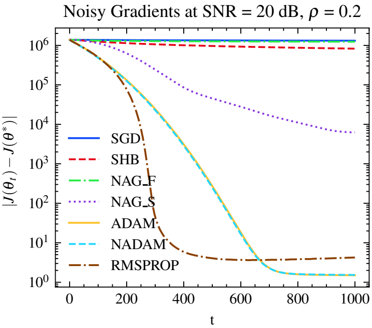
Batch updating with Bernoulli sampling with various rates ρ was tried out. Next slides show the computational results.

IIT Hyderabad

Results with Noisy Gradients



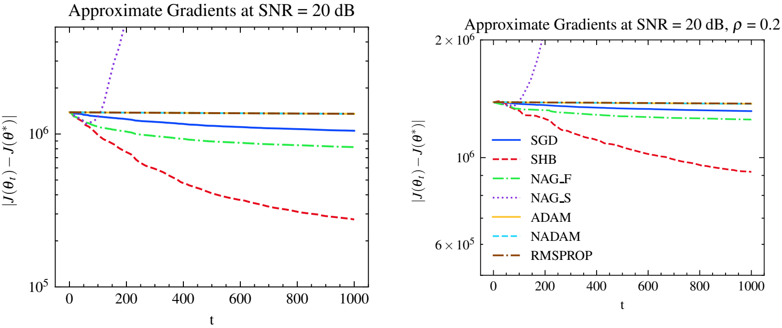
Full update



Bernoulli update

Figure: Comparison of various algorithms using noisy gradients with full and Bernoulli updates

Results with Approximate Gradients



Full update

Bernoulli update

Figure: Comparison of various algorithms using approximate gradients with full and Bernoulli updates

IIT Hyderabad

Some Observations

- With "merely" noisy gradients, ADAM, NADAM and RMSPROP perform the best.
- With Bernoulli updating with just 20% sampling, the performance is comparable to full update.
- However, when *approximate gradients* are used, *all* of these methods diverge badly.
- In contrast, Stochastic Heavy Ball (SHB) method continues to work.
- for Deep NNs, SHB thus seems to be the best method.

Outline

Outline

Fixed Point Problem Formulation

- Suppose $\mathbf{h} : \mathbb{N} \times (\mathbb{R}^d)^{\mathbb{N}} \rightarrow (\mathbb{R}^d)^{\mathbb{N}}$ is a *nonanticipative* family of maps from $(\mathbb{R}^d)^{\mathbb{N}} \rightarrow (\mathbb{R}^d)^{\mathbb{N}}$ with finite memory. Thus $\mathbf{h}(t, \boldsymbol{\theta}_0^\infty)$ depends only on $\boldsymbol{\theta}_{t-\Delta+1}^t$ for each t, for a fixed number Δ.
- Moreover, the dependence is a contraction *in the ℓ_∞-norm.*

$$\|\mathbf{h}(t, \boldsymbol{\psi}_{t-\Delta+1}^t) - \mathbf{h}(t, \boldsymbol{\phi}_{t-\Delta+1}^t)\|_\infty \leq \gamma \|\boldsymbol{\psi}_{t-\Delta+1}^t - \boldsymbol{\phi}_{t-\Delta+1}^t\|_\infty,$$

  for some $\gamma < 1$, for all $t \geq \Delta$, $\forall \boldsymbol{\psi}_0^\infty, \boldsymbol{\phi}_0^\infty \in (\mathbb{R}^d)^{\mathbb{N}}$.
- Therefore, for every sequence $\boldsymbol{\phi}_0^\infty$, the iterations $\mathbf{h}(t, \boldsymbol{\phi}_0^t)$ converge to a unique fixed point $\boldsymbol{\pi}^*$.

Question: How can we find $\boldsymbol{\pi}^*$ when only noisy measurements of $\mathbf{h}$ are available?

Application to RL: Q-Learning.

IIT Hyderabad

Block Asynchronous SA (BASA)

Update scheme:

$$\boldsymbol{\theta}_{t+1} = (\mathbf{1}_d - \boldsymbol{\alpha}_t \circ \boldsymbol{\kappa}_t) \circ \boldsymbol{\theta}_t + (\boldsymbol{\alpha}_t \circ \boldsymbol{\kappa}_t) \circ [\boldsymbol{\eta}_t + \boldsymbol{\xi}_{t+1}],$$

where $\boldsymbol{\eta}_t = \mathbf{h}(t, \boldsymbol{\theta}_0^t)$, $\mathbf{1}_d$ is the vector of all ones, $\boldsymbol{\alpha}_t \in (0,1)^d$ is the *step size vector*, $\boldsymbol{\kappa}_t \in \{0,1\}^d$ is the *update vector*, and $\circ$ denotes the Hadamard (componentwise) product. As before $\boldsymbol{\xi}_{t+1}$ is the measurement error.

Assumptions About the Error

(N1) There exists a finite constant c_1' and a sequence of constants $\{\mu_t\}$ such that

$$\|E_t(\boldsymbol{\xi}_{t+1})\|_2 \leq c_1' \mu_t (1 + \|\boldsymbol{\theta}_0^t\|_\infty), \ \forall t \geq 0.$$

(N2) There exists a finite constant c_2' and a sequence of constants $\{M_t\}$ such that

$$CV_t(\boldsymbol{\xi}_{t+1}) \leq c_2' M_t^2 (1 + \|\boldsymbol{\theta}_0^t\|_\infty^2), \ \forall t \geq 0.$$

A little more general assumptions than earlier.

Choice of Step Size: Global vs. Local Clocks

Distinction first made by Borkar (1998).

For each index $i \in [d]$, define the "counter" process $\{\nu_{t,i}\}$ and its inverse as

$$\nu_{t,i} = \sum_{s=0}^{t} \kappa_{s,i}, \nu_i^{-1}(\tau) := \min\{t \in \mathbb{N} : \nu_{t,i} = \tau\}, \ \forall \tau \geq 1.$$

Then $\nu_i^{-1}(\cdot)$ is well-defined, and

$$\nu_i(\nu_i^{-1}(\tau)) = \tau, \nu_i^{-1}(\nu_{t,i}) \leq t, \nu_i^{-1}(\tau) \leq \tau - 1.$$

Choose a *deterministic* sequence $\{\beta_t\}$. When $\kappa_{t,i} = 1$, if a **global clock** is used, then $\alpha_{t,i} = \beta_t$. If a **local clock** is used, then $\alpha_{t,i} = \beta_{\nu_{t,i}}$.

Assumptions About the Update Process

Assume that there exist constants $r_i > 0, i \in [d]$ such that

$$\frac{\nu_{t,i}}{t} \to r_i \text{ as } t \to \infty, \ \forall i \in [d].$$

Otherwise no assumptions about independence of processes for different indices, or Markovian nature, etc.

Convergence Theorem with Local Clocks

Theorem

Suppose a local clock is used. Suppose that $\{\mu_t\}$ is nonincreasing, and M_t is uniformly bounded, say by M. Suppose in addition that $\beta_t = O(t^{-(1-\phi)})$, for some $\phi > 0$, and $\beta_t = \Omega(t^{-(1-C)})$ for some $C \in (0, \phi]$. Suppose that $\mu_t = O(t^{-\epsilon})$ for some $\epsilon > 0$. Then $\boldsymbol{\theta}_\tau \to \boldsymbol{\pi}^*$ as $\tau \to \infty$ for all $\phi < \min\{0.5, \epsilon\}$. Further, $\|\boldsymbol{\theta}_\tau - \boldsymbol{\pi}^*\|_2 = o(\tau^{-\lambda})$ for all $\lambda < \epsilon - \phi$. In particular, if $\mu_t = 0$ for all t, then $\|\boldsymbol{\theta}_\tau - \boldsymbol{\pi}^*\|_2 = o(\tau^{-\lambda})$ for all $\lambda < 1$.

Convergence Theorem with Global Clocks

Theorem

Suppose a global clock is used. Suppose that β_t is nonincreasing. Suppose in addition that $\beta_t = O(t^{-(1-\phi)})$, for some $\phi > 0$, and $\beta_t = \Omega(t^{-(1-C)})$ for some $C \in (0, \phi]$. Suppose that $\mu_t = O(t^{-\epsilon})$ for some $\epsilon > 0$, and $M_t = O(t^{\delta})$ for some $\delta \geq 0$. Then $\boldsymbol{\theta}_t \to \boldsymbol{\pi}^*$ as $t \to \infty$ whenever

$$\phi < \min\{0.5 - \delta, \epsilon\}.$$

Moreover, $\|\boldsymbol{\theta}_t - \boldsymbol{\pi}^*\|_2 = o(t^{-\lambda})$ for all $\lambda < \epsilon - \phi$. In particular, if $\mu_t = 0$ for all t, then $\|\boldsymbol{\theta}_t - \boldsymbol{\pi}^*\|_2 = o(t^{-\lambda})$ for all $\lambda < 1$.

Outline

Traditional Q-Learning

- Traditional Q-learning is *asynchronous SA*: At time t, only $Q(X_t, U_t)$ is updated.
- Convergence theorems for Q-learning require conditions such as
$$\sum_{t=0}^{\infty} \alpha_t I_{(X_t,U_t)=(x_i,u_j)} = \infty,$$
for each state-action pair (x_i, u_j).
- To ensure the above, it is often assumed that *every policy* results in an irreducible Markov process.
- (Tsitsiklis 2007) Verifying whether *every policy* results in a unichain is NP-hard.
- So we need another set of conditions that are easy to verify.

Batch Q-Learning

- Choose an arbitrary initial guess $Q_0 : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, and m initial states $X_0^k \in \mathcal{X}, k \in [m]$.
- At time t, for each action index $k \in [m]$, with current state $X_t^k = x_i^k$, choose the current action as $U_t = u_k \in \mathcal{U}$, and let the Markov process run for one time step. Observe the resulting next state $X_{t+1}^k = x_j^k$. Then update function Q_t as follows, once for each $k \in [m]$:

$$Q_{t+1}(x_i^k, u_k) = \begin{cases} Q_t(x_i^k, u_k) + \alpha_{t,i,k}[R(x_i, u_k) + \gamma V_t(x_j^k) - Q_t(x_i^k, u_k)], \\ Q_t(x_s^k, u_k), \end{cases}$$

where

$$V_t(x_j^k) = \max_{w_l \in \mathcal{U}} Q_t(x_j^k, w_l).$$

- Repeat.

Step Size Used

Here $\alpha_{t,i,k}$ equals β_t for all i, k if a global clock is used, and equals

$$\alpha_{t,i,k} = \sum_{\tau=0}^{t} I_{\{X_t^k = x_i\}}$$

if a local clock is used.

Convergence Theorem

Theorem

Suppose that each matrix A^{u_k} is irreducible, and that the step size sequence $\{\beta_t\}$ satisfies the Robbins-Monro conditions

$$\sum_{t=0}^{\infty} \beta_t^2 < \infty, \sum_{t=0}^{\infty} \beta_t = \infty.$$

With this assumption, we have the following:

1. *If a local clock is used, then Q_t converges almost surely to Q^*.*
2. *If a global clock is used, and $\{\beta_t\}$ is nonincreasing, then Q_t converges almost surely to Q^*.*

The assumptions are easy to verify!

Outline

Convergence Analysis of Several Optimization Algorithms

- Our theoretical analysis is based on enhancing a well-known theorem known as the Robbins-Siegmund ("almost supermartingale") theorem.
- Our enhancements of the R-S theorem can be used to analyze many popular optimization algorithms currently in use.
- In particular, our methods are readily applicable to block updating as well.
- TUKR & MV have analyzed Polyak's "Heavy Ball" algorithm.
- Others have analyzed ADAM, but only with full coordinate update.
- Analyzing various optimization algorithms using the R-S theorem (with or without block updating) is a promising avenue of research.

Application to RL

Question: Can our approach be extended to Markovian SA?

Challenge: Constructing a suitable error model.

References

- M. Vidyasagar, "Convergence of stochastic approximation via martingale and converse Lya punov methods, *Mathematics of Controls Signals and Systems*, 35, 351–374 (2023).
- T. U. K. Reddy and M. Vidyasagar, "Convergence of momentum-based heavy ball method with batch updating and/or approximate gradients," https://arxiv.org/pdf/2303.16241.pdf (2023)
- Rajeeva L. Karandikar and M. Vidyasagar, Convergence rates for stochastic approximation: Biased noise with unbounded variance, and applications. https://arxiv.org/pdf/2312.02828v2, 2024.
- Rajeeva L. Karandikar and M. Vidyasagar, Recent Advances in Stochastic Approximation with Applications to Optimization and Reinforcement Learning, https://arxiv.org/pdf/2109.03445v5, 2024.

Thank You!

IIT Hyderabad