The Collusion of Memory and Nonlinearity in Stochastic Approximation With Constant Stepsize

arxiv.org/abs/2405.16732

Lucy Huo
ORIE
Cornell

Yixuan Zhang
ISyE
UW-Madison

Yudong Chen
CS
UW-Madison

Qiaomin Xie
ISyE
UW-Madison

November 12, 2024

Stochastic Approximation

- Stochastic Approximation (SA): an iterative method for root-finding and optimization (Robbins and Monro 1951)

$$\theta_{k+1} = \theta_k + \alpha_k g(\theta_k, x_k)$$

Stochastic Approximation

- Stochastic Approximation (SA): an iterative method for root-finding and optimization (Robbins and Monro 1951)

$$\theta_{k+1} = \theta_k + \alpha_k g(\theta_k, x_k)$$

  - SGD: $g(\theta, x)$ noisy gradient estimate of the loss function

Stochastic Approximation

- Stochastic Approximation (SA): an iterative method for root-finding and optimization (Robbins and Monro 1951)

$$\theta_{k+1} = \theta_k + \alpha_k g(\theta_k, x_k)$$

  - SGD: $g(\theta, x)$ noisy gradient estimate of the loss function
  - TD-learning: policy evaluation algorithm in RL

Stochastic Approximation

- Stochastic Approximation (SA): an iterative method for root-finding and optimization (Robbins and Monro 1951)

$$\theta_{k+1} = \theta_k + \alpha_k g(\theta_k, x_k)$$

  - SGD: $g(\theta, x)$ noisy gradient estimate of the loss function
  - TD-learning: policy evaluation algorithm in RL
- Solve for equation $\mathbb{E}_{x\sim\pi}[g(\theta^*, x)] = 0$,
  where π is the stationary distribution of $(x_k)_{k\geq 0}$

Stochastic Approximation

- Stochastic Approximation (SA): an iterative method for root-finding and optimization (Robbins and Monro 1951)

$$\theta_{k+1} = \theta_k + \alpha_k g(\theta_k, x_k)$$

  - SGD: $g(\theta, x)$ noisy gradient estimate of the loss function
  - TD-learning: policy evaluation algorithm in RL
- Solve for equation $\mathbb{E}_{x \sim \pi}[g(\theta^*, x)] = 0$, where π is the stationary distribution of $(x_k)_{k \geq 0}$
- Constant stepsize $\alpha_k \equiv \alpha$
  - Fast initial convergence, easy hyperparameter tuning

θ_k vs. θ^*? Algorithmic implications?

Problem Set-up

$$\theta_{k+1} = \theta_k + \alpha g(\theta_k, x_k)$$

- $(x_k)_{k\geq 0}$ is a Markov chain
  - Uniform ergodicity
    e.g., all irreducible, aperiodic, finite-state Markov chain
  - Reinforcement learning, correlated data
- Strongly convex (non-linear) g + Smoothness
  - L_2-regularized logistic regression
  - Smooth ReLU regression

Main Contribution

- Constant stepsize + Markovian $(x_k)_{k\geq 0}$

$$\theta_{k+1} = \theta_k + \boldsymbol{\alpha} g(\theta_k, x_k)$$

- $(x_k, \theta_k)_{k\geq 0}$ is a time-homogeneous Markov chain

Main Contribution

- Constant stepsize + Markovian $(x_k)_{k\geq 0}$

$$\theta_{k+1} = \theta_k + \boldsymbol{\alpha} g(\theta_k, x_k)$$

- $(x_k, \theta_k)_{k\geq 0}$ is a time-homogeneous Markov chain
- Convergence? How fast is the convergence?
  Weak convergence. Unique limiting stationary distribution.
  Geometrically fast.

Main Contribution

- Constant stepsize + Markovian $(x_k)_{k\geq 0}$

$$\theta_{k+1} = \theta_k + \alpha g(\theta_k, x_k)$$

- $(x_k, \theta_k)_{k\geq 0}$ is a time-homogeneous Markov chain
- Convergence? How fast is the convergence?
  Weak convergence. Unique limiting stationary distribution. Geometrically fast.
- Bias characterization $\mathbb{E}\left[\theta_k - \theta^*\right] = ?$
  - Insights for algorithm design

Main Contribution

- Constant stepsize + Markovian $(x_k)_{k\geq 0}$

$$\theta_{k+1} = \theta_k + \boldsymbol{\alpha} g(\theta_k, x_k)$$

- $(x_k, \theta_k)_{k\geq 0}$ is a time-homogeneous Markov chain
- Convergence? How fast is the convergence?
  Weak convergence. Unique limiting stationary distribution. Geometrically fast.
- Bias characterization $\mathbb{E}\left[\theta_k - \theta^*\right] = ?$
  - Insights for algorithm design
- Existing analysis for constant stepsize:
  - i.i.d. data + non-linear g
    (Dieuleveut, Durmus, and Bach 2020)
  - Markovian data + linear g
    (Huo, Chen, and Xie 2023)

Main Contribution

- Constant stepsize + Markovian $(x_k)_{k\geq 0}$

$$\theta_{k+1} = \theta_k + \boldsymbol{\alpha} g(\theta_k, x_k)$$

- $(x_k, \theta_k)_{k\geq 0}$ is a time-homogeneous Markov chain
- Convergence? How fast is the convergence?
  Weak convergence. Unique limiting stationary distribution. Geometrically fast.
- Bias characterization $\mathbb{E}[\theta_k - \theta^*] = ?$
  - Insights for algorithm design
- Existing analysis for constant stepsize:
  - i.i.d. data + non-linear g (Dieuleveut, Durmus, and Bach 2020)
  - Markovian data + linear g (Huo, Chen, and Xie 2023)

  } $\neq$ Markovian data + non-linear g

Asymptotic Bias Expansion

Theorem 1

For some vectors b_{n}, b_{m}, and b_{c}, that are independent of α, we have the expansion

$$\mathbb{E}[\theta_\infty^{(\alpha)}] = \theta^* + \alpha\Big(b_{\mathsf{n}} + b_{\mathsf{m}} + b_{\mathsf{c}}\Big) + \mathcal{O}\Big(\alpha^{3/2}\Big),$$

where

- b_{n} – *nonlinearity of g (Dieuleveut, Durmus, and Bach 2020)*
- b_{m} – *Markovian correlation of (x_k) (Huo, Chen, and Xie 2023)*
- b_{c} – *Markovian correlation $\times$ nonlinearity*

Implications for Algorithm Design

- Polyak-Ruppert (PR) averaging

$$\bar{\theta}_k := \frac{1}{k/2} \sum_{t=k/2}^{k-1} \theta_t$$

PR-averaging will reduce variance, but not the bias.

Implications for Algorithm Design

- Polyak-Ruppert (PR) averaging

$$\bar{\theta}_k := \frac{1}{k/2} \sum_{t=k/2}^{k-1} \theta_t$$

PR-averaging will reduce variance, but not the bias.

- To reduce bias, use Richardson-Romberg (RR) extrapolation

$$\widetilde{\theta}_k = 2\bar{\theta}_k^{(\alpha)} - \bar{\theta}_k^{(2\alpha)}$$

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\theta}_\infty\right] &= 2\mathbb{E}\left[\theta_\infty^{(\alpha)}\right] - \mathbb{E}\left[\theta_\infty^{(2\alpha)}\right] \\
&= 2\Big(\theta^* + \alpha B^{(1)} + \mathcal{O}(\alpha^{3/2})\Big) - \Big(\theta^* + 2\alpha B^{(1)} + \mathcal{O}((2\alpha)^{3/2})\Big) \\
&= \theta^* + \mathcal{O}(\alpha^{3/2}).
\end{aligned}$$

Numerical Example

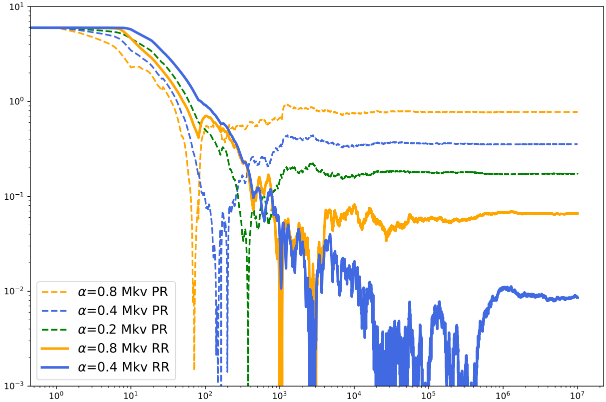

Figure: Presence of Bias in PR and Benefits of RR

Conclusion

- Interplay between Markovian data and the nonlinearity in stochastic approximation (SA) with constant stepsize.
- Practical insights for improving SA algorithms.

Thank You