Non-asymptotic Analysis of Biased Adaptive Stochastic Approximation

Sobihan Surendran, Adeline Fermanian,
Antoine Godichon-Baggioni, Sylvain Le Corff

JdS - 30 May 2024

Introduction: Optimization in Deep Learning

Consider the unconstrained **Optimization Problem:**

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

Gradient Descent (GD):

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

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 $\nabla V(\theta_n)$.

Introduction: Optimization in Deep Learning

Consider the unconstrained **Optimization Problem:**

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

Gradient Descent (GD):

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

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 $\nabla V(\theta_n)$.

In Deep Learning:

- Objective Function: $V(\theta) = \mathbb{E}[\mathcal{L}(F_\theta(x), y)]$.
  - $\Rightarrow$ F_θ: Neural Network with parameters $\theta \in \mathbb{R}^d$ and $\mathcal{L}$: Loss Function.
- **SGD Update:**

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \nabla \mathcal{L}(F_{\theta_n}(x_{n+1}), y_{n+1}).$$

Motivating Examples

- Theoretical analysis of Vanilla SGD relies on **unbiased estimator**.

Motivating Examples

- Theoretical analysis of Vanilla SGD relies on **unbiased estimator**.
- In some applications, only **biased gradient estimators** are accessible.
  - Reinforcement Learning: Policy Gradient and Actor-Critic.
  - Monte Carlo: Importance Sampling and Sequential Monte Carlo.
  - Generative Models (biased objectives): VAE, IWAE, and BR-IWAE.
  - Zeroth-Order Gradient: Adversarial Networks.
  - Bilevel Optimization: Min-Max and Compositional Problems.

Motivating Examples

- Theoretical analysis of Vanilla SGD relies on **unbiased estimator**.
- In some applications, only **biased gradient estimators** are accessible.
  - Reinforcement Learning: Policy Gradient and Actor-Critic.
  - Monte Carlo: Importance Sampling and Sequential Monte Carlo.
  - Generative Models (biased objectives): VAE, IWAE, and BR-IWAE.
  - Zeroth-Order Gradient: Adversarial Networks.
  - Bilevel Optimization: Min-Max and Compositional Problems.
- We present a general framework for analyzing **SGD with biased gradient estimators and adaptive steps** based on biased control.

Table of Contents

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.

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

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

- Special cases: If $A_n = I_d \Rightarrow$ **Stochastic Approximation**.

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.
  - $H_{\theta_n}(X_{n+1}) = \underbrace{\nabla V(\theta_n) + \overbrace{b(\theta_n)}^{\text{bias}}}_{h(\theta_n)} + \underbrace{e_{n+1}}_{\text{noise}}$.
- Special cases: If $A_n = I_d \Rightarrow$ **Stochastic Approximation**.
  - $b(\theta_n) = 0$ and $e_{n+1} = 0 \Rightarrow$ **Gradient Descent**.

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

- 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 gradient estimator.

Examples

- **Adagrad:**
  - Square root of the inverse of the covariance of the gradient.

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

- **RMSProp:**
  - An exponential moving average of the past squared gradients.

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

    where β is the moving average parameter.
- **Stochastic Newton:**
  - A_n = Recursive estimate of the inverse Hessian.

Convergence Results

Convergence Analysis (Informal)

For any $n \geq 1$, let $\gamma_n = C_\gamma n^{-1/2}$ and $R \in \{0, \ldots, n\}$ be some discrete random variable. Under mild assumptions, we have:

$$\mathbb{E}\left[\|\nabla V(\theta_R)\|^2\right] = \mathcal{O}\left(\frac{\log n}{\sqrt{n}} + b_n\right),$$

where b_n corresponds to the bias term.

Convergence Results

Convergence Analysis (Informal)

For any $n \geq 1$, let $\gamma_n = C_\gamma n^{-1/2}$ and $R \in \{0, \ldots, n\}$ be some discrete random variable. Under mild assumptions, we have:

$$\mathbb{E}\left[\|\nabla V(\theta_R)\|^2\right] = \mathcal{O}\left(\frac{\log n}{\sqrt{n}} + b_n\right),$$

where b_n corresponds to the bias term.

- The term $\log n/\sqrt{n}$ arises from classical adaptive step-size methods.
- The term b_n represents the additive bias term, which can be constant or time-dependent.

Application to Adagrad, RMSProp, and Adam

Convergence Analysis

- Assume the smoothness of V and for any $n \geq 1$, let $R \in \{0, \ldots, n\}$ be a uniformly distributed random variable.
- There exists $M \geq 0$ such that for all $n \in \mathbb{N}$, $\|H_{\theta_n}(X_{n+1})\| \leq M$.
- Suppose that for any $n \geq 1$, there exist positive constants α and C_α such that:

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

Then,

$$\mathbb{E}\left[\|\nabla V(\theta_R)\|^2\right] = \mathcal{O}\left(\frac{\log n}{\sqrt{n}} + b_n\right) .$$

Application to Adagrad, RMSProp, and Adam

Convergence Analysis

- Assume the smoothness of V and for any $n \geq 1$, let $R \in \{0, \ldots, n\}$ be a uniformly distributed random variable.
- There exists $M \geq 0$ such that for all $n \in \mathbb{N}$, $\|H_{\theta_n}(X_{n+1})\| \leq M$.
- Suppose that for any $n \geq 1$, there exist positive constants α and C_α such that:

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

Then,

$$\mathbb{E}\left[\|\nabla V(\theta_R)\|^2\right] = \mathcal{O}\left(\frac{\log n}{\sqrt{n}} + b_n\right) .$$

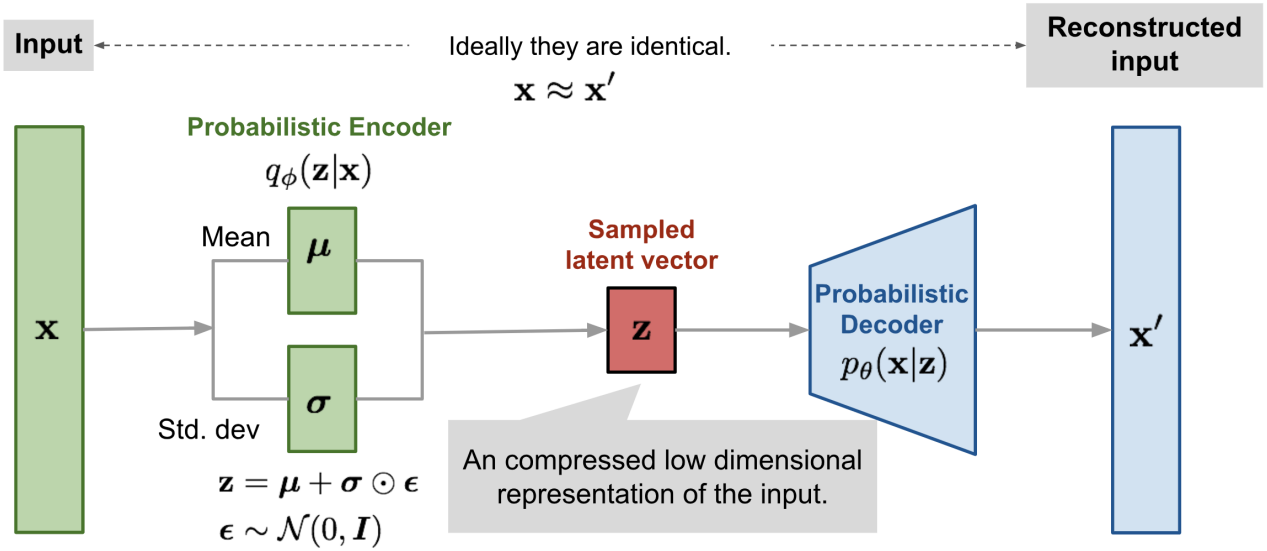
The bias term b_n is given by:

$$b_n = \begin{cases} \mathcal{O}\left(n^{-2\alpha}\right) & \text{if } \alpha < 1/4 , \\ \mathcal{O}\left(n^{-1/2}\right) & \text{if } \alpha > 1/4 , \\ \mathcal{O}\left(n^{-1/2} \log n\right) & \text{if } \alpha = 1/4 . \end{cases}$$

Generative Model: VAE

- **VAE**: A generative model aims to model the distribution of data.
- **Objective**: Maximize the evidence lower bound (ELBO):

$$\log p_\theta(x) = \log \mathbb{E}_{q_\phi(\cdot|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \geq \mathbb{E}_{q_\phi(\cdot|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathcal{L}_{\text{ELBO}}(\theta,\phi;x) \ .$$



Impact of Bias: VAE, IWAE, and BR-IWAE

- **IWAE**: A variant of the VAE that achieves a tighter ELBO:
  $\Rightarrow$ **Bias of gradient estimator** $= \mathcal{O}\left(1/k\right)$.

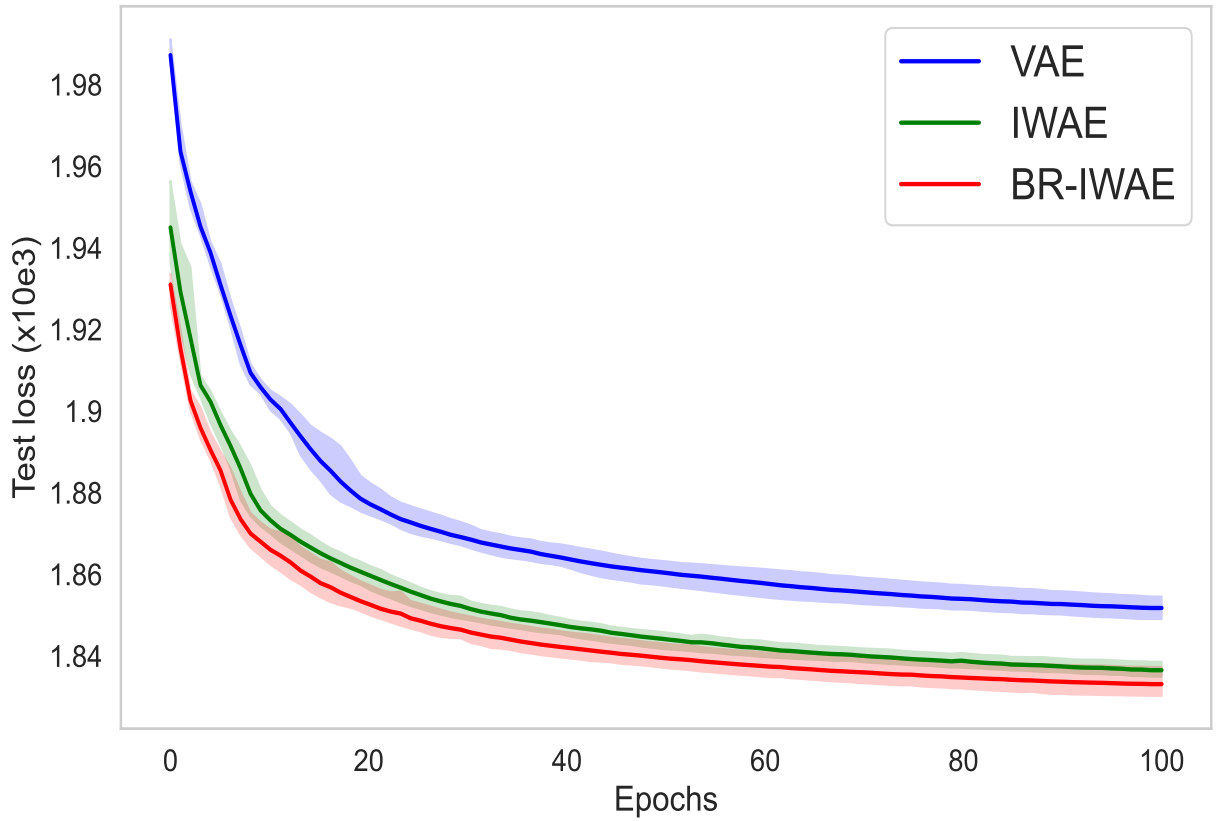


Figure: Negative Log-Likelihood for Different Generative Models on CIFAR-10.

Illustration of our convergence rate in IWAE

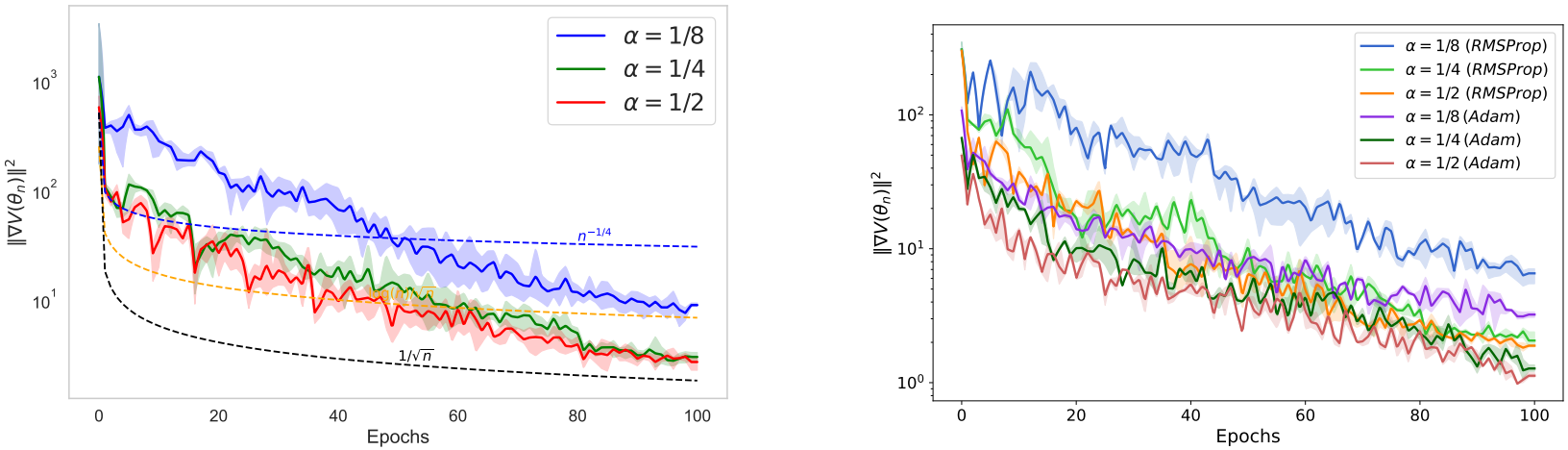


Figure: Value of $\|\nabla V(\theta_n)\|^2$ in IWAE with Adagrad (on the left), RMSProp, and Adam (on the right).

- **The Expected Convergence Rate:**

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

The Impact of Bias over Time

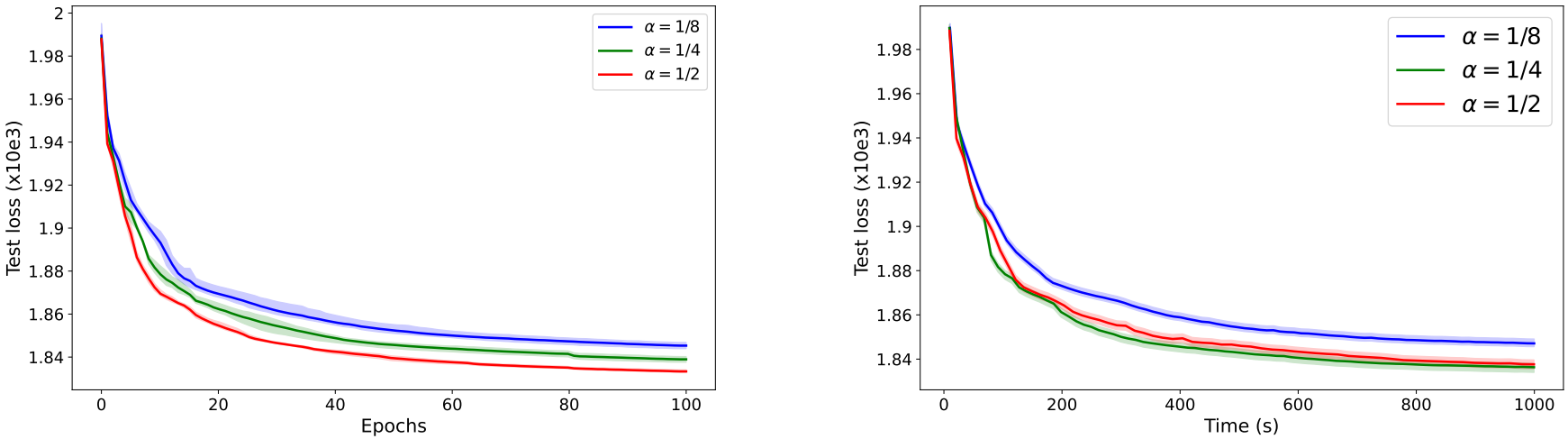


Figure: Negative Log-Likelihood on the CIFAR-10 dataset for different values of α over time (in seconds).

Take home messages

- Convergence rate of **Adaptive Stochastic Approximation** (Adagrad, RMSProp, and Adam):

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

- Crucial choice of an appropriate value α to achieve **fast convergence** without being too **computationally expensive**.

Thank you for your attention!