Non-asymptotic Analysis of Biased Adaptive Stochastic Approximation

Sobihan Surendran[1,2], Adeline Fermanian[2], Antoine Godichon-Baggioni[1], and Sylvain Le Corff[1]

[1]Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, Paris, France

[2]LOPF, Califrais' Machine Learning Lab, Paris, France

SCAN ME

Goal

- ☞ Provide a non-asymptotic analysis of **Stochastic Gradient Descent with biased gradients and adaptive steps** for non-convex smooth functions.
- ☞ Account for a constant and **decreasing bias** over iterations.
- ☞ Application to **Adagrad**, **RMSProp**, and **Adam**.

Introduction

Consider the unconstrained **Optimization Problem:**

$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} V(\theta).$$

In Machine Learning: Objective Function: $V(\theta) = \mathbb{E}[\mathcal{L}(F_\theta(x), y)]$.

$\Rightarrow F_\theta$: Neural Network with parameters $\theta \in \mathbb{R}^d$ and $\mathcal{L}$: Loss Function.

Stochastic Gradient Descent (SGD):

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \widehat{\nabla V}(\theta_n),$$

where γ_{n+1} is the step size and $\widehat{\nabla V}(\theta_n)$ is an estimator of $\nabla V(\theta_n)$.

$\Rightarrow$ Theoretical analysis of Vanilla SGD relies on **unbiased estimator**.

Applications of Biased Gradients

- Reinforcement Learning: Policy Gradient and Actor-Critic.
- Monte Carlo: Importance Sampling and Sequential Monte Carlo.
- Generative Models (biased objectives): VAE and IWAE.
- Bilevel Optimization: Min-Max and Compositional Problems.

♦ Previous works on SGD with biased gradients ([1, 2, 3]):

$$\mathbb{E}\left[\|\nabla V(\theta_n)\|^2\right] = \mathcal{O}\left(n^{-1/2}\log n + bias\right).$$

Adaptive Stochastic Approximation

- **Adaptive Stochastic Approximation:**

$$\theta_{n+1} = \theta_n - \gamma_{n+1} A_n H_{\theta_n}(X_{n+1}), \quad n \in \mathbb{N}.$$

▶ A_n: Sequence of symmetric and positive definite matrices.

▶ $$H_{\theta_n}(X_{n+1}) = \underbrace{\nabla V(\theta_n) + \overbrace{b(\theta_n)}^{bias}}_{h(\theta_n)} + \underbrace{e_{n+1}}_{noise}.$$

- Special cases: If $A_n = I_d \Rightarrow$ **Stochastic Approximation**.
  - $* \; b(\theta_n) = 0$ and $e_{n+1} = 0 \Rightarrow$ **Gradient Descent**.
  - $* \; b(\theta_n) = 0$ and e_{n+1} : zero-mean noise $\Rightarrow$ **SGD** with unbiased estimator.
- **RMSProp and Adam:**

$$A_n = \left[\delta I_d + (1-\beta)\text{Diag}\left(\sum_{k=0}^{n} \beta^{n-k} H_{\theta_k}(X_{k+1}) H_{\theta_k}(X_{k+1})^\top\right)\right]^{-1/2}.$$

Assumptions on Biased Gradients

- **Minimal Assumption:** (extension of [1, 2]) There exist two non-increasing positive sequences $(\lambda_n)_{n\geq 1}$ and $(r_n)_{n\geq 1}$ such that for all $n \in \mathbb{N}$,

$$\mathbb{E}\left[\langle \nabla V(\theta_n), A_n H_{\theta_n}(X_{n+1})\rangle\right] \geq \lambda_{n+1}\left(\mathbb{E}\left[\|\nabla V(\theta_n)\|^2\right] - r_{n+1}\right).$$

- **Mild Assumption in the Case of Bounded Gradients:** There exist $C_\alpha > 0$ and $\alpha > 0$ such that for any $n \in \mathbb{N}$,

$$\tilde{b}_n := \|\mathbb{E}\left[H_{\theta_n}(X_{n+1})\,|\mathcal{F}_n\right] - \nabla V(\theta_n)\| \leq C_\alpha n^{-\alpha}.$$

Convergence Analysis

Theorem: For any $n \geq 1$, let $\gamma_n = C_\gamma n^{-1/2}$, $r_n = C_r n^{-r}$ where $C_\gamma > 0$, $C_r > 0$ and $r > 0$. Let $R \in \{0, \ldots, n\}$ be a uniformly distributed random variable. Under mild assumptions, we have:

$$\mathbb{E}\left[\|\nabla V(\theta_R)\|^2\right] = \begin{cases} \mathcal{O}\left(n^{-1/2}\log n + n^{-r}\right) & \text{if } r < 1/2\,, \\ \mathcal{O}\left(n^{-1/2}\log n + n^{-1/2}\right) & \text{if } r > 1/2\,, \\ \mathcal{O}\left(n^{-1/2}\log n + n^{-1/2}\log n\right) & \text{if } r = 1/2\,. \end{cases}$$

Polyak-Łojasiewicz (PL): For any $n \geq 1$, let $\gamma_n = C_\gamma n^{-\gamma}$ with $C_\gamma > 0$. Under Polyak-Łojasiewicz condition, we have:

$$\mathbb{E}\left[V(\theta_n) - V(\theta^*)\right] = \mathcal{O}\left(n^{-\gamma} + r_n\right).$$

- **I.i.d case.** For an i.i.d. sequence $\{X_n\}$, if $\mathbb{E}[H_{\theta_n}(X_{n+1}) \mid \mathcal{F}_n] = \nabla V(\theta_n)$, the estimator is unbiased. Otherwise, the bias is

$$\tilde{b}_n = \|h(\theta_n) - \nabla V(\theta_n)\|.$$

- **Markov Chain case.** For an ergodic Markov Chain with stationary distribution π, the bias with T samples per step is given by:

$$\tilde{b}_n = \|h(\theta_n) - \nabla V(\theta_n)\| + M\sqrt{\tau_{\text{mix}}/T},$$

where $h(\theta) = \int H_\theta(x)\pi(dx)$ and τ_{mix} is the mixing time.

Application: Stochastic Bilevel Optimization

- **Objective Function:**

$$\min_{\theta \in \mathbb{R}^d} V(\theta) = \mathbb{E}_\xi\left[f(\theta, \phi^*(\theta); \xi)\right] \quad \text{(upper-level)}$$

$$\text{subject to } \phi^*(\theta) \in \operatorname{argmin}_{\phi \in \mathbb{R}^q} \mathbb{E}_\zeta\left[g(\theta, \phi; \zeta)\right] \quad \text{(lower-level)}$$

where f and g are both continuously differentiable, and ξ and ζ are random variables.

- **The gradient of V [4]:**

$$\nabla V(\theta) = \nabla_\theta f(\theta, \phi^*(\theta)) - \nabla^2_{\theta\phi} g(\theta, \phi^*(\theta)) \left[\nabla^2_\phi g(\theta, \phi^*(\theta))\right]^{-1} \nabla_\phi f(\theta, \phi^*(\theta)).$$

Two types of biases: inability to compute $\phi^*(\theta) \;\Rightarrow\; \|\phi_{k+1} - \phi^*(\theta_k)\|^2$.

estimation of $[\nabla^2_\phi g(\theta, \phi)]^{-1} \Rightarrow \|\mathbb{E}[H_k|\mathcal{F}_k] - \nabla V(\theta_k)\|^2$.

Experiments: Importance Weighted Autoencoder

Objective: : Maximize the evidence lower bound (ELBO):

$$\log p_\theta(x) \geq \underbrace{\mathbb{E}_{q_\phi^{\otimes k}(\cdot|x)}\left[\log \frac{1}{k}\sum_{\ell=1}^{k} \frac{p_\theta(x, z^{(\ell)})}{q_\phi(z^{(\ell)} \mid x)}\right]}_{\text{IWAE}} \geq \underbrace{\mathbb{E}_{q_\phi(\cdot|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]}_{\text{VAE}}.$$

Bias Control for IWAE

Theorem: Assume that for all $\theta \in \Theta$, $\|\nabla_\theta \log p_\theta(x, z)\| \leq M$ for some $M > 0$. Then, there exists $C > 0$ such that for all $\theta \in \Theta$ and $\phi \in \Phi$,

$$\left\|\mathbb{E}_{q_\phi^{\otimes k}(\cdot|x)}\left[\widehat{\nabla}_\theta \mathcal{L}_k^{\text{IWAE}}(\theta, \phi; x) - \nabla_\theta \log p_\theta(x)\right]\right\| \leq \frac{C}{k}.$$

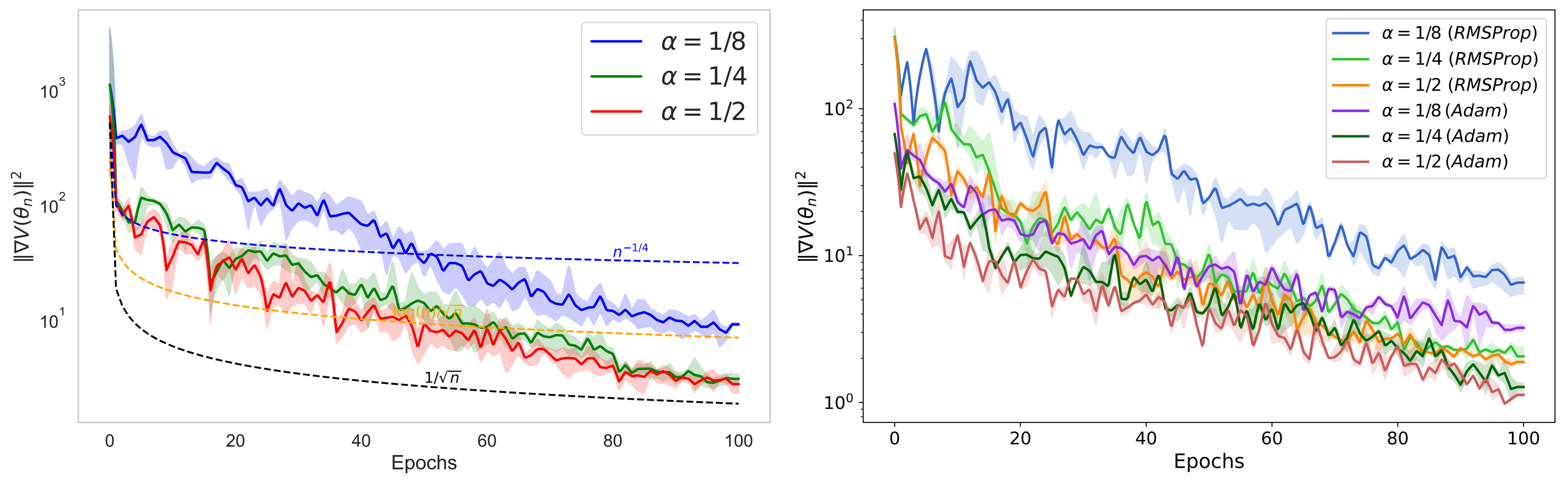

Figure: $\|\nabla V(\theta_n)\|^2$ with **Adagrad** (on the left), **RMSProp**, and **Adam** (on the right).

The Expected Convergence Rate:

$$\mathbb{E}\left[\|\nabla V(\theta_R)\|^2\right] = \begin{cases} \mathcal{O}\left(n^{-1/4}\right) & \text{if } \alpha = 1/8\,, \\ \mathcal{O}\left(n^{-1/2}\log n\right) & \text{if } \alpha = 1/4 \quad \text{and} \quad \alpha = 1/2\,. \end{cases}$$

Conclusion

- ☞ A convergence rate of $\mathcal{O}(n^{-1/2}\log n + b_n)$ for **Adaptive Biased SA** applied to **Adagrad**, **RMSProp**, and **Adam**, under non-convex smooth settings.
- ☞ Improved **linear convergence** rate with **Polyak-Łojasiewicz** condition.
- ☞ Application to **Stochastic Bilevel Optimization** and illustration of our convergence rate with **IWAE**.
- ☞ Crucial choice of an appropriate value α to achieve **fast convergence** without being too **computationally expensive**.

References

[1] Belhal Karimi et al. "Non-asymptotic analysis of biased stochastic approximation scheme". In: *Conference on Learning Theory*. PMLR. 2019, pp. 1944–1974.

[2] Yury Demidovich et al. "A guide through the zoo of biased SGD". In: *Advances in Neural Information Processing Systems*. Vol. 36. 2024.

[3] Ahmet Alacaoglu and Hanbaek Lyu. "Convergence of first-order methods for constrained nonconvex optimization with dependent data". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 458–489.

[4] Tianyi Chen, Yuejiao Sun, and Wotao Yin. "Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 25294–25307.